COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 49 (2025) e70135

ISSN: 1551-6709 online
DOI: 10.1111/cogs.70135

Humans Select Subgoals That Balance Immediate and Future Cognitive Costs During Physical Assembly

Felix J. Binder,[a] Marcelo G. Mattar,[b] David J.Kirsh,[a] Judith E. Fan[c]

[a]*Department of Cognitive Science, University of California San Diego*
[b]*Department of Psychology, New York University*
[c]*Department of Psychology, Stanford University*

Received 3 March 2025; received in revised form 29 September 2025; accepted 1 October 2025

Abstract

From building a new piece of furniture to replacing a lightbulb, people must often figure out how to assemble an object from its parts. Although these physical assembly problems take on many different forms, they also pose common challenges. Chief among these is the question of how to break a complex problem down into subproblems that are easier to solve. What principles determine why some strategies for decomposing a problem are favored over others? Here, we investigate the decisions that people make when considering different *visual subgoals* in the context of attempting to build a series of virtual block towers. We hypothesized that people favor subgoals achieving a balance between how much progress the subgoals would help achieve toward the final goal and how effortful they would be to solve. We tested this hypothesis by defining several computational models of planning and subgoal selection, then evaluating how well these models predicted human planning and subgoal selection behavior on the same problems. Our results suggest that participants rapidly differentiated the computational costs of otherwise similarly ambitious subgoals, and used these judgments to drive subgoal selection. Moreover, our findings are consistent with the possibility that participants were not only sensitive to the immediate computational costs associated with solving the very next subgoal, but also future costs that might be incurred when attempting the rest of the problem. Taken together, these results contribute to our

Correspondence should be sent to Judith E. Fan, Department of Psychology, Stanford University, 450 Jane Stanford Way, Building 420, Stanford, CA 94305, USA. E-mail: jefan@stanford.edu

understanding of how humans make efficient use of cognitive resources to solve complex, grounded planning problems.

Keywords: Planning; Problem solving; Physical reasoning; Task decomposition; Decision-making

1. Introduction

Building new things is a fundamental part of what it means to be human, from smaller artifacts including tools and furniture, to larger structures including houses, cities, and space stations (Hunt et al., 2021). Even in everyday contexts where an individual might be working alone to build something, such as a new bookshelf, there are substantial challenges to overcome. In particular, they have to reason about the dimensions of the planks and how they fit together (Battaglia et al., 2013; Hamrick et al., 2016; Ludwin-Peery et al., 2021; Smith et al., 2018; Ullman et al., 2017). The more complex the artifact one is seeking to build, the more interactions between parts have to be considered. How do humans manage to routinely meet such challenges in their everyday lives (Callaway et al., 2022; Gershman et al., 2015; Griffiths, 2020)?

One simple way people might solve such assembly problems is by focusing on immediate progress: selecting actions that appear to make the most progress toward the overall goal right away. This approach has been well-documented in both real-world and controlled experimental settings (Brooks, 1991; Geffner, 2013; Mattar & Lengyel, 2022; Simon, 1956). For example, when asked to build a block tower by stacking toy blocks, someone using this approach would place each block wherever it makes the most immediate progress, without considering how their current choices affect later parts of the tower. Focusing on immediate progress can be computationally inexpensive and effective, particularly when problems can be solved incrementally and mistakes can be easily corrected. However, this approach becomes less viable in domains where mistakes are costly or irreversible. Consider building a tower out of blocks: the way the foundation is built determines whether the completed tower will be stable, but these dependencies might not be apparent when first placing blocks at the base.

Alternatively, people might adopt a more deliberative approach, expending cognitive effort to plan ahead. Classically, planning is formulated as search over the space of hypothetical future states, accessed by performing a sequence of actions (Fikes & Nilsson, 1971; Kirsh, 2009; Newell & Simon, 1972). However, the computational cost of planning grows exponentially as a function of the size of that search space, with the consequence that many everyday problems seem intractable when searching over sequences of actions alone (Bellman, 1957; van Opheusden et al., 2017). One influential strategy for make planning on complex problems more tractable is hierarchical planning: decomposing the full problem into a sequence of subgoals that are individually easier to plan for and achieve, thereby reducing the total cost of planning (Correa, Ho, Callaway, Daw, & Griffiths, 2023; Huys et al., 2015; Jinnai et al., 2019; Maisto et al., 2015; Sacerdoti, 1974). For example, when constructing a tower out of blocks, a hierarchical planner might first break the problem down into subgoals that entail "building the base," then "constructing the middle portion," then "completing the top." Only after determining this sequence of subgoals would they proceed to plan the specific block placements needed to achieve each one. As such, one shortcoming of typical implementations of hierarchical planning is that they do not consider the computational cost of generating possible problem decompositions in the first place.

An intermediate possibility that strikes a balance between the benefits of taking immediate action and the costs of planning is one where people set only one or a few near-term subgoals to achieve before considering how to solve the remainder of the problem (Agre, 1988; Botvinick, 2012; Hayes-Roth & Hayes-Roth, 1979; Patalano & Seifert, 1997). In other words, only part of the problem might be *within scope* at any point in time, such that decisions about how to decompose a problem are interleaved with actions to solve parts of it, rather than being made in their entirety before taking the first action. For example, when constructing a tower out of blocks, rather than planning the entire sequence of actions to take upfront, a person might instead first focus on constructing the base, then finish doing so before considering which section of the tower to construct next.

Here, we investigate to what degree these various accounts might explain how people approach physical assembly problems. Specifically, we aim to understand how people decide among different possible ways of decomposing these problems to make them more tractable. Toward this end, we conducted a series of experiments wherein participants were presented with a series of block towers to reconstruct from a fixed inventory of blocks. In this virtual environment, participants were unable to move a block once placed. In real-world physical construction settings, undoing an action is often costly, making it more valuable to plan ahead. For instance, one cannot undrill a hole. Solving such block tower problems poses nontrivial challenges for planning, as has been demonstrated in prior work using this task to explore variation in reasoning about physical assembly in humans (Cortesa et al., 2018; Dietz et al., 2019; McCarthy et al., 2020) and artificial agents (Bapst et al., 2019; Sussman, 1975). To investigate which factors people consider when deciding which part of the block tower problem to consider next, we gave participants multiple possible subgoals to choose from and compared their choices to the predictions of various computational models instantiating different subgoal selection strategies.

2. Measuring and modeling human planning costs

The goal of our first study was to measure how much time people needed to formulate plans when presented with a physical assembly problem, and to validate an influential computational model of those planning costs in this task setting. Toward this end, we used a block tower assembly task wherein participants were presented with a series of block towers, and their goal was to exactly reconstruct them by executing a valid sequence of block placements (McCarthy et al. (2020); Fig. 1a). Similar task paradigms have been used to study planning and physical reasoning in both artificial agents (Bapst et al., 2019; Bear et al., 2021; Sussman, 1975) and humans (Cortesa et al., 2018, 2017; Dietz et al., 2019; McCarthy et al., 2020, 2023; Smithwick & Kirsh, 2015).

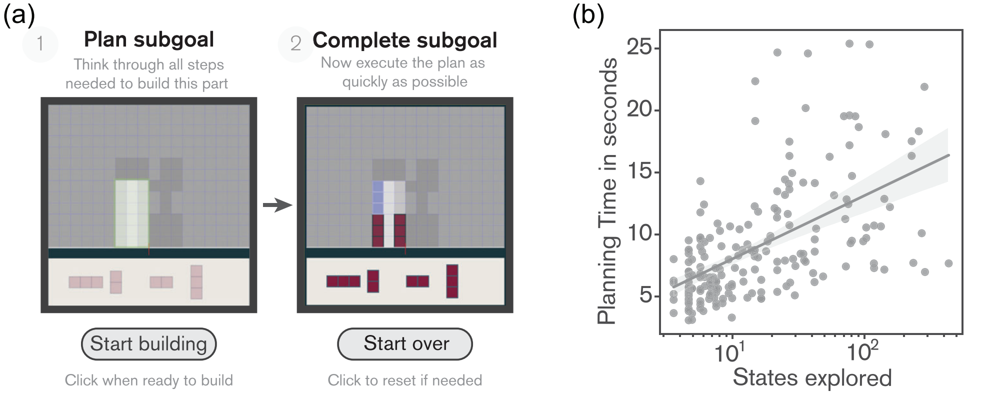

Fig. 1. (a) On each trial, participants were presented with both a block tower and a target subgoal, defined by a rectangular aperture containing part of the tower. In the planning phase of each trial, participants were instructed to take as much time as they needed to come up with a complete plan for solving just that subgoal. Next, in the subgoal completion phase, participants executed the plan under time constraints that made it difficult to perform additional planning. (b) Scatterplot indicating the relationship between the number of states explored by the Best First Search algorithm (x-axis) and the amount of time taken by participants to devise a plan for each subgoal (y-axis). Each dot corresponds to one subgoal embedded within a block tower problem. The predictions of a linear regression model are also shown, with 95% confidence bands.

2.1. Method

2.1.1. Participants

Eighty-six participants (51 male, $M_{\text{Age}} = 38.46$ years) were recruited from Prolific and paid a minimum of $14 per hour. We excluded six participants who navigated away from the study webpage more than five times. In this and all subsequent studies, participants provided informed consent in accordance with the cognizant IRB.

2.1.2. Stimuli

We sought to identify a set of block towers that would enable investigation of how people make plans to complete various subgoals nested within each tower. Toward this end, we procedurally generated 128 stable block tower shapes of varying sizes. From these towers, we initially defined a large set of "initial" subgoals—that is, subgoals that could be built without having to build any other part of the tower first. For each tower, we then generated four pairs of subgoals, yielding a total of 512 pairs of subgoals, where both members of each pair were equated on size, measured by their surface area. Matching the size of subgoals ensured that the choice between subgoals in a pair would be due to differences in how costly it might be to formulate a plan, rather than how much progress toward completing the block tower it would yield.

We used a search algorithm known as Best First Search to estimate how costly each subgoal might be to develop a plan for, as measured in terms of the number of states considered by the search algorithm before a solution is found. We refer to the subgoal with the lower predicted cost as the *easier* one in each pair, and the other one as the *harder* subgoal. To analyze behavior across different subgoal sizes, we divided the subgoal pairs into three size categories: small, medium, and large. Within each size category, we identified the 32 subgoal pairs with the largest difference in planning cost between the easier and harder subgoal. This selection process yielded our final stimulus set of 96 subgoal pairs (192 total subgoals), spanning different sizes and difficulty levels.

2.1.3. *Task procedure*

At the beginning of each session, participants were familiarized with the block tower assembly task by completing a tutorial phase in which they first built a small practice tower without the use of subgoals, then again, but with a predetermined sequence of subgoals. After completing the tutorial, each participant was then presented with a series of 24 subgoals to complete, corresponding to 12 subgoal pairs sampled at random from the full set of 96 subgoal pairs. These 24 subgoals were presented in a randomized sequence, and participants were not told that they would be presented multiple subgoals related to the same block tower problem.

On each trial, participants were presented with a target tower shown as a silhouette, with a specific subgoal region highlighted. Their goal was to solve the presented subgoal by placing blocks to exactly fill the highlighted region of the tower, with no blocks sticking out and no gaps left. The building area was divided into grid cells, with each block taking up one or more cells. Below the building area, participants were presented with an inventory of rectangular blocks that they could use to build the tower. The inventory contained a 1x2 block, a 2x1 block, a 1x3 block, and a 3x1 block (Fig. 1). Each type of block could be used as often as needed. Each block shape could be picked up by clicking on it, then placed within the building area by clicking again. Blocks could only be placed on top of other blocks, or on the ground if there were no other blocks, but never below blocks that had already been placed in the building area.

Visual subgoals were defined as rectangular regions of the building area of any dimensions. When potential visual subgoals were shown to participants, they were displayed as a translucent overlay over the target tower, with the rest of the tower still visible. When solving a subgoal, participants were tasked with accurately recreating just the part of the tower contained within the highlighted subgoal region.

This task environment simulated physical dynamics, including the influence of gravity and collisions between blocks, using the Matter.js physics engine (Brummitt, 2014). If blocks were arranged in an unstable configuration, they would fall and the tower had to be rebuilt from scratch. Due to these physical constraints, participants needed to choose block placements such that at each step throughout the building phase, their construction was stable. In addition, we did not allow blocks to be repositioned or removed once placed. This put pressure on participants to plan ahead: some block placements could make the rest of the tower impossible to complete, even if those consequences were not immediately obvious. For instance, certain block placements could create a single-cell wide gap elsewhere in the tower that could not be filled with any of the available block shapes.

The rationale for presenting matched pairs of subgoals in each session was to make controlled comparisons between subgoals that were otherwise similar, except for the expected computational cost of solving them. So, we prompted participants to solve these subgoals under conditions that enabled us to estimate the computational cost that people incurred to devise a solution. We allowed participants to take as long as they needed to come up with their plan for solving a subgoal, but gave them very limited time to execute their plan. We then used the amount of time participants took before deciding to act as a proxy for how costly that subgoal was to solve. Specifically, each trial was subdivided into two phases: a planning phase and a building phase (Fig. 1a). During the planning phase, participants could take as much time as they needed to come up with a plan to achieve the subgoal, but were unable to move any blocks. Once they were ready, they clicked a button to advance to the building phase. In the building phase, participants had up to 4 s to select and place their first block, and up to 4 s between block placements until they completed the subgoal. If they failed to place the next block within the allotted time, or if the tower became unstable and fell over, the building environment reset and participants had to attempt the subgoal again from scratch. In addition to time pressure, we also gave participants an additional incentive to formulate plans that were likely to be successful. Participants were provided with an initial endowment of 100 points. Each time there was a reset, some of these points would be deducted, which was framed as undesirable. However, participants were neither rewarded nor penalized for the number of points they had retained at the end of the experiment.

2.1.4. Computational model of planning costs

Following prior work, we modeled the process of solving a subgoal as search over a graph of potential actions and their resulting states of the world (Geffner, 2013; Newell & Simon, 1972). A subgoal is completed when a sequence of actions is found that leads from the initial state of the world to a state in which the subgoal is completed. Under this formulation, we approximate the relative cost of completing different subgoals by estimating the number of states that a given search algorithm needs to consider to solve them. In this task setting, we model action-level planning as search over the space of potential sequences of block placements that yield stable configurations at each step.

We focus on *Best First Search* (Dechter & Pearl, 1985; Hart et al., 1968) as our primary candidate for modeling this search process because it has previously shown promise in modeling human problem solving (van Opheusden et al., 2017). Best First Search is a heuristic search algorithm that, like its close relative, the $A*$ algorithm, maintains a list of states to visit, ordered by how promising they are according to a heuristic. On each iteration, Best First Search takes the most promising state from the list of states to explore. If it is a goal state, the algorithm has found a solution and terminates. If not, it adds the states that are reachable by a single action from the current state to the list and then repeats. Ties between equally promising states are broken randomly, which means that Best First Search is not deterministic and can yield different plans on the same problem. To account for this indeterminacy, we ran Best First Search 10 times and averaged the results. Here, we use the percentage of the tower completed as the heuristic criterion, meaning that the most promising state is the one in which the largest fraction of the tower's target shape is covered. Under this criterion, Best First Search is biased toward exploring states where larger blocks are placed and to find solutions that require fewer block placements overall. There was no explicit stopping criterion defined in our implementation of Best First Search. Instead, the algorithm continues to explore until either a solution is found, or it determines that no solution to the subgoal is possible.

Running Best First Search produces two outputs: (1) a sequence of actions that will complete the (sub)goal and (2) the computational cost incurred to arrive at that solution, measured as the number of states that Best First Search explores before finding that solution.

2.2. Results

Overall, we found that participants succeeded in solving these subgoals in a reasonable amount of time. Participants took 8.95 s on average (95% CI: [8.55, 9.34]) to come up with their plan on their first attempt to solve a subgoal. And they were able to succeed within 1.32 attempts (95% CI: [1.28, 1.37]); most often participants solved subgoals on their first try. These results suggest that the subgoals were on average of moderate difficulty: challenging enough to require some thought but not so complex as to be unsolvable within a reasonable timeframe.

We next sought to assess how well the computational cost, as measured by Best First Search, predicted the planning times that our human participants needed for each subgoal (Fig. 1b). We found that the subgoal that Best First Search identified as costlier was also the subgoal that human participants took longer to solve 66.32% (95% CI:[63.29%, 69.30%]) of the time (Fig. 2). To obtain a more graded measure of how well Best First Search could explain human planning times, we also estimated the degree to which variability in computational cost across subgoals estimated by Best First Search could predict variation in human planning times. We estimated the strength of this relationship using the Spearman rank correlation coefficient, which captures the degree to which the rank orderings over subgoals agree with one another, without making strong commitments to the functional form of the relationship between these two variables (i.e., linearity). This analysis revealed a strong and reliable correlation (Spearman's $\rho(188) = 0.56$, bootstrapped 95% CI: [0.50, 0.67], $p < .001$), supporting the use of Best First Search as a proxy for the computational costs that are relevant to how humans approach block tower assembly problems.

We also evaluated several alternative methods for approximating the computational cost of a subgoal. First, we considered *Breadth* First Search (Cormen, 2009), which is a simple planning algorithm that considers all possible sequences of actions in order of their length. It is guaranteed to find the shortest sequence of actions leading to the goal. Since Breadth First Search searches the space indiscriminately, it in practice requires exploring many states to find a solution. While Breadth First Search accounts for some variation in human planning time, it explains less variation than Best First Search (Spearman's $\rho(188) = 0.35$, 95% CI: [0.23, 0.47], $p < .001$). Next, we also considered heuristic features of these subgoals that might conceivably be related to their difficulty, including their size, aspect ratio, number and location of "holes" (i.e., parts of the subgoals that should not be covered by a block), and total area of the tower that consist of holes. We then fit a multiple regression model predicting human planning times from this full set of heuristic features, and found that this model was moderately predictive of human planning time (adjusted $R^2 = .37$), with the number of holes being most predictive among them ($\beta = 7010.97$, $p < .001$). Critically, augmenting

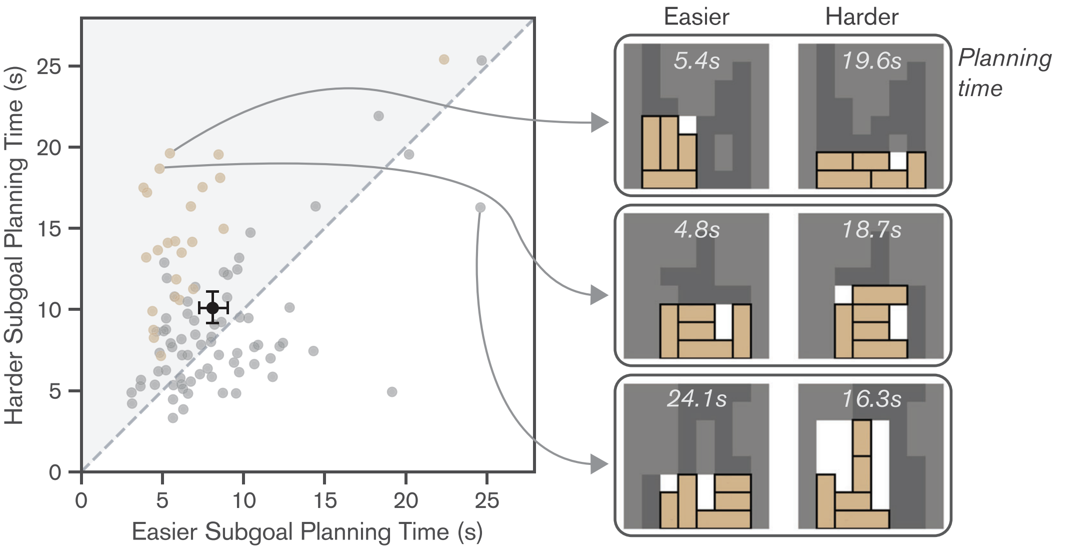

Fig. 2. In a first behavioral experiment, participants were presented with a series of subgoals to solve. Unbeknownst to participants, these subgoals were paired such that both members of each pair were defined over the same tower and had the same surface area, but one was predicted by Best First Search (BFS) to be less costly (Easier) and the other to be costlier (Harder). *Left*: Relationship between the average amount of time participants needed to plan the harder (y-axis) and easier (x-axis) subgoal in each pair. Each dot represents a single subgoal pair, with pairs used in a subsequent subgoal selection experiment rendered in beige. *Right*: Example subgoal pairs. The first two examples represent cases where the predictions of BFS qualitatively match measured human planning times. The left subgoal of the shown pairs is predicted by BFS to be easier to solve; the mean human planning time for each subgoal is inset. The third example represents a case where the predictions of BFS diverged from human planning times.

this heuristic model with the predictions from Best First Search led to gains in explained variance ($\chi^2(1) = 22.54$, $p < .001$). Conversely, augmenting the Best First Search model with heuristic features also led to significant gains in explained variance ($\chi^2(6) = 46.77$, $p < .001$), indicating that both algorithmic search processes and heuristic shortcuts contribute independently to explaining human planning behavior. Taken together, these results validate the use of Best First Search to approximate the computational costs of subgoals in this problem domain, while also demonstrating that humans employ additional strategies not fully captured by the search algorithm alone.

3. Modeling the consequences of different subgoaling strategies on planning costs and performance

Having validated the use of the Best First Search algorithm to estimate the computational cost of solving individual subgoals, our next goal was to formally define a set of hypothetical strategies that people might adopt to select subgoals when solving block tower assembly problems: *No Subgoals*, *Full Decomposition*, *Myopic*, and *Lookahead*. Each of these strategies offers a different way of balancing the tradeoff between the cost of solving the chosen subgoal, the cost required to make this choice in the first place, and the risk of choosing a poor subgoal and getting stuck. We then tested how well each of these strategies explains how people select subgoals.

3.1. Approach

We modeled the selection of subgoals as a resource-rational problem. Resource-rational approaches aim to understand human behavior by assuming that people make optimal use of their limited cognitive resources (Callaway et al., 2022; Griffiths et al., 2015). These approaches acknowledge that while humans may not achieve globally optimal solutions, they can be remarkably efficient at finding good solutions given cognitive constraints, including limited-capacity working memory and attention. In our setting, subgoals are chosen to maximize task performance while minimizing cognitive costs—making the most efficient use of the limited cognitive resources available for planning. In particular, subgoals are chosen to minimize the cognitive cost of finding a solution to the whole problem *given the subgoals*.

3.1.1. Defining subgoals

Formulating a plan to solve an entire block tower problem without first identifying subgoals can be computationally infeasible on problems requiring many actions to complete. Planning can typically be made more tractable by breaking the problems into subgoals. But, given the aim of reducing the action-planning cost for the entire problem, how should the problem be decomposed into subgoals? Assuming that we only care about solving the problem (leaving other considerations such as cost of building materials aside), we want to find a decomposition of the problem into a sequence of subgoals such that the sequence of subgoals jointly completes the problem and minimizes the sum of planning costs across the subgoals in the sequence.

We defined the utility of a subgoal g to be $U_g = P_g - \lambda * C_g$, where P_g is the proportion of the entire problem that the subgoal solves, C_g is the planning cost of solving that subgoal, and λ is the parameter that governs the tradeoff between making rapid progress and avoiding planning costs. When $\lambda = 0$, subgoals are chosen solely based on progress made, while high values of λ lead to choosing subgoals that are very easy to plan. The utility of a sequence of subgoals is the sum of the utilities of its constituent subgoals.

Subgoals in our account were defined visually as rectangular regions of the building area. Completing a subgoal meant completing the part of the tower that was covered by that rectangular region. This differs importantly from defining subgoals as specific target states of the environment. While target states work well for simple tasks like navigation (where they correspond to being in a certain location), they are poorly suited for physical assembly. In complex state spaces like physical assembly, there are too many potential target states to effectively

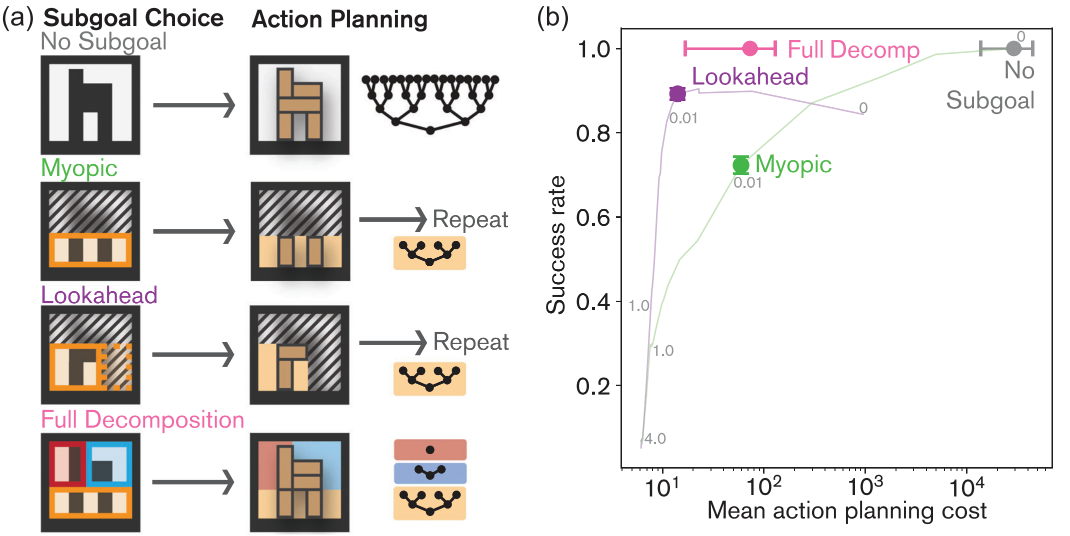

Fig. 3. (a) Three strategies for subgoal selection during physical assembly: *Myopic* (at each turn, only consider the next subgoal), *Lookahead* (at each turn, also consider the subgoal after the one currently being considered), and *Full Decomposition* (break down the entire problem in advance). An example planning graph is shown on the right. (b) The success rate and planning cost for the three strategies. The success rate reflects the proportion of towers that could be perfectly reconstructed under a specific strategy. The planning cost reflects how costly each subgoal is to plan under the subgoals identified under that strategy. For the *Myopic* and *Lookahead* strategies, the relative weight assigned to the value of progress and the cost of planning is set by a weight parameter, λ. The colored discs represent results for $\lambda = 0.01$. The associated curves represent results for other values of λ, indicated by the values adjacent to each curve. Error bars represent bootstrapped 95% confidence intervals.

identify good subgoals. The state of a physical assembly task includes the precise location and orientation of each block—determining these details is the core challenge of the problem itself. By defining subgoals visually, the planner can abstract away from specific block arrangements and focus on the higher-level task of subgoal decomposition. While the rectangular constraint excluded many potential subgoals, this limitation may have provided a useful bias toward valuable subgoals in many assembly scenarios.

In order to make the problem tractable, only sensible subgoals were included. No potential subgoals smaller than the smallest piece (two cells) or larger than 18 cells were allowed. No free floating subgoals (which would be impossible to build) were permitted. Subgoals were not allowed to have empty space on their sides.

3.1.2. Defining subgoal selection strategies

To investigate why one subgoal might be preferable to another at any given point in time, we formalized three general approaches to subgoal selection (Fig. 3a) that vary in how much they account for the rest of the problem: not at all (*Myopic*), partially (*Lookahead*), and completely (*Full Decomposition*).

3.1.2.1. Myopic strategy: The *Myopic* strategy only considers the most ideal next subgoal g_1^*:

$$\text{Myopic:} \qquad g_1^* = \underset{g_1}{argmax}(U_{g_1})$$

The *Myopic* strategy is the simplest of the three. Under this strategy, only the progress achieved by completing the current subgoal and the expected cognitive costs of solving it matter. As such, the *Myopic* strategy is "greedy" in the sense that it will favor more ambitious subgoals to solve next (i.e., P_g is large), so long as they are not too difficult (i.e., C_g is not too large). Once the current subgoal is completed, the subgoal selection process is repeated for the next subgoal and so on. Because the *Myopic* strategy only considers the current subgoal at each time, it can favor decisions that lead to "dead-ends": present commitments that make the rest of the problem more difficult or impossible to solve.

3.1.2.2. Lookahead strategy: The *Lookahead* strategy considers the next d subgoals in sequence:

$$\text{Lookahead:} \qquad g_1^* = \underset{g_1}{argmax}\left(max_{g_2,\ldots,g_d}\left(\sum_{i=1}^{d} U_{g_i}\right)\right)$$

This strategy considers the rest of the problem beyond the current subgoal to a limited degree. Specifically, it makes a decision about the current subgoal such that the subsequent d subgoals also achieve meaningful progress while remaining tractable. Using *Lookahead* generally requires more effort to choose subgoals than the *Myopic* strategy, but could reduce the risk of dead-ends. Here, we set $d = 2$, meaning that the strategy considers the current and the following subgoal. As the number of potential sequences of subgoals grows exponentially with d, considering longer sequences of subgoals at every turn quickly becomes computationally infeasible. At $d = 2$, the *Lookahead* strategy prevents the choice of a subgoal that would lock in an immediate dead-end. At high d, this strategy prevents subgoal choices that lead to dead-ends many subgoals into the future.

3.1.2.3. Full Decomposition strategy: The *Full Decomposition* strategy considers all possible complete decompositions of the problem into a sequence of subgoals $(g_1^*, \ldots, g_n^*)$ that completes the tower, and selects the entire sequence that minimizes the total planning cost[1]:

$$\text{Full Decomposition:} \qquad (g_1^*, \ldots, g_n^*) = \underset{(g_1,\ldots,g_n)}{argmin}\left(\sum_{i=1}^{n} C_{g_i}\right)$$

where $(g_1, \ldots, g_n)$ represents any possible sequence of subgoals that completes the tower, from the first subgoal g_1 to the last subgoal g_n. The *Full Decomposition* strategy is guaranteed to minimize the sum of the planning costs associated with solving each of the subgoals in the winning sequence, at the cost of requiring exhaustive search over the potentially very large set of all possible subgoal sequences to identify the optimal one. To make this computationally tractable, we limited sequences to a maximum length of three subgoals, as exhaustive search over longer sequences was computationally infeasible. Unlike the other two strategies, the *Full Decomposition* strategy only needs to be executed once, as it yields a complete subgoal sequence rather than a single subgoal. As such, no parameter λ is needed to balance the tradeoff between progress and planning cost.

3.1.3. Relationship to other models of task decomposition

This modeling approach extends recently proposed models of resource-rational task decomposition (Correa et al., 2023). Correa et al. (2023)'s framework proposes that people decompose tasks to minimize the overall cost of planning while maintaining task performance. It consists of three nested levels of optimization: (1) action-level planning (to select concrete actions to accomplish a subgoal); (2) subgoal-level planning (which constructs sequences of subgoals that maximize reward while minimizing computational cost); and (3) task decomposition (which selects subgoals based on their value at the subgoal level). Given a set of potential subgoals, this framework specifies how an idealized agent should choose a hierarchical structure that balances task performance with planning costs.

However, one limitation of the formulation in Correa et al. (2023) is that it requires the planner to commit upfront to a sequence of subgoals that completes the entire task. The current modeling approach overcomes that limitation by allowing for *incremental* selection and building of individual subgoals, which enables modeling of planners that look only a few steps ahead rather than ones that decompose the entire problem in advance. While such incrementality might not have been necessary to model navigation tasks where transitions between states are fully reversible, it is highly consequential in the context of block tower assembly, where the set of available subgoals is constrained by earlier decisions to build one part of the tower rather than another.

3.1.4. Simulating consequences of different subgoal selection strategies

In order to understand the consequences that would be expected to follow from the different subgoal strategies, we conducted a computational experiment. Each of these strategies offers a different way of balancing the tradeoff between the cost of solving the chosen subgoal, the cost required to make this choice in the first place, and the risk of choosing a poor subgoal and getting stuck. We conducted simulation experiments to measure how well these strategies performed across a large set of block tower problems varying in size and difficulty. Computational simulations provide a systematic way to understand the tradeoffs between action planning costs, subgoal planning costs, and the risk of getting stuck in dead-ends. While we cannot directly manipulate how human planners approach these problems, these simulations allow us to methodically explore how different subgoal strategies balance these competing factors.

To ensure that we covered the behavior of the strategies across varying levels of complexity, we procedurally generated 500 target towers of varying size and complexity. Target towers were generated by placing 6 to 14 blocks in randomly selected locations such that the tower is stable. From this tower, the outline was extracted and used as the target shape for the subgoals. Note that there are often many possible ways of constructing the outline beyond the sequence of blocks that generated it. For each tower, we ran the planners based on the three planning strategies to generate a sequence of subgoals. The subgoals were then solved using Best First Search. Best First Search is a stochastic algorithm: when two possible states are equally attractive under the heuristic, one is chosen pseudorandomly. To reduce the impact of these pseudorandom choices on the results, we ran each planner 10 times and averaged the results. We report planning costs only on towers that all three strategies could solve, to avoid skewing cost estimates when some planners failed to complete more difficult towers. In the simulation results reported below, we used $\lambda = 0.1$, reflecting a specific balance between prioritizing progress and avoiding planning costs. However, we also explored the consequences of using other values of λ and found that they yield qualitatively similar findings (Fig. 3b).

3.2. Results

As a baseline, we measured how well Best First Search could solve block tower problems without the use of any subgoals. We found that thousands of states needed to be checked, on average, before a solution was found (1.34×10^3 states; 95% CI: [318.07, 2.73×10^3]).[2] Moreover, the cost of finding a solution grew rapidly with the size of the tower: the largest third of towers required considering many more states (1.10×10^5 states, 95% CI: [2.22×10^4, 2.47×10^5]) than did the smallest third (80.99 states, 95% CI: [55.88, 107.53]). These results provide a benchmark for how costly it is to solve these problems by searching directly over sequences of specific actions, and indicate that this strategy quickly becomes prohibitively costly for more complex problems.

The *Full Decomposition* strategy considers all possible sequences of subgoals and selects the one that minimizes the total planning cost, summed over subgoals. As expected, we found that it generated sequences of subgoals where the total number of states explored across all subgoals needed to complete the tower was lower than that of other strategies (8.64 states total, 95% CI: [8.24, 9.03]), providing a lower bound on the computational costs associated with solving a problem through subgoals.

The *Myopic* strategy considers only the utility for the very next subgoal. At $\lambda = 0.1$, we found that it selected sequences of subgoals that led to relatively high total costs for solving each problem (13.58 states; 95% CI: [12.33, 14.82]). Moreover, it also ran into dead-ends for 68.14% of the problems, wherein it chose a subgoal that was itself solvable, but left the rest of the tower impossible to reconstruct with the pieces at hand. Nevertheless, this strategy partly compensates for these higher action-level planning costs and risks of dead-ends by not requiring that costs associated with subgoals beyond the next one be considered.

Finally, the *Lookahead* strategy considers not only the utility of the next subgoal, but also (at least) one potential subgoal following it. While considering the combined utility for multiple subgoals is costlier than considering only one, requiring evaluation of 1141.24 potential subgoals on average compared to the *Myopic* strategy's 111.164329, the *Lookahead* strategy also ends up in dead-ends less often than does the *Myopic* strategy (17.43%) while also keeping the total action planning cost across all chosen subgoals low (9.03 states total; 95% CI: [8.54, 9.52]).

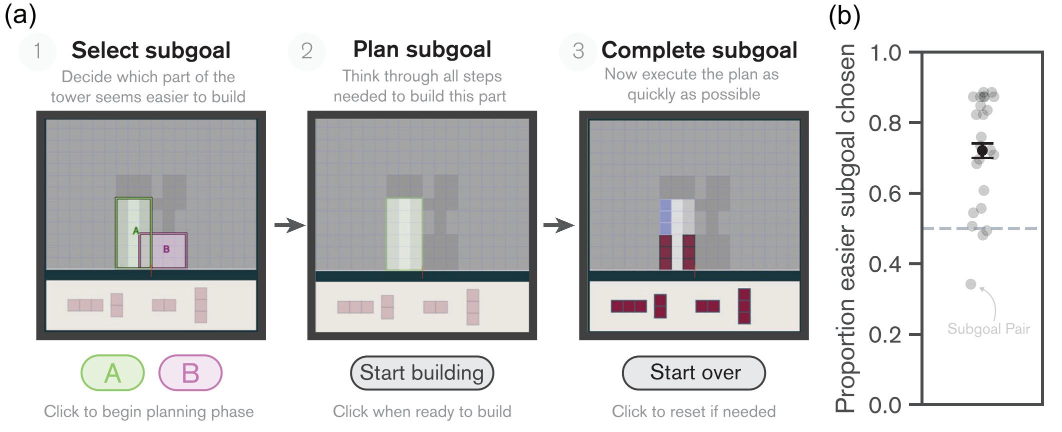

Fig. 4. (a) Participants were presented with two candidate visual subgoals of equal size, one that was estimated to be costlier to solve than the other. (b) Proportion of trials on which participants chose the easier subgoal. Each dot represents the proportion of participants who chose the less costly subgoal in a given pair.

These simulations demonstrate that using subgoals significantly improves the tractability of block tower problems. Beyond simply using subgoals, the specific choice of subgoal strategy matters considerably—affecting not only the cost of solving the problem once subgoals are chosen, but also the effort required to select those subgoals and the risk of encountering dead-ends. In addition, they establish the quantitative impact of adopting each of these strategies on the cost of solving individual subgoals and the cost of choosing which subgoal to pursue next.

We conducted additional experiments exploring the consequences of using higher values of the weight parameter λ across the range from $\lambda = 0$ to $\lambda = 100$, where greater weight is placed on the avoidance of planning costs relative to seeking progress. These experiments revealed that higher values of λ generally lead to the selection of cheaper subgoals, which increases the rate at which any of the above subgoaling strategies ends up in a dead-end (Supplementary Appendix A.1). In the extreme, optimizing strongly for cheaper subgoals increases the total cost incurred choosing subgoals, as these cheaper subgoals are often smaller, and thus more of them are needed to solve the entire problem.

4. What governs human choices over individual subgoals?

The strategies considered in the previous section favor different subgoals based on how much progress it would achieve relative to the estimated computational cost of solving it. We next sought to examine how sensitive people are to these computational costs when deciding which subgoal to pursue next. Toward this end, we presented human participants with pairs of subgoals and asked them to choose which one they would rather solve next (Fig. 4a). In order to isolate the effect of computational cost, the subgoals in each pair were matched with respect to the progress that would be achieved by solving them.

4.1. Method

4.1.1. Participants

Eighty participants (57 male, $M_{\text{Age}} = 33.97$, 1 excluded) were recruited through Prolific and paid a minimum of $14 per hour. No participants from the previous were included in this study.

4.1.2. Stimuli

To study how sensitive participants were to the planning cost of potential subgoals, we presented them with a choice between subgoals that differed in how costly they were estimated to be. We used the same 96 subgoal pairs from the previous study. The subgoals in each pair were matched on size, such that completing either subgoal would yield the same amount of progress toward reconstructing the entire target tower. To determine which subgoal pairs to include in this study, we used the empirical estimates of human planning time we had obtained in our initial study. This procedure yielded 24 pairs of subgoals with the largest gap in planning times between the easier and harder constituent subgoals (Fig. 2; see also Supplementary Appendix A.5).

4.1.3. Procedure

Participants were presented with a pair of matched subgoals and were asked to choose the subgoal that seemed easier to complete quickly. The subgoals were presented as translucent overlays over the target tower (Fig. 4a). After completing the subgoal, participants proceeded to the next trial without having to build the rest of the tower.

4.2. Results

We found that participants reliably chose the *easier* subgoal in each pair (72.63%; 95% CI: [70.57%, 74.68%]), demonstrating that they were sensitive to the actual difficulty of these subgoals, even when the two options yield equal progress (Fig. 4b). Participants deliberated for an average of 6.67 s before making their decision (95% CI: [6.32, 7.07]), which is less time than it had taken participants in the previous experiment to devise a solution for even one of the subgoals (10.27 s, 95% CI: [9.28, 11.22]). That participants made their choices in less time than it takes to solve a subgoal suggests that they were able to determine which subgoal was less costly without coming up with complete solutions for both. However, the longer a participant spent deliberating before making their decision, the more likely they were to choose the easier subgoal ($r(77) = .34$, $p < .001$), consistent with the notion that participants' choices over subgoals might have been generated by partially simulating possible solutions.

5. What governs the choice of subgoals in the context of the entire problem?

We have so far established that people can evaluate the planning costs of individual subgoals and prefer easier ones. One possibility is that the immediate costs associated with the next subgoal are all that people consider at a given time, consistent with the *Myopic* strategy. Alternatively, people might also consider the costs of at least one future subgoal on the path to solving the entire problem, in line with the *Lookahead* and *Full Decomposition* strategies. Our next experiment sought to distinguish between these possibilities by presenting participants with two hypothetical subgoals, one favored under the *Myopic* strategy and the other favored under one of the nonmyopic strategies (i.e., *Lookahead* and *Full Decomposition*). Thus, the *Myopic*-favored subgoal was generally easier to solve, but led participants to incur higher overall computational costs once subsequent subgoals needed to complete the tower were taken into account. Unlike the previous experiment, once participants made their selection, they then had to solve both their chosen subgoal and the rest of the problem, thus making those future costs relevant to the task. To evaluate the degree to which participants consider those future costs, we then compared how well each subgoaling strategy could account for participants' choices by computing the likelihood of their decisions under each strategy. Furthermore, to evaluate the robustness of our findings, we tested two separate sets of subgoal pairs in two independently recruited groups of participants.

5.1. Method

5.1.1. Participants

To assess the robustness of our findings, we conducted this experiment in two samples of participants, using different sets of stimuli that were otherwise generated using an identical procedure. In both versions of the experiment, we recruited 100 participants after exclusions. As before, participants were excluded if they were unable to complete the practice tower or if they navigated away from the online study more than five times. In the first version of the experiment, we recruited 124 participants (63 male, $M_{\text{Age}} = 37.04$) from Prolific, of whom 24 were excluded. In the second version, 140 participants were recruited (91 male, $M_{\text{Age}} = 36.61$), with 40 excluded. Participants were screened to ensure that they did not take part in the other version of the same study. Participants were paid a minimum of $14 per hour.

5.1.2. Stimuli

Having established in the previous study that people are sensitive to the immediate planning costs of their initial subgoal, we now sought to isolate whether they also consider the costs of future subgoals when making their initial choice. To detect this potentially subtler sensitivity to future costs without it being overwhelmed by differences in immediate costs, we deliberately selected pairs of initial subgoals that were approximately matched on both planning cost and progress, but differed in their predicted future planning costs under different subgoal selection strategies. To support a larger search space for identifying informative subgoal pairs, we generated 1280 procedurally generated towers.

5.1.2.1. Assigning values to potential subgoals: To estimate how attractive each initial subgoal was under different strategies, we generated the value for each initial subgoal from the *Myopic*, *Lookahead*, and *Full Decomposition* strategies marginalized over the dynamic

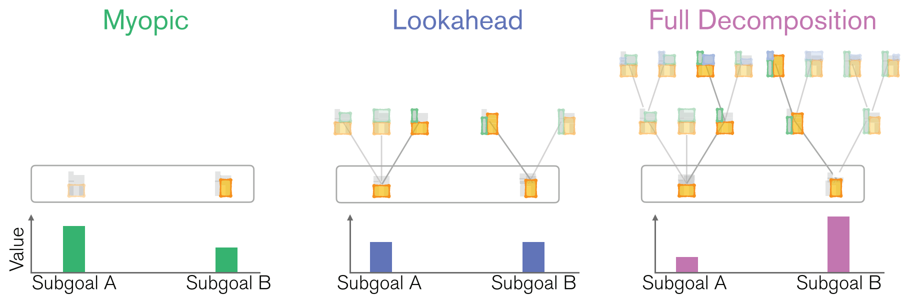

Fig. 6. Schematic showing how the values of the initial subgoal pair are determined by aggregating across sequences of implied future subgoals under different strategies.

range of λ. While the utility of a subgoal (U_g) captures only its immediate progress and cost, the value of an initial subgoal incorporates the implications of that choice for future subgoals under each strategy.

The value of an initial subgoal is calculated differently for each strategy. For the *Myopic* strategy, the value equals the utility of the subgoal itself. This means the *Myopic* strategy would be indifferent between two initial subgoals matched on immediate cost and progress, even if one choice leads to much costlier future subgoals than the other. For the nonmyopic strategies (*Lookahead* and *Full Decomposition*), the value of the initial subgoal is the value of the possible sequences of subgoals that can follow it, as considered by the strategy. The value of a sequence of subgoals is the sum of the values of the subgoals that make up the sequence. Since there are often many potential sequences following an initial subgoal, we took into account the spread over potential sequences by weighing each sequence according to a softmax with temperature = 1 over their values: the more promising a sequence was, the more it contributed to the valuation of the initial subgoal. This softmax weighing of subsequent subgoal sequences captures the intuition that the value of an initial subgoal is influenced by the best possible following sequences, while accounting for noise in the selection of future subgoals (Fig. 6).

5.1.2.2. Selecting subgoal pairs to maximize discriminability: For each set, we chose 12 subgoal pairs where the preferences of the different subgoal selection strategies differed maximally. To ensure that the range of different subgoals was covered, we selected four pairs of small, four pairs of medium, and four pairs of large subgoals. The smallest subgoal contained seven cells of the target shape, the largest 16. Within each size bin, the subgoals were chosen according to two criteria: *diagnosticity* (how much did the nonmyopic and myopic strategies disagree about which of the pair to choose) and *goodness* (were both subgoals a likely choice under the different strategies?). Diagnosticity ensured that the human choice of subgoal was informative about the subgoal selection strategy, while goodness ensured that the subgoals were of the sort that might actually be chosen. If we had only selected subgoal pairs where the strategies disagreed, we might have ended up with subgoals that were both highly inefficient and implausible choices for human participants. *Diagnosticity* is

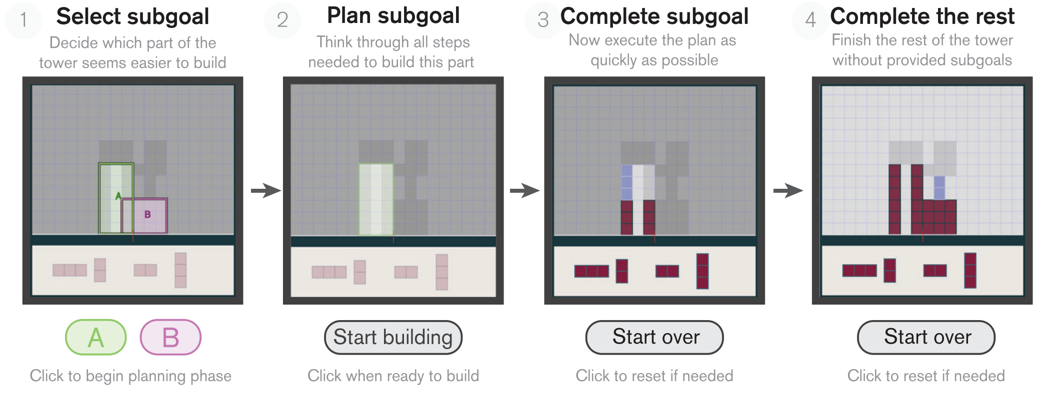

Fig. 5. Participants chose between two candidate subgoals that differed in the difficulty of future subgoals. After completing their chosen subgoal, participants finished building the rest of the tower by iterating between planning and executing the next steps under time pressure.

defined as $|(|M(A) - NM(A)|) - (|M(B) - NM(B)|)|$, where $NM(A)$ is the value of subgoal A under the nonmyopic strategy and $M(A)$ is the value of subgoal A under the *Myopic* strategy. The diagnosticity term is high when one of the strategies prefers subgoal A and the other prefers subgoal B. *Goodness* is defined as $\max(M(A) + NM(A)) + \max(M(B) + NM(B))$. The goodness term ensures that both subgoals are preferred by one of the strategies. See Supplementary Appendix 11 for the values assigned to the subgoals in the chosen subgoal pairs by the different strategies.

Note that while we computed separate values for both the *Lookahead* and *Full Decomposition* strategies, our stimulus selection criteria grouped them together as "non-myopic" (NM) when calculating diagnosticity. This design choice means our stimuli were primarily optimized to distinguish between myopic and nonmyopic planning (i.e., whether people consider future costs at all), rather than to differentiate between different horizons of nonmyopic planning. Consequently, while both *Lookahead* and *Full Decomposition* make distinct predictions about subgoal values, these predictions may be more similar to each other than to the *Myopic* strategy for our selected stimuli.

5.1.3. *Procedure*

Each participant was presented with 12 pairs of subgoals. For each pair, participants were asked to choose between two initial subgoals, then to plan and build their chosen subgoal (Fig. 5). Since this study investigated sensitivity to future costs, participants were required to complete building the rest of the tower to expose them to the consequences of their initial choice. After the initial subgoal, subgoals were not provided. Rather, participants moved between the planning and time-pressured construction phase as often as they wanted—either by clicking a button to move to the planning phase or by letting the 4 s timer expire. Separating planning from building in the free-building phase allowed for the collection of data on the time spent planning versus building the rest of the tower.

5.1.4. Analysis

To understand how well different models explained human behavior, we computed the log-likelihood of human choices under each model. The higher the likelihood that a particular human choice would have been made by the model, the better the model explained human subgoal choice behavior. The model provided a value for each of the two initial subgoals in the pair. We used the softmax over the two values to generate a choice proportion across the two subgoals that the binary choices of participants could be compared to. This comparison yielded a likelihood of the participants' choices under the subgoal selection strategy. To translate from the values to a choice proportion, the softmax temperature T was fitted using a grid search for each model to maximize the log-likelihood of human responses.

To provide a baseline for performance, we included a strategy of indifference. Under this strategy, participants would choose either subgoal with equal probability (50%). The indifferent model provided a floor: any model that explained choices better than random was capturing some aspect of participants' subgoal selection behavior. We also established a ceiling on model performance based on the inherent variability in human choices. Due to individual differences, even an ideal model cannot perfectly predict every participant's choices. This ceiling was determined by using the empirical distribution of participant choices as the predicted probabilities. For example, if 80% of participants selected subgoal A over subgoal B, then the ideal model would predict a choice probability of 0.8 for subgoal A. No model can achieve better predictive performance than using these observed choice proportions as predictions. The upper and lower bounds of the baseline and the floor were computed by resampling participants.

The value of future subgoal sequences depends on λ: for higher λ, the strategies prefer cheaper subgoals over those that make progress. For a model to predict human choices, it needs to account for the preference between avoiding costs and making progress. We fitted the λ for each model to maximize the likelihood of human choices under the model jointly with the softmax temperature described above. See Table 4 in the Supplementary Appendix for the fitted values.

5.2. Results

We evaluated the *Myopic* strategy against both a noise ceiling (i.e., the empirical distribution of participants' choices, representing the best achievable predictive performance) and an indifferent baseline strategy that assigns each subgoal of the pair an equal probability of being chosen (Figure 7A). We found that the *Myopic* strategy performed slightly better than the indifferent baseline in the first sample (log-likelihood ratio: 0.14, 95% CI: [0.05, 0.22]), but this advantage did not generalize to the second sample with different participants and subgoal pairs (log-likelihood ratio: –0.00, 95% CI: [-0.00, -0.00]). In addition, the *Myopic* strategy fell well short of the noise ceiling in both samples, indicating that its weak performance was not merely a reflection of the amount of noise in these data. These two experiments taken together do not provide strong evidence in favor of the *Myopic* strategy.

Next, we compared the two nonmyopic strategies to the *Myopic* strategy. We found that the nonmyopic strategies performed better at explaining participants' choices in both the first sample (log-likelihood ratio: 0.31, 95% CI: [0.13, 0.48]) and second one (likelihood ratio:

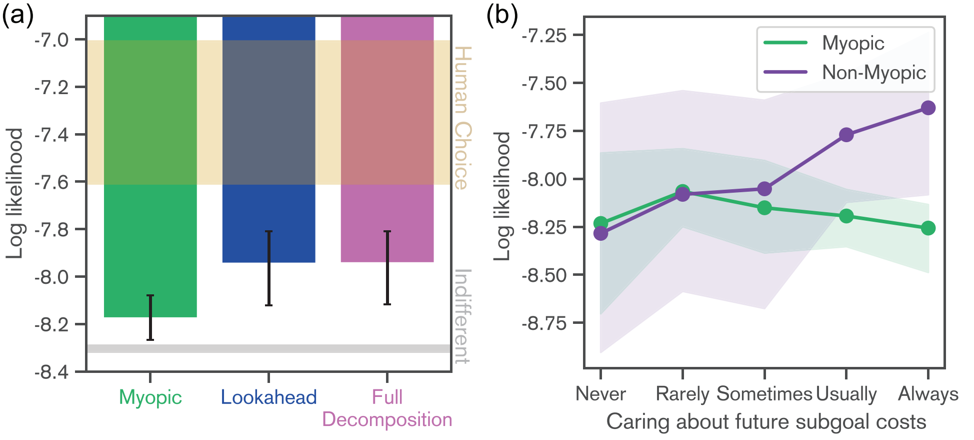

Fig. 7. (a) The mean log-likelihood of a participant's subgoal choices under the different strategies. Participants' subgoal choices are best explained by nonmyopic strategies. The yellow zone ("Human Choice") shows the best achievable log-likelihood when predicting the aggregate choices of all participants. The gray zone ("Indifferent") provides a floor for the explainable variance, showing the expected log likelihood when assuming indifference between the two subgoals. Error bars represent bootstrapped 95% confidence intervals generated by resampling participants. (b) The log-likelihood of myopic and nonmyopic subgoal choices for participants who reported considering future subgoal costs to varying degrees. Shaded areas represent bootstrapped 95% confidence intervals.

0.23, 95% CI: [0.06, 0.38]), providing converging evidence that participants considered both immediate and future costs when making their subgoal selections. However, we found that the nonmyopic strategies also fell short of the noise ceiling in both samples (log-likelihood ratio in first sample: 0.50, 95% CI: [0.39, 0.72]; log-likelihood ratio in second sample: 0.28, 95% CI: [0.27, 0.59]). This gap suggests that while these models of nonmyopic subgoal selection account for a meaningful amount of the explainable variance in human subgoal choices, there are other important factors that would need to be considered to fully capture this behavioral pattern.

When directly comparing the *Lookahead* and *Full Decomposition* strategies, we found mixed evidence for which better explained participants' choices. In the first sample, participants' behavior was explained equally well by both strategies (log-likelihood ratio: –0.01, 95% CI: [−0.16, 0.18]). However, in the second sample the *Lookahead* strategy performed slightly better than *Full Decomposition* (log-likelihood ratio: 0.16, 95% CI: [0.06, 0.30]). This limited ability to distinguish between these two nonmyopic strategies is consistent with our stimulus selection procedure, which prioritized discriminating myopic from nonmyopic planning rather than distinguishing between different horizons of future consideration. As such, we refrain from drawing strong conclusions about the relative performance of *Lookahead* and *Full Decomposition* strategies at this time.

All of the subgoal selection strategies considered so far balance making progress on the overall problem against reducing computational costs. This requires trading off between subgoals that are easy to plan versus those that are more ambitious. To evaluate the degree to which participants are sensitive to both of these factors, we also developed lesioned versions of the subgoal selection strategies. These lesioned models either considered only planning cost ("stingy") or only progress ("greedy"). Neither the "stingy" nor "greedy" versions explained participants' choices as well as the full models (Supplementary Appendix 13), suggesting that participants consider both the potential progress they would make and the potential planning costs they would incur when selecting subgoals.

While our findings so far provide strong evidence that nonmyopic strategies better explain subgoal selection at the group level, different participants might not display these tendencies to the same degree. To explore variation across individuals, we examined their self-reported strategies. Specifically, in a post-study questionnaire, participants indicated the degree to which they explicitly considered future subgoal costs when choosing which subgoal to build next, using a 5-point Likert scale ranging from "never" to "always." We found that while a plurality of participants reported having "usually" or "always" considered future costs (first sample: 57.00%; second sample: 45.54%), a sizable proportion of participants reported "never" or "rarely" considering future costs (first sample: 23.0%; second sample: 26.73%).

For the "never"/"rarely" group, we were unable to reliably measure a difference in performance between the myopic and nonmyopic strategies (first sample likelihood ratio: 0.02, 95% CI: [−0.55, 0.45]; second sample likelihood ratio: –0.12, 95% CI: [−0.37, 0.22]), consistent with the possibility that these participants were employing different strategies altogether that did not rely on consideration of planning costs (Fig. 7b). On the other hand, for those who reported "usually" or "always" considering future costs, the nonmyopic strategies reliably outperformed the *Myopic* strategy (first sample likelihood ratio: 0.49, 95% CI: [0.22, 0.78]; second sample likelihood ratio: 0.35, 95% CI: [0.12, 0.61]). These findings suggest that there is meaningful variation in the degree to which participants' choices are well explained by a nonmyopic model of planning, and that participants are at least somewhat aware of what strategies they use to make their decisions.

In summary, our results suggest that when people are presented with the challenge of decomposing physical assembly problems into efficient subgoals, they are capable of rapidly and accurately judging the relative difficulty of potential subgoals even without fully devising a plan for completing either. Moreover, they often consider not only the immediate cognitive costs associated with completing the very next subgoal, but also future costs associated with solving the remainder of the problem. Finally, our findings highlight opportunities to develop improved models that account for additional variance in human subgoal selection.

6. General discussion

Here, we examined the factors people consider—either implicitly or explicitly—when decomposing a physical assembly problem into subgoals. Specifically, we investigated the degree to which people are sensitive to two key factors: the contribution a subgoal makes to the complete solution, and the computational effort involved in solving it. Our results suggest that participants could rapidly differentiate between otherwise similarly ambitious subgoals and used those judgments to the computationally cheaper ones. Furthermore, their choices indicated consideration of not just the immediate costs of planning, but also future costs, at least to some degree.

Recent work has attempted to model task decomposition under cognitive constraints, assuming that humans select subgoals to balance progress and planning costs. This approach, known as resource-rational task decomposition (Correa et al., 2023), suggests that humans use their cognitive resources efficiently when breaking down complex problems. Using a collection of simple graph-structured navigation tasks, they found evidence supporting this account: people's subgoal choices reflected a balance between making progress and managing planning costs. Our model builds on this framework, extending it to visual subgoals and allowing for the interleaving of planning and acting. One challenge posed by extending resource-rational task decomposition to more complex tasks, such as the physical assembly problems considered in this paper, is that the number of potential states and subgoal sequences grows combinatorially with the size of the problem. Visual subgoals provide a natural way of generating promising candidate subgoals: they are straightforward to define, and they can help reduce planning costs by leveraging the physical locality of actions' effects. However, even with a good prior over potential visual subgoals, finding the optimal sequence becomes intractable when exhaustively considering all options. The interleaved aspect of scoping helps manage this complexity by trading off between finding the most optimal sequence and keeping the cost of choosing subgoals manageable, even as it sacrifices some optimality for speed. The interactive, embedded nature of the visual scoping account, both in time and space, may help explain the puzzling efficiency of human hierarchical planning in complex tasks. Our findings extend Correa et al. (2023)'s results to the more complex domain of block tower construction, validating and generalizing the assumptions of resource-rational task decomposition to more ecologically natural planning tasks.

Our results might help to make sense of other findings in the study of human physical assembly. A recent study found that humans (both adults and children) are biased toward constructing buildings made of physical Lego blocks layerwise (Cortesa et al., 2017, 2018). This layerwise bias can be understood through the lens of visual subgoaling: each layer defines a clear spatial region to focus on, potentially reducing planning costs for the entire problem. However, on the concrete problems used in this study, we found that participants show only a weak or no preference for wider over taller subgoals (first sample: 53.50% chose wider subgoals, $\chi^2(1) = 6.05$, $p = .0139$; second sample: 49.63% chose wider subgoals, $\chi^2(1) = 0.04$, $p = .8405$). Similarly, McCarthy et al. (2023) found that while tower-building trajectories vary between participants, they often pass through similar states. These common states might represent natural visual subgoals that many builders identify as useful intermediate targets. By considering both hierarchical planning through visual subgoals and the search for specific actions, we might be able to better account for these observed patterns in human assembly behavior.

Nevertheless, several limitations of the current study affect our ability to draw stronger conclusions about human planning in physical assembly tasks. First, the computational framework explored here makes strong assumptions that might not fully capture how people approach these problems. While Best First Search was moderately predictive of human planning times, other visual heuristics (e.g., number of "holes" in the tower) explained additional variance not captured by the search algorithm. These findings suggest the value of integrating domain-general models of search and more domain-specific models of perceptual organization to develop improved accounts of human visual subgoaling. Additionally, while Best First Search generates numerically precise estimates of planning costs, it is plausible that people estimate these planning costs less precisely. It would be valuable for future modeling efforts to explicitly account for the amount of uncertainty in people's estimates of planning costs.

Another limitation is that the binary choice paradigm used in our experiments, while useful for direct comparisons between competing hypotheses, offers limited insight into the full range of ways people might spontaneously decompose these problems. Future work might explore more open-ended tasks to elicit subgoals, such as enabling participants to freely define their next subgoal. In addition, future work might make fuller use of implicit behavioral measures, such as the timing of pauses while building, to infer the achievement of subgoals.

A third limitation concerns the untested assumption that people use visual subgoals at all when engaged in physical assembly. Indeed, other recent work on human planning in other environments has found that variation in performance can be explained by search depth rather than the use of subgoals (Van Opheusden et al., 2023; van Opheusden et al., 2017; van Opheusden & Ma, 2019). While our simulations suggest that naive search without visual subgoals is intractable on the block tower task, we required the use of subgoals in our experiments. Moreover, our rectangular visual subgoal definition may be too restrictive—a more flexible way of defining the shape of visual subgoals could better capture how humans naturally decompose these problems. Future work could investigate the task conditions under which people spontaneously use visual subgoals while building towers, and the conditions under which they instead adopt other approaches. Such experiments could, for example, give participants the option of subgoaling, by letting them restrict the amount of the problem they have persistent visual access to, in exchange for additional planning time, to understand how much additional planning time is worth losing visual access to part of the problem.

A fourth limitation of the current study is the incomplete investigation of the contextual factors that might affect the tradeoff value of progress and computational costs, which is captured by a weight parameter, λ, in the current work. While our experiments suggest that varying values of λ can have profound impacts on which subgoals are favored, we did not investigate what factors might influence this balance. Future experiments could fill this gap by manipulating task constraints (e.g., through time pressure or the imposition of working memory load).

A fifth limitation concerns the unresolved differences between the *Lookahead* and *Full Decomposition* hypotheses as potential explanations for subgoal choice in the final behavioral experiment. While our stimulus selection procedure successfully discriminated between myopic and nonmyopic planning strategies, it was not optimized to distinguish between different horizons of future consideration. To obtain more robust and generalizable insights, we conducted two versions of this experiment, with independently generated stimulus sets and groups of participants. However, the gap in performance between the *Lookahead* and *Full Decomposition* models varied enough between these two versions of the experiment that we are unable to definitively conclude that one of these accounts is better than the other. Future work should more systematically disentangle the contributions of variation across assembly problems from those of variation across individual participants, which the current data are not well-equipped to address (because the two versions of the experiments differed along both dimensions simultaneously).

Finally, while the block tower construction task is richer and more interactive than many classical planning problems, it is still far from the complexity of real-world physical reasoning tasks that people perform in their daily lives. Toward closing this gap, future work might take similar experimental approaches to investigate human physical reasoning in contexts where individuals are more fully embedded in the environment and can interact with it in various ways. For example, using three-dimensional environments with more complex interactions would enable investigation of how people choose what to look at and from what point of view, which might have more direct relevance to predicting spontaneous visual subgoaling in more naturalistic physical tasks (Ho et al., 2021).

The classical formalizations of intelligent behavior, such as model-based reinforcement learning, often treat the agent and the world as distinct entities, with cognition formalized as a process internal to the agent and the environment reduced to an input. While successful in many cases, this approach might obscure the role of agent–environment interactions in supporting cognition. The present work attempts to extend a model-based algorithmic lens to understanding how agents can use visual and spatial structure to support efficient planning. The visual subgoaling account proposed here suggests how the spatial embedding of a task can be leveraged to make complex planning tractable. This perspective might inform the development of more efficient embodied artifical intelligence systems, such as robots that reason and plan in a manner similar to humans. It might also provide a framework for understanding how deep reinforcement learning agents generalize or fail to do so when their environment changes (Shah et al., 2022). By developing theories of visual reasoning that bridge the gap between human and machine planning, this work contributes to a more unified understanding of intelligent behavior, one that acknowledges the role of agent–environment interaction in supporting the efficiency and flexibility of human cognition.

Acknowledgments

We thank Will McCarthy, Frederick Callaway, and members of the Cognitive Tools Lab at Stanford University. This work was supported by NSF CAREER award #2436199 and an Office of Naval Research Science of Autonomy award to J.E.F.

Conflict of interest

The authors declare no conflict of interest.

Code and data availability statement

Simulation and model code, as well as human experiment and analysis code, are available at https://github.com/cogtoolslab/Humans-Select-Subgoals-That-Balance-Immediate-and-Future-Cognitive-Costs-During-Physical-Assembly.

Notes

1 Since the sequence of subgoals is guaranteed to complete the tower, the sum of the progress of the subgoals is always 1 and, therefore, not included in the utility calculation.
2 To enable direct comparisons between strategies, planning costs are only reported for the towers for which all subgoaling strategies were able to find a solution.

References

Agre, P. E. (1988). The dynamic structure of everyday life. MIT Artificial Intelligence Laboratory.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K. L., Kohli, P., Battaglia, P. W., & Hamrick, J. B. (2019). Structured agents for physical construction. *arXiv:1904.03177 [cs].*

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bear, D., Wang, E., Mrowca, D., Binder, F., Tung, H.-Y. F., Pramod, R. T., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J. B., Yamins, D. L. K., & Fan, J. (2021). Physion: Evaluating physical prediction from vision in humans and machines. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 15.

Bellman, R. (1957). *Dynamic programming* (1st ed.). Princeton, NJ: Princeton University Press.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Brummitt, L. (2014). Liabru/matter-js.

Callaway, F., Van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, *6*(8), 1112–1125. https://doi.org/10.1038/s41562-022-01332-8

Cormen, T. H. (2009). *Introduction to algorithms* (3rd ed.). Cambridge, MA: MIT Press.

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology*, *19*(6), e1011087. https://doi.org/10.1371/journal.pcbi.1011087

Cortesa, C. S., Jones, J. D., Hager, G. D., & Khudanpur, S. (2018). Constraints and development in children's block construction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 6.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., & Shelton, A. L. (2017). Characterizing spatial construction processes: Toward computational tools to understand cognition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 6.

Dechter, R., & Pearl, J. (1985). Generalized best-first search strategies and the optimality of A*. *Journal of ACM*, *32*(3), 505–536. https://doi.org/10.1145/3828.3830

Dietz, G., Landay, J. A., & Gweon, H. (2019). Building blocks of computational thinking: Young children's developing capacities for problem decomposition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 7.

Fikes, R. E., & Nilsson, N. J. (1971). Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, *2*(3–4), 189–208. https://doi.org/10.1016/0004-3702(71)90010-5

Geffner, H. (2013). Computational models of planning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(4), 341–356. https://doi.org/10.1002/wcs.1233

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883. https://doi.org/10.1016/j.tics.2020.09.001

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. https://doi.org/10.1111/tops.12142

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76. https://doi.org/10.1016/j.cognition.2016.08.012

Hart, P., Nilsson, N., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, *4*(2), 100–107. https://doi.org/10.1109/TSSC.1968.300136

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science*, *3*(4), 275–310.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2021). Control of mental representations in human planning. *arXiv:2105.06948 [cs]*.

Hunt, L. T., Daw, N. D., Kaanders, P., MacIver, M. A., Mugan, U., Procyk, E., Redish, A. D., Russo, E., Scholl, J., Stachenfeld, K., Wilson, C. R. E., & Kolling, N. (2021). Formalizing planning and information search in naturalistic decision-making. *Nature Neuroscience*, *24*(8), 1051–1064. https://doi.org/10.1038/s41593-021-00866-w

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103. https://doi.org/10.1073/pnas.1414219112

Jinnai, Y., Abel, D., Hershkowitz, D. E., Littman, M., & Konidaris, G. (2019). Finding options that minimize planning time.

Kirsh, D. (2009). Problem Solving and Situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 264–306).

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, 101396. https://doi.org/10.1016/j.cogpsych.2021.101396

Maisto, D., Donnarumma, F., & Pezzulo, G. (2015). Divide et impera: Subgoaling reduces the complexity of probabilistic inference and problem solving. *Journal of the Royal Society Interface*, *12*(104), 20141335. https://doi.org/10.1098/rsif.2014.1335

Mattar, M. G., & Lengyel, M. (2022). Planning in the brain. *Neuron*, *110*(6), 914–934. https://doi.org/10.1016/j.neuron.2021.12.018

McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

McCarthy, W. P., Kirsh, D., & Fan, J. E. (2023). Consistency and variation in reasoning about physical assembly. *Cognitive Science*, *47*(12), e13397. https://doi.org/10.1111/cogs.13397

Newell, A., & Simon, H. A. (1972). *Human problem solving*, volume *104*. Englewood Cliffs, NJ: Prentice-Hall.

Patalano, A. L., & Seifert, C. M. (1997). Opportunistic planning: Being reminded of pending goals. *Cognitive Psychology*, *34*(1), 1–36. https://doi.org/10.1006/cogp.1997.0655

Sacerdoti, E. D. (1974). Planning in a hierarchy of abstraction spaces. *Artificial Intelligence*, *5*(2), 115–135. https://doi.org/10.1016/0004-3702(74)90026-5

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). Goal misgeneralization: why correct specifications aren't enough for correct goals.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138. https://doi.org/10.1037/h0042769

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*(2), 101–118. https://doi.org/10.1007/s42113-018-0007-3

Smithwick, D., & Kirsh, D. (2015). Let's get physical: Thinking with things in architectural design. In *CogSci*.

Sussman, G. J. (1975). *A computer model of skill acquisition*. New York; London: American Elsevier; Elsevier.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. https://doi.org/10.1016/j.tics.2017.05.012

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 6.

Van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature 618*(7967), 1000–1005. https://doi.org/10.1038/s41586-023-06124-2

van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences*, *29*, 127–133. https://doi.org/10.1016/j.cobeha.2019.07.002

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1