Belief state-based decision making under uncertainty in a spatial search task

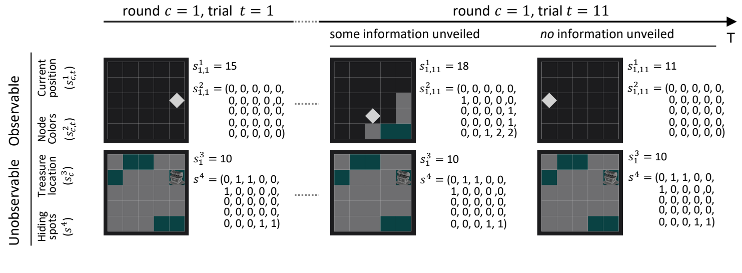
About the author: I, the applicant, am the main contributor to this project. It was my master thesis project and together with my advisor, I am currently preparing the manuscript for publication.

Introduction

Many natural sequential decision scenarios can be conceived as search problems, where the individual pursues the goal to find rewards at unknown locations in an environment with not directly observable reward structures Hills and Dukas [2012]. To maximize their overall reward, humans are required to seek information about their environment (exploration) and use gained knowledge to arrive at locations with high reward densities (exploitation). As exploration typically comes at the expense of time or energy that could otherwise be used for exploitation, they often face a trade-off between pursuing immediate rewards and improving future decisions Cohen et al. [2007], Mehlhorn et al. [2015]. The present work introduces a novel spatial search paradigm (*Treasure Hunt*) that simulates such scenarios, collected experimental choice data of 50 participants, and presents a computational framework that uses neuroscience-inspired artificial agent models that can be tested against human choice data. The agents are formulated as Bayesian decision models Ma [2019] which learn and maintain an internal model of the environment that represents "true" world states by belief states. Newly gained information is incorporated into existing knowledge by Bayesian inference. To balance the value of information and reward the agents add an "information bonus" Gershman [2019], to the expected reward associated with an action. Following the Bayesian framework, we quantify this bonus by means of the expected Bayesian surprise Itti and Baldi [2009], Sun et al. [2011], an information theoretic quantity measuring the effect of new information on the internal representation by means of the divergence between prior and posterior beliefs after updating(ibid.).

Methods

Treasure Hunt task. The goal is to find rewards (treasures) in an environment with unknown (i.e. not observable) location-specific reward structures (hiding spots), but which can be unveiled (i.e. made observable) in the course of the game by choosing informative actions (drilling). On each trial, participants could either *step* on a neighbouring node or *drill* on their position, both at the expense of one move. Drilling changed the node color to green (hiding spot) or grey (no hiding spot). Unveiling and deliberately targeting hiding spots increased participants' chances to find treasures. Fig. 1 shows examples of observable and unobservable states at the beginning and towards the end of one round if the participant either did or did not unveil any information. One run of the game consisted of 10 rounds and in each round, participants had 12 moves to find the treasure.



round c = 1, trial t = 1
round c = 1, trial t = 11
some information unveiled
no information unveiled
T
Observable
Unobservable
Current position ($s^1_{c,t}$)
Node Colors ($s^2_{c,t}$)
Treasure location (s^3_c)
Hiding spots (s^4)
$s^1_{1,1}$ = 15
$s^2_{1,1}$ = (0, 0, 0, 0, 0, 0, 0, 0, 0 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
s^3_1 = 10
s^4 = (0, 1, 1, 0, 0, 1, 0, 0, 0 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1)
$s^1_{1,11}$ = 18
$s^2_{1,11}$ = (0, 0, 0, 0, 0, 1, 0, 0, 0 ,0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 2, 2)
s^3_1 = 10
s^4 = (0, 1, 1, 0, 0, 1, 0, 0, 0 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1)
$s^1_{1,11}$ = 11
$s^2_{1,11}$ = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
s^3_1 = 10
s^4 = (0, 1, 1, 0, 0, 1, 0, 0, 0 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1)

Figure 1: **Task design und state representations.** s^1 is the agent's current position corresponding to nodes $i = 1, ..., 25$. s^2_i denotes each node's color, where each entry s^2_i of the vector can take on the values 0 (black/not unveiled), 1 (grey/not hiding spot) or 2 (green/hiding spot). s^3 denotes the treasure location. s^4_i denotes the hiding spot locations, where each entry s^4_i of the vector can take on the value 0 (not hiding spot) or 1 (hiding spot)

Task model The task is formulated as a partially observable Markov decision process Bertsekas [2005]. The gridworld environment is represented by a graph with 25 nodes and 40 edges. One run is represented by the tuple

$$\mathcal{T} := (T, C, S, A, R, O, p^{s^1_{t+1}, s^3_c, a_t}(r_t), f, g), \tag{1}$$

where $T := 12$ denotes the trials per round and $C := 10$ the rounds per run. $S := s := (s^1, s^2, s^3, s^4)$ denotes the set of states (cf. Fig. 1). $A := \{0, -5, +1, +5, -1\}$ captures the set of actions, $R := \{1, 0\}$ the set of rewards and $O := \mathbb{N}^0_3$ denotes the set of observations. The deterministic reward distribution $p^{s^1_{t+1}, s^3_c, a_t}(r_t)$ specifies a reward probability of 1 if the new position is the treasure location, and zero if else. The functions f and g specifies state transitions and agent's observations, respectively. An agent model interacts with the task by making observations and choosing among a set of available actions.

Agent models. All agents have the general structure represented by the tuple

$$\mathcal{A} := (T, C, S, A, R, O, p_{c,1}(s^3_c, s^4), p^{s^1_{t+1}, a_t}\left(r_t | s^3_c\right), p_{c,t}^{a_{t-1}, s^1_t, s^2_t}\left(o_t | s^3_c, s^4\right)), \tag{2}$$

where T, C, S, A, R and O are defined as in $\mathcal{T}$. $p_{c,0}(s^3_c, s^4)$ denotes the agent's initial belief over the non-observable states. At the start, when no information is unveiled, the initial belief is defined as the discrete uniform categorical distribution over all possible combinations of s^3 and (s^4_i). In all subsequent rounds, the initial belief corresponds to a conditional discrete uniform distribution over all possible s^3 given the posterior marginal belief over s^4 of the last trial of the preceding round $c - 1$.

Intuitively, the agent maintains its belief about hiding spots (s^4) over rounds, while it resets its belief about the treasure location (s^3) based on its maintained hiding spots beliefs. $p^{s^1_{t+1},a_t}\left(r_t|s^3_c\right)$ specifies the agent's uncertainty over the reward. Finally, $p^{a_{t-1},s^1_t,s^2_t}_{c,t}\left(o_t|s^3_c,s^4\right)$ is the deterministic observation distribution, also referred to as likelihood function. The control agents (C1, C2 and C3) are probabilistic and belief state-free, while Bayesian agents (A1, A2 and A3) are deterministic and evaluate the desirability of actions based on their current belief states which they update trial-by-trial.

Belief state updates. The posterior belief is recursively evaluated based on the agent's prior beliefs and the likelihood, in line with the framework of Bayesian inference (c.f. Bertsekas and Tsitsiklis [2008]). Formally,

$$p^{a_{1:t-1},s^1_{1:t},s^2_{1:t}}_{c,1:t}\left(s^3,s^4|o_{1:t}\right)=\frac{p^{a_{t-1}}_{c,t}\left(o_t|s^3_c,s^4\right)p_{c,1:t-1}\left(s^3_c,s^4|o_{1:t-1}\right)}{\sum_{s^3_c}\sum_{s^4}p^{a_{t-1}}_{c,t}\left(o_t|s^3_c,s^4\right)p_{c,1:t-1}\left(s^3_c,s^4|o_{1:t-1}\right)}. \tag{3}$$

All agents employ the same decision rule, that chooses the action with the highest valence. Formally,

$$d(v(\cdot,p_{c,1:t}\left(s^3_c,s^4|o_{1:t}\right))):=\underset{a\in A_{s^1}}{\arg\max}\,v(a,p_{c,1:t}\left(s^3_c,s^4|o_{1:t}\right)). \tag{4}$$

The **exploitative agent A1** seeks to maximize its immediate reward gain. Formally,

$$v_{\text{A1}}\left(a,p_{c,1:t}\left(s^3_c,s^4|o_{1:t}\right)\right)=\sum_{s^3_{nm}\in S^3_{nm,t}}y\left(a,s^1_t,s^3_{nm}\right)\,p_{c,1:t}\left(s^3_c=s^3_{nm}|o_{1:t}\right)\,\mathbb{E}_{p^{s^3_{nm},a}\left(r_t|s^3_c=s^3_{nm}\right)}\left(r_t\right). \tag{5}$$

Action valences are based on the cumulative weighted expected reward over the subset of nodes $s^3_{nm,t}\in S^3_{nm,t}$ that, given the agent's current beliefs, are (a) potential treasure locations, (b) the closest and reachable with the remaining moves and (c), to which action $a\in A_{s^1_t}$ would bring it closer to (evaluated by y). By weighting the expected rewards with the marginal treasure belief, an action's valence is proportional to the agent's uncertainty. In other words, the higher its belief that a node has the treasure, the higher is the value added to the valence of an action that brings it closer to that node. More intuitively, A1 moves towards nodes that it beliefs to be the most probable treasure location, while favouring those that are reachable most quickly.

The **exploitative agent A2** seeks to maximize its information gain and allocates action valences based on the expected Bayesian surprise under its current belief state representation. Formally,

$$v_{\text{A2}}\left(a,p_{c,1:t}\left(s^3_c,s^4|o_{1:t}\right)\right):=\sum_{o_{t+1}}p^{a_{1:t-1},a_t=a,s^1_{1:t},s^1_{t+1}=s^1_t+a,s^2_{1:t}}\left(o_{t+1}|o_{1:t}\right)\times\text{KL}\left(p\left(s^3_c,s^4|o_{t+1}\right)\middle|\middle|p_{c,1:t}\left(s^3_c,s^4|o_{1:t}\right)\right). \tag{6}$$

The first term in Eq. (6) is the posterior predictive distribution and corresponds to the agent's uncertainty over the observations. The second term in denotes the Kullback-Leibler (KL) divergence (Kullback and Leibler [1951]) between the agent's posterior belief in t, and its potential posterior belief in $t+1$ given potential observations. Informally, the KL divergence quantifies the expected information gain resulting from action $a_t=a$ and observation o_{t+1} under its current belief state representation.

The **hybrid agent A3** balances its immediate reward and information gain through a linear combination of both strategies.

$$v_{\text{A3}}\left(a,p\left(s^3_c,s^4|o_{1:t}\right)\right):=v_{\text{A1}}\left(a,p\left(s^3_c,s^4|o_{1:t}\right)\right)+v_{\text{A2}}\left(a,p\left(s^3_c,s^4|o_{1:t}\right)\right). \tag{7}$$

Results

Participants' and agents' behavioral results in the first run are presented in Fig. 2. Both, the exploitative agent A1 and the hybrid agent A3 achieved performances similar to participants, while the explorative agent A2's performance was closer to the control agents (Fig. 2a). Only A2 showed an choice rate of informative actions that was similar to participants, overall trials (Fig. 2b) and in the course of the game (Fig. 2c).

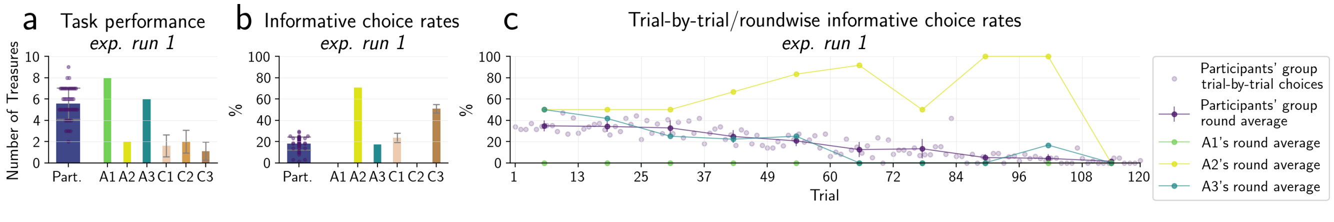


Figure 2: **Participants' and agents' behavioral results in run 1 (a)** Task performance **(b)** Informative choice rates averaged over all trials. **(c)** Trial-by-trial and roundwise average choice rates of informative actions. Results in runs two and three are very similar and thus not depicted here.

Conclusion

The present study introduces a novel experimental and computational framework to investigate information seeking and learning in a spatial search task and how underlying cognitive processes might be realized on a algorithmic level. Participants of the experiment showed a strong tendency to explore at the beginning of a task and gradually shifted towards more exploitation, as more information was accumulated. This choice pattern was best captured by a belief state-based Bayesian agent that evaluates the desirability of actions based on a combination of expected reward and an information bonus that reflects the agent's subjective uncertainty. First, this finding indicates that Bayesian Inference can be used to approximate human learning. Second, this finding supports the notion that humans are guided by their uncertainty and employ a combination of belief state based exploration and exploitation when making sequential decisions in a spatial search problem with uncertainty.

References

Thomas T. Hills and Reuven Dukas. The evolution of cognitive search. *Cognitive search: Evolution, algorithms, and the brain*, pages 11–24, 2012.

Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, May 2007. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2007.2098.

Katja Mehlhorn, Ben R. Newell, Peter M. Todd, Michael D. Lee, Kate Morgan, Victoria A. Braithwaite, Daniel Hausmann, Klaus Fiedler, and Cleotilde Gonzalez. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191–215, July 2015. ISSN 2325-9973, 2325-9965. doi: 10.1037/dec0000033.

Wei Ji Ma. Bayesian Decision Models: A Primer. *Neuron*, 104(1):164–175, October 2019. ISSN 08966273. doi: 10.1016/j.neuron.2019.09.037.

Samuel J. Gershman. Uncertainty and exploration. *Decision*, 6(3):277–286, July 2019. ISSN 2325-9973, 2325-9965. doi: 10.1037/dec0000101.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, June 2009. ISSN 00426989. doi: 10.1016/j.visres.2008.09.007.

Yi Sun, Faustino Gomez, and Juergen Schmidhuber. Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments. In *International Conference on Artificial General Intelligence*, pages pp. 41–51. Springer, March 2011.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, Massachusetts, 3rd edition, 2005. ISBN 978-1-886529-26-7 978-1-886529-08-3.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Probability, Zweite Ausgabe*. Athena Scientific, Belmont, Massachusetts, second edition, July 2008. ISBN 978-1-886529-23-6.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694.