Epistemic drive and memory manipulations in explore-exploit problems

Nicolas Collignon n.collignon@ed.ac.uk
Christopher Lucas clucas2@inf.ed.ac.uk
School of Informatics, University of Edinburgh

Abstract

People often navigate new environments and must learn about how actions map to outcomes to achieve their goals. In this paper, we are concerned with how people direct their search and trade off between selecting informative actions and actions that will be most immediately rewarding when they are faced with new tasks. We find that some people selected globally informative actions and were able to generalize from few observations in order learn new reward structures efficiently. These participants also displayed the ability to transfer knowledge across similar tasks. However, a consistent proportion of participants behaved sub-optimally, caring more about observing novel information instead of maximizing reward. Across four experiments, we present evidence that participants' motivation to explore was influenced by 1) how much they already knew about the underlying task structure and 2) whether their observations remained available. We discuss possible explanations behind people's exploratory drive.

Keywords: active learning; generalization; exploration-exploitation; transfer learning; data-availability;

Introduction

In order to act, plan, and achieve goals, people must learn about their environment and the outcome of possible actions. One reason for human successes in developing new theories and strategies when confronted with new problems is that people are not passive observers. Indeed, children ask informative questions and can adapt their strategies when inquiring about things they don't know (Ruggeri & Lombrozo, 2014), and play with new toys in ways that help them disambiguate uncertain causal relationships and gather information (L. Schulz & Bonawitz, 2007; Cook et al., 2011). The idea that humans learn and interact with their environment by performing intuitive experiments, maximizing information gain, is a popular one (Coenen et al., 2017; Gureckis & Markant, 2012; Nelson, 2005; Gopnik et al., 2004).

In this paper, we are interested in how people learn to select actions that are most rewarding when faced with a sequence of novel but potentially related tasks. We designed experiments to better understand people's exploration and reward maximizing strategies across a sequence of tasks. Do those strategies evolve over time, as they encounter related tasks? Can people transfer structural knowledge and improve their performance by leveraging similarities between tasks? What is the relationship between people's search strategies, their ability to learn and generalize from observations, and how well they do?

When encountering new situations, people are often faced with the decision of either gathering more information about the task to improve the quality of their decision, or choosing an action that has been shown to be rewarding (Hills et al., 2015). A doctor might, for example, want to run more tests to have a better diagnosis for their patient or give them the treatment they believe will best relieve them from their symptoms. To better understand human decision strategies when dealing with the explore-exploit trade-off, Multi-armed Bandits (MAB) have been used extensively. In these experiments, participants have to select between different possible actions (e.g. the arms of a bandit) yielding stochastic rewards, so as to maximize their rewards. In the real world, an essential part of solving problems lies in discovering the underlying structure of the problem, where each action can be represented as a set of continuous and discrete features. In a Contextual MAB (CMAB), there are observable feature that provide information about the arms' reward distributions. Learning how features relate to rewards allows for an efficient representation of the environment, and enables the learner to generalize to new events. Previous studies of human behavior in CMAB problems have shown that people are able to generalize across observations when faced with a large number of options, and make use of uncertainty to direct their search (E. Schulz et al., 2017; Wu et al., 2018; Borji & Itti, 2013). These experiments have assumed the basic structure of the underlying problems to be static, or known in advance. When confronted with unknown task structures, Teodorescu and Erev (2014) showed that people were able to adaptively learn purely exploratory or purely exploitation-oriented policies. However, in their experiment there was no systematic relationship between an option's features and its reward, aside from whether it had been previously explored.

Unlike a CMAB-type task, the tasks we presented to participants were deterministic, meaning that re-selecting an option would always yield the same reward. This was done to ensure a clear distinction between exploration and exploitation in participant decisions. To examine people's ability to use generalization to guide their search we presented them with tasks that contained a large number of choices and a relatively limited number of actions, meaning that generalizing over previous observations is necessary for optimal performance. We chose a simple structure to ensure it would be possible for participants to learn and exploit it when maximizing rewards.

Our first two experiments focus on sequential tasks where participants had no prior information about the underlying reward structure, and where a combination of exploration – to discover task structure and discover optima – and exploitation is necessary to do well. The next two experiments provided participants with training about the reward structures before the task itself. In all of these experiments, we found that some participants selected actions that resolved uncertainty about the underlying structure of the task, and traded off between exploration and exploitation in order to maximize

reward. These participants were also able to transfer knowledge across tasks and gradually improved their performance. We also found a significant number of participants engaged in purely exploratory behavior, consistently preferring to choose novel actions, even when these actions were relatively unrewarding. These results highlight the importance of studying individual differences to better identify the multiple factors that influence human behavior, and of accommodating these differences in models of learning and exploration.

Experiment 1

Across our four experiments participants were given a sequence of grids composed of 9-by-9 arrays of tiles (see Figure 1), with each tile corresponding to a possible action. In this paper, we limit our analysis to the first three grids presented to participants (out of nine), as the latent task structure changed after that point. The grids studied here shared a similar underlying task structure: they had the same kind of relationship between features and rewards, but details of those relationships varied. In our experiment an action consists of selecting an individual tile, which has two features: its horizontal (x), and vertical position (y). Participants had to select tiles to maximize their cumulative rewards over 20 choices in each grid. The task presents a classical explore-exploit tradeoff: Succeeding requires carefully balancing between choosing new tiles to learn about the underlying reward structure or re-selecting tiles that were observed to be rewarding. In Experiment 1, participants received no prior knowledge about the reward structure of the tasks, nor about whether the tasks were related to one another in any way.

We predicted participants would be able to generalize from previous observations and improve by using their growing knowledge of the underlying task structure to select better actions. We measure this by looking at whether participants were able to select more rewarding tiles as they collected more information, and whether they demonstrated confidence in their knowledge by repeatedly selecting (i.e., exploiting) optimal actions. Our second hypothesis was that participants would be able to re-use knowledge across grids, since they shared the same structure, and thus improve their performance from one grid to the next.

We also studied the distance between participants' selections throughout the task to better understand their behavior. Distance between selections is a useful marker of different exploration strategies. For example, participants who seek to reduce uncertainty about the task structure are likely to select tiles that are far apart from each other, as these tend to yield more information about the broad shape of the reward function, in addition to having more uncertain rewards themselves. We call these selections *globally informative* actions. In contrast, participants might sample tiles adjacent to their previous observations, e.g., because they believe they are close to a maximum or because they want to observe local gradients. We call this kind of selection *local search*.

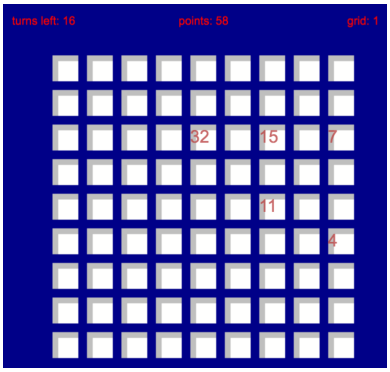

Figure 1: Screenshot of grid presented to participants after 5 observations. Note that in Experiment 1, the rewards disappear shortly after a tile has been selected.

Methods We recruited 79 participants using Amazon's Mechanical Turk service. They received \$0.75-\$1, which was doubled for participants whose final scores were in the top 10 percent. Following the instructions given to participants, we excluded participants whose performance was worse than chance ($n = 3$). We also excluded participants who failed to select more than 2 different tiles on the majority of grids ($n = 5$), as it showed a lack of engagement with the task.

The three grids analysed here used a reward structure where one location (x_m, y_m) was sampled uniformly at random in each grid, and the grid's maximum reward m was sampled from $(\mathcal{N}(\mu = 200, \sigma^2 = 50))$. The reward r for a given tile location (x, y) was exponentially decreasing with its Euclidean distance d from that maximum-reward tile: $r(x,y) = C \cdot e^{-k \cdot d((x,y),(x_m,y_m))}$, rounded to the nearest integer. We chose an exponential relationship between features and rewards to ensure there would be a clear advantage for participants who discovered the maximum-reward tile. We chose a constant $(k = 0.4)$ that led to large differences between the maximum and its closest neighbors while making it unlikely that any tiles would have rewards of zero or one. We used a random maximum reward in order to make it difficult for participants to know they had found the most rewarding tile without knowing the reward structure of the task.

When a tile was selected, the reward was displayed on the tile for 1.5 seconds and added to the cumulative score on the current grid. Participants could re-select tiles they had previously chosen. Participants were given no information about the underlying structure of the grid prior to the task, and were not informed that the tasks were related in any way, apart from a note that there could be patterns behind the rewards.

Results and Discussion For this and all subsequent experiments, we report the normalized scores (between 0 and 1), by dividing each reward by the maximum reward in that grid. We were first interested in seeing whether participants were able to recognize similarities between tasks. We use a general linear model (GLM), with the reward as outcome variable. The turn and grid index were used as predictor variables. Both the turn ($b = 0.02, se = 0.001, p < 0.001$) and the grid ($b = 0.05, se = 0.005, p < 0.001$) were significant factors. Following our hypothesis, participants selected better

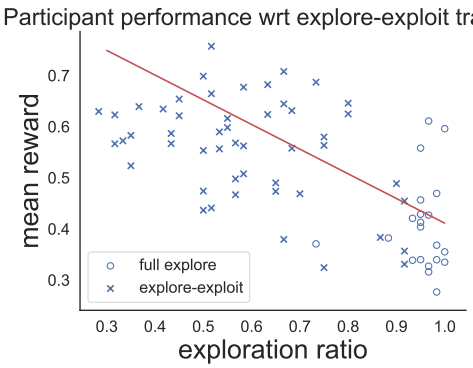

Figure 2: Each point represents a participant. The y-axis is the average reward across all three grids. The x-axis is the proportion of novel selections across all three grids. A value of 1 would mean only selecting new tiles, 0 only selecting the previously-selected tiles.

tiles over time, suggesting that they were able to exploit the underlying reward structure. Participants also improved their performance across grids, suggesting they were able to transfer structural knowledge across tasks (see Figure 3).

As a simple measure of a participant's propensity to explore, we used the proportion of actions that selected a previously-unseen tile ("exploration") versus re-selecting a previously-seen tile ("exploitation"). This distinction is more natural in our tasks than in a traditional stochastic bandit task, as in the latter it can be informative to re-select previously-seen tiles to learn about their reward distributions. There were significant behavioral differences indicated by how people traded off between exploration and exploitation among participants, and in the cumulative rewards they collected ($M = 0.49, SD = 0.30$) (see Figure 2).

Twenty-two participants (31 percent) never re-selected tiles more than twice in the majority of grids. We call these participants *full explore* (FE) participants. We call the other participants (n=49), that traded off exploration and exploitation, *Explore-Exploit* (EE) participants.

EE participants improved across tasks ($b = 0.07, se = 0.006, p < 0.001$) (see Figure 3), supporting our hypothesis that participants who used the underlying task structure to direct their search and maximize reward were able to re-use what they had learned to a new task.

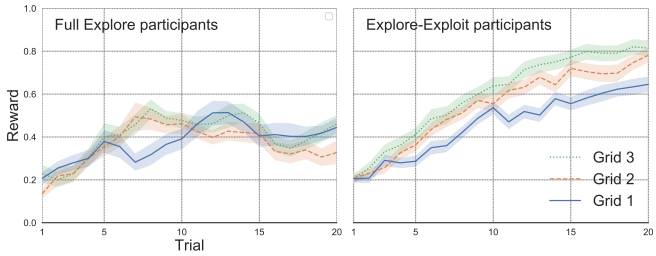

Figure 3: Performance of FE participants (n=22) and EE participants (n=49) in Experiment 1 across all three grids. Error bars in this and all subsequent plots reflect standard errors of the mean.

Across all participants, the proportion of exploratory selections correlated negatively with score ($r(140) = -0.71, p < 0.001$), and FE participants earned lower scores than EE participants ($t(69) = 5.77, p < 0.001, d = 0.15$). Their average scores barely improved from one grid to the next (Figure 3; $b = 0.02, se = 0.008, p = 0.06$).

We used a logistic regression model to evaluate participants' ability to find the maximum across grids. More participants found the maximum as they went on with the grids, hinting that they were better at utilising the underlying task structure ($b = 0.64, se = 0.11, p < 0.001$). Whether participants were engaging in *full exploratory* or *explore-exploit* strategies did not predict if they found the maximum in the tasks ($b < 0.001$). Participants were significantly better than chance at finding the maximua (0.65 of grids, vs. upper bound chance proportion of 0.25; $\chi^2(1, N = 1174) = 188.1, p < 0.001$). Furthermore, participants had overall a strong 'local bias' in their sampling, where they choose tiles close to their last choice more often than chance given the distribution of inter-tile distances ($t(151) = -50.8, p < 0.001, d = -2.34$) (see Figure 4). This suggests that participants engaged in local search strategies, rather than globally informative actions. Both EE and FE groups showed this bias, with adjacent tiles selected in 49% of FE participants' exploratory choices ($SD = 0.17$) and 39% for EE participants ($SD = 0.17$).

In conclusion, Experiment 1 showed that some participants were able to learn the underlying task structure when it was new and traded off between exploration and exploitation to maximize their rewards. These participants transferred knowledge across tasks that shared similarities in their underlying structure. However, a large proportion of participants had a strong tendency to explore in circumstances where exploitation would have yielded much higher scores, preferring unobserved tiles over known tiles with a high reward value. FE participants presented some evidence for learning the underlying structure, but this was not reflected in their score. Why did so many participants adopt such an extreme exploratory policy? One possibility is that they were motivated to learn more about the reward structure, or ensure they had found the maximum possible reward, in line with the inherent curiosity bias observed in people (Kidd & Hayden, 2015; Gottlieb et al., 2013).

We also observed a locality bias in participants' choices. This may have been due to the memory demands of the task. Wu et al. (2018) presented evidence that participants displayed an ability to use generalization to direct their search. Unlike the task used in their study, our task had the rewards disappear after participants selected a tile. Remembering past observations when generalizing might be difficult, and could have led participants to adopt policies that alleviated the complexity of the task. For example, if participants tracked local gradients in rewards and followed increasing rewards, this would only require tracking 2-3 past observations while being less demanding than computing a surrogate model over the general task structure. This would be consistent with the local search strategies exhibited in other domains such as causal learning (Bramley et al., 2015) and category learning (Markant et al., 2016), and the idea that people adapt their high-level strategies to make the most of limited resources (Lieder et al.,

2014). For FE participants, the local bias during exploration could reflect a systematic and memory-efficient policy for exhaustively searching a subset of the tiles for a maximum.

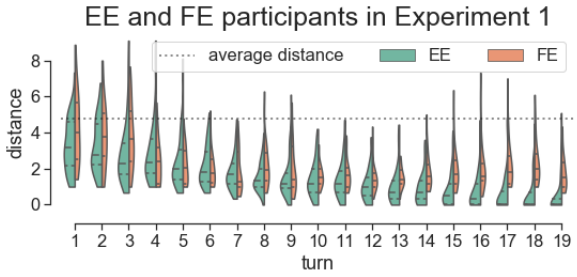

Figure 4: Average distance between selections of EE and FE participants in Experiment 1 presented with quantiles and kernel density estimations. We use Euclidian distance between selections, with 0 counting for a re-selection of the previous click. The dotted line represents the average distance between all tiles in a grid. The shape of the distribution is drawn using a (normal) Gaussian Kernel Density Estimate cut at 0.

In Experiment 2, we presented participants with the same task structure as in Experiment 1, but with changes designed to understand and potentially reduce their strong tendency to explore new tiles. These included persistent indicators of explored tiles' rewards, checks of participants' understanding of the instructions, and different incentives.

Experiment 2

In this experiment, the reward associated with a given tile is displayed continuously once it has been observed. We hypothesized that with participants observations remaining visible, the overall reward pattern would be more evident. We predicted that participants would be able to make more globally informative actions (i.e. exploratory selections would be more distant from each other). Because the underlying structure is made more evident, we also assumed fewer participants would engage in *full exploratory* behavior.

Methods We recruited 72 participants using Amazon's Mechanical Turk service identically to Experiment 1. Participants all received a base payment of $0.75. The reward scheme differed from that in Experiment 1: rather than granting bonuses to the top 10 percent, we gave all participants a bonus proportional to their cumulative score, up a maximum of $0.75. We excluded participants who failed to select more than 2 different tiles on the majority of grids ($n = 4$). In Experiment 2 when a tile is selected by a participant the reward is continuously displayed on the tile and is added to the current cumulative score on the current grid.

In another change from Experiment 1, tiles' rewards were persistently visible after they had been selected, under the logic that it might improve participants' ability to learn the underlying reward structure and increase their ability to find and exploit the maximum. We also added explicit instructions that participants could re-select tiles, and added a pre-task questionnaire to make sure participants understood these instructions. The questionnaire also required participants to understand that their goal was to maximize reward (as opposed to discovering the underlying pattern, or finding the maximum. Participants were not allowed to proceed with the task until they answered all questions correctly.

Results and Discussion Contrary to our predictions that participants would be less prone to *full exploratory* behavior, a significantly larger proportion of participants showed FE behavior in Experiment 2 as compared with Experiment 1 (.47, $n = 32$ vs. .31, $n = 22$; $\chi^2(1, N = 139) = 18.6, p < 0.001$). As in Experiment 1, the proportion of exploratory selections correlated negatively with performance ($r(134) = -0.75, p < 0.0001$). In Experiment 2, EE participants also performed significantly better than FE participants ($t(66) = 9.31, p < 0.0001, d = 0.23$) and improved significantly across tasks ($b = 0.04, se = 0.007, p < 0.0001$), whereas FE participants did not ($b = 0.01, se = 0.006, p = 0.14$).

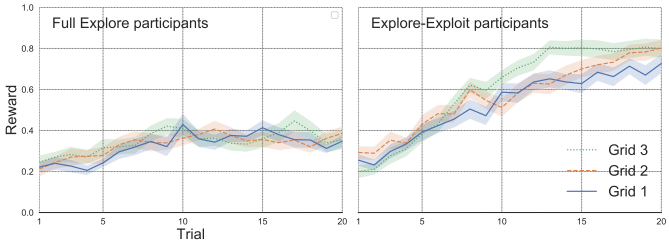

Figure 5: Performance of FE participants ($n = 49.$) vs EE participants ($n = 36$).

To understand the effect of having observations available throughout the task, we compare the performance of EE participants in Experiment 2 (n=36) to the performance of EE participants in Experiment 1 (n=49). Overall, EE participants in Experiment 2 (M=0.58) did slightly better than EE participants in Experiment 1 (M=0.56) ($b = 0.04, se = 0.008, p < 0.001$). This was most pronounced in the first grid ($t(84) = 2.18, p = 0.03, d = 0.08$). We conjecture that EE participants in Experiment 2 learned the reward pattern faster, and EE participants caught up in subsequent grids. This supports the hypothesis that visible observations allowed participants to generalize better, by supporting more global strategies. To test this idea, we looked at the inter-selection distances between the first 5 selections of participants. EE participants' choices in Experiment 2 were more global, with greater distances than EE participants' choices in Experiment 1 ($t(84) = -2.25, p = 0.03, d = 0.66$) (see Figure 6).

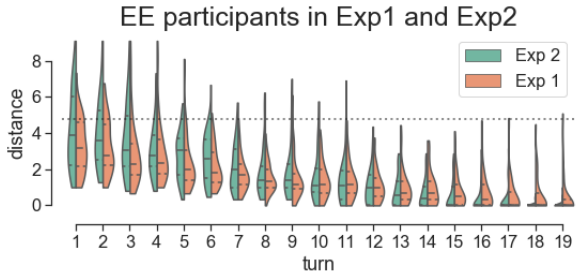

Figure 6: Comparison of distances between selections of EE participants in Experiment 1 and Experiment 2 (see Figure 4 for details). EE participants in Experiment 2 selected more "global" actions (longer distances between selections) during their first actions.

Why did more participants engage in FE behavior in Ex-

periment 2? We conjectured that participants were more motivated to observe rewards for new tiles when previous rewards remained visible, because the overall pattern – and the possibility of better understanding it – might have been more salient to them.

In Experiment 3, we sought to understand why some participants might want to select new tiles almost exclusively, rather than occasionally exploiting what they had learned to earn greater rewards. After Experiment 1, we hypothesized that this might have been due to an intrinsic epistemic drive in participants. Experiment 2 showed that for EE participants were able to leverage visible observations to conduct more global exploration, and led to a better overall performance. However, the observable rewards also seemed to add an incentive for many participants to exclusively choose novel actions, rather than maximising rewards. We hypothesized that this would only be the case for new tasks when participants had no prior knowledge about the underlying reward structure of the tasks, since new observations would not be very informative if participants had a prior about the underlying reward structure.

Experiment 3

We designed Experiment 3 to control explicitly for the potential epistemic drive of FE participants by familiarizing them with the underlying reward structures prior to the task. By making the structure clear to participants prior to the tasks, our primary prediction for Experiment 3 was that fewer participants would engage in FE behavior. We presumed the intrinsic motivation of observing new observations would be attenuated when participants did not gain new information about the task from those observations.

We also hypothesized there would be weaker or no progress across grids since participants would already be familiar with the reward structure when they engage with the first grid. Because of the training, we predicted participants would be more efficient at finding and re-selecting tiles with high values, and would thus perform better overall than in Experiment 1 and 2. Experiment 3 was set up identically to Experiment 2. Participants were told about the underlying pattern and given three practice grids so they could learn the reward structure prior to the task.

Methods We recruited 43 participants using Amazon's Mechanical Turk service, identically to Experiment 2, with the following changes: Participants were only recruited for three grids rather than nine, following the same reward pattern discussed in Experiment 1 and Experiment 2. Because of the shorter duration, participants were paid a base reward of \$0.2. We used a proportionally larger bonus of \$0.6 under the logic that this would further reduce the effects of epistemic drive. Apart from the training grids presented prior to the task, instructions were identical to Experiment 2. During the training, participants were told that each grid had one maximum tile, and the closer a tile is to the maximum the higher the reward. The first training grid had all rewards displayed and participants were instructed to familiarize themselves with the nature of the task. The next two grids were similar to the grids in the actual task (i.e. only observed tiles display reward values) but participants were encouraged to learn the pattern as well as they could. Throughout the task, instructions regarding reward maximisation and the possibility of reselecting tiles were also displayed. We excluded one participant who failed to select more than two different tiles on the majority of grids and one participant who reported not following the instructions upon completing the experiment.

Results and Discussion Surprisingly, 37 percent (15 out of 41) of participants engaged in *Full Exploration* (FE) in Experiment 3. The proportion of FE participants in Experiment 3 was significantly less than the 47 percent we observed in Experiment 2 ($\chi^2(1, N = 109) = 8.82, p = 0.003$), but was nonetheless a higher proportion than anticipated.

As expected, EE participants in Experiment 3 did not improve significantly across grids, since they had been trained extensively on the rule before the assessed task started ($b = -0.01, se = 0.008, p = 0.112$). The average performance of EE participants was significantly better than EE participants in Experiment 2 ($t(61) = 2.29, p = 0.03, d = 0.07$) and EE participants in Experiment 1 ($t(74) = 3.11, p = 0.003, d = 0.09$, suggesting that participants were able to learn the rule during the training and relied on this knowledge when faced with new grids in the task.

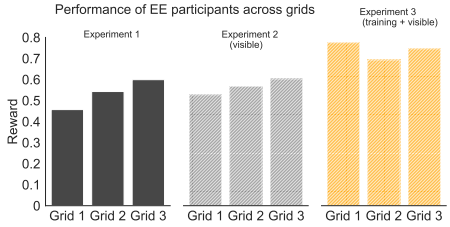

Figure 7: Average performance of EE participants (participants that traded off between exploration and exploitation) across all three grids in Experiment 1, 2 and 3.

To understand the effect of prior knowledge on participants' exploratory patterns, we compared how EE participants explored compared to EE participants in Experiment 2. Participants in Experiment 3 were significantly more locally biased in their initial five selections ($t(359), p < 0.001, d = 1.19$). Participants in Experiment 3 were already familiar with the *Location rule*, and it is probable that they were able to find the maximum by ascending towards the maximum through small incremental steps. EE participants in Experiment 3 had a significantly lower proportion of reselections (0.19 in Experiment 3 vs 0.28 in Experiment 2) ($\chi^2(1, N = 1367) = 17.16, p < 0.001$). Given their higher performance scores, EE participants in Experiment 3 were likely to have a strategy more adapted to the task than in Experiment 2, where participants were still learning the reward structure. Indeed, EE participants in Experiment 2 had a tendency to settle on a sub-optimal tile, finding the maximum tile in 0.62 of grids. EE participants in Experiment 3 took smaller exploratory steps but found the maximum in 0.81 of the grids

($\chi^2(1, N = 185) = 6.69, p = 0.01$).

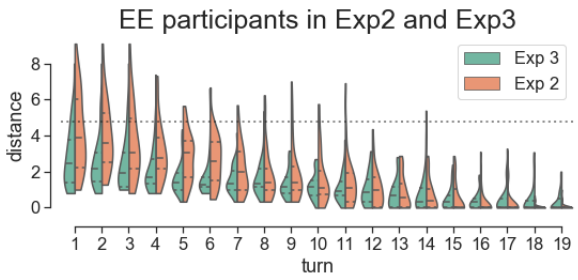

Figure 8: Distance between selections of participants (see Figure 4 for details). EE participants in Experiment 2 had more global observations than EE participants in Experiment 3. This can be explained by that fact that they had no prior knowledge about the task structure.

Contrary to our hypothesis, many participants still engaged in full exploratory behavior. Given this result, we hypothesized that participants might be motivated by observing new rewards rather than learning the underlying reward structure *per se* and that this effect might be emphasized when rewards remain visible after having been observed. Indeed, in Experiment 2, where rewards remained visible, significantly more participants engaged in full-exploratory behavior than in Experiment 1. We designed Experiment 4 to account for these two factors of epistemic motivation: 1) wanting to learn about the underlying reward structure and 2) wanting to attend novel information.

Experiment 4

Experiment 4 followed the design details of Experiment 3, except that rewards were not displayed continuously after they had been selected - they are displayed on the tile and disappear shortly after, like in Experiment 1.

Our main hypothesis for Experiment 4 was that fewer participants would engage in *full exploratory* behavior, since the epistemic reward is attenuated by not having the tiles visible after they have been selected and having training grids prior to the task. We predicted EE participants would perform similarly or slightly worse than in Experiment 3, because of the constraints of not having previous observations visible, but better than in Experiment 1 and 2. We also predicted we would observe little or no transfer effect across grids.

Methods 39 participants were recruited using Amazon Mechanical Turk. One participant was excluded for failing to select more than two different tiles, and one was excluded because their performance was worse than chance.

Results In agreement with our hypothesis, only one participant out of 37 engaged in *Full Exploration*. This was significantly less than in any other experiment. This supports the idea that participants' strategies were driven by an epistemic drive which was twofold:

First, participants were motivated to reveal the underlying reward structure, e.g., reducing the entropy about the structure of the task, or about the location of the maximum. Indeed, participants were less likely to engage in FE behavior in Experiment 4 (known structure and disappearing observations) than Experiment 1 (unknown structure and disappearing observations), and significantly less in Experiment 3 (known structure and visible observations) than Experiment 2 (unknown structure and visible observations).

Second, participants were motivated to observe the outcomes of individual actions. In Experiment 1,2 and 3 a significant proportion of FE participants selected the maximum but consistently opted for selecting novel options rather than re-selecting a previous maximum observation, with a preference for actions that were local to their last one. Participants' drive to select novel actions was enhanced by the fact that information did not need to be kept in working memory. They were less engaged in FE behavior in Experiment 1 (non-visible observations) than Experiment 2 (visible observations), and, similarly, less in Experiment 4 (non-visible observations) than Experiment 3 (visible observations). Though EE participants in Experiment 3 performed slightly better than in Experiment 4, this was not significant ($t(61) = 0.93, p = 0.35, d = 0.04$). Participants in Experiment 4 improved their average performance slightly across tasks ($b = 0.02, se = 0.007, p = 0.02$).

The average distance between the initial five exploratory selections of EE participants was not significantly different in Experiment 3 and Experiment 4 ($t(309) = -0.90, p = 0.37, d = -0.15$). EE participants in Experiment 4 explored significantly more locally than EE participants in Experiment 1 ($t(374) = -2.73, p = 0.007, d = 0.47$). Like in Experiment 3, this supports the hypothesis that participants who were familiar with the underlying structure of the grid were able to find the maximum by taking local exploratory steps until they eventually found the maximum.

Conclusion

In this paper, we focused on the behavioural analysis of participants across four experiments to study how people learn to select rewarding actions in a sequence of novel tasks. We found that some participants were able to learn the underlying structure while balancing exploration and exploitation to maximize their rewards across tasks. They improved their performance from one task to the next by transferring abstract knowledge about their environment. However, consistently across tasks, we observed that a significant proportion of participants engaged in purely exploratory behavior, largely ignoring the reward incentive. We showed that this behavior could be manipulated by controlling the availability of information as the learner selected actions, and by giving prior knowledge before participants engaged with the task. We suggest that people are motivated by two types of epistemic drives: 1) to reduce uncertainty and learn about the structure of the task and 2) to observe new evidence, regardless of its informativeness about the global task structure. The latter was evident when participants continued valuing new actions over maximising rewards, even when they were familiar with the task structure.

Different mechanisms for curiosity have been discussed in the literature, and could be connected to how people learn

in new environments when combined with trying to achieve goals or maximising utility. One such strategy is to entirely dismiss reward feedback, giving rise to a strong novelty drive. This *novelty search* mechanism has been been shown to be very successful in the context of Evolutionary Strategies for tasks with tricky reward functions (Lehman & Stanley, 2011). Some studies have shown that people are biased towards surprise (Gottlieb et al., 2013; Itti & Baldi, 2006). Selecting new actions would make sense under the assumption of possible change, or if one believes that the environment is trying to fool us. Third, the idea of *epistemic actions* could explain part of people's strong drive to select new actions, especially under the constraint of cognitive load, when storing observations is expensive or unrealistic. Epistemic actions refer to actions *in the world* that help solve problems by changing the mental state of the agent, as opposed to performing computations in the head (Kirsh & Maglio, 1994). An example of this behavior is the use of sticky-notes, or of arranging documents in a way that makes it easier to retrieve them rather than by memory alone. In the case of our experiment, observing new information might have been perceived as much cheaper than the possibility of generalizing from few observations.

In our study, we highlight that studying individual differences amongst participants can help us better understand the complex mechanisms at play during active learning in new environments. We hope that by pointing out surprising facets of human behavior, this empirical study can guide the design of better computational models of human learning and exploration. We are currently investigating how computational models of memory, generalization and search (León-Villagrá et al., 2018; Wu et al., 2018; Lucas et al., 2015) can give us further insight into people's representations and strategies when learning in new environments.

References

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In *Advances in neural information processing systems* (pp. 55–63).

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Coenen, A., Nelson, J. D., & Gureckis, T. (2017). Asking the right questions about human inquiry.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers? exploratory play. *Cognition*, *120*(3), 341–349.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, *111*(1), 3.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, *17*(11), 585–593.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Group, C. S. R., et al. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46–54.

Itti, L., & Baldi, P. F. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems* (pp. 547–554).

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, *18*(4), 513–549.

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, *19*(2), 189–223.

León-Villagrá, P., Preda, I., & Lucas, C. G. (2018). Data availability and function extrapolation. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in neural information processing systems* (pp. 2870–2878).

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, *40*(1), 100–120.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4).

Ruggeri, A., & Lombrozo, T. (2014). Learning by asking: how children ask questions to achieve efficient search. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1335–1340).

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2017). Putting bandits into context: How function learning supports decision making.

Schulz, L., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, *43*(4), 1045.

Teodorescu, K., & Erev, I. (2014). On the decision to explore new alternatives: The coexistence of under-and over-exploration. *Journal of Behavioral Decision Making*, *27*(2), 109–123.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915.