Interplay of approximate planning strategies

Quentin J. M. Huys[a,b,1], Níall Lally[c,d], Paul Faulkner[e], Neir Eshel[f], Erich Seifritz[b], Samuel J. Gershman[g], Peter Dayan[h,2], and Jonathan P. Roiser[c,2]

[a]Translational Neuromodeling Unit, Institute of Biomedical Engineering, University of Zürich and Swiss Federal Institute of Technology (ETH) Zürich, 8032 Zurich, Switzerland; [b]Department of Psychiatry, Psychotherapy and Psychosomatics, Hospital of Psychiatry, University of Zürich, 8032 Zurich, Switzerland; [c]Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, United Kingdom; [d]Experimental Therapeutics & Pathophysiology Branch, Intramural Research Program, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892; [e]Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA 90095; [f]Program in Neuroscience and MD-PhD Program, Harvard Medical School, Boston, MA 02115; [g]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [h]Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, United Kingdom

Edited by Larry F. Abbott, Columbia University, New York, NY, and approved December 31, 2014 (received for review July 25, 2014)

Humans routinely formulate plans in domains so complex that even the most powerful computers are taxed. To do so, they seem to avail themselves of many strategies and heuristics that efficiently simplify, approximate, and hierarchically decompose hard tasks into simpler subtasks. Theoretical and cognitive research has revealed several such strategies; however, little is known about their establishment, interaction, and efficiency. Here, we use model-based behavioral analysis to provide a detailed examination of the performance of human subjects in a moderately deep planning task. We find that subjects exploit the structure of the domain to establish subgoals in a way that achieves a nearly maximal reduction in the cost of computing values of choices, but then combine partial searches with greedy local steps to solve subtasks, and maladaptively prune the decision trees of subtasks in a reflexive manner upon encountering salient losses. Subjects come idiosyncratically to favor particular sequences of actions to achieve subgoals, creating novel complex actions or "options."

planning | hierarchical reinforcement learning | memoization | pruning

Humans and other animals often face complex tasks and environments in which they have to plan and execute long sequences of appropriate actions to achieve distant goals. One can represent the space of future actions and outcomes as a tree; such trees grow inordinately (often exponentially) large as a function of the length of the sequence (i.e., the depth of the tree). Rather little is definitively known about how this computational complexity is addressed. Work in the fields of reinforcement learning and artificial intelligence has suggested a number of heuristics that we describe below, namely, hacking, hierarchies, hoarding, and habitization (1–4). Various tasks have been designed to highlight individual heuristics; though how subjects generate and combine them without clear instruction has not been well characterized (however, see refs. 5 and 6).

We previously designed a moderately deep planning problem to elicit a specific heuristic, in this case hacking or pruning of the decision tree (4). However, the task contains many of the elements that make choosing appropriately tricky in general. Thus, we closely examined the nature of, and individual differences between, the performance of subjects, shedding light on the interaction of heuristics in the self-generation of adaptive control when faced with a complex planning problem.

Subjects had to plan a path through a maze so as to maximize their cumulative earnings. On each trial, they were placed in a random state and were asked to plan to a depth of 3, 4, or 5 moves (Fig. 1 *A* and *B*). Because each depth involved a binary choice, planning to depths 3, 4, and 5 corresponded to choosing among a set of 8, 16, or 32 possible sequences. We previously found that the large immediate losses at particular branch points in the tree (the red transitions) encouraged subjects to eliminate possibly lucrative subbranches beneath those points (4). This corresponds to suboptimal pruning or "hacking" of the decision tree (Fig. 1*C*). The analyses presented below show that this was by no means the only strategy subjects used.

Hierarchical task decomposition licenses strategies for reducing computational burdens based on divide and conquer (2, 7) or "chunking" (5, 8–12). The resulting divided problems are easier to conquer because chunks are smaller and ignore aspects of the environment that do not impinge on their domains. The solutions to the subproblems can then be treated as larger-scale actions, often called macroactions or options (1). These simplify solving complex tasks by providing a way of building large decision trees out of smaller numbers of intermediate-sized parts (the macroactions) rather than larger numbers of smaller parts (each individual action). The downside is potential suboptimality. We use the precise form of suboptimality that our subjects exhibited to argue that they hierarchically fragmented the planning problems: Deep problems were solved by concatenating solutions to sequences of shallower problems (e.g., greedily adopting as a depth-5 solution the best depth-3 solution followed by the best remaining depth-2 solution; Fig. 1*D*). We then asked a critical question that has eluded previous approaches to hierarchical control, namely the degree to which the fragmentation is actually computationally advantageous—is the benefit of divide and conquer appropriately realized?

A third, "hoarding" heuristic is known as memoization. If subjects are repeatedly faced with the task of finding a good policy at a state, then rather than building and searching the decision tree each time it is sensible to recall a previous solution and use that. If the previous solution cannot be guaranteed to be correct, then storing and deciding among several past solutions could be wise. When such storage and recall are probabilistic, the heuristic is stochastic memoization. It has been most extensively

Significance

Many problems, particularly sequential planning problems, are computationally very demanding. How humans combine strategies to approximate and simplify these problems is not understood. Using modelling to unpick performance in a planning task, we find that humans are able to exploit the structure of the task to subdivide it and reduce processing requirements nearly optimally. Subtasks are combined in a simple, greedy manner, however, and within subtasks there is evidence of inhibitory reflexes in response to losses.

Author contributions: Q.J.M.H., N.L., N.E., P.D., and J.P.R. designed research; Q.J.M.H., N.L., and P.F. performed research; Q.J.M.H. and S.J.G. contributed new analytic tools; Q.J.M.H., P.D., and J.P.R. analyzed data; and Q.J.M.H., N.L., P.F., N.E., E.S., S.J.G., P.D., and J.P.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 2929.

[1]To whom correspondence should be addressed. Email: qhuys@cantab.net.

[2]P.D. and J.P.R. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1414219112/-/DCSupplemental.

Downloaded by guest on January 4, 2022

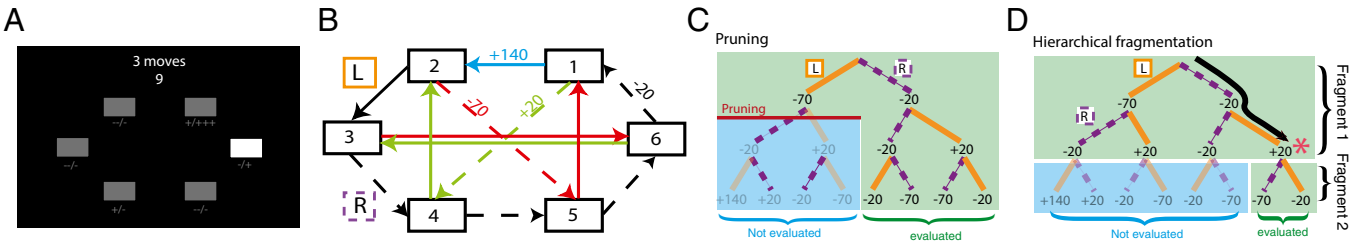

Fig. 1. Task. (*A*) Task display. On each trial, subjects saw six boxes. The bright box indicated the randomly chosen starting location. The number of moves to plan was displayed at the top. During the decision time of 9 s, subjects had to plan between three and five moves. Then, during the input time of 2.5 s, they had to enter their plan as a single sequence of right/left button presses in one go and without immediate feedback as to what state they were currently in or what rewards they had earned in the choice sequence so far. After the entire sequence had been entered, the chosen sequence and the rewards earned were displayed in the order in which they had been entered. Failure to enter a button press sequence of the right length in the given time resulted in a penalty of –200 pence. (*B*) Task structure. Subjects were placed in one of the six boxes ("states") at the beginning of each trial and had to plan a path through the maze that maximized their total outcomes earned. From each state, two successor states could be reached deterministically by pressing either the right (dashed lines) or left (solid lines) key. For example, from state 1, state 4 could be reached by pressing left–left–right. Each transition resulted in a deterministic reward or loss. Red arrows, for instance, denote large salient losses of –70 points. The possible transitions were never displayed on screen. (*C*) Pruning. The decision tree faced by subjects for a depth-3 problem starting in state 3. When encountering one of the large losses (–70, red arrows in *B*) the search along that subtree is terminated. The blue parts of the tree would thereby not be evaluated and thus the cost of computation would be reduced. In this case, pruning leads to a suboptimal sequence appearing as being optimal. (*D*) Hierarchical fragmentation of the same problem. Rather than evaluating the entire depth-3 tree, a 2–1 fragmentation would first search the tree up to depth 2 (large green area), choose a depth-2 sequence (black arrow), and then search the remaining depth-1 tree (bottom right green area). The blue area of the tree is again not evaluated. Optimal choices in the fragmented tree may miss the overall optimal sequence, which in this case would be on the far left of the tree. If a subject emitted the sequence on the far right, this sequence would be more likely under the fragmentation 2–1 than under a nonfragmented tree of full depth 3. The effective "subgoal" corresponding to the target of the first fragment (the end state of the subsequence resulting from the first part of the fragmentation) is indicated by a red asterisk.

investigated in computational linguistics (13–15), and recently imported into decision making (16). Hoarding and hierarchies interact closely: Stored solutions can exactly be considered as macroactions or chunks, and so stochastic memoization can be seen as an answer to another poorly explored question, namely, how hierarchical decompositions arise. In particular, we will see that different subjects fragmented the task in idiosyncratic ways, putatively based on the way that they memoized.

We used both flexible and constrained statistical analyses to examine the use of these heuristics. For the constrained analyses, we stipulated a particular mathematical form for each cognitive process and implemented it in a model that, after its parameters had been fit, reported the likelihood of the subjects' choices. More complex models should provide better fits and were penalized by computing integrated Bayesian information criterion (iBIC) scores, which approximate Bayes factors (4, 17). We tested models including and excluding particular cognitive processes. Those processes present in the model with the best iBIC score were taken as putatively present in subjects' decision making. Importantly, this approach always tests the ability of the hypothesized set of cognitive processes to account for the entire dataset, rather than only hand-selected aspects of the data.

Results

Subjects were trained extensively on both the transition matrix and the rewards associated with each of the transitions. As previously reported (4), subjects "pruned" (Fig. 1*C*) extensively, with outcomes distal to large losses being discounted at a faster rate than outcomes distal to non-large loss outcomes (*Supporting Information*, *Pruning* and Fig. S1). We previously found this to be insensitive to the actual size of the large loss and to reduce earnings overall, and therefore interpreted it as a reflexive "Pavlovian" influence on goal-directed decisions. All findings below incorporate and extend these findings, that is, we correct for them throughout, and ultimately reassess pruning in the light of the more refined analyses of other heuristics. The time pressures imposed in this version of task, and the fact that planning had to be completed before any move was registered, led to few apparent differences in choices relative to the original version (4).

Fragmenting Decisions. We first looked for a very simple type of hierarchy, whereby a smaller initial subproblem of a difficult task is addressed greedily, leaving whatever remains to be solved in turn. It seemed that subjects indeed adopted this approach. For instance, starting from state 1, the optimal path for a depth-3 problem is 1–2–3–4, whereas for a depth-4 problem it is 1–2–5–1–2. However, subjects had a strong predilection for the path 1–2–3–4–2 (Fig. S2, first column, middle row). A similar pattern was qualitatively observed throughout (other panels of Fig. S2). To quantify this statistically, we computed each subject's distribution over depth-3 sequences (cf. Fig. S2) in the depth-3 problems and at the beginning of depth-4 problems and then performed Spearman rank correlation between these distributions over choice sequences. The correlation was on average 0.44. Comparison with correlations obtained from permutations of the distributions revealed that it was individually significant ($P < .05$) for 35/37 subjects (94%). Repeating the analysis by comparing the initial depth-4 sequences of depth-5 choices to the depth-4 choices, the correlation was on average 0.47 and individually significant in all subjects.

Subjects thus seemed to solve harder problems hierarchically, by exploiting the solutions to fragments, which are themselves smaller problems. However, just as there are many different ways of subdividing a complex task into simpler subtasks, there are many ways simpler problem solutions could be substituted into harder problems. To measure directly the hierarchical decomposition and avoid the potential biases in the above simple analysis, we fitted an exploratory model with sufficient flexibility to capture all possible fragmentations (i.e., for each trial, we asked which fragmented decomposition best fitted subjects' actual choices).

In this model, we first examined fragment endpoints (red asterisk in Fig. 1*D*). To do so, we extracted all fragmented choices, that is, we extracted the maximum a posteriori decomposition of each subject's choices, determined the start and endpoints of each fragment, and constructed histograms of endpoints as a function of the fragment origin. Fig. 2*A*, *Left* shows that fragments tended to terminate in state 2 irrespective of the start state (seen as a dark horizontal band), or in the next state around the outer ring of the maze (seen as a dark band below the diagonal). This pattern accounted for 90% of target end states on average (range 0.83–0.99) and was present in all subjects above chance ($P < 10^{-10}$ for each; *Supporting Information*, *Optimal Fragmentation* and Fig. S3). Fragments of length 1 tended to end in the next state along the circle (Fig. 2*A*, *Middle*), whereas fragments of greater length

SEE COMMENTARY
NEUROSCIENCE
Downloaded by guest on January 4, 2022

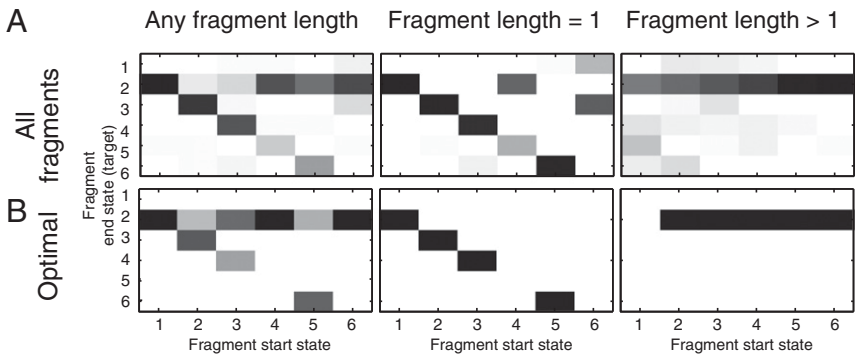

Fig. 2. Fragmentation. (*A*) Fragment endpoint distributions when including all fragments within an individual choice sequence. Each panel shows the distribution of end states for fragments starting in each of the six states (fragment start states). The left column shows the end point distribution when considering all fragments. Fragments starting in state 5 terminated in state 2 or state 6 with high probability. The middle column shows the endpoints of fragments of length one, and the rightmost column the endpoints of fragments of length greater than one. (*B*) Endpoint distribution for fragmentation that achieves the optimal choice at the least computational cost.

terminated in state 2 irrespective of where they started (Fig. 2*A*, *Right*) and typically reaped the big reward when transitioning from state 1. Importantly, this structure was stable through the hierarchical decomposition: It was present when considering only the first fragment at the root of the decision tree and also when considering only the second fragment inside the decision tree (Fig. S4*A*). Finally, we examined the delays between button presses. Button press delays preceding depth-1 fragments were on average 9.7ms longer than the decision time before longer fragments [$t(36) = 2.17$, $P = 0.037$]. Because the computational requirement for shorter fragments is smaller than that for long fragments, this suggests that emitting a short fragment is the result of failed deeper searches. Thus, subjects seemed to build decision trees repeatedly until they found a path through the salient reward (transitioning from state 1 to 2); failing this, they would move one step along the circle (the off-diagonal band) and then reattempt to build a decision tree that would lead them to state 2. This suggests that subjects might have treated state 2 as a subgoal, which makes intuitive sense because this usually corresponds to earning the large reward of +140 when reached from state 1.

The driving force behind fragmentation should be the reduction in computational cost, a simple measure of which is the total number of computations—here largely additions—required to sum the rewards along all paths in a tree. A full evaluation of a depth-3 tree would require a sum of three outcomes for each of the eight arms, that is, $3 \times 8 = 24$ computations, and more generally $d \times 2^d$ computations for a depth-d tree. Using this measure of computational cost, the fragmentation in Fig. 1*D* would require evaluation of a depth-2 and then a depth-1 tree, resulting in $(2 \times 2^2) + (1 \times 2^1) = 10$ computations (i.e., a reduction of the computational cost by more than 50%) (Fig. 1*D*). (The results remain the same when quantifying the cost of using dynamic programming, which is in general more efficient; see *Supporting Information*, *Optimal Fragmentation* for further details.) Given these costs, one can identify, for each state–depth problem, the fragmentation with the least computational cost that still selects the optimal choice. The distribution over fragment endpoints for the optimal decomposition is shown in Fig. 2*B* and matches the empirically inferred distributions well (Fig. 2*A*). Fig. 3 *A* and *B* show that the frequency of fragment depths and the overall distribution over end states also match the optimal distribution well. This suggests that participants chose a hierarchical task decomposition that nearly optimally reduced computational cost. For this task, the near-optimal fragmentation can be interpreted in terms of two organizing principles: one aligned with the perceptually salient transition structure (the ring of states) and the other with the reinforcement structure (the large rewards).

The distribution of fragment endpoints showed substantial structure that was present in every single subject. However, precisely because of the exploratory intention to allow for the discovery of any target pattern, this model (called "baseline + unrestricted fragmentation") enjoyed very many degrees of freedom. It was able to reproduce the detailed choice patterns with high fidelity and accounted for 64% of the variability (Fig. S2, blue line). However, for the same reason, it overfitted the data. When integrating over all fragmentations it had a worse iBIC score than the baseline model without fragmentation (ΔiBIC = 682).

One way to capture the fragmentation structure would be to build a process model (i.e., an account of how subjects actually solve the metacontrol problem of decomposing a problem). This is beyond the scope of the current work. However, in an effort to constrain possible theoretical frameworks for this, we built a reduced model in which the full tree was replaced by the most frequently used fragmented version of the tree for each particular subject. For instance, if the tree starting in state 3 was most frequently decomposed into a tree of depth 2 followed by a depth-1 tree, as displayed in Fig. 1*D*, then this unique fragmentation was assumed to be fixed for the entire experiment for that state and depth. Fixing fragmentations for individuals did not substantially alter the fragment endpoint distributions (Fig. S4*B*) and the optimal pattern was still significantly present in every subject ($P < 10^{-10}$, binomial test). This model ("baseline + restricted fragmentation") improved the fit over the baseline model (Fig. S5 *A* and *B*) and led to a substantial improvement in formal model comparison (Fig. 3*C*). Importantly, this model was as identifiable as the previous simpler models without fragmentation: We fitted each model to surrogate data generated from each of the models and were always able to recover the correct model (*Supporting Information*, *Robustness of Inference* and Table S1).

However, repeating this procedure, but now forcing all subjects to decompose problems in the same manner, produced a worse model (ΔiBIC = 116 compared with baseline model). Thus, fragmentation strategies were stable within individuals, but varied across different subjects. These results suggest that subgoals (the salient reward) and the decomposition of plans are intimately related. They constrain the processes that generate the fragmentation in the first place by identifying salient reinforcements as one central influence.

Stochastic Memoization. The fact that fragmentation is consistent for a given subject but varies between different subjects suggests the possibility that each subject generates one or a few possible decompositions and then sticks with this limited collection. Different subjects could generate and stick with different decompositions. For instance, the subject might initially generate action sequences through a laborious and computationally expensive tree search, but later on simply reuse a past solution. This memoization process (Fig. S6) would make later choices duplicate early fragmentations. One sign of this could be in the temporal evolution of the use of fragments, with those ultimately used most frequently coming to dominate the distribution of fragments slowly. Thus, the distribution over fragments should become more strongly concentrated on a few fragments as the task progresses. We computed the probability of fragments over time in the model "baseline + unrestricted

Downloaded by guest on January 4, 2022

fragmentation." Fig. 3*D* shows that the frequency of the most commonly used fragment increased gradually over time. In marked contrast, the frequency with which all other choice fragments were chosen decayed over time. Fig. 3*E* shows that this results in the entropy over the fragment distribution decaying steeply.

Precisely because the task is too hard to solve perfectly, however, subjects cannot be sure that their previously computed choice sequence really represents the best option. Stochastic memoization refers to probabilistic, as opposed to deterministic, reuse. It is more appropriate when the result of the computation might change if it were recomputed, for instance due to incomplete or error-prone computation. One formalization of such a process is inspired by a method invented in computational linguistics (14, 16) that employs a distribution known as the "Chinese restaurant process" (CRP) (18). A CRP defines a probability distribution consisting of two terms. The first term is proportional to the frequency of past samples, whereas the second is the "base distribution" from which samples are drawn in the first place. Applied to the current problem, this model assumes that the probability of emitting a particular fragment is a weighted sum of two probabilities: the frequency of that particular fragment in the collection of previous choices and the probability that the fragment would be chosen anew if the solution to the problem were recomputed (i.e., the probability under the model baseline + restricted fragmentation). One critical feature of the CRP statistical model is a gradual change, with the choice probability being initially mainly driven by the base model (implying recomputation), and later by subjects' past choices (implying stronger reuse later; see Eq. **2**). This transfer from flexible but costly computation to inflexible reliance on past experience is reminiscent of arguments about the transfer from goal-directed to habitual controllers (3, 19). However, by relying only on which choice was emitted rather than on how good it was, it also differs from certain formalizations of habits (3).

Fig. 3*C* and Fig. S5 show that the CRP addition in the augmented model "baseline + restricted fragmentation + stochastic memoization" outperformed the other models. Fitted reward sensitivity parameters correlated closely with the true reward sizes (mean of 0.994), and the target structure of the fragments was again not substantially altered by the inclusion of stochastic memoization (Fig. S4*C*). Adding stochastic memoization to the model with unrestricted fragmentation also improved all measures of model fit drastically (log likelihood improved for every subject by 43 ± 22; 14% more variance explained; ΔiBIC = 4,162). The same was true when controlling for an increase in the scaling of the reward sensitivities [i.e., an increase in exploitation (20)] over time (ΔiBIC = 86), and the model was also clearly identifiable on surrogate datasets (*Supporting Information*, *Robustness of Inference* and Table S1). Finally, corresponding parameters were highly correlated between all models tested (0.81 ± 0.06), suggesting that parameters captured similar variability in different models.

Pruning. Finally, we considered whether the pruning that we had previously seen using a similar task (4) might have been an artifact of the incomplete analysis of fragmentation and memoization. The baseline model that we fit included two pruning parameters: one that discounted outcomes distal to large losses (γ_S) and another (γ_G) that discounted distal outcomes in a value-independent manner (*Supporting Information*, *Pruning*). However, the relationship between these two parameters was not constrained by the models. We examined the pruning parameters in the (overfitting) model "baseline + unrestricted fragmentation + stochastic memoization" because this model captured 78% of the variance, and hence controlled most strongly for all other processes. Fig. 3*F* shows that the continuation rate after outcomes other than large losses was indistinguishable from 1 (and hence γ_G from zero), arguing that the apparent general discounting factor whereby subjects do not always look to the end of a tree is actually an epiphenomenon of hierarchical decomposition. However, every subject discounted outcomes distant to large losses more steeply than other distant outcomes ($1-\gamma_S < 1-\gamma_G$ for 37/37 subjects). That is, pruning remained a powerful effect even when controlling for fragmentation and reuse as much as possible. Given that most fragments were short enough to be computed fully (Fig. 3*B*; however, note that owing to the length of the fragments the actual number of choices being part of longer fragments is higher), this also underscores our previous contention that pruning is a Pavlovian and reflexive response to aversive outcomes (4).

Intelligence Quotient. It has been suggested that subjects' ability to decompose problems into larger chunks is a key ingredient of intelligence (21). The correlation between mean fragment length and verbal intelligence quotient (IQ) measured by a reading test was not significant ($\rho = 0.08$, $P = 0.64$).

Discussion

Our results suggest that humans naturally decompose problems in a way that efficiently trades computational cost for performance;

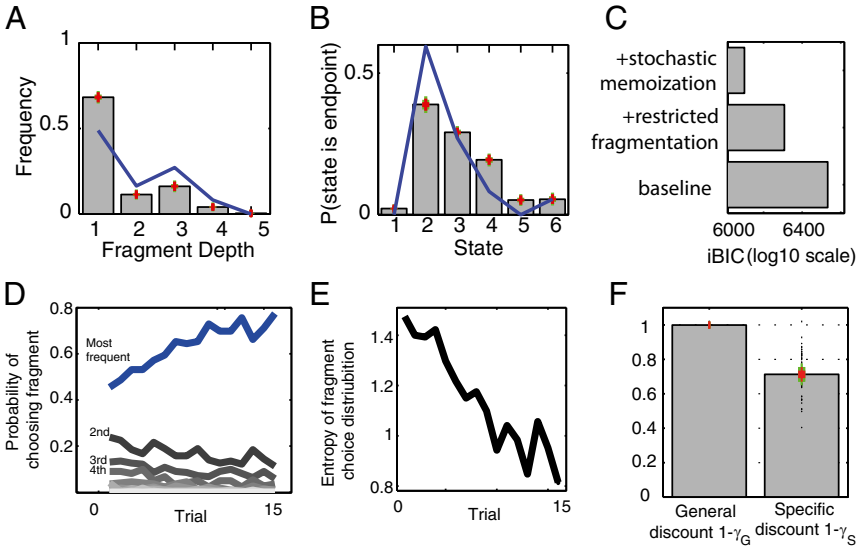

Fig. 3. Fragment characteristics. (*A*) Distribution over inferred fragment lengths. (*B*) Overall distribution over fragment endpoints. State 2 is the most frequent endpoint. Blue lines in *A* and *B* show the distributions for the optimal fragmentations. (*C*) Nested model comparison. Each bar shows the group-level iBIC score for one model, when adding additional cognitive processes. (*D*) Over time, only the most frequently used fragment increases in frequency, whereas all others decay and are used less frequently. (*E*) The entropy of the distribution over fragments used falls nearly linearly over time. (*F*) Discount factors (within fragments). An outcome lying x transitions ahead is multiplied by $1-\gamma$ a total of $x-1$ times. For outcomes lying distant to large losses ("specific pruning") $1-\gamma_S$ is substantially smaller than 1, implying robust discounting. In contrast, for outcomes distant to non-large loss outcomes ("general pruning"), $1-\gamma_G$ is indistinguishable from 1 for every subject, meaning that these are not downweighted within fragments. Thus, subjects search to the end of the fragment but show a strong tendency to stop the search at large losses even within the fragments ($1-\gamma_S < 1$).

SEE COMMENTARY
NEUROSCIENCE
Downloaded by guest on January 4, 2022

that the fragmentation of the task and the search strategy are shaped both by salient rewards and by salient perceptual features of the task; that subjects initially generate action sequences using a tree-search process, but later rely on a stored representation thereof; and that whenever subjects perform a mental search of a decision tree they have a tendency to prune the tree upon encountering salient losses. These features operate cumulatively: Subjects concurrently use multiple approximations when solving planning tasks. We were able to find sufficient evidence in favor of relatively complex models, potentially because of the high accuracy exhibited by the subjects—Fig. S2, bottom row, shows that even at depth 5 subjects rarely made very poor choices, even given the relatively tight time constraints.

Fragmentation. The fact that the decomposition achieved in this task is so close to the optimum (Fig. 3 *A* and *B*) is striking, particularly because finding an optimal fragmentation is typically more difficult than finding an optimal path. We were not able to address directly how this was achieved (i.e., to build a process model) but rather fitted all possible fragmentations to identify the relevant features.

One possibility is that the fragmentation arises from the exact sequence in which parts of the decision tree are searched. A hint comes from the suggestion that subjects search for the large reward, accept whatever path they find that leads them there, and then start recomputing from there onward. If they fail to find a path to the large reward, they move one step along the circle and try again. One immediate prediction from this is borne out: Fragments that only move one step along the circle often represent a failure of the search and should take longer to produce than longer fragments, even though they involve less computational cost. Thus, if it is this strategy that drives subjects' internal search through the decision tree, then the optimality of the decomposition hinges directly on the relationship between the subgoal subjects aim for and the likelihood that this subgoal is on the optimal path. In the present task, the large reward subgoal was often on the optimal path. The fact that subjects rely so strongly on the salient reward in defining the subgoal bears some resemblance to the impact the large losses have. We have previously argued that the losses induce pruning in a reflexive, approximate, Pavlovian manner (4), and it may be that the selection of subgoals follows similar rules, as opposed to being derived in an adaptive manner from a clever insight into the task structure and optimality of various decompositions (see also ref. 21). It would certainly be highly instructive to alter the reward matrix such that this is not the case any more—one would expect the optimality of the decomposition to then break (see Fig. S7 for a predicted fragment endpoint distribution for a simple alteration to the reward matrix).

One might compare the hierarchical decomposition that we observed with those that have been studied in frameworks that explicitly set out to study hierarchical control (rather than to study pruning, which was our original target). A central construct in those tasks is that of a functional bottleneck—a state that makes a worthy subgoal because many paths have to flow through it by virtue of its position in state space (22–27). In our task, no state has this status—all states are equally connected—it is the reward structure that licenses the particular fragmentation.

Against our expectations (21) we did not find a correlation between verbal IQ and the average length of fragments. This is possibly because the Wechsler Test of Adult Reading (WTAR) used here is more a measure of verbal than fluid IQ. Alternatively, optimality in this task implies efficient selection of paths at the least computational cost. It is conceivable that the prediction should have been the opposite: that participants with higher IQ should have smaller average fragment length but achieve similar outcomes. A mixture between these two effects may explain the current null finding.

Memoization and Option Generation. One important insight that has engendered extensive research in decision making over the past decade is the distinction between goal-directed decisions and cached habits. Whereas the former suffer the sort of computational complexities that justify fragmentation, the latter suffer from a requirement for sampling from the world, or a model thereof: Instead of thinking through the future, the consequences of choices are experienced, and these experiences are cached to determine future choices (3, 28). Memoization is explicitly a form of caching, exhibiting the signature characteristic that if the environment suddenly changes (for instance via outcome devaluation or contingency degradation) the cached values will remain the same, and so control based on them will look maladaptive.

One common formalization of cached habit is in terms of state-action or $\mathcal{Q}$ values (29), which estimate the long-run utility that would be accrued from a state given a particular first action. A similar process might be applied to entire action sequences (6). However, these values depend on the depth of the problem, or subproblem, being solved, and because this is not fixed in the current problem these approaches provide little traction. Instead, stochastic memoization invites consideration of what might be a simpler form of cached habit, namely, a fixed sequence of actions (5, 11), which is like a macroaction or option (1). The equivalent of progressive habitization arises naturally from Eq. **2**, because the probability of performing a whole new tree search (i.e., sampling from the base measure) decreases with the number of relevant trials so far. However, note that our form of stochastic memoization was independent of reward (i.e., memoization did not depend on the actual quality of the solution produced; it is therefore more like a refined form of choice kernel); indeed, it is known that nonhuman primate choices, for instance, depend substantially on their own past choices, above and beyond the rewards associated with the decisions (30, 31). Similar arguments have been made for human choices in a variety of tasks and settings (32, 33) and have been argued to be under dopaminergic (34) and serotonergic (35) control.

Materials and Methods

Participants. We recruited 41 healthy volunteers (21 female; 23.3 ± 3.7 y) via the University College London psychology subject pool. They were screened for past and present psychiatric disorders (including drug and alcohol abuse) with the Mini International Neuropsychiatric Inventory (36). Subjects with past or current axis I diagnosis were excluded (one participant was excluded owing to previous substance dependence). Subjects completed the WTAR (37) (mean = 111, SD = 4.2) to assess IQ. The study was approved by the University College London Graduate School Ethics Committee. Subjects provided written, informed consent and were remunerated based on performance, up to a maximum of £40.

Task. The task is described in Fig. 1 and was adapted for functional MRI (fMRI) from one described in detail elsewhere (4) and programmed in Cogent 2000 (www.vislab.ucl.ac.uk/Cogent), a stimulus presentation toolbox for MATLAB (version 7.1). The fMRI results will be reported elsewhere, and we here report analyses only of the behavior of the same 37 subjects included there. Subjects were first extensively trained on the transition and reward matrix and all passed a test. Each of 90 trials of the main experiment began in a random starting state, but the combination of starting state and depth were biased such that in 60 trials it was optimal to transition through large losses, whereas in 30 trials the optimal path did not involve transition through a large loss. As part of the training, subjects had performed 32 trials that matched those of the main experiment, but in 18 of which there was no time restriction. The analyses presented here include these training trials. The experiment contained additional "restricted plan" trials, where subjects chose between two predefined paths, as a control condition for the fMRI analysis. These trials were not analyzed here.

Fragmentation. This model subdivided action sequences. The probability of a sequence $\mathbf{a}$ was represented as the product of the probability of K fragment action sequences:

$$p(\mathbf{a}|s,d,\mathcal{Q}^P) = p\left(\mathbf{a}^{(1)}|s,d\right)\prod_{k=2}^{K} p\left(\mathbf{a}^{(k)}|\mathbf{a}^{(k-1)},s,d,\mathcal{Q}^P\right), \quad [1]$$

where s and d denote start state and overall depth. The probabilities depend on the value $\mathcal{Q}^P$ from the baseline model that includes pruning and loss

Downloaded by guest on January 4, 2022

sensitivity (*Supporting Information, Pruning*). The (k)'th fragment $\mathbf{a}^{(k)}$ starts where the $(k-1)$'th fragment $\mathbf{a}^{(k-1)}$ ends, hence the dependence of fragment $\mathbf{a}^{(k)}$ on fragment $\mathbf{a}^{(k-1)}$. A sequence of length d can be subdivided into fragments of lengths 1 to d in 2^{d-1} different ways. For instance, a sequence of length 3 could be composed of three sequences of lengths 1, a single sequence of length 3, a sequence of length 2 followed by a length-1 sequence, or a sequence of length 1 followed by a sequence of length 2. Because the identity of the particular fragmentation used by a subject is not known, this needs to be integrated out. On each step of the group fitting procedure (4) we applied an expectation-maximization procedure to each individual subject to infer both parameters of the base model and the fragmentation used on each particular trial.

Stochastic Memoization. This model allowed for the reuse of fragments. Each entire action sequence was again subdivided as above into k segments. The probability of generating the particular segment $\mathbf{a}^{(k)}$ was the sum of two components. The first component was the probability if it was recomputed (i.e., the probability assigned to it by the baseline model; see *Supporting Information, Pruning*). The second component was proportional to how frequently that particular fragment action $\mathbf{a}^{(k)}$ had been emitted in that particular state up to that point. That is, the probability of emitting a fragment $\mathbf{a}^{(k)}$ was formalized as a Dirichlet process with the choice probability distribution from the baseline model $p(\mathbf{a}^{(k)}|\mathcal{Q}^p)$ serving as the base measure. Let $n_{sd(k)}$ be the total number of times a subject has emitted a fragment of depth $d(k)$ in state s so far, and $n_{sd(k)}(\mathbf{a}^{(k)})$ the number of times the subject chose the fragment $\mathbf{a}^{(k)}$. The probability of an action is then a sum of two components, weighted by a parameter α:

$$p\left(\mathbf{a}^{(k)}|s,d(k),n\right)=\frac{n_{sd(k)}\left(\mathbf{a}^{(k)}\right)}{n_{sd(k)}+\alpha}+\frac{\alpha}{\alpha+n_{sd(k)}}p\left(\mathbf{a}^{(k)}|\mathcal{Q}^p\right). \quad [2]$$

As $\alpha\to\infty$, only the second factor involving $p(\mathbf{a}|\mathcal{Q}^p)$ remains and this model reduces to the previous "pruning" model. However, as $\alpha\to 0$, the probability distribution becomes dominated by the past choices: Whichever fragment $\mathbf{a}^{(k)}$ was most frequently chosen up to that point is most likely to be chosen again. Hence, α serves as a measure of how strongly past choices determine current choices. Furthermore, as n grows with time, the second term vanishes whereas the first term remains $\mathcal{O}(1)$. Thus, over time, this model assumes that subjects rarely reevaluate the tree by computing $\mathcal{Q}^p$, but rather mostly sample from their past choices proportionally to the past choice frequency. Because the identity of the fragmentation is not known, inference involves a sum over all possible fragmentation histories. Because this is not tractable, we approximate inference with a Viterbi-like scheme, where at each trial the most likely fragmentation is assumed to have been chosen and the history terms n updated accordingly.

Model Fitting and Model Comparison. We applied a nested model comparison strategy. We start from the simplest model and always evaluate whether additional model complexity is warranted by computing approximate Bayes factors (the integrated group-level BIC scores; see ref. 4) for each model. All models were fitted using MATLAB version 8.0 (MathWorks). We used the parallel processing toolbox and the function fminunc. All parameters were transformed to lie on the real line for inference. Models had the following number of parameters: lookahead, 1; discount, 2; pruning, 3; pruning + loss, 6; pruning + loss + restricted fragmentation, 24; and pruning + loss + restricted fragmentation + reuse, 25. See *Supporting Information, Robustness* for an assessment of the robustness of this approach.

ACKNOWLEDGMENTS. This study has appeared in abstract form and was funded by a British Academy grant to J.P.R. Q.J.M.H. received funds from the German Research Foundation and N.L. from a Wellcome Trust–National Institutes of Health (NIH) studentship. N.E. is supported by a Sackler Fellowship in Psychobiology and NIH Grants T32GM007753 and F30MH100729. E.S. is supported by the Swiss National Science Foundation, the University of Zürich, and the Neuroscience Centre Zürich and P.D. by the Gatsby Charitable Foundation.

1. Sutton RS, Precup D, Singh S (1999) Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif Intell* 112:181–211.
2. Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition* 113(3):262–280.
3. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8(12):1704–1711.
4. Huys QJM, et al. (2012) Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Comput Biol* 8(3):e1002410.
5. Dezfouli A, Balleine BW (2012) Habits, action sequences and reinforcement learning. *Eur J Neurosci* 35(7):1036–1051.
6. Dezfouli A, Balleine BW (2013) Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLOS Comput Biol* 9(12):e1003364.
7. Dietterich TG (2000) Hierarchical reinforcement learning with the MAXQ value function decomposition. *J Artif Intell Res* 13:227–303.
8. Jog MS, Kubota Y, Connolly CI, Hillegaart V, Graybiel AM (1999) Building neural representations of habits. *Science* 286(5445):1745–1749.
9. Koechlin E, Ody C, Kouneiher F (2003) The architecture of cognitive control in the human prefrontal cortex. *Science* 302(5648):1181–1185.
10. Badre D (2008) Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn Sci* 12(5):193–200.
11. Ostlund SB, Winterbauer NE, Balleine BW (2009) Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *J Neurosci* 29(25):8280–8287.
12. Logan GD, Crump MJC (2010) Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science* 330(6004):683–686.
13. Michie D (1968) Memo functions and machine learning. *Nature* 218:19–22.
14. O'Donnell TJ, Goodman ND, Tenenbaum JB (2009) Fragment grammars: Exploring computation and reuse in language. Technical Report MIT-CSAIL-TR-2009-013 (Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA).
15. O'Donnell TJ (2015) *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage* (MIT Press, Cambridge, MA).
16. Wingate D, Diuk C, O'Donnell T, Tenenbaum J, Gershman S (2013) Compositional policy priors. Technical Report MIT-CSAIL-TR 2013-007 (Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA).
17. Kass R, Raftery A (1995) Bayes factors. *J Am Stat Assoc* 90(430):773–795.
18. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *J Am Stat Assoc* 101(476):1566–1581.
19. Keramati M, Dezfouli A, Piray P (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Comput Biol* 7(5):e1002055.
20. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
21. Bhandari A, Duncan J (2014) Goal neglect and knowledge chunking in the construction of novel behaviour. *Cognition* 130(1):11–30.
22. Hirtle SC, Jonides J (1985) Evidence of hierarchies in cognitive maps. *Mem Cognit* 13(3):208–217.
23. McGovern A, Barto AG (2001) Automatic discovery of subgoals in reinforcement learning using diverse density. *Proceedings of the Eighteenth International Conference on Machine Learning* (Morgan Kaufmann, San Francisco), pp 361–368.
24. Wiener JM, Mallot HA (2003) 'Fine-to-coarse' route planning and navigation in regionalized environments. *Spat Cogn Comput* 3:331–358.
25. Şimşek Ö, Wolfe AP, Barto AG (2005) Identifying useful subgoals in reinforcement learning by local graph partitioning. *Proceedings of the 22nd International Conference on Machine Learning* (Assoc for Computing Machinery, New York), pp 816–823.
26. Opsahl T, Agneessens F, Skvoretz J (2010) Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc Networks* 32:245–251.
27. Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM (2013) Neural representations of events arise from temporal community structure. *Nat Neurosci* 16(4):486–492.
28. Huys QJM, Guitart-Masip M, Dolan RJ, Dayan P (2015) Decision-theoretic psychiatry. *Clin Psychol Sci*, in press.
29. Watkins CJCH (1989) Learning from delayed rewards. PhD thesis (Cambridge Univ, Cambridge, UK).
30. Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* 84(3):555–579.
31. Seo H, Barraclough DJ, Lee D (2007) Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cereb Cortex* 17(Suppl 1):i110–i117.
32. Camerer C, Ho TH (1998) Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation. *J Math Psychol* 42(2/3):305–326.
33. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6):1204–1215.
34. Rutledge RB, et al. (2009) Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *J Neurosci* 29(48):15104–15114.
35. Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R (2012) Serotonin selectively modulates reward value in human decision-making. *J Neurosci* 32(17):5833–5842.
36. Sheehan DV, et al. (1998) The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 59(Suppl 20):22–33, quiz 34–57.
37. Wechsler D (2001) *Wechsler Test of Adult Reading Manual* (The Psychological Corp, San Antonio, TX).

SEE COMMENTARY
NEUROSCIENCE
Downloaded by guest on January 4, 2022