Cognitive Systems Research 88 (2024) 101297

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

From human explanations to explainable AI: Insights from constrained optimization

Inga Ibs [*,1], Claire Ott[1], Frank Jäkel, Constantin A. Rothkopf

Centre for Cognitive Science & Institute of Psychology, TU Darmstadt, Alexanderstraße 10, Darmstadt, 64283, Hesse, Germany

ARTICLE INFO

Action editor: Mog Stapleton

Keywords:
Explanations
Explainable AI
Constrained optimization
Linear programming
Complex problem solving
Microworlds

ABSTRACT

Many complex decision-making scenarios encountered in the real-world, including energy systems and infrastructure planning, can be formulated as constrained optimization problems. Solutions for these problems are often obtained using white-box solvers based on linear program representations. Even though these algorithms are well understood and the optimality of the solution is guaranteed, explanations for the solutions are still necessary to build trust and ensure the implementation of policies. Solution algorithms represent the problem in a high-dimensional abstract space, which does not translate well to intuitive explanations for lay people. Here, we report three studies in which we pose constrained optimization problems in the form of a computer game to participants. In the game, called Furniture Factory, participants manage a company that produces furniture. In two qualitative studies, we first elicit representations and heuristics with concurrent explanations and validate their use in post-hoc explanations. We analyze the complexity of the explanations given by participants to gain a deeper understanding of how complex cognitively adequate explanations should be. Based on insights from the analysis of the two qualitative studies, we formalize strategies that in combination can act as descriptors for participants' behavior and optimal solutions. We match the strategies to decisions in a large behavioral dataset (> 150 participants) gathered in a third study, and compare the complexity of strategy combinations to the complexity featured in participants' explanations. Based on the analyses from these three studies, we discuss how these insights can inform the automatic generation of cognitively adequate explanations in future AI systems.

1. Introduction

Applications of Artificial Intelligence (AI) in the real world face ethical concerns and a lack of trust. Explainable AI is often seen as a remedy. In the development of trusted AI systems, there are two approaches to explainable AI. The first one uses interpretability as a criterion for how well the algorithm's decisions can be understood, and the other aims for explicit explanations of decisions. As a foundation for the latter, Miller (2019) proposes to use human explanations as a point of reference. The biases and expectations that humans have for explanations are hoped to provide a good basis for evaluating and improving automatically generated explanations. Often, explainable AI systems are designed by engineers who start with their own intuitions and with what is technically possible within an existing AI system. The explanations are then evaluated in subsequent user studies (Rong et al., 2023). In this paper, we start from empirical observations on how users explain their solutions to a class of problems before we implement an explainable AI system — which is then done in a companion paper (Ott & Jäkel, 2024).

As an example, take a real-world optimization problem: Planning a country's transition to renewable energy. Experts build complex models of the energy system and use different scenarios for which optimization and planning algorithms return the best course of action (Barbosa et al., 2021). While experts who develop such models have some understanding of the problem and the solutions, the model's behavior can surprise them and help them to improve their understanding. In fact, traditionally, an important reason why scientists and engineers build models and simulations in the first place is to manage the complexity of real-world problems. Ideally, these models are as complex as necessary to be realistic but also as simple as possible to be understandable. Increasingly, however, solutions to very complex real-world problems can only be generated by complex algorithms performing lengthy and cumbersome computations.

How can we best support humans in understanding complex problems with complex solutions that were found by complex algorithms? Obviously, explanations can have various functions and depend on the

[*] Corresponding author.
E-mail address: inga.ibs@tu-darmstadt.de (I. Ibs).
[1] Both authors contributed equally to this paper.

https://doi.org/10.1016/j.cogsys.2024.101297
Received 31 July 2023; Received in revised form 6 February 2024; Accepted 3 October 2024
Available online 18 October 2024

explainee, their prior knowledge, and their specific questions. Here, we will focus on people who have a vested interest in understanding a problem and have to make decisions based on their understanding (or advise the people who have to actually make the decisions). Explaining the algorithm for finding a solution can help build trust, but often will not help people understand the solution better. The goal of explainable AI cannot be to explain all the details of an algorithm or how exactly a solution was found. After all, we use computers because we cannot and do not want to deal with all these details ourselves. Hence, human experts who try to understand complex problems will try to build intuitions by simplifying and abstracting a problem as much as they can to reduce complexity. Instead of finding an optimal solution to a problem, they may use approximations and heuristics. Instead of giving a complete explanation for a solution, they may settle for plausible arguments and leave the rest to a computer.

The challenge of generating cognitively adequate explanations in AI is thus twofold. First, we have to translate precise computations on data structures that are convenient for computers to representations that humans can understand and would use in their own heuristic reasoning. Second, we need to provide explanations with a complexity that humans can handle. It will not always be possible to do so without losing something in the translation, but the hope is that it can often be done in a way that the explanation is satisfactory for the explainee — keeping in mind that people often suffer from the illusion of explanatory depth (Rozenblit & Keil, 2002).

In this paper, we will study how people solve constrained optimization problems. In particular, we will describe the mental representations and heuristics they use to solve such problems. In addition, we will elicit explanations for how people think they solve these problems and how they evaluate the solutions. We also analyze the complexity of people's heuristics and explanations. We argue that these analyses can form the basis for future explainable AI systems. In a companion paper, we present a first prototype for such a system (Ott & Jäkel, 2024).

1.1. Constrained optimization and explainable AI

We decided to study constrained optimization problems because they frequently occur in real-world applications. Like other optimization problems, they consist of an objective function and a set of variables. The goal is to choose the values for the variables such that the objective function takes its optimal value (either the minimum or the maximum, depending on the problem). In addition, constrained optimization problems have constraints that have to be fulfilled, hence the name. In the energy transition example, the variables of interest are how many power plants should be built or decommissioned, the objective function is the cost, and the constraints are to stay below the CO2 targets, and at the same time meet the energy demand. In general, in addition to energy systems, constrained optimization tools have long been used to solve many types of difficult, but structured optimization problems, such as those arising in transportation or resource distribution (Dantzig & Thapa, 1997). When these optimization problems are formalized as Linear Programs (LPs) or Mixed Integer Linear Programs (MIPs), well-studied and well-understood algorithms exist to find an optimal solution. However, understanding these algorithms does not mean that the optimal solution or the original problem are easy to understand. Moreover, due to the widespread use of these algorithms, people who do not understand the algorithms might still need to understand the problems that they want to solve and the solutions that an algorithm provides.

With the success of neural networks, which are essentially black boxes, explainable AI has become an important sub-field of AI (Adadi & Berrada, 2018). But the term *explainable AI* is now used more generally for AI algorithms that come with explanations, and it is recognized that also classical AI methods often require explanations. Still, some readers might not classify numerical solvers for optimization problems as AI algorithms, even though they are the backbone of many AI applications and the historical roots of AI and operations research overlap considerably. However, as they are among the most important algorithms that are used in applied decision analysis, there is a practical need to explain the solutions that they generate. This has long been recognized for Linear Programming and, therefore, dedicated AI systems have been developed that explain LPs (Greenberg, 1983, 1993). Interestingly, humans are quite adept at solving constraint satisfaction problems as long as the problems are sufficiently small (finding the optimal solution is more difficult for them, though). This opens up the possibility to develop AI algorithms that mimic the heuristic strategies that people use when they try to solve constrained optimization problems in order to generate explanations that will be meaningful for them.

1.2. Overview

We will first introduce the basic paradigm that we use for our behavioral studies, the Furniture Factory, in Section 2. We designed two tasks from one constrained optimization problem, formulated as an MIP, where one is dynamic and participants have to make consecutive decisions that effect later states, while the other task is more focused on how participants explore and evaluate alternative solutions. In Section 3 we analyze concurrent explanations that were elicited in a study based on the exploration task. The goal was to identify the representations and heuristics that people use to explain their decisions while solving constrained optimization problems. We also elicited post-hoc explanations in a second study based on the sequential-decision making version of the task in Section 4. This was done to validate the representations and heuristics that were found in the first study (Section 3). We then formalize general versions of the heuristic strategies that people use. These formalized strategies can be used as components of automatically generated explanation. We investigate how well combinations of these strategies describe the behavioral data and compare their complexity to the participants' explanations. In Section 5 we match these strategies against another, more extensive dataset from a third study with more diverse participants. On this dataset, we investigate in more detail how well the heuristic strategies match situations in which participants agree on one decision.

2. Furniture factory

We developed a new paradigm, the Furniture Factory, that was specifically designed to investigate how people solve and explain attribute-rich constrained optimization problems.

The Furniture Factory is based on an example problem featured in Chvatal (1983, p. 102–105). The furniture manufacturing problem describes a company with two branches, one factory branch specializing on chairs and tables and the other one on beds and bookcases. These items of furniture require different amounts of wood, metal, and different amounts of time in specific workshops. The objective in this problem is to maximize profit by determining the number of pieces of furniture to build under the constraint that the total amount of resource costs for the furniture items does not exceed the available resources. Fig. 1 shows the structure of the Furniture Factory. This problem is highly connected, in the sense that the optimal value of one variable depends on multiple constraints and the values of other variables. The paradigm can be used in experimental tasks involving lay people (regarding their experience with furniture production and optimization), since domain knowledge is not required to solve this problem.

The Furniture Factory is formalized as a MIP with the aim to find a solution vector $x \in \mathbb{N}_0^4$ with the properties:

$$\begin{aligned} \text{maximize} \quad & \sum_{f \in F} c_f \cdot x_f \\ \text{s.t.} \quad & a_{r:} \cdot x \leq b_r \end{aligned}$$

for all resources $r \in R$ and furniture $f \in F$.

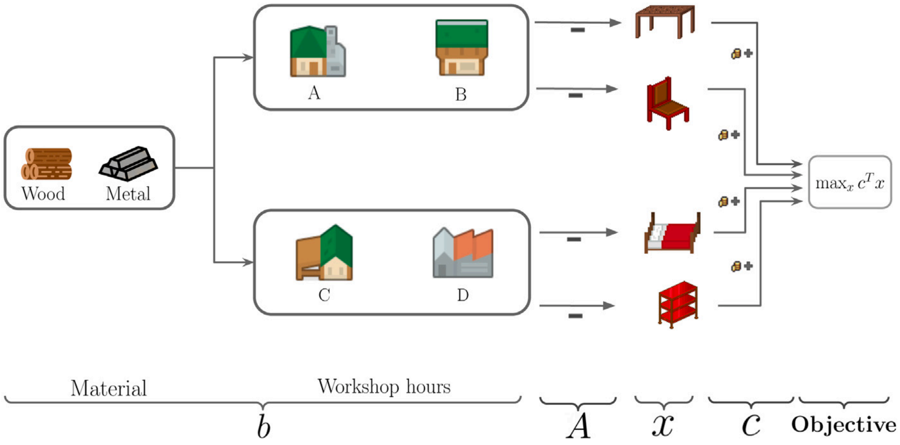


Fig. 1. Overview of the structural design of the Furniture Factory paradigm.

matrix $A \in \mathbb{Z}^{6\times4}$ contains all material and time costs, the vector $b \in \mathbb{Z}^6$ contains the available resources and the vector $c \in \mathbb{Z}^4$ contains the profit for each furniture item. Column and row vectors of A are denoted with $a_{:f}$ and $a_{r:}$. The formalization of the problem within the framework of MIPs allows the systematic variation of the task space by altering the cost functions and constraints of individual problems. Optimal solutions from the starting state and intermediate steps in trajectories can be fully computed in a reasonable time. We use well-conditioned problems, i.e., problems whose integer solution is close to the optimal solution of the LP-relaxation because ill-conditioned problems are harder to solve computationally (Pisinger, 2005).

2.1. Adaptation for problem-solving studies

Our studies aim to explore how humans solve constrained optimization problems and explain their solution approaches. To better understand how people approach these problems, we designed two different experimental tasks based on the Furniture Factory. To elicit concurrent explanations, we translated the paradigm to an exploration task, in which participants had to explore the solution space of the Furniture Factory in order to find the optimal solution (Section 3). To elicit post-hoc explanations with which we can compare the concurrent explanations, we translated the paradigm to a *sequential decision-making* task, in which participants were tasked to build furniture items sequentially (Section 4). Whereas the exploration task allows backtracking and comparing solutions, the sequential decision-making task encourages planning ahead before a decision is made. Generally, the task design aims to make the tasks accessible and solvable, so people can perform close to optimal.

To keep participants engaged and motivated, especially during online experiments, the tasks were posed as computer games (Frodl, 2021). Our games use a classic pixel-map design and some moving parts to indicate activity. We also use feedback, such as stars and medals, for good and optimal solutions. We added narrative elements to enhance the user experience.

To compose trials for the studies, we define different problem instances, which vary in furniture-associated profits (c) and the number of resources given at the start of a trial (b). In each trial, participants have to solve different problem instances of the Furniture Factory. The different trials are explained within the game as different months in which the furniture factory has to be managed. Across months, i.e. across trials, demand and supply fluctuate.

2.2. Defining instances

To gain insights into the solution approaches that participants employ, the trials in the experimental task should be engaging and non-trivial. Therefore, the difficulty of the task is an important factor, which can be manipulated using a variety of experimental parameters, including time constraints, support tools, and the number of decisions the participant has to make. The domain of constraint optimization problems provides a rich base for different problem instances, even in the same problem domain. It provides a complex environment with many interdependent problem parameters defined by the cost function c and the constraints (A, b), from which many problem instances can be created.

Although they are easily manipulable, the number of constraints of the Furniture Factory makes it difficult to set the trial parameters with an analytic approach. A critical factor for task difficulty is the size of the decision space, which in the Furniture Factory depends on the number of furniture items that can be built from the available resources given the costs in resources for each item. However, changing either the resource cost (A) or the availability of resources (b) effects the overall number of furniture that can be built. Therefore, manipulating these parameters can result in a straightforward problem, with only a few decisions, or an extremely complex problem with a vast problem space. By considering the connections of various parameter manipulations and the fact that small linear programs – like those we need for our experiments – can be solved quickly on a modern computer, we can search systematically for desirable task parameters.

In designing the different problem instances for the trials, the search space for suitable parameters was reduced by determining the size of the decision space by fixing the resource cost parameters A and the availability of workshop hours (parts of b) for all instances. Fixing these values followed the narrative of the experiment since the costs for the production of items should not change, and we assume that workshop capacities do not change for the entire management time either. Manipulating the profit function and the available material allowed us to construct different non-trivial problem instances. We also wanted to encourage participants to think intuitively about the problem space and find qualitative strategies. Hence, we did not allow the use of notes or calculators and chose the numbers to make the computations manageable.

To systematically search combinations of problem instances that we can use in the experiments, we defined different markers for promising problem instances based on their optimal solutions. In particular, we wanted some variation in the furniture items that occur in the optimal solution of each individual instance and across all instances. This should limit priming effects and stop participants from simply building the same items in different trials.

To evaluate the difficulty of each problem instance, three ‘naive’ agents were used to estimate a lower baseline for performance. The goal is to find instances in which the baseline is not too close to the optimal solution, which ensures that the most naive strategies are not always the best. The *expensive material agent* only builds the items that require the most material at a given step. The *immediate profit agent* builds the maximum amount of the item with the highest profit. Finally, the *margin profit agent* sums the cost of an item and subtracts it from its profit. If any of those naive agents yield close to optimal performance in all instances, the instance combination is deemed too easy.

We chose twelve problem instances for the three studies discussed in the following sections based on these factors (variability in items for the optimal solution over trials and difficulty for naive agents). The amount of resources was doubled for the exploration study that allowed backtracking to increase the task’s difficulty.

3. Eliciting representations and heuristics with concurrent explanations

In a first study, we gathered qualitative data in the form of concurrent explanations and quantitative data in the form of solution trajectories using the exploration task design of the Furniture Factory. In the following, we analyze the form and complexity of solution strategies featured in explanations elicited from participants while they solved constrained optimization problems. The main purpose of this first study was to generate hypotheses about the representations and heuristics that people use by directly eliciting explanations.

3.1. Method

3.1.1. Setup

The elicitation study was conducted as a computer game and took place in a university lab (in contrast to the other studies reported in this paper that were conducted online). Screenshots of the game are shown in Fig. 2. Participants explored the space of possible solutions using four control sliders until they were satisfied with their solution. The final solution is defined by the number of beds, bookcases, tables, and chairs that the participant wants to build in one trial of the game. In the game, different trials were labeled as different months. All relevant information for solving the task was given and was always present on the screen. This information included the number of units of resources still available, the resources necessary to build each item of furniture (its cost), the value of a finished furniture item (its profit), the current trial number, and the total return over all trials, i.e., the cumulative value. To motivate participants, a tally of optimal solutions was shown in the form of stars.

Participants had to manage the Furniture Factory for seven trials, i.e., seven months, that corresponded to seven different problem instances, including one tutorial. If a participant attempted to move a slider to build more items than are possible with the available resources, they were informed about the insufficient resources and the slider was blocked at the highest possible value. Starting from the second trial, a newspaper page was shown at the beginning of each trial with the new net profit for each furniture item, the available wood and metal, and an overview of the performance so far (Fig. 2 b). After each trial, participants were also shown their solution and an optimal solution, i.e., the number of chairs, tables, beds, and bookshelves, which would have brought the most profit in this trial. Participants were assigned to one of two groups. Both groups received the same seven problem instances, albeit in two different orderings to counterbalance the potential effect of learning. Before the seven trials, all participants were introduced to the mechanics of the problem, the control sliders, and the objective of the game through a tutorial.

3.1.2. Verbal reports

Verbal reports are a well-established method to collect data about the strategies that participants use to solve a problem (Ericsson & Simon, 1993; Newell & Simon, 1972). An advantage over questionnaires is that questions do not need to be tweaked to certain situations, which might bias participants. However, there are several concerns with the common think-aloud methodology, and about asking people about their problem solving strategies more generally. The most important concern is that asking people to think aloud might change their behavior, i.e., the method might be reactive. It is known that people can be nudged to think more analytically about a task (Chi et al., 1994) or use more metacognition (Berardi-Coletta et al., 1995), depending on how exactly they are instructed. While, in general, think-aloud instructions do not seem to be reactive, other methods for eliciting verbal reports are indeed reactive and often lead to an increase in performance (Fox et al., 2011). This is particularly true for methods that ask participants to explain their reasoning. We take this as an advantage because we want to study successful problem solving and therefore explicitly ask participants to explain their actions (Jäkel & Schreiber, 2013). Still, constantly talking might change how participants play the game and might take the fun out of the task.

Importantly, however, we are interested in how people explain their behavior when solving constrained optimization problems in order to lay the groundwork for automatically generated and cognitively adequate explanations for future AI systems. Hence, we very directly asked participants to give reasons for their behavior. Even if the reasons that people give should turn out to be not completely consistent with their behavior (which we will also check below in the two follow-up studies), their explanations are still valuable for us as we want to leverage their explanations for explainable AI.

Concretely, participants were instructed to explain each action and why they did it or what their goal was. There also was a reminder to ‘please remember to justify all of your actions out loud’ in the newspaper. If people did not speak for some time, they were reminded to do so by the experimenter, who was sitting in the room during the whole experiment but out of sight. After the last two trials, participants were asked to reflect on their solution and how they rate their solution.

All utterances were recorded together with a screencast and analyzed using QualCoder (Curtain, 2021). The audio recordings were transcribed and analyzed. Since the aim was to identify concrete representations and strategies from the explanations, a coding scheme comprising categories of attributes and operations for similar utterances was defined. This coding scheme was iteratively refined to find the most precise categorization of these utterances. Two coders (two research assistants) labeled all utterances according to the resulting scheme, and we use all codes where both coders agreed. The labels we are interested in here relate to the representation used by the participants to explain their reasoning. They fall in the categories of *assessing* and *comparing* different attributes, like the available materials, or current costs and profits of furniture items. Utterances that could be ascribed to concrete heuristic strategies were grouped by categories defining the basis of these strategies (details of the strategies will be explained below).

3.1.3. Participants

The study was conducted with twelve participants, however the data from one participant had to be excluded due to technical difficulties during the experiment. This resulted in a total of 11 datasets from five female and six male participants. Participants were recruited among students of psychology and cognitive science who received partial course credit for their participation. Participation was voluntary and participants gave informed consent. The study was approved by the local ethics committee.

A short questionnaire was given to the participants, in which they had to indicate their gaming experience and their age and gender. Their age ranged from 18 to 29 with seven participants stating their age between 18 and 24 and four between 25 and 29. Six participants

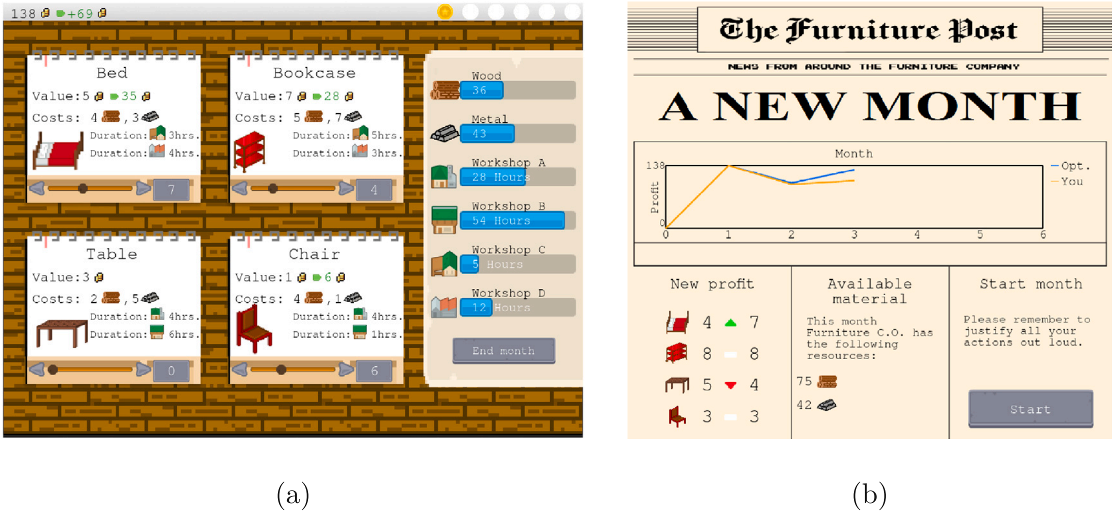


Fig. 2. The exploration task as computer game. (a) The main interface with all task relevant information and sliders to change the built furniture. (b) Display at the beginning of each trial, to indicate changes in net profit for the furniture and available material.

described their gaming experience as playing 'no games in my spare time' and five participants stated they played at 'irregular intervals'. The language of the instructions and the participants' reports was their native language, German. All prompts and utterances will be translated in the following.

3.2. Results

3.2.1. Performance

We use Percentage Under Optimal (PUO) as a measure of performance of the participants, measuring how much worse a solution is compared to the optimal value. This value is calculated as

$$\left(\frac{\text{participant score}}{\text{optimal score}} - 1\right) \cdot 100.$$

The mean PUO over all participants and trials was close to optimal, with −2.0 (SD: 3.18). However, there was no participant who found the optimal solution in all six trials and performance varied across different participants, as shown in Fig. 3 (a).

3.2.2. Analysis of the explanations

To solve the task, participants needed to evaluate the different furniture items and decide which items should occur more often in the solution and which items less often. We are interested in these evaluations and comparisons: Which and how many furniture items are compared at the same time, and which and how much of the available information is used for these comparisons.

Example 3.1. Throughout this paper, we will use the situation shown in Fig. 2 (a) as a running example. Formally, the situation is represented by six state attributes (available resources, i.e., available wood, metal, and available time in the different workshops), which can be found in vector b and for each item, five furniture-specific non-zero attributes (profit, cost of wood, metal, and required time in the different workshops) found in c and A. In this state, the attributes have the following values (b) with R = {wood, metal, workshop A, workshop B, workshop C, workshop D}, and needed resources (A) and costs (c) of the furniture F = {table, chair, bed, bookcase}:

$$A = \begin{pmatrix} 2 & 4 & 4 & 5 \\ 5 & 1 & 3 & 7 \\ 4 & 4 & 0 & 0 \\ 6 & 1 & 0 & 0 \\ 0 & 0 & 3 & 5 \\ 0 & 0 & 4 & 3 \end{pmatrix} \qquad \begin{aligned} c &= (3, 1, 5, 7) \\ b &= (36, 43, 28, 54, 5, 12). \end{aligned}$$

There are five furniture-specific attributes (profit, cost of wood, metal, and required time in the different workshops) and six state attributes (available resources, i.e., available wood, metal, and available time in the different workshops).

From the analysis of the utterances, it is clear that the participants did not refer to all these attributes at once when they explained how they decided which item to build next. Instead, at any point in time they only focused on a subset of attributes. Three prominent heuristic strategies were observed that focus mostly on one aspect of the problem.

For the first strategy, participants focus on available resources and depleting them evenly. This strategy does not take profit into account, and therefore uses an incomplete representation of the problem. Utterances like 'I am now trying to get wood and metal to about the same level' and 'And just try to really pay attention to my time so that I use it up evenly' are indicators for this balancing strategy. Participants predominantly focus either on balancing material or time, so we differentiate between *material oriented* and *time oriented* strategies. Participants tend to use one of two methods to balance their values. The first is to build the item whose cost resembles the ratio of the resources most closely. In Example 3.1, such a strategy focused on material, i.e. metal and wood, prefers bookcases because the cost-ratio of 5 : 7 is most similar to the resource-ratio of 36 : 43. The other method is building the item with the most considerable difference in the two imbalanced resource-costs to close the gap quickly. In Example 3.1, this strategy prefers tables because the costs of 2 wood and 5 metal will close the gap between available wood and metal most quickly.

The second heuristic is entirely based on profits (vector c) and ranks the items by their immediate profit regardless of their cost. This strategy is reflected in utterances which we coded as *profit oriented*, like 'To produce as much of the most expensive item as possible, and then adapt the other furniture items to it'. This strategy prefers bookcases in Example 3.1 as long as there are enough resources to build them because they bring a profit of 7.

In a third heuristic strategy, participants estimated or calculated some ratio of costs and profits like 'A bed needs 7 resources and gives 7, a chair 5 and 3, that is, better build beds than chairs, right?'. We subsumed these utterances under *cost–benefit oriented* strategy. There are many ways in which some costs can be compared to the items' profit. For example, the profit can be divided by the sum of the costs for metal and wood. In Example 3.1, this leads to beds being preferred as they give 5 profit for a cumulative material cost of 7, which is a bigger ratio than for the other items. However, many other

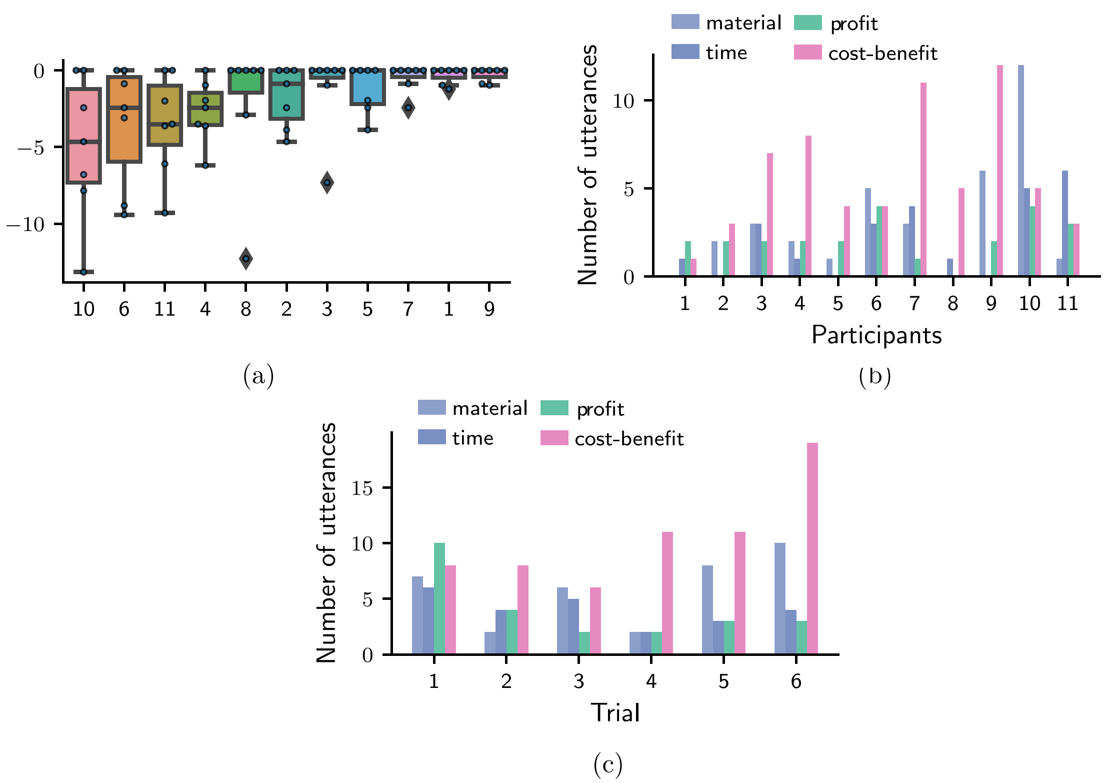


Fig. 3. Performance scores and strategy frequencies in the concurrent explanation experiment. (a) Percentage Under Optimal scores grouped by participants. The participants were ordered by their mean performance scores. (b) Amount of different comparison strategies mentioned by participants. The colors indicate the basis of the strategies (profit corresponds to profit oriented, etc.). (c) Amount of different comparison strategies by trial.

instantiations of this strategy are possible, which we will discuss later in the section. Fig. 3 (b) shows how often different strategies were coded in the utterances of different participants. While some participants did not mention either a material-oriented, time-oriented (balancing) or a profit-oriented strategy, all participants recognized the need to evaluate furniture items with regard to their cost and benefit. In Fig. 3 (c), the frequency of utterances for the individual trials is shown. Whereas in the first trial, the strategies are mentioned similarly often, in the later trials, participants mention the profit oriented strategy less, and cost–benefit and material balancing strategies become more prevalent.

The heuristic strategies uttered here can be used as components for explanations, whenever they apply. However, we also need to include more complex reasoning. To keep explanations understandable, it is important to know which complexities are still useful and what information is meaningful for participants. To gain more insights into these questions, we analyzed the participants' cost–benefit considerations in more detail. We concentrate only on cost–benefit considerations, since they are the most diverse utterances and give further insights into structure of the decision strategies. On average, participants compare 2.1 furniture items (SD: 0.8), e.g., chairs against beds. In this comparison, they involve 3.2 attributes (SD: 1.0). For example, they consider the costs as well as the general availability of resources. Fig. 4 (a) shows how state attributes (the general availability of the resources) and furniture-specific attributes (the required resources for each item and its profit) are combined to compare different numbers of furniture items. The most common combination is comparing two furniture items along three furniture-specific dimensions (these are often profit and either two material costs or two time costs). As an estimate of how complex a comparison is, we compute its complexity as |furniture items| · |furniture-specific attributes| + |state attributes|. This measure should roughly reflect the number of pieces of information that a participant needs to hold in working memory to make the comparison. This results in heuristics of complexity 6 as the most common complexity for cost–benefit considerations. This is around the expected magical number of 7 ± 2 (Miller, 1956). However, participants sometimes also consider far more complex comparisons, often when more furniture items are compared and more state attributes are involved (see Fig. 4 (b)). This is possible because they make these comparisons sequentially and therefore do not overwhelm their working memory.

Taking a closer look at the different components of an explanation – like the furniture items, state attributes, and furniture-specific attributes – shows that some are combined more often in comparisons than others. We use normalized point-wise mutual information, defined as

$$pmi_{xy} = \frac{-1}{\ln(p(x,y))} \ln\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

for occurrences of any combination of explanation components. Fig. 5 shows that most components have a negative value, indicating that they occur less often together than expected if they were independent. This matches our observation that participants focus on specific aspects of the problem at a time. Material resources and costs have especially low values when compared to time cost and resource, indicating that most comparison rules use either time- or material-based information. Metal and wood costs as well as resources have larger positive values, indicating that they frequently occur together. The same applies for beds and bookcases, as well as for tables and chairs. As mentioned before, people often focused on comparing two furniture items, and these values show that there is a preference for comparing items that are produced in the same workshops (see Fig. 1).

3.3. Conclusion

So far, we have observed that participants find optimal or close-to-optimal solutions for the exploration task of the Furniture Factory. This suggests that our participants use structured solution approaches to the problem because the problem instances were designed in a way that naive agents do not achieve optimal performance. By analyzing the

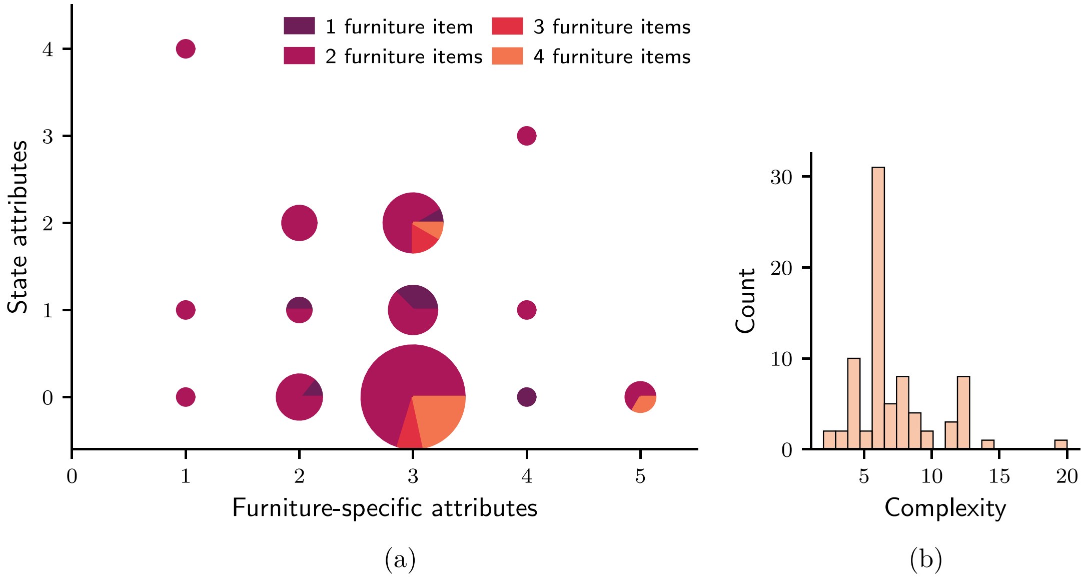
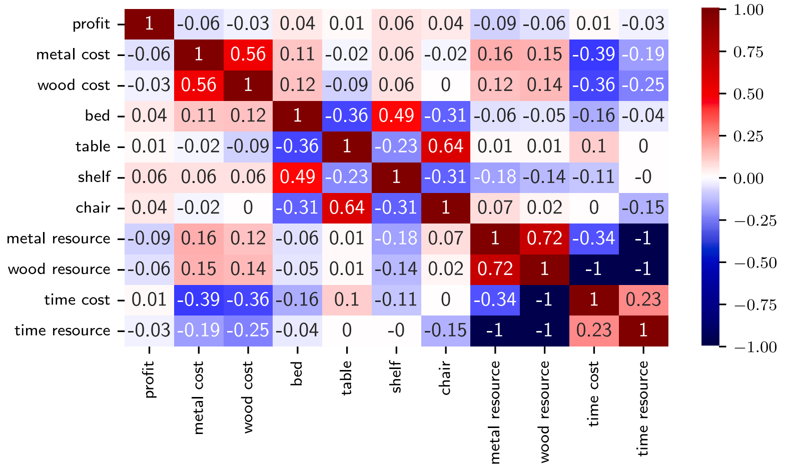


Fig. 4. Different complexity measures of cost–benefit evaluations. (a) Used number of state attributes like the available wood or metal compared to furniture-specific attributes like profit or cost. The pie charts indicate how many furniture items were compared at the same time. (b) Frequencies of the complexity of strategies and calculations mentioned calculated as |furniture items| · |furniture-specific attributes| + |state attributes|.



	profit	metal cost	wood cost	bed	table	shelf	chair	metal resource	wood resource	time cost	time resource
profit	1	-0.06	-0.03	0.04	0.01	0.06	0.04	-0.09	-0.06	0.01	-0.03
metal cost	-0.06	1	0.56	0.11	-0.02	0.06	-0.02	0.16	0.15	-0.39	-0.19
wood cost	-0.03	0.56	1	0.12	-0.09	0.06	0	0.12	0.14	-0.36	-0.25
bed	0.04	0.11	0.12	1	-0.36	0.49	-0.31	-0.06	-0.05	-0.16	-0.04
table	0.01	-0.02	-0.09	-0.36	1	-0.23	0.64	0.01	0.01	0.1	0
shelf	0.06	0.06	0.06	0.49	-0.23	1	-0.31	-0.18	-0.14	-0.11	-0
chair	0.04	-0.02	0	-0.31	0.64	-0.31	1	0.07	0.02	0	-0.15
metal resource	-0.09	0.16	0.12	-0.06	0.01	-0.18	0.07	1	0.72	-0.34	-1
wood resource	-0.06	0.15	0.14	-0.05	0.01	-0.14	0.02	0.72	1	-1	-1
time cost	0.01	-0.39	-0.36	-0.16	0.1	-0.11	0	-0.34	-1	1	0.23
time resource	-0.03	-0.19	-0.25	-0.04	0	-0	-0.15	-1	-1	0.23	1

Fig. 5. Matrix of the normalized point-wise mutual information of two atoms being present in an utterance coded as cost–benefit. Red numbers indicate a more frequent co-occurrence than expected if atom mentions were independent, blue numbers indicate a less frequent co-occurrence than expected.

protocols of the elicited verbal reports, we obtained several insights about the representations and heuristic strategies used in the explanations of our subjects. Participants described profit-based, balancing, and cost–benefit oriented strategies. Utterances coded as balancing or cost–benefit-based strategies often consider either material or time resources only. The complexity of the strategies in the utterances was analyzed as well. As a result of the reduced representations, most strategy utterances had a complexity of 6, but much higher complexities were observed in the concurrent explanations. Thus, the strategies and representations identified in the explanations and the complexity analysis can provide a point of reference for components of automatically generated explanations.

4. Validating representations and heuristics in post-hoc explanations

While the previous experiment was designed to elicit concurrent explanations by participants to investigate the representations and heuristics involved in finding solutions to Furniture Factory problems, it is clear that people do not solve constrained optimization problems the way MIPs solvers do. In the first experiment, people could backtrack. However, in most real-world scenarios, people will not be able to backtrack and, clearly, explanations are easier to understand when they have a sequential structure that does not require a lot of backtracking. Hence, in order to investigate whether the problem representations and heuristic strategies elicited in the first study generalize to the sequential scenario, we devised a second study. In this study, participants had to sequentially build individual items of furniture without being allowed to backtrack, i.e., without being able to change previous decisions. This contrasts with the task used in the first study, where participants could backtrack by changing the sliders back to a previous setting and where they could thus explore the whole solution space before making a final decision on what to build. Solving the sequential decision-making problem requires more systematic planning with decision strategies based on the current state and future states of the system. This necessity of planning to solve the problem well helps us to elicit informative goal-oriented explanations, since decisions during a trial are not based on exploratory actions but should be based on strategic

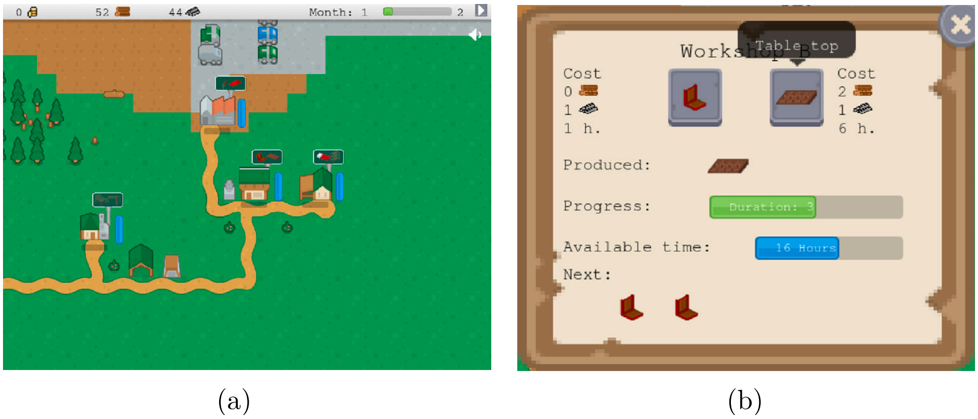


Fig. 6. Interface for the game of the sequential decision-making task. (a) Main interface with the map of different workshop buildings. Actions and information are accessed by clicking on the different buildings. (b) Workshop B where table tops and chair backs can be built with the indicated resources. Items can be queued to be built as soon as the previous items are finished.

reasoning. In this study with the sequential decision-making task, we infer problem-solving strategies from participants' explanations (similar to the previous study) but not concurrently. Instead, we ask participants for post-hoc explanations.

4.1. Method

4.1.1. Experimental setup

The sequential decision-making task based on the Furniture Factory was posed to participants in the form of an online game, for which the interface is shown in Fig. 6. Given the initial resources, a participant could decide to build parts of furniture items, which lead to a new state of the environment in which the resources used in the current decision were deducted from the available resources. A trial ended when the resources were insufficient to build any additional furniture items. Then, the next trial started with a new set of resources. The objective was again to maximize the profits in each trial by building the optimal amount of each furniture item.

In addition to the change into a sequential decision problem, the paradigm was 'gamified' further to motivate participants to solve the task. Instead of deciding to build a whole piece of furniture at once, the decision was broken down into smaller steps where participants had to build all the parts that were needed for each item. For example, a chair needed four legs that needed to be built individually. To build one furniture item, the participant had to click on two workshop buildings and build the respective parts, which together were automatically assembled into one item. Building the parts required time. The time was proportional to the cost of workshop hours associated with it — 2–3 s per part, which resulted in 10–20 s in total for a full furniture item. Together with a time limit to play a full trial (3 min) these dynamics were intended to motivate participants to plan their actions in advance and develop systematic strategies to solve the problem. The real-time nature of the task also added to the participants' excitement about the game.

The experiment was given to the participants in the form of an online game, which was implemented in JavaScript and hosted on a Sosci-Survey Server (Leiner, 2019). Building decisions were logged during the experiment. The study was conducted in German, and all the instructions and questions mentioned here are translations. Participants had to complete a questionnaire for demographic information (age, gender) and game experience and were then instructed to play the game in an interactive tutorial. The game started with a training trial with the same problem parameters as the problem given in the tutorial. Subsequently, they had to complete 11 trials, each involving a different problem instance, defined by a unique set of resources and profit values associated with the furniture items.

To elicit post-hoc explanations from the participants for their solutions, questions were posed during the game by fictional characters. Until trial six, a supervisor at the company first explained the game to the participants and then asked them to evaluate their solutions after every trial. Then, a new coworker was introduced who asked the player to explain their strategies to him. To elicit explanations, the coworker asked 'What should I do to find the best possible solution?' after trial eight and 'How do I find out what to build in each month?' after trial ten. To ensure comparability between the answers on the questionnaire, the order of the instances featured in the trials was not randomized.

4.1.2. Participants

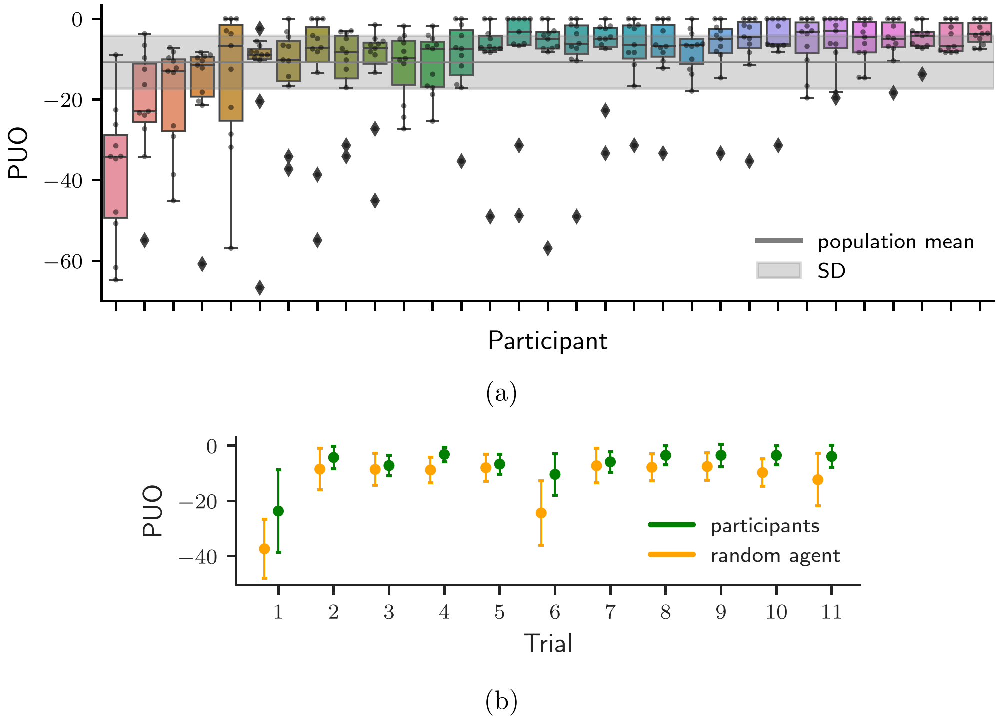
31 participants took part in the study, with 21 being female and 10 male. The age of the participants ranged from 18 to 59 years. Out of the total, 22 participants were between the ages of 18 and 24, 6 were between 25 and 29, into each of the age ranges of 30 to 39, 40 to 49, and 50 to 59 fell one participant. Regarding their gaming experience (using a predefined scale), 21 participants stated that they played computer games at irregular intervals, with a maximum frequency of once per month. The other 10 participants stated that they did not play computer games in their spare time. The study was advertised to students and employees of the university, and students studying psychology or cognitive science received partial course credit for participation. Participants gave informed consent and the study was approved by the local ethics committee.

4.2. Results

4.2.1. Performance

The mean performance under optimal for the 11 trials was −10.78 (SD: 6.54). Generally, the aggregated scores for all trials by each subject show variability in the performance scores, as shown in Fig. 7 (a). Note, however, that due to the non-randomized trial order, the effect of the linear program parameters cannot be disentangled from possibly confounding learning effects. To investigate if participants were able to handle the interface and formed strategies about their moves, the completeness of the solutions was analyzed. If at the end of the trial more furniture items could have been built, the solution was classified as incomplete, otherwise complete. On average, with exclusion of the training trial, 71.84% of the solutions were complete (SD: 9.02). The mean performance under optimal for the complete solutions was −6.47 (SD: 7.59).

Besides the comparison with an optimal agent, comparing the complete solutions of a group with the expected performance of a random agent can provide insights into the reasoning quality of a group. The performance of the group in the experiment (see Fig. 7 (b)), shows



(a)

(b)

Fig. 7. Performance of the participants in the sequential decision-making task without backtracking. (a) Percentage Under Optimal scores grouped by participants. Participants are ordered by their mean performance scores. The gray line corresponds to the population mean over all trials, and the shading to the standard deviation. (b) Average performance of participants (only complete solutions) compared to a random agent for each trial. The error bars represent the standard deviations.

that the participants performed better than random ($p < 0.001$, one-sided binomial sign test) in each trial apart from the training trial. For calculating the expected value of a random agent, all possible action trajectories have to be considered and the probabilities of each step need to be accumulated. At the beginning of a trial, all furniture items are available and have a probability of being chosen of 0.25 while later in a trial some items might not be a valid option due to a lack of resources, which changes the probabilities for the other items. We disregard the decomposition into parts.

4.2.2. *Formal descriptions of explanations*

The explanations that were elicited during the game were coded by three people, two of the authors and a research assistant, and analyzed using Qualcoder (Curtain, 2021) with a similar coding scheme as in the elicitation study (see Section 3). We used all codes that were used by at least two coders. Each participant gave two explanations for their strategies, one after the 8th and one after the 10th trial. Similar to the utterances in the previous study, participants mentioned balancing, ranking by profit (e.g., 'Determine which item makes the most profit and build some of it first') and determining cost–benefit ratios (e.g., 'It is important to produce furniture that is as resource-efficient as possible. This means that the ratio of the number of resources consumed to the selling price should be as good as possible') as primary strategies. Thus, the heuristic strategies that were mentioned in the concurrent explanations of the previous study (Section 3) are also used in post-hoc explanations for finding the optimal solution here. Fig. 8 (a) shows that keeping the available material balanced was mentioned most often, followed by concentrating on profit. Balancing the available time and weighing costs against benefits were also mentioned in about 30% of the explanations.

When focusing on balancing resources, some participants chose furniture items whose cost resembled the ratio of the available resources, as can be seen in utterances like 'build the furniture that most closely corresponds to the ratio.' Others preferred items with the most difference in their costs, like described in the utterance 'Pay attention to the extent to which this production imbalances your materials, and then produce either tables or chairs that require material that is strongly one-sided.' However, most explanations remain too vague to differentiate these two heuristics, so we count them together under balancing.

We want to know how complex these general explanations can be, how many pieces of information are involved, and how the different strategies are combined. Fig. 8 (b) shows that general explanations often contain either two or all six different state attributes (i.e., the available resources: metal, wood, and the workshop capacities). They also often include two or three furniture-specific attributes (i.e., the profit or the required resources of an item). We assume that general explanations are always about all four furniture items at once, which leads to higher complexity values than in the previous study, as can be seen in Fig. 8 (c).

Fig. 9 (a) shows that in statements about combinations of strategies, all combinations were mentioned at least once, however, only profit- and time-based strategies occur together more frequently. Fig. 9 (b) takes a closer look at which strategies occur with what information. We see the expected positive correlations between strategies and their main components (e.g., the time-based balancing heuristic co-occurs frequently with utterances about the time resources in the workshops), while most other relations are neither particularly frequent nor rare.

Based on the analysis of the verbal reports obtained in the previous study (discussed in Section 3) and the explanations given in this study, we derived general abstract versions of the heuristic strategies that people use: *immediate profit*, *balancing*, *gap reduction*, and a general form of *cost–benefit* strategy. A formal definition of each strategy is given in Box 10. We decided to formalize these strategies to capture people's intuitions in a way they might use them if they had plenty computational resources, not to be most similar to their actual cognitive processes. Obviously, participants will not actually compute exact cost–benefit ratios when the numbers are crooked. They will usually round, approximate, and resort to rough estimates. But we believe the formal strategies in Box 10 capture the essence of the strategies without these complications. Hence, they form a good basis

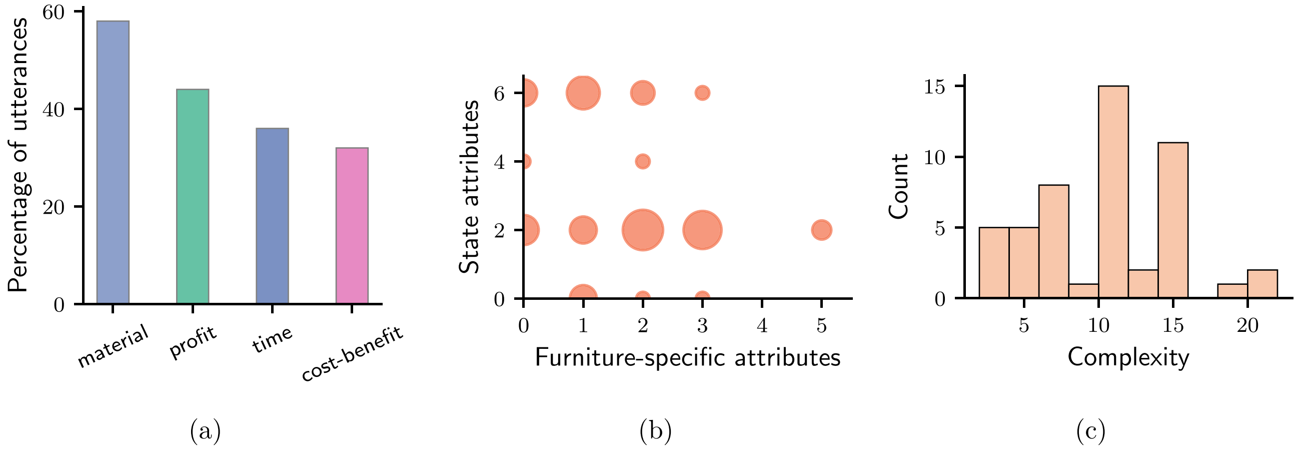


Fig. 8. Complexity measures for the strategies featured in post-hoc explanations. (a) Percentage of different strategies mentioned. (b) Number of state attributes that were used compared to furniture-specific attributes, like profit or cost. (c) Frequency of the overall complexity of strategies mentioned by the participants calculated as 4 · |specific information| + |state attributes|.

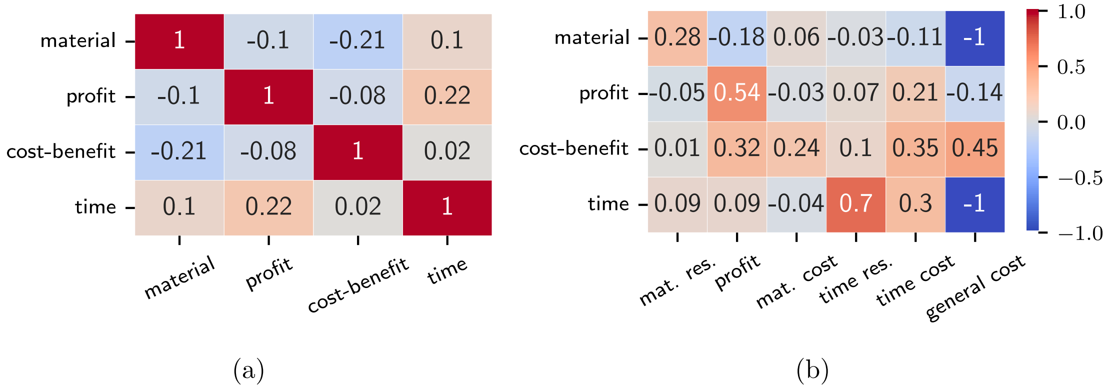


Fig. 9. Matrices showing the normalized point-wise mutual information of two classes of utterances co-occurring. Positive values (red) indicate a more frequent co-occurrence than expected of independent values, negative values (blue) indicate a less frequent co-occurence. (a) Co-occurrence of strategy utterances that are either coded as material, time, profit, or cost oriented strategies. (b) Co-occurrence of the strategy utterances together with information statements.

as components for cognitively adequate explanations for constrained optimization solutions.

We define *immediate profit* (see green box in Box 10) as building the item with the maximal profit in the set of items that can be built (feasible set), corresponding to a maximum complexity of 1 furniture-specific attribute × 4 items. This strategy is easy to compute, and the reasoning for choosing the item that yields the most profit is straightforward. Of course, this strategy only captures some of the problem and is usually not optimal on its own, but it can be a good starting point. This strategy ignores costs and available resources as mentioned in Section 4. In Example 3.1 immediate profit leads to the following values $V : V(\text{chair}) = 1$, $V(\text{table}) = 3$, $V(\text{bed}) = 5$ and $V(\text{bookcase}) = 7$, which leads to bookcases being the preferred choice.

Explanations for cost–benefit strategies (see pink box in Box 10) range in their concreteness and which aspects of the problem are included. A thorough evaluation should include the available resources as well as the costs. If a resource is rare, costs for this resource should weigh more than costs of a resource, that is available in abundance. Therefore costs are divided by the available resource values. The *cost–benefit* strategy evaluates furniture items based on their profit divided by the sum of its normalized costs, which leads to a complexity of 5 furniture-specific attributes × 4 items + 6 state attributes. The table has the best cost–benefit ratio in Example 3.1, when material and time costs are considered and are normalized using the resources. This notion of cost–benefit reflects that more metal than wood is available, and more time is spent in workshops A and B than in workshops C and D.

While *immediate profit* focuses only on furniture profits, some strategies focus solely on available resources and the according costs. Available resources define which building choices are still available, and a balanced use can lead to more available choices in later decisions. To formalize the heuristics that balance resources, we use two forms, each capturing this constraint-oriented approach. The first one, which we simply call *balancing* (see purple box in Box 10), evaluates items based on the similarity of their cost ratio to the available resource ratio. A strategy that can also be used for more than two values at a time is to calculate the cosine of the angle, between the resource vector and the cost vector. The minimal angle, and therefore maximum cosine value, is between the two vectors whose ratios most resemble each other. The complexity of this strategy depends on the number of constraints that are considered. Since participants indicated that they either consider time or material, we distinguish between material-based and time-based *balancing*. For *material balancing*, the complexity corresponds to 2 specific attributes × 4 items + 2 state attributes (required and available wood and metal resources). In Example 3.1, bookcases have a cost ratio for material most similar to the available wood and metal.

The complexity of *time balancing* corresponds to 2 specific attributes × 4 items + 4 state attributes. Remember that there are four workshops and, hence, the available time in each workshop are the relevant state attributes. Although in theory, therefore, four specific attributes are considered here, two of the attributes are irrelevant for the comparison (time costs are zero because each item only needs time in two workshops) for two items respectively which leads to a reduced complexity.

Formal Strategies

Take the item with the maximum value V according to a strategy:

Immediate Profit Complexity: 4

$$V(f) = \begin{cases} c_f, & \text{if } f \in F' \\ 0, & \text{otherwise} \end{cases}$$

Balancing (material or time) Complexity: 10 or 12

$$V(f, S) = \begin{cases} \frac{b|_S \cdot a_{:f}|_S}{\left\|b|_S\right\| \left\|a_{:f}|_S\right\|} & \text{if } f \in F' \\ 0, & \text{otherwise} \end{cases}$$

Cost-Benefit Complexity: 26

$$V(f) = \begin{cases} \frac{c_f}{\sum\limits_{r \in |R_0|} \frac{a_{r,f}}{b_r}}, & \text{if } f \in F' \\ 0, & \text{otherwise} \end{cases}$$

Gap Reduction (material or time) Complexity: 10 or 12

With $min = \operatorname{argmin}(b|_S)$ and $max = \operatorname{argmax}(b|_S)$ for all material or time resources from which an item can still be built.

$$V(f, min, max) = \begin{cases} a_{max,f}, & \text{if } f \in F' \text{ and } a_{min,f} = 0 \\ \frac{a_{max,f}}{a_{min,f}}, & \text{if } f \in F' \text{ and } a_{min,f} > 0 \text{ for all } f \in F' \\ 0, & \text{otherwise} \end{cases}$$

F' is the set of items that can still be built, R_0 is the set of resources that are bigger than zero and S is a subset of resources, either the set of materials or the set of workshop times. We denote vectors reduced to the relevant parts defined by subset S with $v|_S$.

Box 10.

Since it is not clear which form the actual balancing of the resources has from the statements of the participants, we also considered a second form of balancing, *gap reduction*. *Gap reduction* (see orange box in Box 10) considers the resources with the maximum and minimum value and chooses the item with the highest cost ratio of the two. This approach can be especially useful if the goal is to balance large differences in available resources quickly or to use the remaining resources most sufficiently. This strategy – as the previous one – can be material- or time-based and has a complexity of 2 specific attributes × 4 items + 2 state attributes if it is material-based or 2 specific attributes × 4 items + 4 state attributes if it is time-based, since first the relevant state attributes have to be compared to find the most and least restrictive resources and then the items are compared by their ratio of costs of these two resources. Building a table in Example 3.1 is the fastest way to diminish the gap between wood and metal or between remaining time in workshops C and B.

We test these formalizations of the strategies by matching their choices to the participants' trajectories. Fig. 10 (a) shows the similarity scores for the sets of strategies that participants used in their explanations for each instance. We define similarity as the number of steps in a trajectory in which the participants' choice matches the choice of at least one of the strategies in the set, divided by the length of the trajectory. Importantly, we do not claim that high similarity scores indicate that the matched strategies derived here are unambiguous models of what the participants are doing, but rather a description of their decisions that matches the representation of human explanations, as we derived the strategies from them. The mean similarity between the formal strategies in the sets the participants indicated, and their choices is 0.62 (SD:0.23) for both trials. Apparently, the choices of the participants either do not entirely agree with the formal forms of the strategies derived here, or their explanations do not cover the full strategy set. Fig. 10 (b) shows the similarity of each individual strategy in the set for the two trials for which the participants provided post-hoc explanations. Since from the participants' explanations it is not clear which kind of balancing strategy – *gap reduction* or *balancing* – they use, if they indicated balancing as their strategy, we used both forms in the set to match the trajectory. *Immediate profit*, *cost–benefit* and the two material-oriented strategies match on average 0.37–0.55 proportions of the trajectories individually. The two time-oriented strategies have only a mean similarity score of 0.17. These results show that each formal strategy describes at least some portion of the trajectory in these instances.

Although the agreement of the statements with the formal strategies for the two trials after which they were elicited is not perfect, the formalized strategies can provide a good description of individual solution steps if they match a high proportion of choices because they are derived from participants' explanations. Fig. 10 (c) shows the similarity values for the set of all formal strategies for each participant aggregated over the 11 trials. The mean similarity score over all trajectories is 0.911. Fig. 10 (d) shows the distribution of complexity of the strategy sets for all trajectories that were matched completely (similarity score = 1). If multiple strategies are combined in the set, their complexity scores are summed up. Although the complexity of the formal strategies is generally higher than the complexity of the cost–benefit comparison statements shown in Fig. 8, the mean complexity of 23, explanations with the same complexity were also observed in the explanation from the participants.

4.3. Conclusion

We obtained written explanations from participants in a sequential decision-making task derived from the Furniture Factory. Strategies mentioned in these explanations were similar to heuristics used in participants' step-wise reasoning in the previous elicitation study (Section 3). We formalized general forms of these strategies and confirmed their validity on the choices that participants made in this study. Since the strategies could match a high proportion of participants' choices and contain heuristics that the participants mentioned in their explanations, these strategies can be used as general descriptions for problem-solving trajectories for constrained optimization problems.

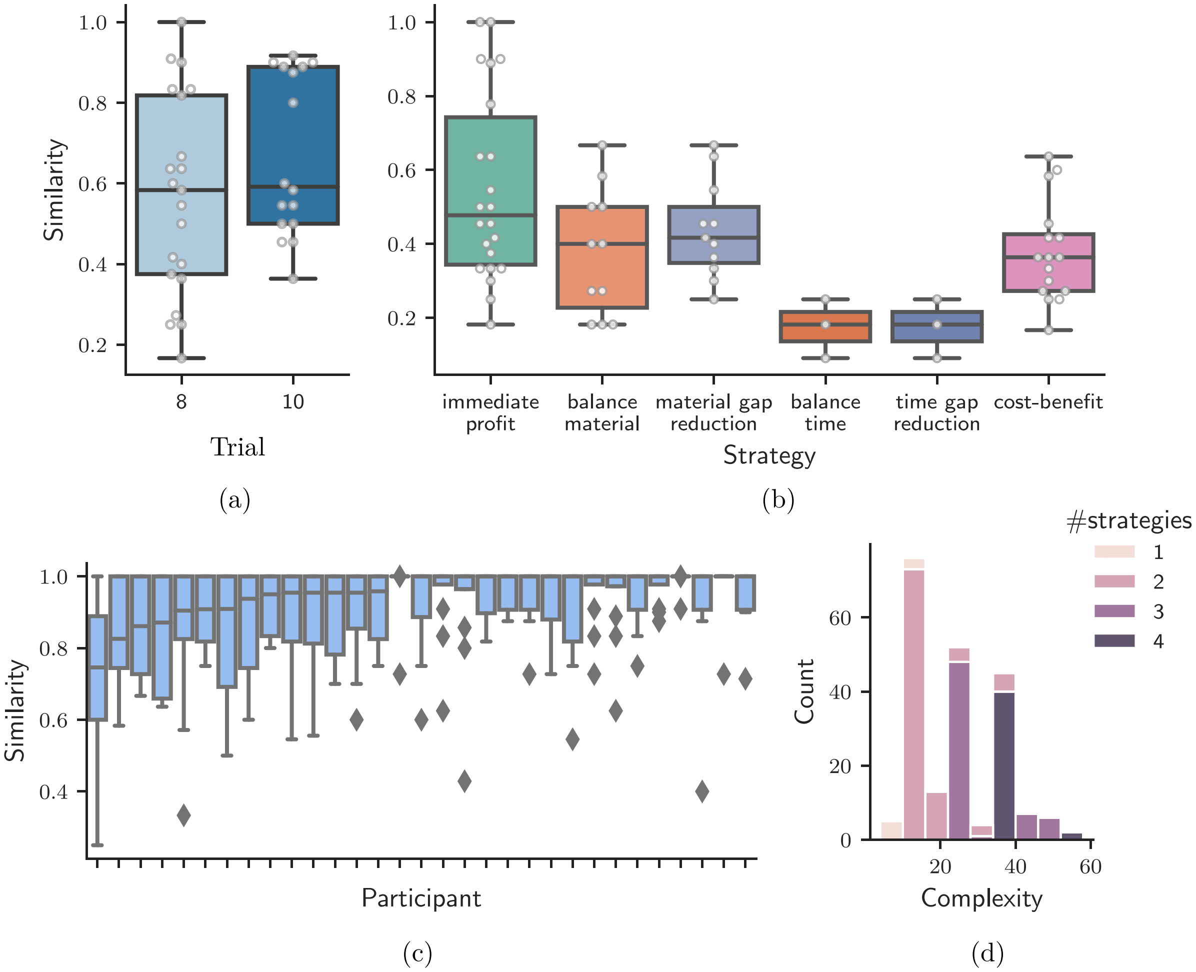


Fig. 10. Similarity scores and complexity of strategies matched to participant trajectories. (a) Similarity between strategies indicated in elicited explanations and matched combinations of formalized strategies for the trials before the prompt. (b) Similarity score for each formalized strategy based on the proportion of each participant's trajectory they match individually. (c) Similarity between combinations of formalized strategies and the trajectories, grouped by participant and ordered by median similarity. (d) Frequency of complexity values for the combinations of strategies that matched trajectories fully (similarity of 1). The different numbers of strategies represented in the matched combination is distinguished by color.

5. Matching representations and heuristics to behavioral data

To investigate how well our formalized strategies match participants' actions and to analyze when certain strategies are used, we carried out a third experiment. We used the online platform (Prolific, 2014) (versions of the years 2021–2022), which offers a diverse participant pool across, among others, many nationalities and educational backgrounds. In the sequential decision-making paradigm, the space of possible actions is vast and after a few steps, different participants encounter different states of the problem. However, with a larger behavioral data-set, we can analyze quantitatively how often participants take the same decisions and how often different strategies agree. The aim of this last study is to confirm the strategies that we found in the qualitative analyses of the previous two studies on a larger and more diverse sample of participants and provide more quantitative details on their use. We also improved the experimental design.

5.1. Method

5.1.1. Experimental setup

The study was conducted using the game that was already used in the validation study discussed in Section 4, but with slight adaptations. We translated the questionnaire and the game to English in order to reach more potential participants online. We decided to shorten the overall time of the task by giving each participant only six problem instances to solve. We also wanted to test for learning effects by varying the order of the instances. The set of possible instances consisted of the same 12 instances used in Section 4. To have all of them occur in an equal number, we had four different groups, each consisting of 6 instances such that each instance occurred in two groups. The groups were balanced for various criteria. Each group had at least one instance with two possible optimal solutions. Another aspect was the optimal strategy for solving a problem, i.e., the groups should have about the same number of instances solvable with an immediate profit or cost–benefit strategy. Instances that cannot be solved optimally by only one of these strategies were distributed as well. To avoid that participants always build the same furniture items, group composition was tested for the similarity of the optimal solutions in each group. To make the analysis of trajectories simpler, parts which previously had to be built multiple times for one furniture item (e.g., four legs for a table), were joined into a bundle, such that only one click was needed to build them. We also included medals into the feedback to show participants whether they had found the optimal solution (gold), a solution within 5% (silver) or 10% (bronze) of the optimal value.

Participants were given a link leading them from the (Prolific, 2014) website to a Sosci-Survey Server (Leiner, 2019). After completing the tutorial, participants had to complete six trials of the game. Participants were asked to rate their solution in the last four trials. All the

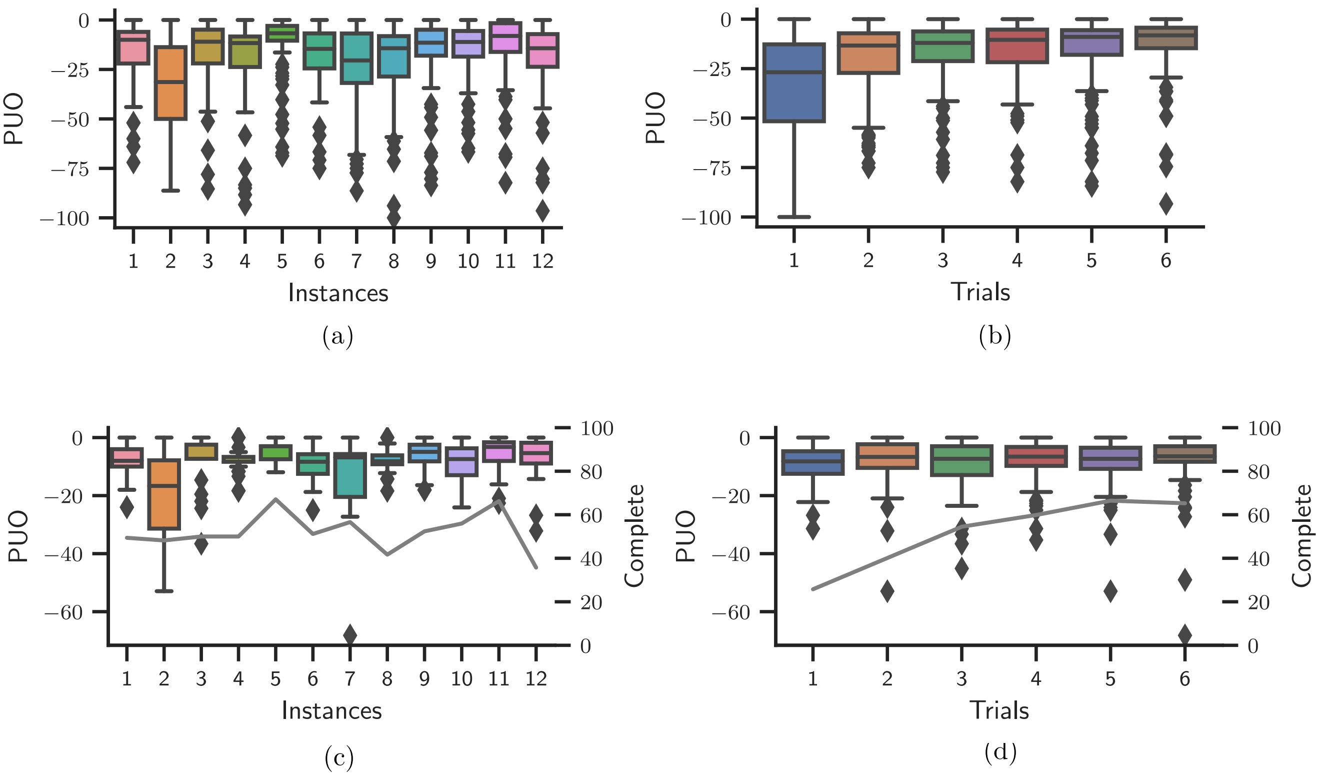


Fig. 11. Percentage Under Optimal scores for the validation dataset. (a) Scores grouped by instances. (b) Scores grouped by trials. (c) Scores grouped by instances for only complete solutions. (d) Scores grouped by trials for only the complete solutions. The gray lines in (c) and (d) show the percentage of complete solutions.

other settings were analogous to the previous study. During the whole experiment, all actions within the game were logged.

5.1.2. *Participants*

207 Participants were recruited among (Prolific, 2014) users to participate in an online experiment run via the online platform. The only selection criterion were fluent English skills. They were reimbursed for their time and could also gain bonus payments for each optimal or good (PUO ≥ 90%) solution. All participants gave informed consent and were informed that they could withdraw at any moment during the study and still be reimbursed for their time. The local ethics committee gave approval for this study.

From this original set of participants, 40 were excluded because they did not provide any complete solution, an indicator that they did not grasp the task, resulting in a dataset of 167 participants. In this dataset, 110 participants indicated that they are male and 57 indicated that they are female. The participants' age ranged from 18 to over 60 years, with 52 participants being between the age 18 and 24, 42 participants between 25 and 29, 43 participants in the ranges 30 to 39, 20 in the range of 40 to 49, 9 in the range 50 to 59 and one participant over 60.

Participants stated their gaming experience on a discrete scale, from 'I don't play computer games' to 'I play computer games regularly (1x week)'. 92 participants stated that they play games regularly at least once per week, 43 participants 2–3 times a month, 19 participants that they play at irregular intervals, maximally once per month. The other 13 participants stated that they do not play computer games in their spare time. After completing the game, participants were asked if they had any further comments, and many stated spontaneously that they enjoyed the task and that they found the challenge engaging (56 of 167).

5.2. *Results*

5.2.1. *Performance*

The performance-under-optimal scores across trials and instances are shown in Fig. 11. The mean PUO for the 6 trials was −18.72 (SD: 19.13). To investigate further if participants were able to handle the interface and formed strategies about their moves, the completeness of the solutions was analyzed. If leftover resources and spare parts at the end of the trial could have been used to construct more furniture items, the solution was classified as incomplete. On average, with exclusion of the training trial, 52.0% of the solutions were complete. If only complete solutions are considered, the mean PUO was −8.51 (SD: 8.34). Fig. 11 (c) shows that some instances seem more difficult than others. The learning effects in the performance-under-optimal score (Fig. 11 (b)) vanish when only considering complete solutions (Fig. 11 (d)).

5.2.2. *Strategy similarity*

To investigate how well the strategies that we derived from the qualitative data of the previous studies (see Box 10) can match the decisions in a bigger dataset, we investigate the similarity of the formal strategies on the trajectories obtained here. Trajectories with less than five steps are excluded for this analysis because the similarity score for them is not meaningful as a few steps can be easily matched with a high score but do not capture systematic patterns within their trajectories. This reduces the set of considered trajectories from 1002 to 884 trajectories. The mean similarity score of 0.87 (SD:0.18) for all the considered trajectories is slightly worse than on the smaller dataset from the previous study (see Section 4, Fig. 10). Fig. 12 (a) shows the aggregated similarity scores by instance. The mean similarity values for participants show a high variability, with the maximum mean similarity score for one participant being 1 and the minimum mean score being 0.5 (SD: 0.1). Between instances, we also observe a difference in how well the strategies can match the trajectories on average. Instances 2, 6, 8, 10, 12 have a mean similarity over 0.9, the highest being instance 2 with 0.96. Instances 3 and 4 however are matched the worst with 0.74 and 0.77 respectively. The similarity scores show

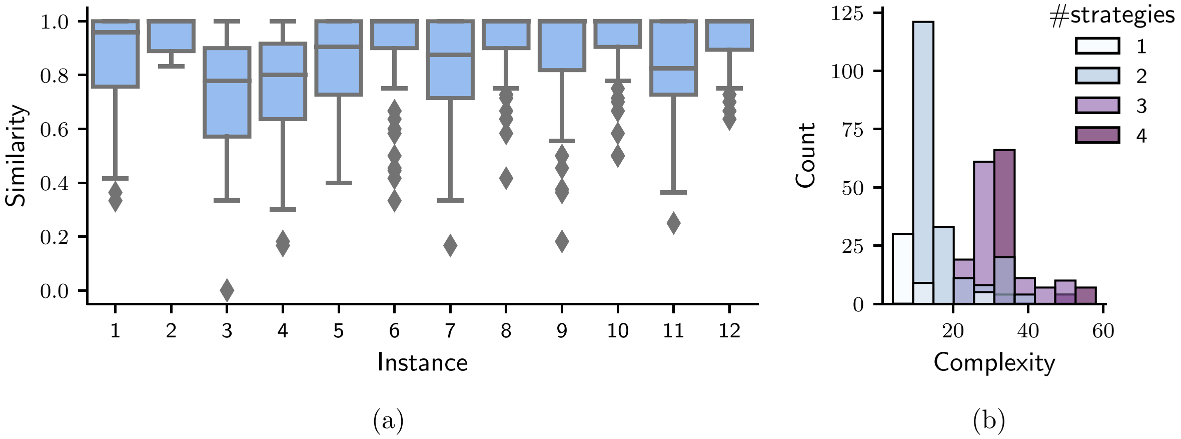


Fig. 12. Similarity scores and complexity of strategies matched to participant trajectories from the prolific dataset. (a) Similarity values for combinations of strategies for participant trajectories grouped by instance. Only trajectories longer than 5 steps are considered. (b) Frequency of complexity values for the combinations of strategies that matched trajectories fully (similarity of 1).

that the formalized strategies are consistent with many of the choices participants made. Matching strategies represent a good description for the choice, although we do not claim that they accurately model the thought process of the participant in each step. Fig. 12 (b) shows the distribution of complexity values for the best matching strategy combinations. Although many trajectories can be described perfectly with strategy combinations that have a complexity score within the range of the complexity of the elicited post-hoc explanations (see Section 4), some trajectories can only be matched with more than three strategies, resulting in a high complexity.

5.2.3. Decision point analysis

Decision paths in instances of the Furniture Factory can be represented as directed graphs $G = (V, E)$. The set of vertices in the graph V corresponds to all states (i.e. a specific set of resources and full furniture items) that do not violate the instance's constraints. Edges between vertices correspond to transitions between states via one building action (e.g., building a chair). An example graph for one problem instance based on the data from the validation study (Section 4) and the dataset obtained with prolific can be seen in Fig. 13 (a). The left-most vertex is the starting state, where all the resources in this trial are still available. The graph representation allows the comparison of the size of the decision space between instances, and to map the trajectories of the participants onto graphs and analyze the variability in their decisions. We disregard the decomposition into parts for this analysis.

To analyze in which states participants might prefer one strategy over another, we look at vertices in the graph where their decisions cluster. The intuition behind this analysis is that if we observe patterns in collective decision-making, it is important that components in explanations can describe them; hence the formal strategies should match those descriptions. We approximate a level of agreement for each vertex based on the choice that most participants in the state took. The more likely it is that this choice was taken randomly by the number of participants that took it, the lower the level of agreement.

For this analysis, we combine the trajectories from the previous validation study (Section 4) and the trajectories from the dataset obtained here. Each vertex that we consider corresponds to one state in an instance and has, therefore, a specific set of resources and profit information for the items. The resulting dataset comprises 7299 vertices with their most frequently taken decision.

We calculate an approximation for how surprising the most frequently taken decision in a specific state is, given by the number of participants that took this decision, the overall number of participants that had to make a choice and the available choices in this state. We call this measure the *level of agreement*; for formal details of the determination of this proxy, see Appendix A. It can take values between zero and one, where zero corresponds to no surprise, for example, in states where participants did not agree at all, where only one participant had to decide, or where only one choice was available. The higher the level of agreement, the more surprising the most prominent choice of participants. The most prominent choices with a high level of agreement are of particular interest since their strategy matches are more expressive because they are likelier to be not random.

Fig. 13 (a) gives the trajectory graph for instance 2. Vertices with a level of agreement bigger than 0 are colored and scaled by their value. As illustrated by the graph, a large proportion of vertices visited by participants have a level of agreement of 0 (gray lines) between the participants. Most of these vertices with low level of agreement were part of only one trajectory or are vertices in which only one choice is available (due to resource constraints). These are excluded from further analysis. The distribution of the level of agreement for the remaining 1594 vertices (21% of the original dataset) is shown in Fig. 13 (b). 513 vertices in the set have a value of 0. The remaining 1081 vertices (14.81% of the original dataset) comprise the dataset on which we base our following analysis on, since they might provide information on when participants collectively prefer a specific choice.

Figs. 13 (c) and 13 (d) shows the number of states with high participant agreement in each instance and for each step. The states are similarly distributed over each instance, however are more likely to be in the earlier steps of the trajectories. This is in accordance with our expectation, since it is more likely that participants encounter similar states at the beginning of an instance than later.

Fig. 13 (e) shows the distribution over the number of strategies which match the choices in states with high agreement between the participants in the dataset. 18.7% of the choices are not matched by any strategy. However, 57.4% of the decisions are matched by a small number (one or two) of the strategies. For states in which many strategies (three or more) predict the same choice the high level of agreement could be interpreted as a result of multiple ways to arrive at the same decision, therefore the similarity scores for the strategies in these states are not highly indicative of the kind of decisions which are matched.

To investigate which strategies are most valuable as descriptors for decisions in which participants agree, we analyze the occurrence of the strategy combinations. Fig. 14 shows the number of decisions each strategy and each (co-occurring) strategy set match in the dataset. *Balance material*, *material gap reduction* and *immediate profit* match most often, whereas the two time-based strategies match only very few decisions individually. The most frequently co-occurring sets of strategies are *immediate profit* together with *cost–benefit*, as well as the

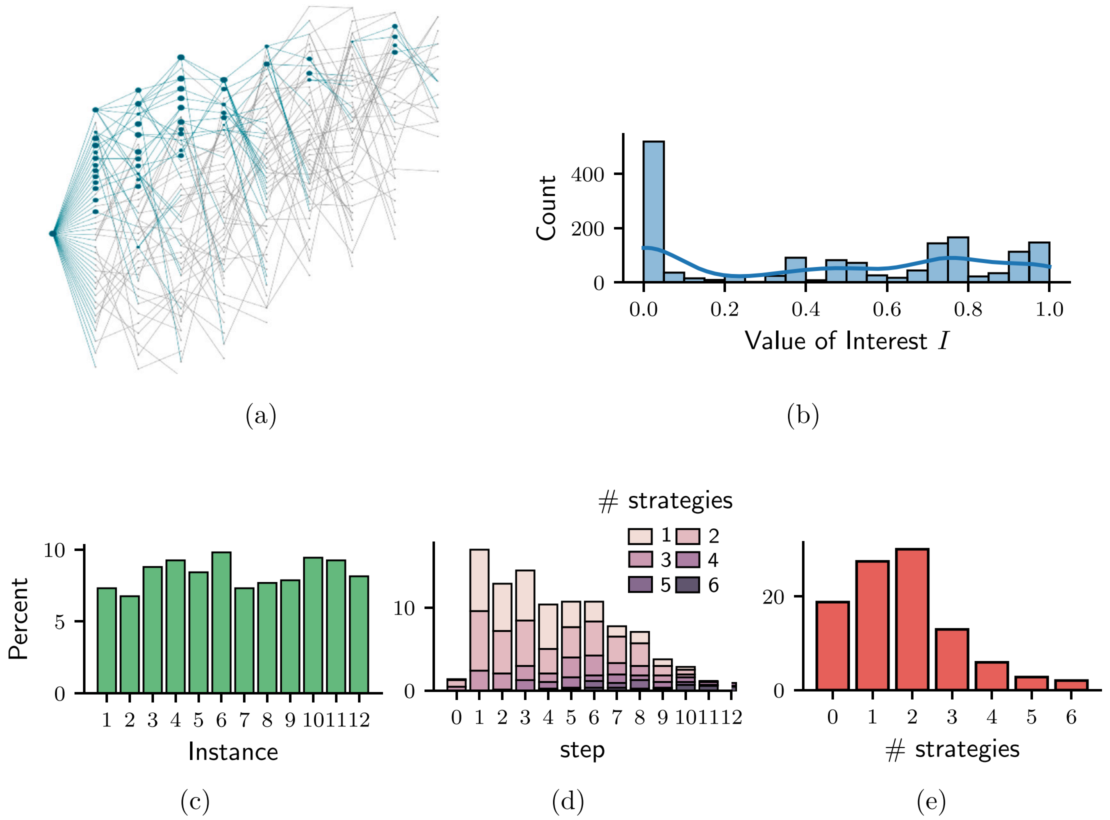


Fig. 13. Description of the decision points of interest. (a) Graph representation of the participants' trajectories, for instance 2. Vertices with an agreement higher than zero are colored blue and scaled by their level of agreement. (b) Frequency of different values of level of agreement for all non-trivial vertices. (c) Percentages of instance occurrence in the decision point set. (d) Percentages of step occurrences in the decision point set, colored by number of strategies in the combination sets matching the points. (e) Percentages of matches per number of strategies in sets matching the decision points.

two material-based strategies. In the set of decision points which could be matched only to one strategy, a difference in temporal prevalence can be observed (see Appendix B). While *balance material* occurred more at the beginning of a problem-solving trajectory, *material gap reduction* was used more towards the end, and *immediate profit* was used throughout.

5.3. Conclusion

Improving the study design and collecting data from a more extensive and diverse set of participants allowed us to gain a deeper understanding through more quantitative analyses. We validated the formal strategies derived from the explanations in the two previous studies by analyzing how well the strategies can match the observed decisions on the complete dataset. Especially in states in which participants collectively show a strong preference for one specific decision, the good matching scores of the formal strategies show that they are promising components for automatically generated explanations. We described the most prominent strategies matching these decisions: *Balance material*, *material gap reduction*, and *immediate profit* and showed how the prevalence of each strategy changed over time for points which can be matched to one singular strategy.

6. Discussion

We conducted three behavioral studies based on the same constrained optimization problem, the Furniture Factory (Section 2), to analyze how humans solve complex problems and how they explain their solutions. This new paradigm, which captures many aspects of general constrained optimization problems, was implemented in two tasks and posed as computer games, as it is often done in research on complex problem solving and dynamic decision-making in microworlds (Gonzalez et al., 2005). Verbal reports were collected concurrently in a study using an exploration task (Section 3). This was done to elicit the different heuristic strategies that participants mention in their explanations. In a second validation study (Section 4), where we used a sequential decision-making setting, written post-hoc explanations from participants were analyzed. The components of the concurrent explanations that were observed in the exploration setting in the first elicitation study (Section 3) were also observed in the post-hoc explanations in the second validation study and combined to more complex descriptions of strategies. Based on the insights of these two studies, we formalized general versions of these strategies and were able to verify that they matched the observed behavior. To investigate how well these strategies can describe solution trajectories in general, we conducted and analyzed a third and larger study with more diverse participants (Section 5) and were able to confirm that the formal strategies can indeed match the behavior of the participants.

In this paper, we studied human explanations for solving complex problems. From these explanations, we can derive insights about the complexity of human explanations and their components. More concretely, we have observed that humans use reduced representations of the problems for their decisions, i.e., they mostly focus on subsets of the available and relevant information. However, they do not limit their overall approach to the application of only one strategy, and instead use sets of different strategies to find solutions to constrained optimization problems. This is also reflected in the complexity scores associated with their explanations. Although for the simplest heuristics the complexity score is limited (most of the complexity scores lie between 1 and 10), since participants also reduced their decision set for individual decisions, the complexity score for the full explanations can also reach

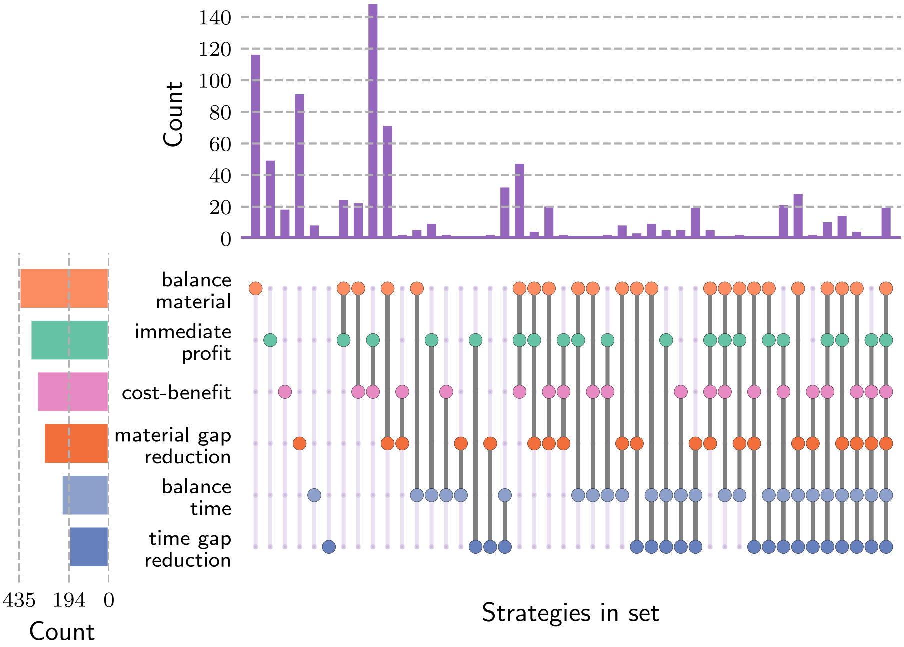


Fig. 14. Occurrences for the different sets of strategies. The horizontal barplot on the left indicates how often a strategy matched decisions in the overall set, the counts on the upper axis show the matches for each set, symbolized by the colored points below.

higher values (up to 22). Even though more complex sets of formalized strategies might be needed to match optimal solutions or solutions found by participants, the complexity values show that humans can consider complex sets of strategy descriptions in a full explanation. These insights are very valuable for explainable AI: Automatically generated explanations for constrained optimization problems should probably follow the same principles as human explanations for these problems.

It is essential for human-centered explainable AI methods that the recipient can understand the features used in the explanations. However, for many explainable AI systems, it is unclear if the features selected by the researcher or engineer match the representations the explainee has of the system. One step that many advocate for is to validate the features used in explainable AI methods with user studies (Hoffman et al., 2023; Rong et al., 2023; Wells & Bednarz, 2021). In contrast, for tasks where it is unclear from the start what these features could be and how the users represent the tasks at hand, like constrained optimization problems, insights from human studies have to inform the development of explainable AI methods (Miller, 2019). For example, Ehsan et al. (2019) propose a process to elicit natural language rationales from users while they perform a task in order to train a system that automatically produces rationales that match the language and the features used by humans. Though our elicitation process is similar, we obtain human strategies that can be used as building blocks for a variety of explanation generation methods. Furthermore, the strategies presented here – *immediate profit*, *balancing*, *gap reduction* and *cost–benefit* – are all applicable to constrained optimization problems in general, not just the Furniture Factory.

As LPs are constrained optimization problems, the strategies formalized here can be used to generate cognitively adequate explanations for LPs. And since the strategies formalized here match the set of human decisions so well, and we have observed that many solutions from participants for our problem instances are (close to) optimal, they are a promising starting point to be used directly in automated explanations. The complexity of the computations for the individual building blocks increases with problems of more complex structure based on a higher number of constraints and more actions. However, illustrating their effect on simple examples can provide a basic intuition. Therefore, although they might be more complex, their abstraction remains similarly expressive. We do not expect these building blocks to be an exhaustive set of explanations for problems of every structure or complexity, but they can be combined with other approaches to form explanations. An example of such a combination is described in a companion paper (Ott & Jäkel, 2024). The tool proposed there uses the strategies that were formalized here as building blocks for explanations. The tool uses various methods to simplify and structure LPs. It also generates step-wise explanations of the optimal solution using these and similar strategies as reasons. The strategies were generalized to capture resource allocation problems similar to the ones shown here, as well as requirement satisfaction problems with a focus on minimizing cost while satisfying given requirements. The resulting explanations will be tested for quality in future work.

Additionally, the strategies that we formalized here can be combined with other explainable AI methods. Since the strategy descriptions can be matched without access to the solver, they can be a basis for solver-agnostic post-hoc solution explanation algorithms. As the strategies represent the relation of different solution parts to the structure of the problem, they could also be applied to explain solutions of black-box solvers. Although descriptions derived from the strategies map to individual parts in a solution, they can be used to enrich algorithms for global explanations similar to policy summarization approaches in reinforcement learning, where summaries of algorithm behavior are enriched with information about features used for individual decisions (Huber et al., 2021). In future works, we will leverage this intuition and investigate the formalized strategies as a foundation for post-hoc solution descriptions. By decomposing the formal strategies into individual logical attributes, we plan to search the space of logical descriptions of optimal trajectories built out of combinations of these logical attributes. The iterative search allows searching for the description with minimal necessary complexity.

The current study also raises a number of new questions, particularly about participants' exploration behavior and their evaluation of solutions (Ott et al., 2024). Answering this question could also provide more insights into which aspects of optimal solutions are harder or easier to follow, or why a solution is recognized as optimal or not. This is especially important when explanations are used to build trust. One important problem remains: How do participants decide when to use which strategy? Due to the complexity of the problems, features that trigger strategies are hard to identify. Future directions of research should also investigate when people use more simple or more complex strategies, and what features of a state might trigger the use of specific strategies. Since the focus of this study was to find generalized descriptions of behavior which match participants' descriptions, the data obtained here is not sufficient to completely model the reasoning process of participants. A better understanding of these triggering attributes could further improve explanations constructed using different strategies as components.

Interestingly, some strategies identified in this study were also found, in similar form, in other studies investigating how people solve constrained optimization problems. Murawski and Bossaerts (2016) compare the behavior of participants in their study of the knapsack problem with a greedy strategy, analogous to the *cost–benefit* strategy discussed here. Some of the arithmetic strategies discussed for the vehicle routing problems in Kefalidou and Ormerod (2014) are also similar. However, not all constrained optimization problems are best described by the strategies formalized here. Optimization problems with a spatial representation, like the vehicle routing problem or the classic problem of the traveling salesperson, are often represented spatially by humans in their descriptions and are therefore not captured well by the arithmetic strategies used here (Dry et al., 2012). Also, the set of building blocks for explanations might have to be adapted for problems with very different dynamics. For example, for problems involving uncertainty, the set of explanatory building blocks has to be augmented to capture the reasoning of trading the risks of an action with its possible benefits.

It is important to point out that – contrary to what is usually done in problem-solving research – we explicitly asked participants to explain their problem-solving strategies (cf. Jäkel & Schreiber, 2013). When we translated these explanations into formal strategies, these strategies matched a fair proportion of the decisions that we could observe in the behavioral data. But in some cases these matches were far from perfect, and we did not collect data that would allow us to match explanations and behavior on a trial-by-trial, or even decision-by-decision basis. Hence, we do not want to make any strong claims about what people really do when they solve constrained optimization problems. For explainable AI, it may not even be relevant to understand precisely how people solve a problem. It is probably more relevant to understand how people explain their solutions and strategies, even if their explanations do not necessarily match up perfectly with their behavior. It is still interesting though that people's explanations and their behavior are relatively consistent with each other.

Also, dependent on the context and the explainee, it can be necessary to provide other types of explanations, for example, causal or solver-dependent explanations, which illustrate the inner workings of the solution algorithms. Experts in constrained optimization certainly require a different explanatory level than laypeople. However, this study provides valuable insights into which representation humans have of strategies of constrained optimization, which we can leverage for explainable AI.

Taken together, we showed how behavioral studies can be used to examine human explanations in complex problem solving. We see this as a necessary step towards incorporating human thought processes in the automatic generation of explanations by AI systems. In general, such a cognitive approach to explainable AI will be crucial if the goal is to provide cognitively adequate explanations. As explained in the introduction, this is hard to achieve because the representations that AI algorithms use to solve problems are usually very different from the human representations of the same problems. Also, we want AI systems to perform all the computations that are too hard for humans to follow in order to achieve super-human performance. Hence, the best we can hope for is to make the solutions that an AI system computes plausible for humans. Furthermore, especially in more complex tasks, which are either not typically solved by humans or only solved using intuition, it is difficult to elicit good explanations. Our combination of eliciting explanations, validating strategies and matching behavioral data proved very fruitful in this regard. In order to generate an explanation for the solution of an algorithm, we need to consider human reasoning and leverage insights from human explanations for the problems in question. Cognitively adequate explainable AI needs to leverage human explanations. Here, we have made first steps in this direction for the important class of constrained optimization problems.

MIP Mixed Integer Linear Program

LP Linear Program

PUO Percentage Under Optimal

CRediT authorship contribution statement

Inga Ibs: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Claire Ott:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Frank Jäkel:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Constantin A. Rothkopf:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

This work was funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) in the project 'PlexPlain: Erklärende KI für komplexe Lineare Programme am Beispiel intelligenter Energiesysteme' [grant number 01IS19081] and by the Hessian Ministry of Higher Education, Research, Science and the Arts (Hessisches Ministerium für Wissenschaft und Kunst) and its LOEWE research priority program 'WhiteBox' [grant number LOEWE/ 2/13/519/03/06.001(0010)/77].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank Eric Frodl for programming the initial experiments in the form of a game and Jan Mackensen for adapting the experiment to the specific studies. Also, we would like to thank Sabine Dotterer for conducting and transcribing the concurrent explanation study and Jaron Schäfer and her for their help with labeling the data. We would also like to thank the anonymous reviewers for their constructive criticism.

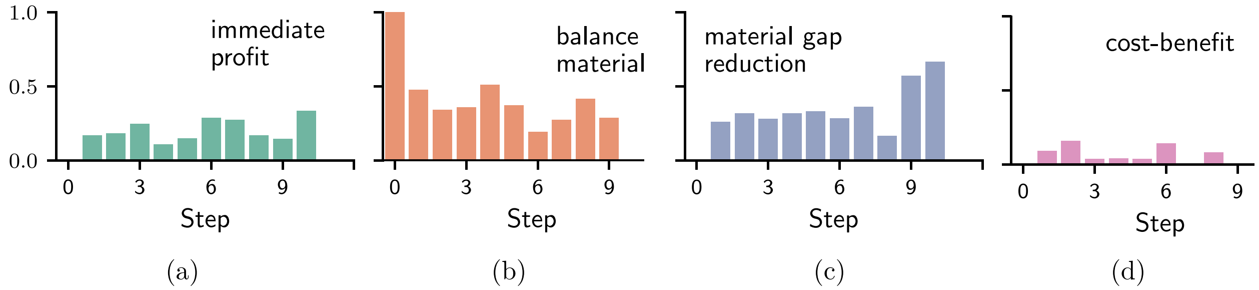


Fig. 15. Matches for each strategy in each step. The counts are normalized by the number of vertices in each step for which the most prominent choice was matched by only one strategy.

Appendix A. Decision point identification

We calculate a proxy for how surprising the most frequently taken decision is given the available choices, the number of participants n in the state and the number of participants n_c that followed the choice c. The maximum of the n_c determines the most frequently taken choice $c_{\max}$ that was taken $n_{c_{\max}}$ times. We approximate the theoretical distribution for the observed value $n_{c_{\max}}$ under the hypothesis that choices are made uniformly at random by n participants and repeatedly sampling from this process 1000 times. We call the resulting random variable $s_{\max}$. From this approximated distribution, we derive the probability that the maximum cluster size is at least as big as the observed value $n_{c_{\max}}$. This probability is called $P_v(s_{\max} \geq n_{c_{\max}})$ under the hypothesis that all participants decide uniformly at random. The level of agreement we assign to each vertex in the set of interest is then $1 - P_v(s_{max} \geq n_{c_{\max}})$.

Appendix B. Prevalence of matched strategies

Fig. 15 shows the frequency of matching decisions for each strategy for each step, normalized by the overall number of decision points in the set of points where only one strategy matched. In the set of states where only one strategy matches one decision, *balance material* is the most prominent match for choices in the early steps in the trajectory. The occurrence of the choices that match only *immediate profit* fluctuates, same as *cost–benefit*. However, choices agreeing only with *material gap reduction* occur more in later steps of the trajectories.

Data availability

Data will be made available on request.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

Barbosa, J., Ripp, C., & Steinke, F. (2021). Accessible modeling of the German energy transition: An open, compact, and validated model. *Energies*, *14*, 8084. http://dx.doi.org/10.3390/en14238084.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., & Rellinger, E. R. (1995). Metacognition and problem solving: A process-oriented approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 205–223.

Chi, M. T., De Leeuw, N., Chiu, M. H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477. http://dx.doi.org/10.1207/s15516709cog1803_3.

Chvatal, V. (1983). Linear programming. In *Series of books in the mathematical sciences*. W. H. Freeman.

Curtain, C. (2021). Qualcoder 2.5 [computer software]. https://github.com/ccbogel/QualCoder.

Dantzig, G. B., & Thapa, M. N. (1997). Linear programming. In *Springer series in operations research and financial engineering*. New York: Springer-Verlag, http://dx.doi.org/10.1007/b97672.

Dry, M., Preiss, K., & Wagemans, J. (2012). Clustering, randomness, and regularity: Spatial distributions and human performance on the traveling salesperson problem and minimum spanning tree problem. *The Journal of Problem Solving*, *4*, http://dx.doi.org/10.7771/1932-6246.1117.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 263–274).

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*, 316–344.

Frodl, E. (2021). *The furniture company: building games to measure human performance in optimization problems* (Bachelor thesis).

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*, *21*, 273–286.

Greenberg, H. (1983). A functional description of ANALYZE: A computer-assisted analysis system for linear programming models. *Association for Computing Machinery. Transactions on Mathematical Software*, *9*, 18–56. http://dx.doi.org/10.1145/356022.356024.

Greenberg, H. J. (1993). How to analyze the results of linear programs—part 1: Preliminaries. *Interfaces*, *23*, 56–67.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, *5*, Article 1096257.

Huber, T., Weitz, K., André, E., & Amir, O. (2021). Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, *301*, Article 103571.

Jäkel, F., & Schreiber, C. (2013). Introspection in problem solving. *Journal of Problem Solving*, *6*, 20–33.

Kefalidou, G., & Ormerod, T. C. (2014). The fast and the not-so-frugal: Human heuristics for optimization problem solving. *Cognitive Science*, *36*.

Leiner, D. (2019). Sosci survey. https://www.soscisurvey.de.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.

Murawski, C., & Bossaerts, P. (2016). How humans solve complex problems: The case of the knapsack problem. *Scientific Reports*, *6*, 34851. http://dx.doi.org/10.1038/srep34851.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ott, C., Ibs, I., Jäkel, F., & Rothkopf, C. (2024). Furniture Factory: Human exploration strategies in optimization problem solving (Manuscript in preparation).

Ott, C., & Jäkel, F. (2024). Simplifex: Simplifying and explaining linear programs. *Cognitive Systems Research 88*, 101298. http://dx.doi.org/10.1016/j.cogsys.2024.101298.

Pisinger, D. (2005). Where are the hard knapsack problems? *Computers & Operations Research*, *32*, 2271–2284. http://dx.doi.org/10.1016/j.cor.2004.03.002.

Prolific (2014). https://www.prolific.co.

Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2023). Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

Wells, L., & Bednarz, T. (2021). Explainable AI and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in Artificial Intelligence*, *4*, Article 550030.