# Modelling Individual Differences in Exploratory Strategies: Probing into the human epistemic drive

**Nicolas Collignon**
School of Informatics
University of Edinburgh
n.collignon@ed.ac.uk

**Christopher Lucas**
School of Informatics
University of Edinburgh
clucas2@inf.ed.ac.uk

## 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 novel tasks. We examine how memory constraints and prior knowledge affect this drive to explore by studying the exploratory strategies of people across four experiments. We find that some people were able to learn new reward structures efficiently, selected globally informative actions, and could transfer knowledge across similar tasks. However, a significant proportion of participants behaved sub-optimally, prioritizing collecting new information instead of maximizing reward. Our evidence suggests this was motivated by two types of epistemic drives: 1) to reduce uncertainty about the structure of the task and 2) to observe new evidence, regardless of how informative they are to the global task structure. The latter was most evident when participants were familiar with the task structure, hinting that the drive to gather knowledge can be independent of learning an abstract representation of the environment. This was not the case when observations did not remain visible to participants, suggesting that participants may adapt their exploratory strategies not only to their environment but also to the computational resources available to them. Our initial modelling results attempt to explain the different cognitive mechanisms underlying human exploratory behaviour across tasks, and are able to capture and explain systematic differences across conditions and individuals.

**Keywords:** active learning; generalization; exploration-exploitation; heuristics; transfer learning;

## 1 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 [1], and play with new toys in ways that help them disambiguate uncertain causal relationships and gather information [2, 3]. The idea that humans learn and interact with their environment by performing intuitive experiments, maximizing information gain, is a popular one [4, 5, 6, 7].

In this work, 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 perform?

When faced with 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 [8]. 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 yielding stochastic rewards, so as to maximize 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), each arm has a set of features that may be informative of the arm's reward distribution. Learning how features relate to rewards allows for an efficient representation of the environment, and enables the learner to generalize to new events.

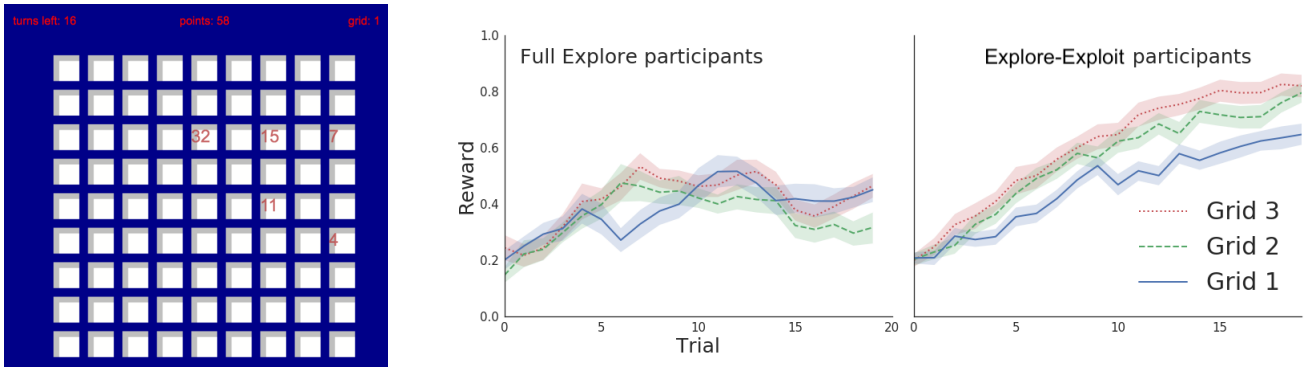
We report on two experiments where people have to find rewarding actions in a sequence of tasks, and where the reward structure is unknown. We compare them to cases where participants were trained on the reward structures prior to the task. We find evidence that some participants selected actions that resolve 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. Conversely, a significant proportion of participants engaged in pure exploratory behavior, consistently preferring to attend novel information rather than maximizing rewards. We highlight the importance of studying individual differences when studying human learners and identify independent factors of epistemic drive that guide human exploration.

## 2 Experiment 1

We designed our initial experiment to look at how participants adapt to change of reward structure, and detect similarities between tasks, with the hypothesis that people's behavior would be well accounted by Bayesian models. What we find instead is that, contrary to previous studies, the behavior of many participants deviated from those models' predictions.

To better understand this phenomenon, we focused on the first three tasks each participant completed, which shared a similar underlying reward structure. Participants were given a sequence of grids of 9x9 tiles, with each tile corresponding to a possible choice. Participants had to select tiles to maximize their cumulative rewards over 20 choices in each grid. This presents a classical explore-exploit trade-off: Succeeding in the task 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 each grid, contextual features (x,y) predicted for rewards. When a tile is selected, the reward is displayed for a short period of time and is added to the cumulative score on the current grid. Participants were given no information about the underlying structure of the grid prior to the task, apart from the fact that there may be patterns behind the rewards across tiles. In the game, it is possible to re-select a tile repeatedly, and contrary to traditional bandit tasks, rewards were deterministic for any given tile. This was done to ensure actions were distinctly either exploratory or exploitative (as opposed to a stochastic case, where one could re-select an option to learn about it's volatility.).

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, we observed that a large proportion of participants had a strong tendency to over-explore, preferring unobserved tiles over known tiles with a high reward value. Twenty-two participants (31 percent) never re-selected tiles more than twice in any of the grids. We call these participants *Full eExplore* (FE) participants. We call the other participants (n=49), that traded off exploration and exploitation, *Explore-Exploit* (EE) participants. We plot the performance of EE and FE participants across all three grids in Figure 1. Further,




(a) Game screenshot

(b) Experiment 1 performance

Figure 1: (a) The grid presented to participants after 5 observations. Note that in Experiment 1, the rewards disappear shortly after a tile has been selected. (b) Performance of FE participants (n=22) and EE participants (n=49) in Experiment 1 across all three grids. The plotted confidence interval corresponds to the standard error (ci=68%).

participants had overall a strong 'local bias' in their sampling. Both EE and FE groups showed this bias, with adjacent tiles selected in 49% of FE participants' exploratory choices and 39% for EE participants.

To explain the large proportion of FE participants, we hypothesized participants may have been driven by wanting to learn more about the reward structure and collect information. This would be consistent with the local search strategies exhibited in other domains such as causal learning [9], category learning [10], or more generally with people's inherent curiosity bias [11, 12]. We hypothesized that this would only be the case for new tasks when participants still had something to learn about the underlying reward structure of the tasks.

## 3 Experiment 2

Experiment 2 was identical to Experiment 1, but with the reward displayed continuously once a tile has been observed. We added comprehension questionnaires and changed the reward scheme to rule out the alternative explanations about participants' extreme exploratory behavior. We hypothesized that with participants observations remaining visible, the overall reward pattern would be more evident. Thus, participants would be more likely to re-select tiles with high values and perform better than in Experiment 1. Because the underlying structure was more evident, we also assumed fewer participants would engage in *full exploration* behavior, since their curiosity drive would be less pronounced. We also hypothesized that participants would be able to make more globally informative actions (i.e. exploratory selections would be more distant from each other).

Against our expectations, participants were overall more prone to engage in *full exploratory* behavior than in Experiment 1. It could be that participants were further motivated to collect more observations when they remained visible, as the pattern might have been more salient to them and allowed them to learn better. Following our hypothesis that visible observations allowed participants to generalize better, EE participants in Experiment 2 had more global exploratory selections at the beginning of each grid. This could explain their better average performance on the first grid when compared to those in Experiment 1.

## 4 Experiment 3

In Experiment 3, we tried to understand the large proportion of participants that engaged in full exploratory behavior. After Experiment 1, we hypothesized that this might have been due to an intrinsic epistemic drive in participants. We controlled for several alternative hypotheses, such as memory constraints, unclear instructions, or reward incentives, but this led to more participants engaging in pure exploratory behavior. We designed Experiment 3 to control explicitly for the potential epistemic drive of FE participants. To do this, we explicitly instructed participants about the relationship between a tile's location and the corresponding reward, 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 *full exploratory* behavior, since the epistemic reward would be largely attenuated. We also hypothesized there would be weaker or no progress across grids since participants would already be familiar with the reward structure from the first grid. With this 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 except from the addition of a training step where participants were given one practice

grid where all the rewards were continuously displayed, then two further practice grids, similar to the actual task grids, so that they could learn the underlying pattern prior to performing the task.

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 have been emphasized by the fact that rewards remained visible after having been selected once. Indeed, in Experiment 2, where rewards remained observable, significantly more participants engaged in full-exploratory behavior than in Experiment 1. We designed Experiment 4 to account for both factors of epistemic motivation: 1) wanting to learn about the underlying structure or the location of the maximum, and 2) wanting to observe novel information.

## 5 Experiment 4

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.

In agreement with our hypothesis, only one participant out of 37 engaged in *Full Exploration*. This was significantly less than in any other condition. 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. 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, with a preference for actions that were local to their last one

Participants' drive to reduce local uncertainty was enhanced by the fact that information became available once it has been observed once. They were engaged less in FE behavior in Experiment 1 (non-visible observations) than Experiment 2 (visible observations), and less in Experiment 4 (non-visible observations) than Experiment 3 (visible observations).

## 6 Computational Modelling: Initial Results

We are currently investigating how computational models of memory, generalization and search can give us insight into people's representations and strategies when learning in new environments. Besides the important differences across experiments, we are also interested in investigating the differences in behaviour of participants from the same experimental condition. People's explore-exploit strategies have been shown to carry significant differences across individuals [13]. More generally, advances in statistical and modelling tools has led to an increased interest understanding qualitative differences in how people think and act [14].

We outline briefly the different components used in our model to capture different mechanisms of human behavior. To model directed search, we use the predictions of Gaussian Process (GP) with an RBF Kernel. We take a fully Bayesian treatment of the GP kernel hyperparameters, as presented in [15]. GPs have been successful in explaining human function learning phenomena [16, 17], unifying conflicting theories about how humans learn functions. More recently they have also been applied to study decision making in multi-armed bandit problems [18]. We define a greedy weight component that assigns a probability weight to reselect the currently maximum known value. To account for the local bias observed in participants, we use the inverse Manhattan distance (IMD) to the last observation and fit with a softmax temperature parameter to individual participants. We also add a negative weight on previous observations and a random exploratory term (uniform probability for all observations). Models are fit to individual participants by using a Differential Evolution algorithm to maximise the maximum likelihood function. We use an L1 penalty on all weight parameters and an exponential penalty on the local-bias temperature parameter for more interpretable models. We map the resulting models in Figure 2 to highlight clusters of behaviours across all four experiments. Table 1 presents the parameters of cluster centroids obtained after running a Gaussian Mixture Model over all participants, as plotted in Figure 2. The results show that we can obtain interpretable parameters that are consistent with observed the participant behaviors.

## 7 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

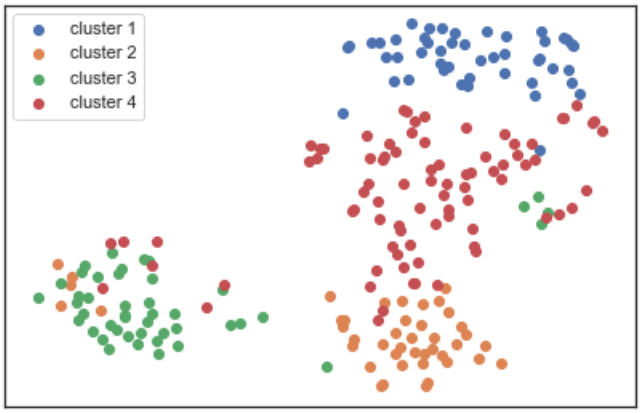



Figure 2: t-SNE visualisation of MLE parameters for individual participants across all 4 experiments. Clusters are obtained via a Gaussian Mixture Model. Cluster centroids are reported in Table 1.

| | $\alpha$ directed search ($E[x]$ under the GP) | $\beta$ global search ($\sigma^2$ under GP) | Greedy weight (reselecting max-known) | local-bias weight | local-bias temperature | dampen previous observations | random exploration |
|---|---|---|---|---|---|---|---|
| cluster 1 | 0.3 | 0.06 | 0.22 | 0.03 | 75.13 | 0.15 | 0.24 |
| cluster 2 | 0.02 | 0 | 0.12 | 0.35 | 26.15 | 0.3 | 0.21 |
| cluster 3 | 0.1 | 0 | 0.08 | 0.51 | 7.21 | 0.03 | 0.28 |
| cluster 4 | 0.22 | 0.04 | 0.19 | 0.19 | 1.57 | 0.14 | 0.18 |

Table 1: Parameters of cluster centroids of Gaussian Mixture Model. Weight parameters are normalised (i.e. all but the local-bias temperature). These results show that we can obtain interpretable parameters that are consistent with observed the participant behaviors. E.g. Cluster 1 corresponds to EE participants with global exploration, cluster 3 corresponds to FE participants with a strong local bias.

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. 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.

## References


[1] Azzurra Ruggeri and Tania Lombrozo. Learning by asking: how children ask questions to achieve efficient search. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 1335–1340, 2014.

[2] Laura Schulz and Elizabeth Baraff Bonawitz. Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, 43(4):1045, 2007.

[3] Claire Cook, Noah D Goodman, and Laura E Schulz. Where science starts: Spontaneous experiments in preschoolers? exploratory play. *Cognition*, 120(3):341–349, 2011.

[4] Anna Coenen, Jonathan D Nelson, and Todd Gureckis. Asking the right questions about human inquiry. 2017.

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

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

[7] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[8] Thomas T Hills, Peter M Todd, David Lazer, A David Redish, Iain D Couzin, Cognitive Search Research Group, et al. Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, 19(1):46–54, 2015.

[9] Neil R Bramley, David A Lagnado, and Maarten Speekenbrink. Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3):708, 2015.

[10] Douglas B Markant, Burr Settles, and Todd M Gureckis. Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, 40(1):100–120, 2016.

[11] Celeste Kidd and Benjamin Y Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460, 2015.

[12] Alison Gopnik. Explanation as orgasm. *Minds and machines*, 8(1):101–118, 1998.

[13] Mark Steyvers, Michael D Lee, and Eric-Jan Wagenmakers. A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, 2009.

[14] Daniel J Navarro, Thomas L Griffiths, Mark Steyvers, and Michael D Lee. Modeling individual differences using dirichlet processes. *Journal of mathematical Psychology*, 50(2):101–122, 2006.

[15] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[16] Christopher G Lucas, Thomas L Griffiths, Joseph J Williams, and Michael L Kalish. A rational model of function learning. *Psychonomic bulletin & review*, 22(5):1193–1215, 2015.

[17] Eric Schulz, Josh Tenenbaum, David K Duvenaud, Maarten Speekenbrink, and Samuel J Gershman. Probing the compositionality of intuitive functions. In *Advances In Neural Information Processing Systems*, pages 3729–3737, 2016.

[18] Charley M Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12):915, 2018.