Computational characterization of human sequential decision making under uncertainty

Model-free, model-based, exploitative and explorative strategies

DISSERTATION

zur Erlangung des akademischen Grades
Doktorin der Naturwissenschaften
(Dr. rer. nat.)

am Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

vorgelegt von
Dipl.-Psych. Lilla Horvath

Berlin, 2021

Erstgutachter: Prof. Dr. Dirk Ostwald

Zweitgutachter: Prof. Dr. Hauke R. Heekeren

Datum der Disputation: 17. Mai 2021

Summary

In life, many decision-making problems are complicated because agents - biological and artificial alike - typically can not directly observe all aspects of their environments. Moreover, consequences of the agents' actions in terms of reward gain typically unfold over time. The aim of this dissertation is to computationally characterize how humans tackle such problems from two perspectives.

The first perspective is to identify if decisions are governed in a model-free or a model-based fashion; while for model-free strategies it is sufficient to have access to some instantaneous reward-related information or the reward history, model-based strategies require representations of the statistical regularities of the environment. The second perspective is to identify if decisions are governed in a purely exploitative or a combined exploitative-explorative fashion; while purely exploitative strategies only seek to harness the knowledge about the environment, combined explorative-exploitative strategies also seek to accumulate knowledge about the environment.

In Chapter 1 of this dissertation, I present an agent-based modeling framework suitable to decompose correlates of human sequential decision making under uncertainty with respect to both perspectives. This framework capitalizes on partially observable Markov decision processes terminology, heuristics, belief states and dynamic programming, as well as standard statistical inference approaches to connect models and data. In Chapters 2 and 3, I put the agent-based modeling framework into use and investigate human participants' strategies in novel bandit and multistep tasks, respectively. In both tasks, I provide behavioral evidence for model-based strategies. Further, I demonstrate that the model-based strategy conforms to a combined explorative-exploitative agenda in the bandit task. By contrast, I show that in the multistep task, the model-based strategy conforms to a purely exploitative agenda, which is neurally enabled by the orchestrated activity in a distributed network of cortical and subcortical brain regions. In Chapter 4, I embed these findings within the broader discussion they contribute to, outline how the arbitration between different strategies could be organized and describe possible extensions of the agent-based modeling framework.

In summary, by adopting an agent-based modeling framework, this dissertation provides evidence for a predominantly model-based nature of human sequential decision making under uncertainty. In addition, by showing that exploitation is not always complemented by exploration, this dissertation highlights that humans can flexibly adjust their strategies, thereby meeting the ever-changing demands of life.

Zusammenfassung

Viele Entscheidungsprobleme im Leben sind dadurch kompliziert, dass sowohl biologische als auch künstliche Agenten typischerweise nicht alle Aspekte der Umgebung unmittelbar observieren können. Zudem entfalten sich die Konsequenzen von Aktionen hinsichtlich des Belohnungsgewinns erst im Laufe der Zeit. Das Ziel dieser Dissertation ist es aus zwei Blickwinkeln komputational zu erfassen, wie Menschen solche Probleme angehen.

Der erste Blickwinkel versucht zu identifizieren, ob Entscheidungen auf Basis einer modellfreien oder modellbasierten Art getroffen werden; während es für modellfreie Strategien ausreichend ist Zugang zu momentanen belohnungsbezogenen Informationen oder zur Belohnungsgeschichte zu haben, benötigen modellbasierte Strategien Repräsentationen von den statistischen Regelmäßigkeiten der Umgebung. Der zweite Blickwinkel versucht zu identifizieren, ob Entscheidungen auf Basis einer rein exploitativen oder kombiniert exploitativ-explorativen Art getroffen werden; während rein exploitative Strategien nur darauf abzielen, sich das Wissen über die Umgebung zu Nutze zu machen, zielen kombinierte explorativ-exploitative Strategien auch darauf ab, Wissen über die Umgebung anzusammeln.

In Kapitel 1 dieser Dissertation stelle ich ein agentenbasiertes Modellierungsframework vor, das ermöglicht, Korrelate humaner sequentieller Entscheidungsfindung unter Unsicherheit in Bezug auf beide Blickwinkel zu zerlegen. Dieses Framework basiert auf der Terminologie partiell-observierbarer Markov Entscheidungsprozesse, Heuristiken, Bayes'scher Zustandsrepräsentation und dynamischer Programmierung sowie klassischen statistischen Inferenzansätzen um Modelle und Daten zu verknüpfen. In Kapiteln 2 und 3 setze ich das agentenbasierte Modellierungsframework ein um die Strategien humaner Teilnehmer in neuartigen Bandit- beziehungsweise Mehrschritt-Aufgaben zu untersuchen. In beiden Aufgaben erbringe ich Nachweise für den Einsatz modellbasierte Strategien auf der Verhaltensebene. Des Weiteren demonstriere ich, dass die modellbasierte Strategie in der Bandit-Aufgabe einer kombinierten explorativ-exploitativen Agenda entspricht. Im Gegensatz dazu zeige ich, dass die modellbasierte Strategie in der Mehrschritt-Aufgabe einer rein exploitativen Agenda entspricht, die neuronal durch die orchestrierte Aktivität eines verteilten Netzwerks kortikaler

und subkortikaler Hirnregionen unterstützt wird. In Kapitel 4 bette ich diese Ergebnisse in die breitere Diskussion ein, stelle dar, wie eine Auswahl verschiedener Strategien erfolgen könnte und beschreibe mögliche Erweiterungen des agentenbasierten Modellierungsframeworks.

Zusammenfassend zeigt diese Dissertation durch die Anwendung eines agentenbasierten Modellierungsframeworks, dass die sequentielle Entscheidungsfindung unter Unsicherheit bei Menschen vorwiegend modellbasierter Natur ist. Durch den Nachweis, dass exploitative Strategien nicht immer durch explorative Strategien ergänzt werden, hebt die Dissertation darüber hinaus hervor, dass Menschen ihre Strategien flexibel anpassen können, um den sich ständig ändernden Anforderungen des Lebens gerecht zu werden.

Acknowledgements

Many people have provided immeasurable help over the last years in putting together this dissertation.

First and foremost, I would like to thank my PhD advisor Dirk Ostwald. Your clarity of thought, scientific rigor and dedication to open science have invaluably shaped my approach to research and pushed me to strive for a deep understanding as well as a precise reporting of computational characterization of human decision making.

I am also especially grateful to my second PhD advisor, Hauke Heekeren, whose door was always open to provide support. I greatly benefited from the scientific discussion about our projects and I am deeply thankful for your advice on how to maneuver my path in science.

Soyoung Park, Julia Rodriguez Buritica and Peter Mohr have kindly argreed to be members of my PhD committee, which I am most grateful for.

I would like to thank Michael Milham for providing me with the opportunity to perform a study at the Nathan Kline Institute, which formed the basis of the work presented in the second chapter of this dissertation. Your enthusiasm for science and unwavering optimism made my time in New York truly enjoyable. I am also grateful to Stanley Colcombe and Shruti Ray for their support in carrying out this study. I sincerely thank Philipp Schwartenbeck for countless inspiring discussions about the experimental setup and data analysis.

I am thankful to Ralph Hertwig for providing the infrastructure at the Max Planck Institute for Human Development critical to realize the study upon which the work described in the third chapter of this dissertation is based. Rui Mata helped planning this study and Loreen Tisdall helped with data collection, for which I am very grateful.

I would like to express my heartfelt gratitude to Wei Ji Ma for believing in me and being beyond supportive throughout a rocky decision process. I feel extremely fortunate to have had the opportunity to get to know you and your lab. I owe another special thank you to Andreas Horn for the endless encouragement. You were one of the firsts to show me how much fun science can be and your positive attitude never ceases to inspire me.

I hugely benefited from discussions with members of the Center for Cognitive

Neuroscience Berlin. I particularly want to thank Lisa Velenosi, Pia Schröder and Kathrin Tertel. Your insightful and witty way of thinking, with respect to science and beyond, made all the difference in the long PhD days. I would also like to extend this thank you to Yuan-hao Wu and Miro Grundei for our lunch and coffee breaks that I valued so much.

Thank you Daniela Satici-Thies for the immense help with all my administrative problems and somehow always finding a way to make seemingly impossible ends meet. I am also grateful for receiving generous support by an Elsa-Neumann-Scholarship during the first three years of my PhD.

Last, but not least, I would not be writing this acknowledgement marking the end of my PhD if it were not for my friends and family. I am wholeheartedly grateful for your constant encouragement and patience. I especially would like to thank Miriam Dowe for listening and making me listen, Lisa Graaf for reminding me of what life has to offer outside of PhD even during its most intense phases, Judit Varga for running with me to the finish line and Petra Vancsura for cheering us all along. I will forever be indebted to my parents, Lilla Kardos and Istvan Horvath, who have always and unconditionally been there for me. Finally, no words would do justice to express how grateful I am to Benjamin Ostendorf for *everything* and to the little person who has spent every second of the last months with me for being the most wonderfully independent company I could ever wish for.

Contents

1 | General introduction

To reach certain goals in life, we often have to make a sequence of decisions. Consider, for example, a gambler playing on a slot machine in a casino or a high school student with an aspiration to become an astrophysicist. In formal terms, in both examples the goal of the decision-making agent can be described as trying to choose actions as to maximize its cumulative reward: The gambler tries to pull the best lever on each turn to win as much money as possible; the high school student tries to make the best career choice in each situation to secure their dream job. To pull the best lever or make the best career choice, both the gambler and the high school student would need to be omniscient about their environment (e.g., the precise probability with which the pulling of a lever returns a certain reward or the precise expectations of a college admission committee). This requisite, however, defeats the purpose of gambling and is unrealistic when it comes to pursuing a career goal.

As demonstrated by these examples, most sequential decision-making problems are complicated by uncertainty (Bach & Dolan, 2012; Bach, Hulme, Penny, & Dolan, 2011; Glimcher & Fehr, 2013; Ma & Jazayeri, 2014; Rao, 2010; Vilares & Kording, 2011; Yoshida & Ishii, 2006). In the face of uncertainty, two key questions arise (Dayan & Daw, 2008). The first question concerns the environmental components agents draw on to evaluate actions. An influential dichotomy in this regard is the model-free versus model-based distinction (e.g., Collins and Cockburn, 2020; Daw, Niv, and Dayan, 2005; Fischer, Bourgeois-Gironde, and Ullsperger, 2017; Korn and Bach, 2018; D. A. Simon and Daw, 2011; Speekenbrink and Konstantinidis, 2015). Broadly put, model-free decision making assumes that agents directly evaluate actions based on the reward history or some presently available reward-related information. In contrast, model-based decision making assumes that action evaluation is governed by the agents' own representation of the statistical regularities of their environment. The second question is whether agents evaluate actions in a purely exploitative fashion or combine exploitation with exploration (e.g., Cohen, McClure, and Yu, 2007; Daw, O'Doherty, Dayan, Seymour, and Dolan, 2006; Schwartenbeck et al., 2019; Wilson, Geana, White, Ludvig, and Cohen, 2014). Exploitative decision making assumes that agents try to maximize their cumulative reward based on what they know about the environment at the time of their decision. A combined explorative-exploitative perspective, on the other hand, assumes that agents also try to improve their knowledge about the environment and therefore take into account the amount of information they can gain by choosing a certain action.

In this dissertation, I computationally characterize human sequential decision making under uncertainty along these two questions in two tasks that share central features with the examples introduced above. Specifically, in Chapter 2, I study the behavioral strategies human participants employ in a bandit task, which captures situations similar to the example of the gambler. In Chapter 3, I investigate the computations human participants perform to solve a multistep task - which captures situations similar to the example of the high school student - on behavioral and neural levels.

In this introduction, I conceptually situate these two empirical chapters within relevant theories. To this end, I follow Marr's levels of analysis (Marr, 1982) and thereby introduce the agent-based modeling framework (Ostwald, 2020a) which I adopt in the empirical chapters to uncover the computational underpinnings of the applied sequential decision-making strategies. I conclude the introduction by giving a brief overview of the remainder of the present dissertation.

1.1 Marr's levels of analysis and agent-based modeling

In his work about visual perception, Marr (1982) proposed a conceptual framework consisting of three hierarchical levels to systematically study, understand and discuss the brain and its functions. Marr's framework has since been an inspiration to cognitive scientists and neuroscientists in guiding scientific inquiry (see, for example, Hauser, Fiore, Moutoussis, and Dolan, 2016; Niv and Langdon, 2016), and has recently also gained attention in machine learning research (Hamrick & Mohamed, 2020). According to Marr, on the first 'computational' level the problem at hand is to be formally defined. On the second 'algorithmic' level, alternative solutions as to how the problem can be tackled are to be described. Finally, on the third 'implementation' level, plausible ways for a (neural) system to realize these alternative solutions are to be considered.

Agent-based modeling as outlined by Dirk Ostwald (Ostwald, 2020a) and adopted in Chapters 2 and 3 offers a formal framework to investigate human sequential decision making under uncertainty. In its current form, this framework consists of three building blocks that can be readily mapped onto Marr's levels of analysis as follows: The first building block is the task model, which corresponds to a probabilistic formulation of the choice environment. Rooted in probabilistic optimal control theory, agent-based modeling adopts the terminology of partially observable Markov decision processes in the definition of the task model (e.g., Bäuerle and Rieder, 2011; Bertsekas, 2000; Puterman, 2014). By representing the choice environment and thereby specifying the problem to be solved, this building block parallels the computational level in Marr's framework. The second building block is the set of agent models. These models capitalize on Bayesian inference, dynamic programming, heuristics and reinforcement learning to formulate various strategies that can be used to solve the problem (e.g., Dayan and Daw, 2008; Gigerenzer, Todd, and the ABC Research Group, 1999; Hassabis, Kumaran, Summerfield, and Botvinick, 2017; Ma, 2019; Rao, 2010; Russell and Norvig, 2010; Sutton and Barto, 2018; Wayne et al., 2018; Wiering and van Otterlo, 2014). In essence, the second building block of agent-based modeling exhausts the requirements of Marr's algorithmic level. However, to evaluate the plausibility of the agent models in light of human participants choice data, the agent models have to be statistically embedded (e.g., Daunizeau et al., 2010; Farrell and Lewandowsky, 2018). The ensuing set of behavioral models constitutes the third building block of agent-based modeling. Although as it currently stands, agent-based modeling does not directly specify a building block that maps onto Marr's implementation level, methods such as model-based general linear modeling (GLM) of functional magnetic resonance imaging (fMRI) data can be considered for this purpose (Friston & Dolan, 2010).

In line with Marr's framework, in the following, I first give a formal description of the sequential decision-making problem under uncertainty by introducing the general task model architecture in Section 1.2. Here, I also highlight how this can be tailored to bandit and multistep tasks studied in detail in Chapters 2 and 3, respectively. In Section 1.3, I then introduce the general agent model architecture and review its variations capturing strategies in terms of the dichotomies model-free versus model-based, and exploitation versus exploration. Additionally, I here also describe the general behavioral model architecture. Finally, in Section 1.4, I present the model-based GLM for fMRI approach applied in Chapter 3 to identify the network of brain regions enabling the realization of algorithmic solutions as formulated by the agent models.

1.2 The sequential decision-making problem under uncertainty

In 1957, Richard Bellmann introduced the theory of Markov decision processes (MDPs) suitable to model a wide range of real-world sequential decision-making problems (Bellman, 1957). Ever since, the theory of MDPs has been paramount in operations research and extended to accommodate uncertain[1] conditions (Bäuerle & Rieder, 2011; Bertsekas, 2000; Bertsekas & Tsitsiklis, 1996; Lovejoy, 1991; Puterman, 2014; Sutton & Barto, 2018; Wiering & van Otterlo, 2014). The ensuing theory of partially observable Markov decision processes (PoMDPs) offers a formal language to describe the scope of the sequential decision-making problem under uncertainty as well as a principled way to derive the optimal solution. In this section, I draw on PoMDPs theory to introduce the general task model architecture, detail how this model differs for bandit and multistep tasks, and - as a prelude to the next section - I lay out the notion of optimal solution. To this end, I throughout rely on the literature listed in this paragraph and complement it with further resources wherever appropriate.

1.2.1 The general task model architecture

In general terms, the model of a task formally captures the agent's choice environment using mathematical sets and probability distributions.

The first set component of the task model is the set of time points denoting the epochs at which the agent may interact with the choice environment. In the standard case and as is assumed throughout this dissertation, this set is discrete and finite. At each time point, the choice environment has a certain configuration. The second set component of the task model - the set of states - represents all possible values of these configurations. Crucially, certain aspects of the environmental configurations may be overt, constituting the directly observable part of the state, while others may only be imprecisely signalled, constituting the not directly observable part of the state. The values the imprecise signal can take on form the third set component of the task model, the set of observations.[2] In each state, the choice environment allows certain

[1] The term uncertain has been used to refer to choice environments, in which the dynamics, such as the state transitions, are stochastic. Yet, it has also been reserved to signify choice environments, in which some components, such as the state or the reward dynamics, are only partially observable (e.g., Knight, 1921; Russell and Norvig, 2010). As will become apparent in the remainder of this section, in this dissertation I adopt this latter conceptualization.

[2] If the imprecise signalling is due to disturbances in sensory processing (cf. Bach and

actions to be undertaken by the agent. The set of all actions presents the fourth set component of the task model. The last essential set of the task model comprises all numerical rewards the choice environment may generate.

How the elements of these sets relate to each other can be described with the observation, reward and state transition probability distributions of the task model. Concretely, the observation probability distribution encapsulates the dynamics between states and observations by specifying how the former gives rise to the latter. The reward and state transition probability distributions specify how state-action pairs lead to rewards and new states, respectively.

1.2.2 Task model variations

This general architecture allows for considerable flexibility to match the specifics of different tasks. Two classes of tasks that are often used to study sequential decision making under uncertainty are bandit and multistep tasks.

Bandit tasks

Bandit tasks, which were first systematically discussed by Robbins (1952), capture choice environments in which actions are not interdependent. More specifically, in bandit tasks, choosing an action in a given state does not have an effect on the next state but only on the immediately accrued reward. Such tasks are thus well suited to model, for example, treatment allocation in clinical trials or gambling (e.g., Berry and Fristedt, 1985; Brand, Woods, and Sakoda, 1956; Bubeck and Cesa-Bianchi, 2012; Cohen et al., 2007; Dayan and Daw, 2008; Gabillon, Ghavamzadeh, and Lazaric, 2012; Speekenbrink and Konstantinidis, 2015; Whittle, 1988). As delineated above, in the case of gambling, the agent (gambler) chooses from a finite set of actions (pulls one of the levers) and receives a reward (monetary gain or loss) as dictated by the reward probability distribution. Then, the agent again faces the same set of actions and the process gets repeated until the time horizon is reached (game is finished). A crucial and inherent aspect of bandit tasks is that the reward structure of the environment is not directly observable. Depending on the bandit task at hand, this can either be formulated as a part of the state being not directly observable, or as parameters of the reward probability distribution being not directly observable. In both cases, the reward returned to the agent conveys noisy information

Dolan, 2012), the set of observations can instead be considered internal to the agent. However, in the sequential decision-making problems studied in this dissertation no such disturbances are assumed and I therefore conceive the set of observations as a part of the task model.

about the not directly observable component of the task model and thus, in this sense, rewards serve as observations. Of course, to conceptualize rewards in terms of observations, rewards have to be observable. In life, however, this may not always be the case. In Chapter 2, I introduce a bandit task suitable to model choice environments in which the reward is observable only following certain actions. Given the symmetrical reward structure adopted in this task, its not directly observable nature is captured by the state, whose value changes over time. Such switching state bandit tasks require the specification of state transitions, which - per definition - are independent of actions.

Multistep tasks

In contrast to bandit tasks, multistep tasks capture choice environments in which actions are interdependent. That is, in multistep tasks, actions do not only affect the immediate rewards but they also affect the next state and thereby future rewards (e.g., Daw, Gershman, Seymour, Dayan, and Dolan, 2011; Dayan and Daw, 2008; Korn and Bach, 2018; Lehmann et al., 2019; Schrittwieser et al., 2020; D. A. Simon and Daw, 2011; Wayne et al., 2018). The above introduced example of the high school student presents a choice environment that can be modeled in terms of a multistep task: In a given state (e.g., at an interview with the college admission committee) the agent (high school student) chooses an action (e.g., highlights her keen interest in black holes), receives an immediate reward (e.g., bonus points) according to the reward probability distribution and enters a new state (e.g., gets accepted to the program) according to the state transition probability distribution. In the new state, the agent is presented with a new set of actions to choose from, each action producing different immediate rewards and new states. Uncertainty may pervade multistep tasks, for instance, if part of the state is not directly observable (Bach & Dolan, 2012; Dayan & Daw, 2008; Rao, 2010; Yoshida & Ishii, 2006). As a specific example, consider the high school student again. At the interview, the exact expectations of the college admission committee might only be imprecisely signalled by the members' subtle reactions. In Chapter 3, I introduce a multistep task embedded in the spatial domain suitable to model similar choice environments.

1.2.3 The notion of optimal solution

On the basis of the task model, the theory of PoMDPs offers a principled way to identify normative sensible decisions.[3] The central presumption thereby is that the ultimate goal of an agent is to maximize the cumulative obtained reward.

To achieve this goal, agents have to choose the sequence of actions for which the expected sum of rewards over the time points is maximal. This optimal action sequence can, in principle, be found by applying dynamic programming, which capitalizes on the recursive scheme of the Bellman equation (Bellman, 1957). In its standard form, the Bellman equation states that the optimal action in a given state maximizes the sum of the expected immediate reward and the optimal value of the expected next state, which corresponds to the maximum expected sum of rewards that can be obtained starting from the expected next state.

Even if all aspects of the state are directly observable (i.e, the choice environment can be described in terms of MDPs), applying dynamic programming can be computationally costly, for example, in multistep tasks with large state spaces and time horizons, such as chess (e.g., Bellman, 1961; Huys et al., 2012; van Opheusden, Galbiati, Bnaya, Li, and Ma, 2017).[4] It yet becomes even more computationally costly if some aspects of the state are only imprecisely signalled by the observations (i.e, the choice environment can be described in terms of PoMDPs), as is the case in both the bandit and multistep tasks studied in detail in Chapters 2 and 3 of this dissertation. This is because under such circumstances, the optimal action has to be evaluated with respect to the agent's subjective uncertainty about the state, i.e., the belief state. In other words, in the Bellman equation as formulated above, states have to be replaced by belief states. Fundamental to this replacement is that just like states, belief states satisfy the Markov property, which prescribes that in the choice environment the past is independent of the future given the present. The Markov property also implies that at a given time point the belief state - formally a probability distribution over states given past actions and observations - can

[3] In operations research, some scholars (e.g., Bäuerle and Rieder, 2011; Puterman, 2014) explicitly differentiate between partially observable Markov decision *processes* and partially observable Markov decision *problems*; they apply the former term when specifying a quantitative model of the problem at hand and the latter term when combining this quantitative model with the optimality criterion.

[4] This anyways existing difficulty possibly explains why in human decision neuroscience research the experimental study of multistep tasks has so far largely focused on scenarios without state uncertainty.

be computed recursively on the basis of the belief state, action and observation at the previous time point using Bayes' rule (Dayan & Daw, 2008; Kaelbling, Littman, & Cassandra, 1998; Rao, 2010; Russell & Norvig, 2010).

Beside the computational load posed by the combination of dynamic programming and state inference, it may also easily exhaust the memory space. Thus, while under certain simplifying conditions optimal solutions can be attained[5], most real-life problems modeled in terms of PoMDPs necessitate approximations (Berry & Fristedt, 1985; Dayan & Daw, 2008; Rao, 2010; Russell & Norvig, 2010). Furthermore and most importantly for the purpose of this dissertation, given the cognitive capacity limits of biological agents, such approximations present themselves suitable to be adopted by humans (Gershman, Horvitz, & Tenenbaum, 2015; Griffiths, Lieder, & Goodman, 2015; H. Simon, 1957).

1.3 Approximate solutions and their behavioral plausibility

A large variety of algorithmic methods exists to obtain approximate solutions. Of these, a class of methods inherits from the above outlined normative scheme and uses concepts from Bayesian inference and dynamic programming (Bäuerle & Rieder, 2011; Bertsekas, 2000; Bertsekas & Tsitsiklis, 1996; Ma, 2019; Puterman, 2014; Rao, 2010; Sutton & Barto, 2018; Wiering & van Otterlo, 2014; Yoshida & Ishii, 2006). In decision neuroscience research, these methods are usually referred to as model-based, because they rely upon the defining probability distributions of the task model. In contrast, for model-free methods it is sufficient to have knowledge of only the overt set components of the task model (e.g., Collins and Cockburn, 2020; Daw et al., 2005; Dayan, 2012; Dayan and Daw, 2008; Korn and Bach, 2018; Speekenbrink and Konstantinidis, 2015). These methods come from heuristic decision making (Gigerenzer et al., 1999; Tversky & Kahneman, 1974) and reinforcement learning (RL; Bertsekas and Tsitsiklis, 1996; Rao, 2010; Sutton and Barto, 2018; Wiering and van Otterlo, 2014) and operate on the basis of instantaneous information about or previous experience with rewards.[6] Another important perspective to classify methods is

[5] For example, optimal solutions to stationary bandit tasks can be derived for finite (Bellman, 1956; Berry & Fristedt, 1985) and infinite (Gittins & Jones, 1974) time horizons.

[6] In artificial intelligence research, approximate methods that borrow from the PoMDPs theory and therefore necessitate knowledge about the probability distributions of the task model are sometimes termed model-based RL methods. Correspondingly, the term model-free

the distinction between exploitation and exploration-exploitation. Exploitative methods are solely guided by the perspective of reward gain based on the accumulated knowledge about the choice environment. Explorative-exploitative methods, conversely, are also guided by the perspective of information gain to advance their knowledge about the choice environment (Berry & Fristedt, 1985; Bertsekas & Tsitsiklis, 1996; Cohen et al., 2007; Dayan & Daw, 2008; Schwartenbeck et al., 2019; Sun, Gomez, & Schmidhuber, 2011; Sutton & Barto, 2018; Wiering & van Otterlo, 2014). Despite the apparent differences between methods, the structural requirements imposed on the agents adopting them have some key commonalities (Russell & Norvig, 2010). Therefore, in what follows, I first present the general agent model architecture and then detail its variations in terms of the dichotomies model-free versus model-based and exploitation versus exploration. I close this section by describing the general behavioral model architecture, which formalizes the embedding of the agent models into a statistical framework.

1.3.1 The general agent model architecture

The word agent originates from the Latin agere, which means to do. Accordingly, central to agents is that they perceive their environment, on the basis of which they act as to reach their goal. This suggests that one part of the agent model has to specify the agent's representation of the task model. The other part, in turn, has to specify how the agent draws on this representation to evaluate the actions and make decisions (Russell & Norvig, 2010).

As discussed in detail below, the agent's copy of the task model can vary greatly. Some methods only require the agent to represent the overt set components of the task model, i.e., time points, directly observable part of states, observations, actions and rewards. Others also require representations of the possible values of the not directly observable part of state and the probability distributions of the task model. Given that the probabilistic representations are internal to the agent, they are to be conceived as subjective uncertainties, even if the corresponding probability distributions of the task model are overt (cf. Ma, 2019).

On the basis of its task representation, the agent evaluates the actions, which is formalized in terms of the valence function, and makes a decision, which is formalized in terms of the decision function. Similar to the value function of

is used for RL methods that do not necessitate such knowledge (e.g., Schrittwieser et al., 2020; Sutton and Barto, 2018; Wiering and van Otterlo, 2014).

the PoMDPs theory, the valence function assigns a number to each action. This number is, however, not the optimal value of the action but an approximation thereof and can therefore be considered as a measure of the action's subjective desirability as viewed by the agent. Depending on the valence function, the agent may need to apply Bayesian inference and form a belief state. Thus, in this case, the agent model also has to comprise the specification of the agent's initial subjective uncertainty about the state. Drawing on the valences, the decision function adjudicates between actions, implementing either a stochastic or a deterministic valence maximizing scheme.

1.3.2 Agent model variations

Model-free/model-based and exploitative/explorative methods assume certain characteristic configurations of the agent's task representation, valence and decision functions, and, consequently, the above introduced general architecture.

Model-free versus model-based

Common to model-free methods is that agents do not need knowledge about the task model beyond the overt set components. Constrained by the simplicity of such task representations, all model-free methods directly allocate valences to actions. Yet, an abundance of different ways exists to do this. Inspired by heuristic decision making, a simple yet often efficient way is to allocate action valences based on the latest reward-related information, conveyed, for example, by observations (Dayan, 2012; Gigerenzer & Gaissmaier, 2011; Korn & Bach, 2018; Robbins, 1952; Wilson & Collins, 2019). Another way was originally described by the decision neuroscientists Rescorla and Wagner (1972) and further developed in RL research under the name temporal difference learning. The key aspect of these model-free methods is that an action's valence depends on the associated reward history, where an arbitrary constant learning rate controls the extent to which the latest experience is taken into account (Bertsekas & Tsitsiklis, 1996; Glimcher & Fehr, 2013; Wiering & van Otterlo, 2014; Wilson & Collins, 2019).

In contrast to model-free methods, model-based methods require that the agent maintains representations of all sets and the probabilistic dependencies between their elements. Together with the agent's initial belief state, the ensuing complete task model is put into use to probabilistically infer the state and to allocate valences to actions by considering their consequences with respect to future states, observations and rewards. To this end, model-based methods employ some combination of Bayesian inference and approximate dynamic programming. Approximations can, for instance, be implemented by limiting the time horizon considered or by using a heuristic to evaluate possible next belief states (Bertsekas & Tsitsiklis, 1996; Geffner & Bonet, 1998; Huys et al., 2012; Korf, 1990; van Opheusden et al., 2021; Wiering & van Otterlo, 2014). As exemplified by this latter approach, model-free methods may complement model-based methods. Applying temporal difference learning in the space of belief states is another example of such a mixture method (Babayan, Uchida, & Gershman, 2018; Dayan & Daw, 2008; Rao, 2010; Starkweather, Babayan, Uchida, & Gershman, 2017).

Exploitation versus exploration

Exact solution to a problem modeled in terms of PoMDPs yields an optimal balance between exploitation and exploration. To approximate the optimal balance, model-free as well as model-based methods across the entire spectrum, from purely exploitative to purely explorative, have been proposed (Bertsekas & Tsitsiklis, 1996; Cohen et al., 2007; Dayan & Daw, 2008; Schwartenbeck et al., 2019; Wiering & van Otterlo, 2014).

Purely exploitative methods disregard the perspective of information gain. Instead, at each time point, they harness the knowledge about the choice environment acquired through previous interactions and allocate action valences from the perspective of reward gain. This can be done both in a model-free way, relying, for example, on a reward-related heuristic, or in a model-based way, evaluating, for example, the belief state-weighted expected reward (Knox, Otto, Stone, & Love, 2012; Lee, Zhang, Munro, & Steyvers, 2011; Speekenbrink & Konstantinidis, 2015). Crucial thereby is that the agent adopts a valence maximizing deterministic decision function so that the action with the highest exploitative valence is realized.

Correspondingly, purely explorative methods have to ensure that the action with the highest explorative valence is realized and they therefore also require a valence maximizing deterministic decision function. Yet, opposite to purely exploitative methods, these methods seek to improve their knowledge about the choice environment and thus allocate action valences from the perspective of information gain. Two commonly applied measures of information gain are the frequentist upper confidence bound (Auer, Cesa-Bianchi, & Fischer, 2002) and the expected Bayesian surprise (Itti & Baldi, 2009; Ostwald et al., 2012; Sun et al., 2011). The former expresses information gain in a model-free fashion based on the extensiveness of the action's associated reward history, whereas the latter expresses information gain in a model-based fashion based on the shift in the belief state.

By combining the valences of purely explorative and purely exploitative methods, explorative-exploitative methods take both information gain and reward gain into account (Chakroun, Mathar, Wiehler, Ganzer, & Peters, 2020; Gershman, 2018; Navarro, Newell, & Schulze, 2016; Wilson et al., 2014; Zhang & Yu, 2013). These methods either implement a valence maximizing deterministic decision function or a stochastic decision function. The rational behind using a stochastic decision function is that information may also be gained 'by chance', i.e., through adding some noise to the action selection process. This, in turn, suggests that explorative-exploitative methods may also be formed by taking the valences of a purely exploitative method and using a stochastic decision function - with constant (e.g., ϵ-greedy (Sutton & Barto, 2018; Wiering & van Otterlo, 2014) or softmax operation (Reverdy & Leonard, 2015)) or belief state-dependent (Thompson sampling[7]; Thompson, 1933) noise. While belief state-dependent noise assumes a model-based method, constant noise can also be added to the exploitative valences allocated by a model-free method.

In Chapters 2 and 3, I computationally characterize human participants choice behavior with respect to the dichotomies model-free versus model-based and exploitation versus exploration in bandit and multistep tasks, respectively. To this end, I use agent models that implement model-free purely exploitative, model-based purely exploitative, purely explorative and exploitative-explorative methods. More specifically, the model-free purely exploitative agents of both model spaces rely on reward-related heuristics. Their model-based counterparts evaluate the belief state-weighted expected reward or employ belief state-based heuristic real-time dynamic programming. In contrast, the model-based purely explorative agents are guided by the expected Bayesian surprise. Finally, the model-based exploitative-explorative agents perform linear convex combinations of the valences allocated by the model-based purely exploitative and purely explorative agents and apply valence maximizing deterministic decision functions.

[7] Thompson sampling is traditionally formulated as allocating valences based on random draws from the belief state and subsequently using a valence maximizing deterministic decision function.

1.3.3 The general behavioral model architecture

In order to computationally characterize human participants choice behavior by means of the agent models, for each agent model a corresponding behavioral model has to be formulated. The behavioral model specifies an embedding of the agent model into a statistical inference framework (Daunizeau et al., 2010; Farrell & Lewandowsky, 2018). In Chapters 2 and 3, I follow a standard procedure to accomplish this and nest the agent's valence function in a softmax operation (Reverdy & Leonard, 2015).

To probabilistically translate between the action valences internal to the agent and the action observable by the experimenter, the exponential softmax operation evaluates the action valences in relation to one another. Thereby, a parameter controls the extent to which the probabilities reflect the difference in the action valences: The lower the parameter value, the higher the probability that the experimenter observers the action with the higher action valence. As a consequence, this parameter can be interpreted as post-decision (or observation) noise.

Of note, as alluded to above, in many decision neuroscience studies the softmax operation is commonly applied as a stochastic decision function to form explorative-exploitative agents (e.g., Chakroun et al., 2020; Daw et al., 2006; Dezza, Angela, Cleeremans, and Alexander, 2017; Gläscher, Daw, Dayan, and O'Doherty, 2010; Hauser et al., 2014; Speekenbrink and Konstantinidis, 2015). In these studies, behavioral models are usually not additionally formulated and the parameter of the softmax operation is interpreted as a tendency for random exploration. In the agent-based modeling framework adopted in this dissertation, the agent models and behavioral models are throughout explicitly separated. This is to highlight that in contrast to operations and artificial intelligence research, in decision neuroscience the agent models are used to explain experimentally acquired human data - and therefore have to be statistically embedded.

1.4 Neural implementation of alternative solutions

Beyond evaluating their behavioral plausibility, decision neuroscience seeks to answer how different agent models might be realized by the neural system (Dayan & Daw, 2008; Glimcher & Fehr, 2013; Niv & Langdon, 2016; Sutton &

Barto, 2018). Neural data can stem from different modalities, ranging from single cell recordings (e.g., Costa and Averbeck, 2020; Schultz, Dayan, and Montague, 1997; Starkweather et al., 2017) to fMRI (e.g., Chakroun et al., 2020; Daw et al., 2011; O'Doherty, Dayan, Friston, Critchley, and Dolan, 2003), and thus, many approaches linking Marr's algorithmic level with the implementation level can be considered. One of the most popular approaches is to analyze fMRI data obtained from human (or other primate) participants simultaneously with the behavioral data using model-based GLM (Friston & Dolan, 2010). In the current section, I describe the model-based GLM for fMRI approach with an emphasis on ways it can be integrated with the agent models discussed in the previous section.

1.4.1 Model-based GLM for fMRI

In the analysis of fMRI data, applying the statistical inference framework of GLM is a standard technique to localize cognitive processes in the brain (Huettel, Song, & McCarthy, 2009; Ostwald, 2020b). Typically, this analysis proceeds as follows. First, the spatially logged time-series data acquired from a single participant are modeled using multiple linear regression design, where each regressor (of interest) represents a certain type of experimental event. Then, the parameter estimates are combined with contrast weight vectors and evaluated on the group-level, using, for example, one-sample t-tests. Every ensuing statistical parametric map informs about the brain regions specialized for the cognitive process associated with the respective experimental events.[8] To establish the functional anatomy of algorithmic methods, in model-based GLM for fMRI, the participant-level design matrix additionally comprises parametric regressors representing sequences of latent quantities produced by these methods (Friston & Dolan, 2010). Consequently, model-based GLM for fMRI can readily accommodate agent models implementing model-free/model-based and exploitative/explorative methods and thereby connect the above outlined approximate solutions with neural realization.

To form parametric regressors, usually a basis regressor - such as the trial regressor modeling the events pertaining to the state-observation-action-reward tetrad per time point - is subjected to agent model-based quantities. A key latent quantity derived from an agent model on a trial-by-trial basis is the chosen action valence according to the participant's previous interactions with the choice environment (e.g., Chakroun et al., 2020; Daw et al., 2006; Korn

[8] Of course, the exact interpretation depends on the applied contrast weight vector.

and Bach, 2018; D. A. Simon and Daw, 2011). In several decision neuroscience studies, this quantity is expressed relative to the valence of the other available actions, which relates to the notion of choice conflict (e.g., Boorman, Behrens, Woolrich, and Rushworth, 2009; Shenhav, Straccia, Cohen, and Botvinick, 2014). Another tradition is to derive quantities expressing some difference between two trials. For instance, in the case of an agent model adopting temporal difference learning, the so called reward prediction error between the old and new valences of the chosen action can be considered (e.g., Daw et al., 2011; Doll, Duncan, Simon, Shohamy, and Daw, 2015; Fischer et al., 2017; Gläscher et al., 2010; Rao, 2010; D. A. Simon and Daw, 2011). Bayesian surprise is another such quantity, which is computed as the divergence between the prior and posterior belief states and provides a readout of a model-based agent's state inference (e.g., Fischer et al., 2017; Gijsen, Grundei, Lange, Ostwald, and Blankenburg, 2020; Itti and Baldi, 2009; O'Reilly, Jbabdi, Rushworth, and Behrens, 2013; Ostwald et al., 2012; Schwartenbeck, FitzGerald, and Dolan, 2016).

After identifying the agent model best accounting for participants' choice data in a multistep task, in Chapter 3, I map the network of brain regions supporting its architecture. To this end, I analyze the fMRI data collected from each participant using model-based GLM. Concretely, to evaluate the neural correlates of the combination of state inference and exploitation by means of heuristic real-time dynamic programming as implemented by the group-favored agent model, the latent quantities Bayesian inference and chosen action valence are employed.

1.5 Overview of the dissertation

In this dissertation, I computationally characterize - on behavioral and neural levels - how humans make sequential decisions under uncertainty in two tasks that capture central aspects of daily choice environments. In doing so, I focus on answering whether the applied strategies reflect model-free or model-based and exploitaive or exporative-exploitative processes. To accomplish this, I rely on an agent-based modeling framework capitalizing on PoMDPs terminology, heuristics, belief states and dynamic programming, as well as standard statistical inference approaches connecting models and data.

In Chapter 2, human sequential decision making under uncertainty is behaviorally studied in an information-selective reversal bandit task. In contrast to previous bandit tasks, in the task introduced in this chapter, reward observations are not available for each action, forcing the decision maker to explicitly evaluate the benefit of exploration against the benefit of exploitation. The results show that in such choice environments, humans employ a model-based exploitative-explorative strategy as captured by an agent model seeking to maximize a convex combination of the belief state-weighted expected reward and expected Bayesian surprise.

While investigated theoretically, the empirical study of strategies used in multistep tasks with partially observable states remains elusive. To address this, in Chapter 3, behavioral and fMRI data collected from human participants on a novel spatial search task are analyzed. Similar to the results of Chapter 2, the behavioral data is best accounted for by a model-based agent implementing Bayesian inference. The belief state, however, is put into use in a purely exploitative fashion, as captured by a heuristic real-time dynamic programming algorithm. The results of model-based GLM for fMRI demonstrate that the latent quantities Bayesian surprise and chosen action valence underlying this strategy are represented in a large network of cortical and subcortical brain regions.

Of note, as also indicated in the List of manuscripts included at the end of the dissertation, the work presented in both empirical chapters is under preparation for publication and can be read as self-contained.

Chapter 4 concludes this dissertation by synthesizing the main findings of Chapters 2 and 3, and discussing them in a broader context. Finally, as an outlook, I outline theoretical and empirical questions arising from these findings.

1.6 References

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2-3), 235–256.

Babayan, B. M., Uchida, N., & Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature Communications*, *9*(1), 1–10.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586.

Bach, D. R., Hulme, O., Penny, W. D., & Dolan, R. J. (2011). The known unknowns: Neural representation of second-order uncertainty, and ambiguity. *Journal of Neuroscience*, *31*(13), 4811–4820.

Bäuerle, N., & Rieder, U. (2011). *Markov decision processes with applications to finance*. Springer Science & Business Media.

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *16*(3/4), 221–229.

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

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, N.J: Princeton University Press.

Berry, D. A., & Fristedt, B. (1985). Bandit problems: Sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, *5*, 71–87.

Bertsekas, D. P. (2000). *Dynamic programming and optimal control* (2nd edition). Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming.* (Vol. 3). Athena Scientific.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733–743.

Brand, H., Woods, P. J., & Sakoda, J. M. (1956). Anticipation of reward as a function of partial reinforcement. *Journal of Experimental Psychology*, *52*(1), 18–22.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv.*

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, *9*, e51260.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). 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.

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, 1–11.

Costa, V. D., & Averbeck, B. B. (2020). Primate orbitofrontal cortex codes information relevant for managing explore–exploit tradeoffs. *Journal of Neuroscience*, *40*(12), 2553–2561.

Daunizeau, J., Den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2010). Observing the observer (i): Meta-bayesian models of learning and decision-making. *PLoS ONE*, *5*(12), e15554.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, *22*(6), 1068–1074.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Dezza, I. C., Angela, J. Y., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific Reports*, *7*(1), 1–13.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fischer, A. G., Bourgeois-Gironde, S., & Ullsperger, M. (2017). Short-term reward experience biases inference despite dissociable neural correlates. *Nature Communications*, *8*(1), 1–14.

Friston, K., & Dolan, R. J. (2010). Computational and dynamic models in neuroimaging. *NeuroImage*, *52*(3), 752–765.

Gabillon, V., Ghavamzadeh, M., & Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in neural information processing systems 25 (NIPS 2012)*, 3212–3220.

Geffner, H., & Bonet, B. (1998). Solving large POMDPs using real time dynamic programming. *Proc. AAAI fall symp. on POMDPs.*

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

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

Gigerenzer, G., Todd, P., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* Oxford University Press.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gijsen, S., Grundei, M., Lange, R. T., Ostwald, D., & Blankenburg, F. (2020). Neural surprise in somatosensory bayesian learning. *bioRxiv.*

Gittins, J., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sarkadi, & I. Vincze (Eds.), *Progress in statistics* (pp. 241–266). North-Holland.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain.* Academic Press.

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

Hamrick, J., & Mohamed, S. (2020). Levels of analysis for machine learning. *arXiv.*

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hauser, T. U., Fiore, V. G., Moutoussis, M., & Dolan, R. J. (2016). Computational psychiatry of ADHD: Neural gain impairments across Marrian levels of analysis. *Trends in Neurosciences*, *39*(2), 63–73.

Hauser, T. U., Iannaccone, R., Ball, J., Mathys, C., Brandeis, D., Walitza, S., & Brem, S. (2014). Role of the medial prefrontal cortex in impaired decision making in juvenile attention-deficit/hyperactivity disorder. *JAMA Psychiatry*, *71*(10), 1165–1173.

Huettel, S., Song, A., & McCarthy, G. (2009). *Functional magnetic resonance imaging*. Oxford University Press, Incorporated.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol*, *8*(3), e1002410.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1), 99–134.

Knight, F. H. (1921). *Risk, Uncertainty and Profit.* Houghton Mifflin Co.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2*, 398.

Korf, R. E. (1990). Real-time heuristic search. *Artificial Intelligence*, *42*(2-3), 189–211.

Korn, C. W., & Bach, D. R. (2018). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications*, *9*(1), 1–15.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019). One-shot learning and behavioral eligibility traces in sequential decision making. *eLife*, *8*, e47463.

Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, *28*(1), 47–65.

Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron*, *104*(1), 164–175.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* Henry Holt; Co., Inc.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

Niv, Y., & Langdon, A. (2016). Reinforcement learning with Marr. *Current Opinion in Behavioral Sciences*, *11*, 67–73.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

O'Reilly, J. X., Jbabdi, S., Rushworth, M. F., & Behrens, T. E. (2013). Brain systems for probabilistic and dynamic prediction: Computational specificity and integration. *PLoS Biol*, *11*(9), e1001662.

Ostwald, D. (2020a, April 6). *Agent-based behavioral modeling* [Webinar]. Retrieved January 2, 2021, from https://www.youtube.com/watch?v=CM4B7veAV00

Ostwald, D. (2020b). *The general linear model 20/21* [Lecture notes]. Retrieved February 23, 2021, from https://www.ewi-psy.fu-berlin.de/einrichtungen/arbeitsbereiche/computational_cogni_neurosc/teaching/The_General_Linear_Model_20_211/The_General_Linear_Model_20_21.pdf

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177–188.

Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming.* John Wiley & Sons.

Rao, R. P. (2010). Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, *4*, 146.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasky (Eds.), *Classical conditioning II* (pp. 64–99). Appleton-Century-Crofts.

Reverdy, P., & Leonard, N. E. (2015). Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, *13*(1), 54–67.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

Schultz, W., Dayan, P., & Montague, R. P. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Schwartenbeck, P., FitzGerald, T. H., & Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, *125*, 578–586.

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *8*, e41703.

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience*, *17*(9), 1249–1254.

Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*, *31*(14), 5526–5539.

Simon, H. (1957). *Models of man; social and rational.* Wiley, New York.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*(2), 351–367.

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, *20*(4), 581.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011). Planning to be surprised: Optimal bayesian exploration in dynamic environments. In J. Schmidhuber, K. R. Thórisson, & M. Looks (Eds.), *Artificial general intelligence* (pp. 41–51). Springer Berlin Heidelberg.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, *39*, 1254–1259.

van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., Li, Y., & Ma, W. J. (2021). Revealing the impact of expertise on human planning with a two-player board game. *PsyArXiv*.

Vilares, I., & Kording, K. (2011). Bayesian models: The structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, *1224*(1), 22.

Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., Rae, J., Mirowski, P., Leibo, J. Z., Santoro, A., et al. (2018). Unsupervised predictive memory in a goal-directed agent. *arXiv*.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 287–298.

Wiering, M., & van Otterlo, M. (2014). *Reinforcement learning: State-of-the-art*. Springer Publishing Company, Incorporated.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, *50*(5), 781–789.

Zhang, S., & Yu, A. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in Neural Information Processing Systems*, *26*.

2 | Belief state-based exploration and exploitation in an information-selective reversal bandit task

2.1 Introduction

Uncertainty is an inherent part of real-life sequential decision making (Bach & Dolan, 2012). Humans often face new and changing environments without being able to directly observe the underlying structure. Consequently, in their quest to maximize the obtained reward, humans have to alternate between exploration and exploitation (Cohen, McClure, & Yu, 2007; Dayan & Daw, 2008; Schwartenbeck et al., 2019; Sutton & Barto, 2018). Exploration refers to action choices that maximize information gain (or, equivalently, minimize uncertainty), and thus advance the knowledge about the structure of the environment. Exploitation refers to action choices that maximize reward gain by harnessing the accumulated knowledge.

A standard testbed to study sequential decision making under uncertainty is the bandit paradigm (Berry & Fristedt, 1985; Robbins, 1952). Two variants of the bandit paradigm have been widely adopted to model real-life explore or exploit problems (Bubeck, Munos, & Stoltz, 2009; Hertwig & Erev, 2009; Sutton & Barto, 2018; Wulff, Mergenthaler-Canseco, & Hertwig, 2018). In both variants, in each trial the deciding agent chooses between a finite set of actions with different expected reward values and observes a reward with a probability specific to the chosen action. While the actions' expected reward values are not directly observable, the agent can estimate them by integrating information from reward observations. The difference between the two variants stems from their respective goals. In the first variant, the goal is to maximize the reward in the final trial. The number of trials preceding the final trial is self-determined by the agent. In contrast, in the second variant, the goal is to maximize the cumulative reward across all trials. Crucially, as a result, in the first variant the reward observation confers information but no reward in all but the final trial. This variant - termed *pure exploration* (Bubeck et al., 2009) or *sampling*

(Hertwig & Erev, 2009) paradigm - thus raises the question as to the extent of exploration by means of the number of trials preceding the final trial in which the accumulated knowledge can be exploited (Ostwald, Starke, & Hertwig, 2015). In the second variant, the reward observation confers both information and reward in each trial. This variant - termed *exploration-exploitation* (Sutton & Barto, 2018) or *partial-feedback* (Hertwig & Erev, 2009) paradigm - thus raises the question of how to strike a balance between exploration and exploitation in each trial. Numerous tasks exist related to either paradigm. For example, the 'observe-or-bet' task (Blanchard & Gershman, 2018; Navarro, Newell, & Schulze, 2016; Tversky & Edwards, 1966) offers an interesting extension of the pure-exploration/sampling paradigm. Similar to the pure-exploration/sampling paradigm, the agent can self-determine the number of pure exploratory actions. However, in contrast, instead of a single action with economic consequence, the agent can take as many as they wish and can also switch back to exploration at any time. To keep exploration and exploitation separated as in the pure-exploration/sampling paradigm, the reward is not observable in the trials with an economic consequence.

A plethora of real-life sequential decision-making problems can be modeled with the pure exploration/sampling and exploration-exploitation/partial-feedback paradigms as well as with related tasks such as the observe or bet task described above. However, these are not suited to model a class of naturalistic problems, in which each available action yields certain reward, but only some yield also information. As an example, consider a patient with high blood pressure. When a new and potentially more effective drug is introduced, the patient can choose between (1) trying out the new drug under medical supervision, where the blood pressure is closely monitored or (2) continuing the old drug without medical supervision. The first option confers both reward (blood pressure in optimal range or not) and information, while the second option confers only reward but no information. Even if the old drug was proven effective in the past, given that the blood pressure can change over time, it might be beneficial for the patient to choose the first option over the second. Situations of this type are similar to the ones modeled with the exploration-exploitation/partial-feedback paradigm in that each action has an economic consequence. Therefore, to maximize the cumulative reward, humans have to balance between exploration and exploitation for each decision. Importantly, however, in these situations information is detached from reward for a subset of actions, akin to the pure exploration/sampling and observe or bet scenarios. Consequently, they arguably pose a more pronounced exploration-exploitation dilemma, because humans are forced to explicitly evaluate the benefit of information gain against the benefit of reward gain.

The goal of this work is to characterize human sequential decision making in such problems. To this end, we introduce a novel information-selective reversal bandit task, which shares key characteristics with the classical symmetric two-armed reversal bandit task (e.g., Bartolo and Averbeck, 2020; Costa, Dal Monte, Lucas, Murray, and Averbeck, 2016; Gläscher, Hampton, and O'Doherty, 2009; Hauser et al., 2014), but in which information is randomly withheld for either the action with the high or the low expected reward value. To formalize different sequential decision-making strategies, we propose a set of agent-based computational models (Russell & Norvig, 2010). In our modeling initiative, we capitalize on recent results showing that one way humans balance between exploration and exploitation is to add an 'information bonus' to the value estimate of an action, which reflects the associated uncertainty (e.g., Gershman, 2018, 2019; Lee, Zhang, Munro, and Steyvers, 2011; Wilson, Geana, White, Ludvig, and Cohen, 2014; Wu, Schulz, Speekenbrink, Nelson, and Meder, 2018). Specifically, we formulate Bayesian agents that represent subjective uncertainty about the structure of the environment in the form of a belief state. The Bayesian agents use the belief state to make either exploitative (i.e., value estimate maximizing actions), explorative (i.e., information bonus maximizing actions), or hybrid explorative-exploitative (i.e., combined value estimate and information bonus maximizing) actions. Notably, we adopt a Bayesian treatment of exploration and quantify the information bonus as the expected Bayesian surprise (Itti & Baldi, 2009; Ostwald et al., 2012; Sun, Gomez, & Schmidhuber, 2011). In addition to the Bayesian agents, we also formulate belief state-free agents that implement simple strategies, such as the 'win-stay-lose-switch' strategy (Robbins, 1952). Upon validating our modeling initiative, we provide empirical evidence for a belief state-based hybrid explorative-exploitative strategy based on choice data from 24 participants. In summary, we demonstrate that, in scenarios where every action has an economic consequence but only some have also an epistemic consequence, humans are guided by their subjective uncertainty to resolve the exploration-exploitation dilemma.

2.2 Methods

2.2.1 Experimental methods

Participants. Young adults were recruited from the Nathan Kline Institute Rockland Sample (NKI-RS), a community-ascertained and comprehensively characterized participant sample of more than 1000 individuals between 6 and 85 years of age (Nooner et al., 2012). We initially intended to enroll individuals from the lower and upper ends of the attention deficit hyperactivity disorder (ADHD) spectrum because we were interested in the relationship between ADHD symptoms and behavioral strategies in our task. Yet, the final sample of 24 individuals (12 female, 23 right-handed, age range: 18-35 years, mean age: 24.5 years, standard deviation age: 5.5 years) represented the mid-range of the ADHD spectrum. Moreover, individuals were only invited if they had no lifetime history of severe neurological or psychiatric disorder. We therefore treated the group of participants as a healthy sample and did not conduct analyses to relate ADHD symptoms to task behavior. For additional details about the recruitment and sample characteristics, please refer to Supplementary Material A.1.

Procedure. The study consisted of a one-time visit of 3.5 hours to the Nathan Kline Institute for Psychiatric Research (Orangeburg, NY, US) and was approved by the local Institutional Review Board. After providing written informed consent, participants were first requested to fill out a series of questionnaires measuring symptoms of ADHD and other mental disorders. Next, participants received detailed written instructions about the information-selective reversal bandit task and were encouraged to ask any clarification questions. Please refer to Supplementary Material A.2 for the instructions provided to the participants. To familiarize participants with the task they next completed a test run of the task on a desktop computer. Finally, participants completed two experimental task runs in a Magnetic Resonance Imaging (MRI) scanner, while behavioral, eye tracking and functional MRI data was acquired. Note that in the current work, we only report results from the analysis of the behavioral data acquired during MR scanning. The visit ended with the participants receiving a reimbursement of $100 (see below for details).

Experimental design. We developed a symmetric two-armed reversal bandit task, in which the available actions were not only associated with varying expected reward values but also with varying information gains (information-selective reversal bandit task, Figure 2.1a). More specifically, on each task trial participants could decide between the actions of choosing a square on the right of a computer screen versus choosing a triangle on the left of the screen, or, between the actions of choosing a square on the left versus choosing a triangle on the right of the screen. Depending on the shape chosen, the action was either lucrative and returned a reward of +1 with a probability of 0.85 and a reward of -1 with a probability of 0.15 or detrimental and returned a reward of +1 with a probability of 0.15 and a reward of -1 with a probability of 0.85. Depending on the laterality of the shape chosen, the action was also either informative and the returned reward was revealed to the participant, or it was non-informative and the returned reward was not revealed to the participant. Specifically, following an informative action an image of a moneybag was displayed for the reward of +1 and an image of a crossed-out moneybag was displayed for the reward of -1. In contrast, following a non-informative action an image of a question mark moneybag was displayed for both the reward of +1 and -1. Importantly, while the actions' lucrativeness was not directly observable and could only be inferred from the revealed rewards, the actions' informativeness was directly observable throughout the experiment. In particular, for half of the participants the right screen side was associated with the informative action and the left screen side was associated with the non-informative action. For the other half of the participants the coupling between screen side and action informativeness was reversed. As a visual reminder, the informative and non-informative sides were also indicated by black and grey backgrounds, respectively. Note that we use the terms informative side and non-informative side in accordance with the action definitions. Similarly, we will hereinafter also use the terms lucrative shape and detrimental shape for simplicity.

The experiment consisted of two runs of 80 trials each. In half of the trials choosing the square was lucrative and choosing the triangle was detrimental. In the other half, choosing the square was detrimental and choosing the triangle was lucrative. We pseudo-randomized the sequence of lucrative shapes such that choosing a certain shape was lucrative for 17-23 consecutive trials upon which the actions' lucrativeness reversed. This yielded a total of three shape lucrativeness reversals (or equivalently, four blocks of trials without a reversal) in a run (Figure 2.1b). Furthermore, we also pseudo-randomized the trial-by-trial sequence of choice options (e.g. choice between the square on the informative side or the triangle on the non-informative side) with two constraints. First, a certain choice option combination occurred for a maximum of five consecutive trials. Second, in 50% of the trials in which the square was lucrative the square was presented on the informative side (and the triangle on the non-informative side) while in the other 50% the square was presented on the non-informative side (and the triangle on the informative side). The same applied to those trials in which the triangle was lucrative. This way we did not only counterbalance the shape-side combinations but also ensured that participants faced a choice between a “lucrative and informative” or “detrimental and non-informative” action (trial type I, abbreviated as L $\cap$ I or D $\cap$ N) in half of the trials. Accordingly, in the other half of the trials participants faced a choice between a “lucrative and non-informative” or “detrimental and informative” action (trial type II, abbreviated as L $\cap$ N or D $\cap$ I; Figure 2.1a). Importantly, for a consistent experiment history across participants, we generated the sequence of lucrative shapes and choice options prior to the study and used the identical trial sequence for all participants.

Participants were encouraged to maximize the cumulative sum of returned rewards across all trials. As an incentive, participants were informed that in addition to a standard reimbursement of $70 for partaking in the study, they would receive a bonus up to $30 depending on their final balance at the end of the second run of the information-selective reversal bandit task. They were not further informed about the balance-bonus conversion rate. In effect, however, all participants were payed the full bonus of $30 as instructed by the Institutional Review Board.

Trial design. Each trial started with the presentation of the two available choice options and participants were given a maximum of 2.5 seconds to indicate their choice (Figure 2.1c). If they responded within the time window the border of the chosen shape turned white to signal the recording of their choice. The duration of this feedback signal depended on the response time such that the choice options and feedback were presented for a total of 3 seconds. Next, a post-choice fixation cross was presented for 3-5 seconds. This was followed by the image representing the respective choice outcome (moneybag, crossed-out moneybag, question mark moneybag) with a presentation duration of 3 seconds. Before a new trial commenced, an inter-trial fixation cross was displayed for 3-5 seconds. If participants did not respond within the time window the message 'too slow' appeared for 0.5 seconds followed by an inter-trial fixation cross, a reward of -1 was automatically registered to their account and the next

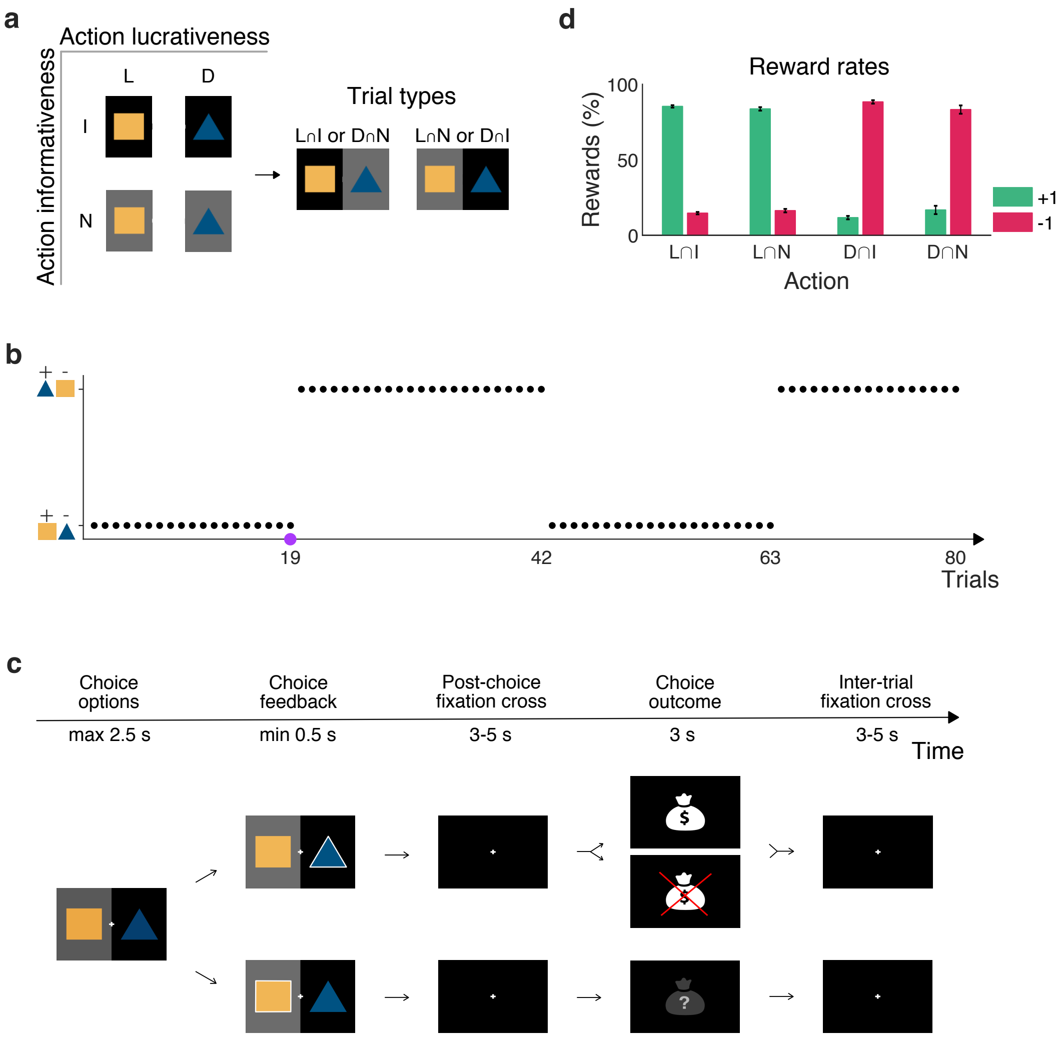

Figure 2.1. Information-selective reversal bandit task. a Experimental design. The actions differed in terms of lucrativeness (lucrative (L) or detrimental (D)) and also in terms if informativeness (informative (I) or non-informative (N)). The former was associated with shapes (square or triangle) and the latter was associated with sides (black or grey). On a given trial, the choice options image represented a choice between either a lucrative and informative or detrimental and non-informative action (trial type I; L ∩ I or D ∩ N) or a lucrative and non-informative or detrimental and informative action (trial type II; L ∩ N or D ∩ I). Note that we here depict the design for those trials in which the square was lucrative (e.g. the nineteenth trial in the first run, see panel **b**). **b** Run design. Every 17 to 23 trials the reward probabilities associated with the shapes reversed. Here, the reversal times of the first run are shown. For example, on trial 19 (marked with a purple dot) the square was lucrative, i.e. choosing the square returned a reward of +1 with a probability of 0.85 and a reward of -1 with a probability of 0.15 and the triangle was detrimental, i.e choosing the triangle returned a reward of +1 with a probability of 0.15 and a reward of -1 with a probability of 0.85. This reversed on trial 20 and choosing the triangle became lucrative and the square became detrimental. A run consisted of 80 trials. **c** Trial design. Participants could indicate their choice within 2.5 seconds of the choice options onset. If they chose the shape on the black side the returned reward was revealed (top). If they chose the shape on the grey side the returned reward was not revealed (bottom). Assuming that this example shows the nineteenth trial of the first run, the options represent a choice between L ∩ N action or D ∩ I action. **d** Quality assurance: Normalized histogram of the online sampled rewards. The reward rates of +1 (green bars) and -1 (red bars) are shown as functions of action. They were consistent with the underlying discrete categorical distributions and only varied with respect to the actions' lucrativeness but not informativeness.

trial commenced. Notably, while the sequences of lucrative shapes and choice options were generated prior to the experiment, the fixation cross duration times and the returned rewards were sampled online as participants interacted with the task. Specifically, the fixation cross duration times were sampled uniformly from an interval of 3 to 5 seconds. The reward values +1 and -1 were sampled from discrete categorical distributions with probabilities 0.85 and 0.15 for the lucrative action and with probabilities 0.15 and 0.85 for the detrimental action, respectively. As shown in 2.1d the rewards returned to the participants followed the underlying distributions.

2.2.2 Descriptive analyses

We evaluated nine summary choice rates for every participant. In particular, we first evaluated overall and trial type-specific valid choice rates. These were defined as the number of valid action choices on all trials, on type I trials and on type II trials divided by the number of all trials, of type I trials and of type II trials, respectively. For example, by design there were 80 trials of type I. If a participant missed to make a valid choice on one of these trials the trial type I valid choice rate was 79/80. We then evaluated the choice rates of the lucrative and informative, lucrative and non-informative, detrimental and informative and detrimental and non-informative actions. These choice rates were computed by dividing the number of respective actions by the number of valid choices of the corresponding trial type. Consequently, the action choice rates of a given trial type were symmetrical, i.e. they summed up to 100%. For example, if a participant made 79 valid action choices on type I trials of which 65 were lucrative and informative and 14 were detrimental and non-informative the lucrative and informative action choice rate was 65/79 and the detrimental and non-informative action choice rate was 14/79. In addition, we evaluated the choice rates of the lucrative actions and the informative actions. These were computed by dividing the sum of the number of lucrative and informative and lucrative and non-informative actions and the sum of the number of lucrative and informative and detrimental and informative actions by the number of valid choices on all trials, respectively. For example, if a participant made 65 lucrative and informative and 58 lucrative and non-informative action choices of the 159 valid choices made on all trials the lucrative action choice rate was 123/159. The individual summary choice rates were then averaged across participants to obtain group summary choice rates and the standard error of the mean (SEM) was evaluated.

In addition to the summary choice rates, we also evaluated trial-by-trial choice rates. Specifically, we computed group trial-by-trial lucrative and informative, lucrative and non-informative, detrimental and informative, and detrimental and non-informative action choice rates. To this end, for every trial we divided the number of respective actions by the number of valid choices on the trial over participants. As a given trial belonged to one of the two trial types it either had lucrative and informative and detrimental and non-informative action choice rates or lucrative and non-informative and detrimental and informative action choice rates. Consequently, in accordance with the summary action choice rates, the choice rates of each trial were symmetrical. For example, by design the first trial of the first run was of type I for every participant. If on this trial 18 participants chose the lucrative and informative action, 5 chose the detrimental and non-informative action and 1 missed to make a valid choice, then the lucrative and informative action choice rate of this trial was 18/23 and the detrimental and non-informative action choice rate was 5/23. Finally, for each trial between two reversals we computed the average group trial-by-trial lucrative and informative and lucrative and non-informative action choice rates across the eight blocks. Note however, that as the trial sequence was pseudo-randomized the average between reversals group trial-by-trial choice rates of a particular trial were computed based on different number of data points. For example, of the eight first trials three were of type I and thus had a group trial-by-trial lucrative and informative action choice rate, while five were of type II and thus had a group trial-by-trial lucrative and non-informative action choice rate. In addition, note that as the number of trials between two reversals varied, there were less than eight 18th to 23rd trials.

2.2.3 Model formulation

Task model To render the task amenable to computational behavioral modelling, we first formulated a model of the task using concepts from the theory of partially observable Markov decision problems (Bertsekas, 2000). Specifically, we represent an experimental run by the tuple

$$M_{\text{Task}} := \left(T, S, A, R, O, p^{s_t^1, a_t}\left(r_t\right), f, g\right), \tag{2.1}$$

where

- T denotes the number of trials, indexed by $t = 1, ..., T$.

- $S := \mathbb{N}_2 \times \mathbb{N}_2$ denotes the set of states $s := (s^1, s^2)$. The first state component s^1 encodes the lucrative shape. Specifically, on trial t, s_t^1 takes on the value 1 if the square is lucrative and takes on the value 2 if the triangle is lucrative. From the perspective of the agent, s^1 is not directly observable. The second state component s^2 encodes the available actions. Specifically, on trial t, s_t^2 takes on the value 1, if the agent can choose between the square on the informative side or the triangle on the non-informative side. If on trial t the agent can choose between the square on the non-informative side or the triangle on the informative side, s_t^2 takes on the value 2. From the perspective of the agent, s^2 is directly observable.

- $A := \{A_1, A_2\}$ denotes the set of state-dependent action sets. Specifically, depending on the observable state component s_t^2 on a given trial t the available actions are either $A_1 := \{1, 4\}$ or $A_2 := \{2, 3\}$ for $s_t^2 = 1$ or $s_t^2 = 2$, respectively. If the available action set is A_1, then the agent can choose between $a = 1$, which corresponds to choosing the square on the informative side or $a = 4$, which corresponds to choosing the triangle on the non-informative side. If the available action set is A_2, then the agent can choose between $a = 2$, which corresponds to choosing the square on the non-informative side or $a = 3$, which corresponds to choosing the triangle on the informative side.

- $R := \{-1, +1\}$ denotes the set of rewards r.

- $O := \mathbb{N}_3$ denotes the set of observations o. $o = 1$ encodes the image of the crossed-out moneybag, $o = 2$ encodes the image of the moneybag and $o = 3$ encodes the image of the question mark moneybag.

- $p^{s_t^1, a_t}(r_t)$ is the state- and action-dependent reward distribution. For each combination of $s^1 \in S^1$ and $a \in A_{s^2}$, the state- and action-dependent reward distribution conforms to a discrete categorical distribution over r_t with probability parameters listed in the first panel of Table 2.1. As an example, consider $s^1 = 1$ (square is lucrative) and $a = 1$ (square on the informative side chosen). In this case, a reward of -1 is returned with a probability of 0.15 and a reward of $+1$ is returned with a probability of 0.85. On the other hand, if $s^1 = 2$ (triangle is lucrative) and $a = 1$ (square on the informative side chosen), the reward probabilities are reversed.

- f is the state evolution function, which specifies the value the state s_t takes

s_t^1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
a_t	1	1	2	2	3	3	4	4	1	1	2	2	3	3	4	4
r_t	-1	+1	-1	+1	-1	+1	-1	+1	-1	+1	-1	+1	-1	+1	-1	+1
$p^{s_t^1,a_t}(r_t)$	0.15	0.85	0.15	0.85	0.85	0.15	0.85	0.15	0.85	0.15	0.85	0.15	0.15	0.85	0.15	0.85

a_t	1	1	2	2	3	3	4	4
r_t	-1	+1	-1	+1	-1	+1	-1	+1
o_t	1	2	3	3	1	2	3	3

s_t^1	1	1	2	2
s_{t+1}^1	1	2	1	2
$p\left(s_{t+1}^1 \vert s_t^1\right)$	0.9625	0.0375	0.0375	0.9625

Table 2.1. Formal task components Upper table shows the state- and action-dependent reward distribution $p^{s_t^1,a_t}(r_t)$, middle table shows the observation function g and lower table shows the action-independent state transition distribution $p\left(s_{t+1}^1|s_t^1\right)$.

on at trial t,

$$f : \mathbb{N}_T \to S, t \mapsto f(t) := s_t. \tag{2.2}$$

f is defined in a tabular form and corresponds to the sequence of lucrative shapes and choice options presented to all participants (cf. Supplementary Material A.3).

- g is the observation function

$$g : A \times R \to O, (a, r) \mapsto g(a, r) := o \tag{2.3}$$

as defined in the second panel of Table 2.1. For the informative actions $a = 1$ and $a = 3$, g is injective: The reward $r = -1$ is mapped onto the observation $o = 1$, corresponding to the image of the crossed-out moneybag, while the reward $r = +1$ is mapped onto the observation $o = 2$, corresponding to the image of the moneybag. For the non-informative actions $a = 2$ and $a = 4$, g is a not injective: Both rewards $r = -1$ and $r = +1$ are mapped onto the observation $o = 3$, corresponding to the image of the question mark moneybag.

Agent models. We designed five agent models denoted by A1, A2, A3, C1, C2 to account for the putative cognitive processes underlying participants' choices. Before we introduce the individual characteristics of these agents we first represent the general structure of an agent interacting with an experimental run. This takes the form of a tuple

$$M_{\text{Agent}} := \left(T, S, A, R, O, p\left(s_1^1\right), p\left(s_{t+1}^1|s_t^1\right), p^{a_t}\left(o_t|s_t^1\right), p^{a_t}\left(r_t|s_t^1\right)\right), \tag{2.4}$$

where

- T, S, A, R and O are defined as the corresponding sets of the task model M_{Task}.

- $p\left(s_1^1\right)$ denotes the initial agent belief state, which specifies the agent's subjective uncertainty over the non-observable state component s_1^1 at trial $t = 1$. $p\left(s_1^1\right)$ is defined in terms of the discrete categorical distribution

  $$p(s_1^1 = 1) = 0.5 \text{ and } p(s_1^1 = 2) = 0.5. \tag{2.5}$$

  As $p(s_1^1)$ is fully parameterized by specifying $p(s_1^1 = 1)$ we hereinafter also represent the initial belief state with the scalar $b_1 := p(s_1^1 = 1)$.

- $p\left(s_{t+1}^1|s_t^1\right)$ is the state-state transition distribution, which specifies the agent's subjective uncertainty over the non-observable state component s_{t+1}^1 at trial $t+1$ given the non-observable state component s^1 at trial t. More specifically, for each $s^1 \in S^1$, the state-state transition distribution corresponds to a discrete categorical distribution over s_{t+1}^1 with probability parameters listed in the third panel of Table 2.1. Note that the trial-by-trial state transitions are probabilistic because from the perspective of the agent a reversal in the shapes' lucrativeness could happen between any two trials. This is in contrast with the state evolution from the task perspective, which - given the apriori defined sequence of lucrative shapes - is deterministic (eq. 2.2). Crucially, participants were informed that a reversal would happen 1-4 times in a run but they were not further informed about the approximate number of trials without a reversal. Therefore, we equipped the agent with a constant reversal probability of 0.0375, which reflects the true reversal frequency in a run (there were 3 reversals across the 80 trials). For example, if $s_t^1 = 1$ (square is lucrative) the agent allocates the probability of 0.9625 that on the next trial s_{t+1}^1 again takes on the value 1 (square is lucrative) and the probability of 0.0375 that its value changes to 2 (triangle is lucrative).

- $p^{a_t}\left(r_t|s_t^1\right)$ is the action-dependent state-conditional reward distribution, which specifies the agent's subjective uncertainty over the reward r_t given the non-observable state component s^1 and action a at trial t. More specifically, for each combination of $s^1 \in S^1$ and $a \in A_{s^2}$, the action-dependent state-conditional reward distribution defines a discrete categorical distribution over r_t with probability parameters corresponding to

$$p^{a_t=a}\left(r_t = r|s_t^1 = s^1\right) := p^{s_t^1=s^1,a_t=a}\left(r_t = r\right). \tag{2.6}$$

Notice that the only difference between the agent's action-dependent state-conditional reward distribution and the task's state- and action-dependent reward distribution is that for the former the state is conceived as a random variable, while for the latter the state is conceived as a parameter. We equipped the agent with the true reward emission probabilities to reflect the task instructions. In particular, participants were truthfully informed that choosing the lucrative shape would return a reward of +1 with a high probability and a reward of -1 with a low probability and, that choosing the detrimental shape would return a reward of +1 with a low probability and a reward of -1 with a high probability.

- $p^{a_t}\left(o_t|s_t^1\right)$ is the action-dependent state-conditional observation distribution, which specifies the agent's subjective uncertainty over the observation o_t given the non-observable state component s^1 and action a at trial t. In detail, for each combination of $s^1 \in S^1$ and $a \in A_{s^2}$, the action-dependent state-conditional observation distribution corresponds to a discrete categorical distribution over o_t with probability parameters resulting from transforming the distribution of r_t by the observation function g. Formally,

$$p^{a_t=a}\left(o_t = o|s_t^1 = s^1\right) := \sum_{\{r|g(a,r)=o\}} p^{a_t=a}\left(r_t = r|s_t^1 = s^1\right). \tag{2.7}$$

For the informative actions $a \in \{1, 3\}$, it thus follows that

$$p^{a_t=a}\left(o_t = 1|s_t^1 = s^1\right) = p^{a_t=a}\left(r_t = -1|s_t^1 = s^1\right) \tag{2.8}$$

and

$$p^{a_t=a}\left(o_t = 2|s_t^1 = s^1\right) = p^{a_t=a}\left(r_t = +1|s_t^1 = s^1\right). \tag{2.9}$$

For non-informative actions $a \in \{2, 4\}$, on the other hand, it follows that

$$p^{a_t=a}\left(o_t = 3|s_t^1 = s^1\right) = p^{a_t=a}\left(r_t = -1|s_t^1 = s^1\right) + p^{a_t=a}\left(r_t = 1|s_t^1 = s^1\right) = 1. \tag{2.10}$$

As an example, consider the case $s^1 = 1$ (square is lucrative) and $a = 1$ (choose square on the informative side). The agent allocates the same probabilities to observing either the image of the crossed-out moneybag or the image of the moneybag as to obtaining a reward of -1 or $+1$, respectively. Alternatively, if for example $s^1 = 1$ (square is lucrative) and $a = 4$ (choose triangle on the non-informative side) the agent allocates a probability of 1 to observing the image of the question mark moneybag.

Bayesian agents (A1, A2 and A3) The Bayesian agents maintain a belief state, which subserves their action choice. Specifically, the distributions $p\left(s_1^1\right)$, $p\left(s_{t+1}^1|s_t^1\right)$ and $p^{a_t}\left(o_t|s_t^1\right)$ of M_{Agent} induce an action-dependent joint probability distribution $p^{a_{1:T-1}}\left(s_{1:T}^1, o_{1:T-1}\right)$. This allows for the recursive evaluation of the belief state $p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right)$ on trial t given the history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$ as

$$p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) = \frac{\sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}{\sum_{s_t^1}\sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}, \tag{2.11}$$

with the prior belief state given by $p\left(s_1^1\right)$ on trial $t = 1$. For a derivation of eq. 2.11, please refer to Supplementary Material A.4. Intuitively, the Bayesian agents thus update their belief state in a trial-by-trial fashion based on the observation made after choosing a shape on either side and by accounting for a reversal in the shapes' lucrativeness. In our implementation of the belief state update we represented the distributions $p\left(s_1^1\right)$, $p\left(s_{t+1}^1|s_t^1\right)$ and $p^{a_t}\left(o_t|s_t^1\right)$ with stochastic matrices and evaluated the belief state using matrix multiplication in order to optimize computational time (cf. Supplementary Material A.5).

Based on their belief state representation, the Bayesian agents then decide for an action based on a combination of an action valence function, which evaluates the desirability of a given action in the light of the agent's current belief state, and a decision function, which selects the maximal desirable action as the action to issue. Specifically, the scalar representation of the belief state

$$b_t := p^{a_{1:t-1}}\left(s_t^1 = 1|o_{1:t-1}\right) \tag{2.12}$$

constitutes the basis for action evaluation by means of an action valence function

$$v : A \times [0,1] \to \mathbb{R}, (a,b) \mapsto v\left(a,b\right). \tag{2.13}$$

As detailed below, the exact forms of the valence function differ between agents A1, A2, and A3. However, to realize an action, all Bayesian agents pass the evaluated action valences to a maximizing decision rule of the form

$$d : \mathbb{R} \times [0, 1] \to A_{s^2}, v(\cdot, b) \mapsto d(v(\cdot, b)) := \arg\max_{a \in A_{s^2}} v(a, b). \tag{2.14}$$

On every trial, the Bayesian agents thus choose the action with the highest valence.

A1: The belief state-based exploitative agent Agent A1 uses its belief state to maximize the immediate reward gain. To this end, agent A1 uses an action valence function that allocates action valences based on the action-dependent expected reward under the current belief state,

$$v_{\mathrm{A1}}(a, b) := b\mathbb{E}_{p^a\left(r_t|s_t^1=1\right)}(r_t) + (1 - b)\,\mathbb{E}_{p^a\left(r_t|s_t^1=2\right)}(r_t). \tag{2.15}$$

The panels of Figure 2.2a visualize the A1 valences for actions $a \in A_1$ (choose square on the informative side or triangle on the non-informative side; left panel) and $a \in A_2$ (choose square on the non-informative side or triangle on the informative side; right panel) as functions of the belief state b. The expected reward is

$$\mathbb{E}_{p^a(r_t|s_t)}(r_t) = 0.85 \cdot -1 + 0.15 \cdot 1 = -0.7 \tag{2.16}$$

for choosing the detrimental shape and

$$\mathbb{E}_{p^a(r_t|s_t)}(r_t) = 0.85 \cdot 1 + 0.15 \cdot -1 = 0.7 \tag{2.17}$$

for choosing the lucrative shape. Consequently, the more certain A1 is that a given shape is lucrative (as b gets closer to 0 or 1 from 0.5) the higher the belief state-weighted expected reward for choosing that shape and accordingly, the lower it is for choosing the other shape. As the belief state-weighted expected reward is irrespective of the side of the shape, in the case of both sets of available actions A1 allocates valences without taking the actions' informativeness into account.

A2: The belief state-based explorative agent Agent A2 explores its belief state to maximize the immediate information gain. To this end, on each trial $t \in \mathbb{N}_T$ A2 allocates a valence to each available action $a \in A_{s^2}$ based on the expected Bayesian surprise (Itti & Baldi, 2009). Formally,

$$v_{\mathrm{A2}}\left(a,b\right) := \sum_{o\in O} p^{a_{1:t-1},a}\left(o_t = o|o_{1:t-1}\right) KL\left(p^{a_{1:t-1},a}\left(s^1_{t+1}|o_{1:t-1},o\right)\middle|\middle|p^{a_{1:t-1}}\left(s^1_t|o_{1:t-1}\right)\right), \tag{2.18}$$

where

$$p^{a_{1:t-1},a}\left(o_t|o_{1:t-1}\right) = bp^a\left(o_t|s^1_t = 1\right) + \left(1-b\right)p^a\left(o_t|s^1_t = 2\right) \tag{2.19}$$

is the posterior predictive distribution and

$$KL\left(p^{a_{1:t-1},a}\left(s^1_{t+1}|o_{1:t-1},o\right)\middle|\middle|p^{a_{1:t-1}}\left(s^1_t|o_{1:t-1}\right)\right) = b^{a,o}\ln\left(\frac{b^{a,o}}{b}\right) + \left(1-b^{a,o}\right)\ln\left(\frac{\left(1-b^{a,o}\right)}{\left(1-b\right)}\right) \tag{2.20}$$

is the Kullback-Leibler (KL) divergence. The former specifies the agent's subjective uncertainty over the observation o_t given action a in trial t and the history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$. For a derivation of eq. 2.19, please refer to Supplementary Material A.4. For implementational details, please refer to Supplementary Material A.5. The latter corresponds to the Bayesian surprise. Specifically, it quantifies the shift between the agent's belief state b at trial t and the simulated belief state $b^{a,o}$ at trial $t+1$ that would result after action a and observation o in trial t. The panels of Figure 2.2b visualize the A2 valences for actions $a \in A_1$ (choose square on the informative side or triangle on the non-informative side; left panel) and $a \in A_2$ (choose square on the non-informative side or triangle on the informative side; right panel) as functions of the belief state b. Choosing the shape on the non-informative side does not deliver reward information. Therefore, the expected Bayesian surprise-based A2 valence is always higher for the informative action, irrespective of the agent's belief state. Yet, the difference between the informative and non-informative action valences depends on the belief state. Specifically, in contrast to A1, the more uncertain A2 is about the lucrative shape (as b gets closer to 0.5 from 1 or 0) the larger the difference between the valences and thus the stronger the agent's preference for the informative action.

A3: The belief state-based hybrid explorative-exploitative agent Agent A3 combines the choice strategies of A1 and A2 and uses its belief state to maximize the combination of immediate reward gain and information gain. Formally, in each trial $t \in \mathbb{N}_T$ for each available action $a \in A_{s^2}$, A3 evaluates its action valences based on the convex combination of the action valences of agents A1 and A3

$$v_{\mathrm{A3}}(a,b) := \lambda v_{\mathrm{A1}}(a,b) + (1-\lambda)\, v_{\mathrm{A2}}(a,b), \tag{2.21}$$

where $\lambda \in [0,1]$ is the weighting parameter.

The panels of Figures 2.2c and d visualize the A3 valences for actions $a \in A_1$ (choose square on the informative side or triangle on the non-informative side; left panels) and $a \in A_2$ (choose square on the non-informative side or triangle on the informative side; right panels) as functions of the belief state b for λ values 0.5 and 0.25, respectively. For $\lambda = 1$ the valences of agent A3 correspond to the valences of A1 and for $\lambda = 0$ they correspond to the valences of A2. For λ values of the interval $]0,1[$, the strategy of A3 is a mixture of A1 and A2: For less extreme belief values (b close to 0.5) A3 allocates a higher valence to choosing the shape on the informative side even if the agent allocates a lower probability to that shape being lucrative. This shows the contribution of A2. For more extreme belief state values (b close to 0 or 1) A3 allocates a higher valence to choosing the shape with the higher probability to be lucrative even if the action is non-informative. This shows the contribution of A1. Note, however, that a λ value of 0.5 should not be understood as A3 resembling 50% the strategy of A1 and 50% the strategy of A2. The reason for this is that A3 applies a convex combination of A1 and A2 valences and they have different ranges. Therefore, while for $\lambda = 0.5$ the valences of A3 primarily reflect the contribution of A1 (Figure 2.2c), the contribution of A2 becomes evident for $\lambda = 0.25$ (2.2d).

Control agents C1 and C2 The control agents C1 and C2 rely on heuristic choice strategies. Because C1 and C2 do not represent a belief state, their action valence function is a function of a only,

$$v : A \to \mathbb{R}, a \mapsto v\,(a)\,. \tag{2.22}$$

To realize an action on trial $t \in \mathbb{N}_T$, both agents use a probabilistic decision rule. Specifically, C1 and C2 directly translate the action valences into action and observation history-dependent choice probabilities.

C1: The belief state-free random choice agent C1 is the simplest agent and may be considered a cognitive null model. This agent does not have an optimization aim based on which it could differentiate between actions.

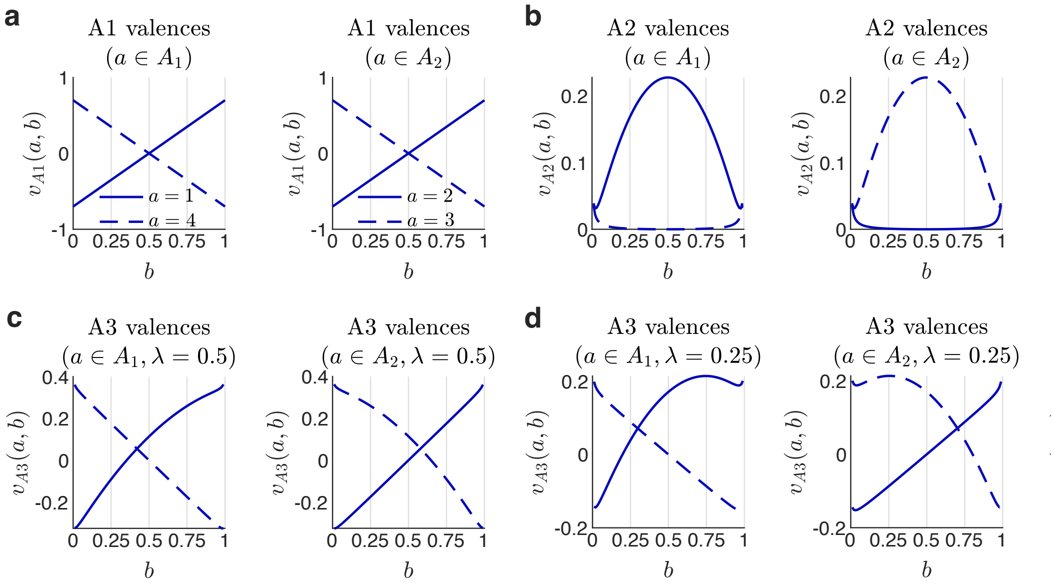

Figure 2.2. Action valences of the Bayesian agents. a Action valences of agent A1 as functions of the belief state. A1 allocates action valences based on the belief state-weighted expected reward. As the expected rewards for choosing the lucrative or detrimental shape are constant, the more extreme the agent's belief that a given shape is lucrative the higher the valence it allocates to choosing the corresponding shape and the lower the valence it allocates to choosing the other shape. The valences of A1 do not depend on the actions' informativeness and therefore the two panels are identical. **b** Action valences of agent A2 as functions of the belief state. A2 allocates action valences based on the expected Bayesian surprise, which is higher for the informative action than for the non-informative action and therefore the two panels are converse. The less extreme the agent's belief that a given shape is lucrative the larger the difference. **c-d** Action valences of agent A3 with $\lambda = 0.5$ and $\lambda = 0.25$ as functions of the belief state. A3 allocates action valences based on the convex combination A1 and A2 action valences. The higher the value of λ the more the valences of A3 resemble the valences of A1 and correspondingly, the lower the value of λ the more the valences of A3 resemble the valences of A2.

Therefore, C1 allocates equal valences to all available actions $a \in A_{s^2}$,

$$v_{\text{C1}}(a) := \frac{1}{|A_{s^2}|} = 0.5. \tag{2.23}$$

C2: The belief state-free win-stay-lose-switch agent Agent C2 aims to maximize the immediate reward without a belief state. To this end, C2 adopts a heuristic win-stay-lose-switch strategy (Robbins, 1952). Specifically, on each trial $t \in \mathbb{N}_T$ C2 considers which shape to choose based on previous observations that signal the reward value. This agent does not take the sides into account when allocating action valences. Formally, on trial $t = 1$ the strategy of C2 corresponds to

$$v_{\text{C2}_1}(a) := 0.5 \text{ for all } a \in A_{s_1^2}. \tag{2.24}$$

Then, on trials $t = 2, 3, ..., T$ agent C2 allocates action valences according to

$$v_{\mathrm{C2}_t}(a) := \begin{cases} 0, & \text{if } o_{t-1} = 1 \text{ and } a \in \mathcal{A}_{t-1} \text{ or } o_{t-1} = 2 \text{ and } a \notin \mathcal{A}_{t-1} \\ 1, & \text{if } o_{t-1} = 2 \text{ and } a \in \mathcal{A}_{t-1} \text{ or } o_{t-1} = 1 \text{ and } a \notin \mathcal{A}_{t-1} \\ v_{\mathrm{C2}_{t-1}}(a), & \text{if } o_{t-1} = 3 \end{cases}, \tag{2.25}$$

where $\mathcal{A}$ denotes the set of actions of choosing a given shape and thus $\mathcal{A} := \{1, 2\}$ for the actions choose square and $\mathcal{A} := \{3, 4\}$ for the actions choose triangle. In words, C2 allocates equal initial action valences, because on trial $t = 1$ no previous observations are available and therefore the agent can not differentiate between actions. Then, on trials $t = 2, 3, ..., T$ the valences C2 allocates depend on the observation on trial $t-1$. Specifically, if on trial $t-1$ the choice of a shape results in the observation $o = 1$, i.e. the image of the crossed-out moneybag, then on trial t the agent allocates a valence of 0 to choosing the same shape and a valence of 1 to choosing the other shape. In contrast, if on trial $t-1$ the choice of a shape results in the observation $o = 2$, i.e. the image of the moneybag, then on trial t the agent allocates a valence of 1 to choosing the same shape and a valence of 0 to choosing the other shape. Crucially, if on trial $t-1$ the choice of a shape results in the observation $o = 3$, i.e. the image of the question mark moneybag, the value of the returned reward is not signalled to the agent and therefore on trial t C2 relies on its valence allocation scheme from trial $t-1$. That is, the valence C2 allocates to choosing a given shape on trial t corresponds to the valence the agent allocated to choosing that shape on trial $t-1$.

2.2.4 Model evaluation and validation

Data analysis models To evaluate the agent models in light of the participants' data, we first embedded the agent models into a statistical framework to account for post-decision noise. In particular, for agents A1, A2, A3 and C2 we formulated the data analysis models by combining the agent-specific valences with the softmax decision rule (Reverdy & Leonard, 2015). Specifically, we defined the probability of action a given the history of actions $a_{1:t-1}$ and observations $o_{1:t-1}$ as

$$p^{\tau}\left(a_t = a | a_{1:t-1}, o_{1:t-1}\right) := \frac{\exp\left(\tau^{-1} v(a, \cdot)\right)}{\sum_{\tilde{a} \in A_{s2}} \exp\left(\tau^{-1} v(\tilde{a}, \cdot)\right)}, \tag{2.26}$$

where for each agent $v(a,\cdot)$ is substituted with the agent-specific valence function. Parameter $\tau \in \mathbb{R}_{>0}$ encodes the level of post-decision noise: The lower the value of τ the higher the probability that the action with the higher valence is realized and thus the lower the post-decision noise. Notably, as agent C1 allocates equal action valences throughout, for any τ value the softmax decision rule would return uniform probabilities. Therefore, for this agent a softmax decision rule is not necessitated and the data analysis model corresponds to

$$p\left(a_t = a|a_{1:t-1}, o_{1:t-1}\right) = p\left(a_t = a|a_{1:t-1}\right) := v_t\left(a\right). \tag{2.27}$$

Parameter estimation We used a maximum log likelihood (ML)-based approach to estimate the parameters of the data analysis models of A1, A2, A3 and C2 based on the experimentally acquired data. Specifically, we assumed conditionally independently and identically distributed actions and thus for each participant defined the log likelihood function of each model as

$$l_N : \Theta \to \mathbb{R}, \theta \mapsto l_N(\theta) := \ln \prod_{n=1}^{N} p^{\theta}\left(a_n|a_{1:n-1}, o_{1:n-1}\right) = \sum_{n=1}^{N} \ln p^{\theta}\left(a_n|a_{1:n-1}, o_{1:n-1}\right). \tag{2.28}$$

Note that we here replaced t with n and T with N. The reason for this is to emphasize that we only considered trials with a valid choice and thus $n \in \mathbb{N}_N$ denotes the participant's nth valid trial. For agents A1, A2 and C2 the log likelihood l_N is a function of parameter τ of the softmax decision rule and thus, for these agents $\theta := \tau$. For A3, the log likelihood l_N is additionally a function of the weighting parameter λ and thus, for this agent $\theta := (\tau, \lambda)$. For every agent and participant we estimated the free parameters by maximizing the log likelihood function l_N using Matlab's constrained nonlinear optimization routine *fmincon* (Byrd, Gilbert, & Nocedal, 2000; Byrd, Hribar, & Nocedal, 1999; Waltz, Morales, Nocedal, & Orban, 2006). We set the boundary constraints to 0.01 and 2.5 for τ and to 0 and 1 for λ. The initial values were chosen random from a continuous uniform distribution between the boundary constraints. To mitigate the risk of finding local instead of global maxima we repeated the parameter estimation procedure 10-times and recorded the parameter estimates of the repeat with the highest maximum log likelihood. Note that as the data analysis model of C1 does not have free parameters, for this agent the log likelihood function l_N is specified directly.

Model comparison To compare the models, we first computed the Bayesian Information Criterion (BIC; Schwarz, 1978) for each agent and participant as

$$\mathrm{BIC} = l_N(\hat{\theta}) - \frac{k}{2} \ln N, \tag{2.29}$$

where k is the number of free parameters. The BIC scores of all agents and participants were subsequently entered for random-effects Bayesian model selection as implemented in the *spm_BMS* function in the SPM toolbox to obtain protected exceedance probabilities (www.fil.ion.ucl.ac.uk/spm/; Rigoux, Stephan, Friston, and Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, and Friston, 2009). These indicated the group-level probability that a particular model was more likely than any other model of the model space. In addition, we computed a pseudo-r^2 statistic ρ (McFadden, 1973) for each participant and agents A1, A2, A3 and C2 as

$$\rho = 1 - \frac{l_{N_{\mathrm{Agent} \neq \mathrm{C1}}}(\hat{\theta})}{l_{N_{\mathrm{C1}}}} \tag{2.30}$$

to express the variance of a participant's choices explained by a strategy that differentiates between actions as implemented by agents A1, A2, A3 and C2 compared to a non-differentiating strategy as implemented by agent C1.

Model recovery We conducted model recovery analysis to validate our modelling initiative and test if we can reliably distinguish the agent-specific data analysis models from each other. Specifically, for every model we first generated synthetic data and then evaluated all models based on these data using ML-based approach as described above. The data corresponded to the agent-specific actions on 160 trials, with trial sequence identical to the participants' trial sequence and agent action-dependent observations as given by the observation function g and the state- and action-dependent reward distribution $p^{s_t^1, a_t}(r_t)$. For agent C1 data were generated directly. For agents A1, A2, A3 and C2 data were generated with τ parameter values between 0.05 and 2.5 with an increment of 0.05 to evaluate if we can identify each model for different levels of post-decision noise across the τ parameter space used for estimation. In addition, for A3 data were generated with parameter values $\lambda \in \{0.1, 0.25, 0.3, 0.5, 0.7, 0.9\}$ to probe if it can be distinguished from agents A1 and A2, for low to high values of the weighting parameter. With each data generating model and parameter value we simulated 24 data sets, the same number as the participant sample size. Based on each synthetic data set we evaluated the maximum log likelihood of all models and conducted model comparison analysis as described above. We repeated this procedure 10-times and computed the average protected exceedance probabilities across repeats.

Parameter recovery To test if we can reliably estimate the free parameters of the data analysis models of A1, A2, A3 and C2 across the parameter space as defined by the estimation boundary constraints, we conducted parameter recovery analysis. To this end, we used the synthetic ML parameter estimates obtained during model recovery analysis. More precisely, for a given data generating model and simulation parameter we computed the average ML parameter estimates of the same model across the group of 24 synthetic data sets and 10 repeats and compared it with the simulation parameter.

Winning model validation To validate the most plausible model (Wilson & Collins, 2019), we generated synthetic data sets with it for each participant using the participant's ML parameter estimates. Like in the model recovery analysis, the data corresponded to the agent's actions on 160 trials, with trial sequence identical to the participants' trial sequence and the agent's action-dependent observations as given by the observation function g and the state- and action-dependent reward distribution $p^{s_t^1, a_t}(r_t)$. We then entered each synthetic participant data set for descriptive analyses as described above. Specifically, we computed the same nine summary choice rates as for the participants and repeated the data generation and summary choice rates evaluation 100-times. We then computed the averages across simulation repeats and entered them for synthetic group-summary choice rates evaluation. Furthermore, we also evaluated synthetic group trial-by-trial choice rates as well as their between reversals averages based on the simulation repeats-averaged agent actions.

In addition and finally, we performed model and parameter recovery analysis of the winning model with the participants' ML parameter estimates to ensure that the model can reliably account for the empirically acquired data. We used the same recovery procedures as described above with the exception that within a single repeat each of the 24 data sets was generated with a different participant's ML parameter estimates. Note that to avoid confusion, we hereinafter use the subscript $_p$ to refer to empirical (participant) parameter estimates.

2.3 Results

2.3.1 Descriptive results

Participants completed the majority of trials with an overall valid choice rate of 97.86% ± 0.62. There was no difference in the number of missed choices with respect to trial types: The valid choice rate on type I trials was 97.97% ± 0.62 and the valid choice rate on type II trials was 97.76% ± 0.68. On type I trials, the majority of action choices were lucrative and informative (87.45%±1.53) while only a few were detrimental and non-informative (12.55%± 1.53). The difference between the choice rates on type II trials was less pronounced: 66.01% ± 2.28 of the action choices were lucrative and non-informative while 33.99%±2.28 were detrimental and informative. Summed over informative and non-informative action choices, the lucrative action choice rate was 76.74%±1.7, whereas summed over lucrative and detrimental action choices, the informative action choice rate was 60.74% ± 0.92. Notably, participants made significantly more lucrative choices, if the lucrative action was also informative (and the alternative detrimental and non-informative) compared to lucrative choices, if the lucrative action was also non-informative (and the alternative detrimental and informative; two-sided paired sample t-test, $t(23) = 11.55, p < 0.001$). Together, these results suggest that while participants' choices were primarily guided by the actions' lucrativeness, participants also took the action's informativeness into account.

We next evaluated group trial-by-trial choice rates. As shown in Figure 2.3a, on the majority of trials the lucrative action choice rates prevailed over the detrimental action choice rates. This was more pronounced for the trial-by-trial lucrative and informative action choice rates. Crucially, both the average lucrative and informative and the average lucrative and non-informative action choice rates showed an overall increase between two reversals (Figure 2.3b). This indicates that participants gradually resolved their uncertainty about the currently lucrative shape. Moreover, although the average lucrative and informative action choice rate was larger than the average lucrative and non-informative action choice rate on all trials between two reversals (with the exception of the 18th trial), their difference decreased slightly between the first trials after and the last trials before a reversal (2.3c). This suggests that with decreasing uncertainty about the currently lucrative shape participants took the actions' informativeness less into account.

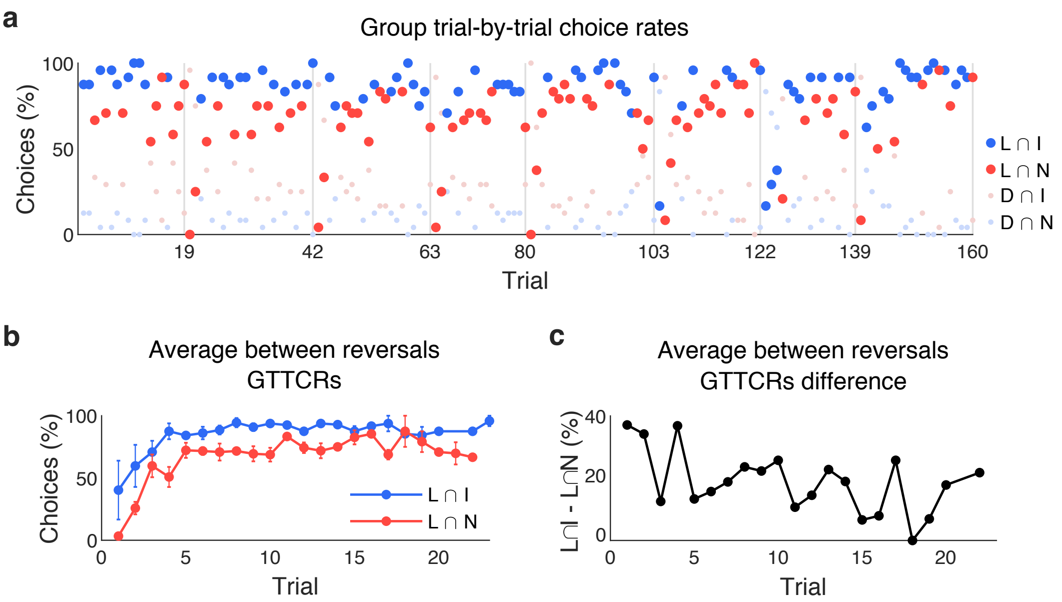

Figure 2.3. Descriptive results. a Group trial-by-trial lucrative and informative action, lucrative and non-informative action, detrimental and informative action and detrimental and non-informative action choice rates. The light grey vertical lines at the annotated trials represent the trials after which a reversal occurred (in addition, the lines at t = 80 and t = 160 mark the end of the first and second run, respectively). The choice rates marked by light and dark hues of the same color were symmetrical. On most trials the lucrative action choice rates prevailed. This was more pronounced for the trial-by-trial lucrative and informative action choice rates. **b** Average between reversals group trial-by-trial lucrative and informative action and lucrative and non-informative action choice rates. Both choice rates increased over the trials between two reversals. **c** Average between reversals group trial-by-trial lucrative and informative action and lucrative and non-informative action choice rates difference. The difference decreased between the first trials after and the last trials before a reversal (compare for instance the distinct spikes at trials four and 18). Note that the acronym GTTCRs in the titles of **b** and **c** stands for group trial-by-trial choice rates. Error bars display the SEM.

2.3.2 Modeling results

Model and parameter recovery results We validated our modeling initiative by conducting recovery analyses of the agent-specific data analysis models and their free parameters across the entire parameter space. Figure 2.4 summarizes the results of the model recovery analyses. For each data generating model the corresponding subplot shows the protected exceedance probabilities of each data evaluation model. For the data generating models of C2, A1, A2 and A3 these probabilities are shown as functions of the post-decision noise parameter τ used for data generation. As shown in Figure 2.4a for data generated with C1 the protected exceedance probability was maximal for C1, which indicates that the data analysis model of C1 is identifiable. For data generated with C2 and A1 the protected exceedance probabilities were maximal for C2 and A1, respectively, for all values of τ. This indicates that the data analysis models of C2 and A1 are identifiable for low to high levels of post-decision noise. For data generated with A2 the protected exceedance probabilities were maximal for A2 for τ values up to 0.35 and for C1 for larger values of τ. This indicates that A2 is identifiable for low but not for high levels of post-post-decision noise. For data generated with A3 with $\lambda = 0.25$ the protected exceedance probabilities were maximal for A3 up to $\tau = 0.25$, after which A1 and then C1 exceeded. This indicates that, similarly to A2, A3 with $\lambda = 0.25$ is identifiable for low but not for high levels of post-decision noise. With increasing noise the data is better accounted for by A1 and eventually by C1.

Notably, for data generated with A3, model recovery depend not only on the post-decision noise parameter τ but also on the weighting parameter λ. As shown in Figure 2.4b for λ values up to 0.5 the protected exceedance probabilities were maximal for A3 for small τ values $(0.25, 0, 2$ and 0.1, respectively). Then, for $\lambda = 0.1$, A2 and then C1 prevailed, while for $\lambda = 0.3$ and $\lambda = 0.5$, A1 and then C1 prevailed. Due to the increasing similarity between the valences of A3 and A1 (cf. Figure 2.2c-d), for λ values larger than 0.5 the protected exceedance probability profiles shifted towards that of A1. More precisely, for $\lambda = 0.7$ and $\lambda = 0.9$ the protected exceedance probabilities were maximal for A1 up to $\tau = 1.9$ and $\tau = 2.4$, respectively, and for C1 for larger τ values. Together, consistent with the interpretation of the weighting parameter, these results imply that the data analysis model of A3 is identifiable for low to medium λ values for low levels of post-decision noise. Otherwise, A3 can not be distinguished from A1 or A2 and eventually, from C1.

The results of the parameter recovery analyses are visualized in Figure 2.5. Here, model-specific ML parameter estimates are displayed as functions of the post-decision noise parameter τ. The recovered ML parameter estimates $\hat{\tau}_{C2}$, $\hat{\tau}_{A1}$, $\hat{\tau}_{A2}$ and $\hat{\tau}_{A3}$ were consistent with the simulation parameters $\tau_{C2} = \tau_{A1} = \tau_{A2} = \tau_{A3}$ for small values. Otherwise, the parameters were first over- then underestimated. As shown in Figure 2.5a this bias was subtle for C2 and A1 and only affected large τ values (between approximately 1.5 and 2.5) while it was more pronounced for A2 and A3 (with $\lambda = 0.25$) and affected medium to large τ values (between approximately 0.5 and 2.5). These results are consistent with the model recovery results: For large post-decision noise, data generated with any model starts to resemble a random choice strategy and therefore the recovered parameter estimate for τ reaches asymptote.

Figure 2.5b shows the parameter recovery results for data generated with

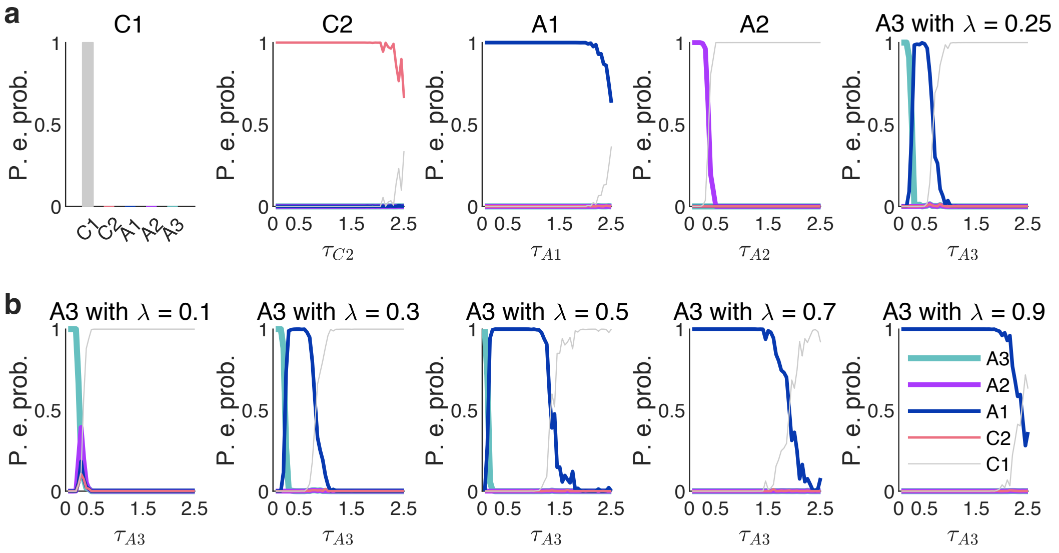

Figure 2.4. Model recovery results. **a** Model recovery results for all agents. Each subplot pertains to a data generating agent and shows the protected exceedance probabilities of the data analysis models evaluated on datasets of the data generating agent. For data generated with agent C1, the protected exceedance probability was maximal for C1. For data generated with agents C2, A1, A2 and A3, the protected exceedance probabilities depend on the value of the post-decision noise parameter τ used for data generation. Agents C2 and A1 are recoverable up to high levels of post-decision noise. Agents A2 and A3 (with $\lambda = 0.25$) are recoverable for low levels of post-decision noise. **b** Model recovery results for agent A3 with different λ values. For data generated with agent A3, the protected exceedance probabilities also depend on the value of the weighting parameter λ used for data generation. Agent A3 is recoverable up to medium λ values for low levels of post-decision noise.

A3 with different values of λ. For small values of τ, λ was reliably recovered across the parameter space, with a slight underestimation for $\lambda = 0.9$. Yet, for medium to large τ values the ML parameter estimate $\hat{\lambda}$ was biased. These findings are also in line with the model recovery results: First, the deflation effect for $\lambda = 0.9$ and small values of τ shows that for large values of the weight parameter A3 is indistinguishable from A1 and thus the estimate $\hat{\lambda}$ reaches asymptote. Second, the bias in $\hat{\lambda}$ for medium to large values of τ again shows that with increasing post-decision noise data generated with A3 and any weight starts to resemble a random choice strategy and thus λ can not be reliably identified.

Model comparison results Upon validating our modeling initiative we evaluated and compared the agent-based models in light of participants' data. For 18 of the 24 participants the BIC score was maximal for agent A3. Accordingly, the group cumulative BIC score was maximal for this agent showing that A3 explained participants' choices the best (Figure 2.6a left panel). Moreover, the group-level protected exceedance probability was larger than 0.99

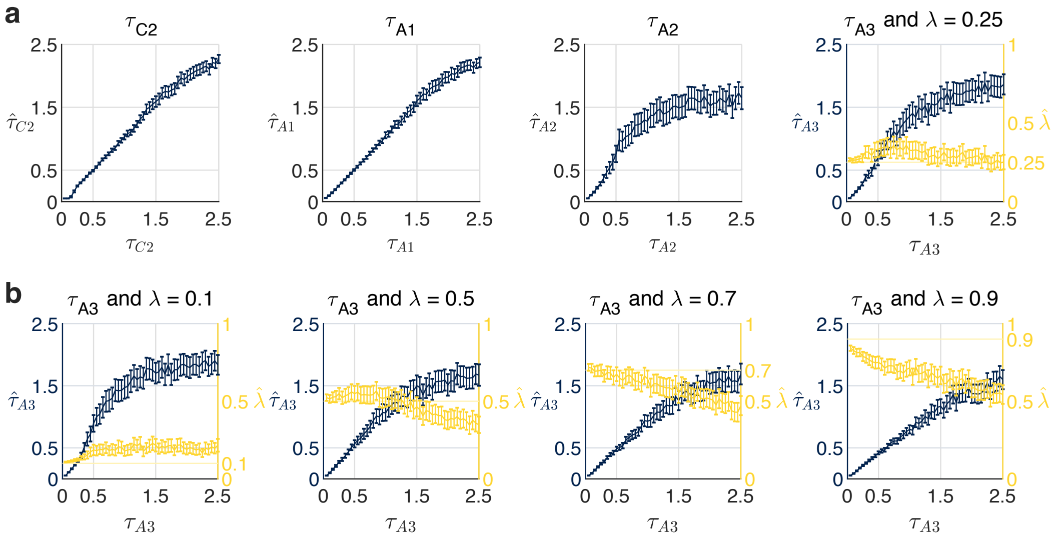

Figure 2.5. Parameter recovery results. **a** Parameter recovery results for all agents with free parameters. For every agent, the corresponding subplot shows the ML parameter estimates as functions of the post-decision noise parameter τ used for data generation. The post-decision noise parameter of agents C2 and A1 is recoverable from small to medium values. The post-decision noise parameter of agents A2 and A3 (with $\lambda = 0.25$) is recoverable for small values. The bias in $\hat{\tau}_{C2}$, $\hat{\tau}_{A1}$, $\hat{\tau}_{A2}$ and $\hat{\tau}_{A3}$ shows that for a sufficiently high level of post-decision noise, all agents emulate agent C1, and the estimates consequently reach asymptote. **b** Parameter recovery results for agent A3 with different λ values. The weight parameter λ of agent A3 is recoverable for small τ values, except for $\lambda = 0.9$. This shows that for large λ values agent A3 emulates agent A1 and, consequently, $\hat{\lambda}$ reaches asymptote. For medium to large values of the post-decision noise parameter, $\hat{\lambda}$ is a biased estimate. Error bars display the average SEM across repeats.

for A3 (Figure 2.6a right panel), which supports the conclusion that the most frequently applied strategy among the group of participants was the strategy implemented by A3. In addition, the pseudo-r^2 statistic returned a considerably high value for the quality of model fit: On average, A3 explained 56.73% $\pm$ 4.6 of the participants choices.

Winning model validation results To assess the behavioral validity of the winning A3 model, we generated synthetic data with it with each participant's ML parameter estimates $\hat{\theta}_{p,A3}$ and computed the same summary and trial-by-trial choice rates as for the participants. Consistent with the empirical results, most synthetic action choices were lucrative and informative with a rate of 84.98% $\pm$ 1.29, followed by significantly fewer lucrative and non-informative synthetic actions with 66.35% $\pm$ 1.96 (two-sided paired sample t-test, $t(23) = 11.59, p < 0.001$). Furthermore, as shown in Figure 2.6b the between reversals trial-by-trial dynamics of the synthetic actions exhibited a very similar pattern to that of the participants (cf. Figure 2.3b-c). Specifically, while both the average lucrative and informative and the average lucrative and non-informative action choice rates increased between two reversals (left panel), their difference decreased moderately between the first trials after and the last trials before a reversal (right panel). Altogether, these results lend a high face validity to the most plausible A3 model.

As a last step we conducted model and parameter recovery analyses of the winning A3 model and the participants' ML parameter estimates $\hat{\theta}_{p,A3}$ to evaluate the model's reliability in explaining the experimentally acquired data. Participants varied moderately with respect to both $\hat{\tau}_{p,A3}$ and $\hat{\lambda}_{p,A3}$. Specifically, $\hat{\tau}_{p,A3}$ ranged from 0.035 to 0.377 with an average of 0.124 ± 0.014 and $\hat{\lambda}_{p,A3}$ ranged from 0.044 to 0.622 with an average of 0.274 ± 0.027. As suggested by the results of the recovery analyses across the entire model and parameter space, for comparable τ and λ values the model of A3 and its free parameters are reliably recoverable. Indeed, based on data generated with A3 and $\hat{\theta}_{p,A3}$ both the model (Figure 2.6c) and parameter recovery (Figure 2.6d) analyses were successful. These results confirm the reliability of the best fitting A3 model.

2.4 Discussion

In this work, we addressed the question of how humans make sequential decisions if all actions bear economic consequences but only some deliver also information. By collecting participant choice data on an information-selective reversal bandit task, we demonstrated that in such situations, humans balance between exploratory and exploitative actions depending on their level of uncertainty. To arrive at this conclusion, we applied a comprehensive set of descriptive and agent-based computational modeling analyses (Russell & Norvig, 2010), including model evaluation based on *relative* (i.e., cumulative BIC score, protected exceedance probability and pseudo-r^2 statistic) as well as *absolute* (similarity between empirical and synthetic choice patterns) measures (Wilson & Collins, 2019). Formally, the behaviorally most plausible strategy was captured by a Bayesian agent that assessed the desirability of an action by applying a convex combination of the expected Bayesian surprise (Itti & Baldi, 2009) and the expected reward under its belief state. A series of recovery analyses validated our modeling initiative and established the robustness and reliability of our results.

We deem the key contributions of this work to be threefold: We introduced

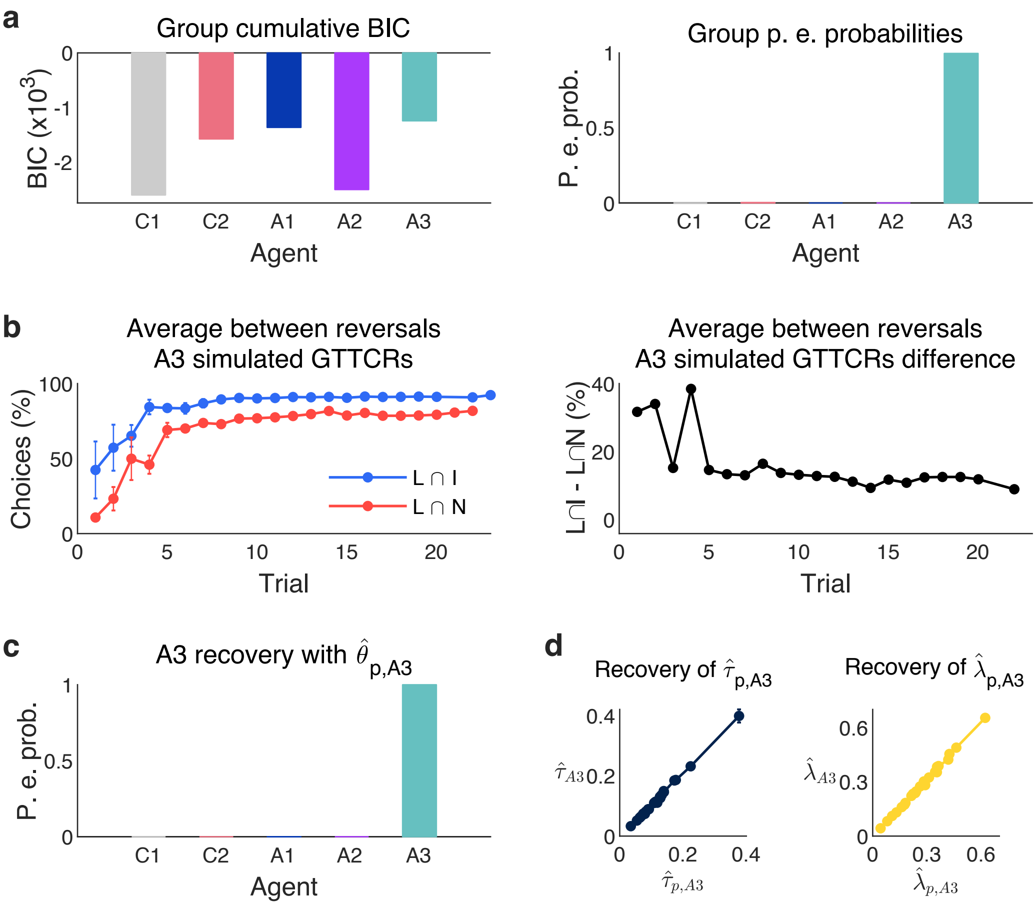

Figure 2.6. Computational modeling results. a Model comparison results. Both the group cumulative BIC scores (left panel) and the protected exceedance probabilities (right panel) were maximal for the agent A3 indicating that this model explained participants' choice data the best. **b** Model A3 validation results. Average between reversals group trial-by-trial lucrative and informative action and lucrative and non-informative action choice rates (left panel) and their difference (right panel) computed based on synthetic data sets generated with A3 and $\hat{\theta}_{p,A3}$. The patterns closely resemble those observed in the participants' data. **c** Model recovery result based on data generated with A3 and $\hat{\theta}_{p,A3}$. The protected exceedance probability was maximal for A3 indicating that the winning model was identifiable. **d** Parameter recovery results based on data generated with A3 and $\hat{\theta}_{p,A3}$. Both the post-decision noise parameter estimates $\hat{\tau}_{p,A3}$ (left panel) and the weighting parameter estimates $\hat{\lambda}_{p,A3}$ (right panel) were reliably recoverable.

a novel information-selective reversal bandit task, we proposed and thoroughly validated an agent-based modeling framework, and we provided evidence for uncertainty-guided exploration-exploitation. In the following, each of these contributions is discussed in turn.

As the first of our contributions, we introduced an information-selective reversal bandit task suitable to model a class of canonical sequential decision-making problems, in which information about the conferred reward is not available for every action. As already mentioned in Section 2.1, previous research primarily employed pure-exploration/sampling and exploration-exploitation/partial-feedback paradigms to study sequential decision making under uncertainty (Bubeck et al., 2009; Hertwig & Erev, 2009; Sutton & Barto, 2018; Wulff et al., 2018). The pure-exploration/sampling paradigm (Bubeck et al., 2009; Hertwig & Erev, 2009; Ostwald et al., 2015) models sequential decision-making problems in which an action either confers information or reward. In the classical variant of this paradigm, an action that confers reward automatically terminates the task. In an extended variant, referred to as the observe-or-bet task (Blanchard & Gershman, 2018; Navarro et al., 2016; Tversky & Edwards, 1966), the deciding agent can freely switch between the 'observe' action that confers information and the 'bet' actions that confer reward. Specifically, the observe action returns information about the expected reward value of the bet actions, but it does not return reward. The bet actions return rewards according to their associated reward distributions but no information. Similar to the bet actions, in our task one of the available actions confers only reward but no information. However, in our task, the other available action does not only confer information, as the observe action does, but it also confers reward. Therefore, while exploration and exploitation are temporally separated in the observe or bet task, they have to be balanced simultaneously in our task. In this regard, our task is similar to the exploration-exploitation/partial-feedback paradigm (Hertwig & Erev, 2009; Sutton & Barto, 2018).

The classical variant of the exploration-exploitation/partial-feedback paradigm is the multi-armed bandit task (Berry & Fristedt, 1985). In the multi-armed bandit task, the deciding agent chooses between a set of 'arms'. Akin to our task, each arm confers reward according to its associated reward distribution, and, in contrast to our task, each arm confers also information about its expected reward value. A drawback of this design is that the co-occurrence of reward and information evokes a confound between an action's value estimate and the associated uncertainty: As people tend to favor the action with the higher value estimate, for that action the associated uncertainty becomes smaller - simply because for that action more reward observations were made. This makes it difficult to uncover uncertainty-guided exploration in the standard multi-armed bandit task (Dezza, Angela, Cleeremans, & Alexander, 2017; Gershman, 2018; Wilson et al., 2014). Our task circumvents this problem by adopting a symmetrical reward structure of the actions: The probability of the positive reward for the lucrative action is identical with the probability of the negative reward for the detrimental action. Likewise, the probability of the negative reward for the lucrative action is identical with the probability of the positive reward for the detrimental action. This way, each reward observation following the informative action confers the same amount of information about the expected reward value of both the lucrative and detrimental action. Furthermore, as in each trial information is randomly coupled with either the lucrative or the detrimental action, our task arguably evokes a more marked exploration-exploitation dilemma than the multi-armed bandit task.

As the second of our contributions, we proposed and thoroughly validated an agent-based modeling initiative. This initiative consisted of belief state-based agents formalizing subjective uncertainty-based exploitative, explorative and hybrid explorative-exploitative strategies as well as belief state-free agents formalizing simple random choice and win-stay-lose-switch strategies. The belief state-based agents implement Bayesian update to infer the not directly observable structure of the task environment, i.e., which is the lucrative and which is the detrimental action. While optimal Bayesian learning may seem to be a strong assumption, it has been shown to approximate human learning reasonably well in comparable non-stationary tasks, such as the two-armed reversal bandit task (Hampton, Bossaerts, & O'Doherty, 2006) or non-stationary versions of the observe or bet task (Blanchard & Gershman, 2018; Navarro et al., 2016) and the multi-armed bandit task (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Speekenbrink & Konstantinidis, 2015). In addition, by including belief state-free agents in our model space, we also accounted for simple strategies that do not require Bayesian update. Of these, the win-stay-lose-switch agent adopts a well established effective strategy to approach similar bandit problems (Robbins, 1952).

The three belief state-based agents implement their respective strategies by following different optimization aims. In particular, the belief state-based exploitative agent seeks to maximize the belief state-weighted expected reward. The belief state-based explorative agent seeks to maximize the expected Bayesian surprise. The belief state-based hybrid explorative-exploitative agent seeks to maximize the convex combination of these two quantities. Belief state-weighted expected reward is a natural quantity to formally capture immediate reward gain and thus myopic exploitation (Sutton & Barto, 2018). Expected Bayesian surprise is one of many quantities that have been proposed to formally capture immediate information gain and thus myopic exploration (Schwartenbeck et al., 2019). As alluded to in Section 2.1, we here opted for Bayesian surprise due to its putative representation in the human neurocognitive system

(Gijsen, Grundei, Lange, Ostwald, & Blankenburg, 2020; Itti & Baldi, 2009; Ostwald et al., 2012).

Importantly, we would like to note that our definition of exploration pertains to a form of exploration that is generally referred to as 'directed exploration' (Gershman, 2018, 2019; Wilson et al., 2014). This term is used to distinguish information gain maximizing exploration from 'random exploration'. Random exploration is a form of exploration that achieves information gain by implementing a stochastic strategy, i.e. making stochastic choices based on the actions' reward value estimate. While there are more principled ways such as Thompson sampling (Thompson, 1933), random exploration is commonly accounted for by the softmax decision rule (Reverdy & Leonard, 2015). Notably, we here did not link the softmax decision rule to random exploration. Instead, we used it to embed the agents into a statistical framework in order to account for post-decision noise. This way, we clearly separated the deterministic strategies implemented by the agents and the probabilistic agent-based data analysis models. In future work, we aim to broaden our model space by considering agents that adopt random exploration. Crucially, we argue that this step will require a statistical framework that allows to reliably partition the variability of choice data into components relating to the agent's stochastic strategy and to post-decision noise (cf. Ostwald, Kirilina, Starke, and Blankenburg, 2014).

As the third and final of our contributions, we provided clear evidence for uncertainty-guided exploration-exploitation in our task: As uncertainty decreased, participants' choices were less influenced by the prospect of information gain and more influenced by the prospect of reward gain. Our finding is consistent with the behavior in the observe or bet task. In the first empirical study using the observe or bet task, Tversky and Edwards (1966) found that participants explored more, i.e., chose the observe action more frequently, if they (falsely) believed that the underlying environment was dynamic, i.e., the lucrative and detrimental bet actions reversed over time. While Tversky and Edwards (1966) did not relate this result to the notion that dynamic environments promote uncertainty, which, in turn, promotes exploration, in a recent study, Navarro et al. (2016) formally tested this hypothesis. By modeling participants' choices in both static and dynamic versions of the observe or bet task, they demonstrated that switches between the exploratory observe action and the exploitative bet actions were mediated by uncertainty.

Our result is also in line with recent findings from studies employing multi-armed bandit tasks. Work by several groups showed that when controlling for the value estimate-uncertainty confound, behavior in static two-armed bandit tasks reflects an uncertainty-dependent combination of exploration and exploitation (Dezza et al., 2017; Gershman, 2018, 2019; Wilson et al., 2014). Notably, however, consistent with the notion that the value estimate-uncertainty confound has the potential to mask directed exploration, findings from earlier studies not accounting for this confounding effect are less conclusive. For example, Zhang and Yu (2013) also found evidence for a belief state-based explorative-exploitative strategy in static four-armed bandit tasks. In contrast, Daw et al. (2006) did not find evidence for such a strategy in analyzing choices in a dynamic four-armed bandit task with changing action values. While the finding from Daw et al. (2006) is contrary to our finding, acknowledging that value estimate and uncertainty are not confounded in our task, these two findings can be reconciled.

In conclusion, in the current work we introduced a task that models a subset of real-life sequential decision-making problems that has been neglected previously: problems, in which information about the conferred reward is not available for every action. Importantly, this task allows to investigate a pronounced form of simultaneous exploration and exploitation processes without introducing the value estimate-uncertainty confound. We proposed an agent-based modeling framework formalizing various sequential decision-making strategies in our task and discussed how this framework may be extended in the future. Together, by analyzing participants' choices in our task using descriptive and agent model-based methods, we provide orthogonal evidence for an uncertainty-guided balance between exploration and exploitation in human sequential decision making under uncertainty.

2.5 Data and code availability

Data formatted according to the Brain Imaging Data Structure (Gorgolewski et al., 2016) and code implementing all analyses are hosted on Open Science Framework (Nosek et al., 2015) and are available at https://osf.io/vdmah/?view_only=a581d36aa8944464a7a81578b79afa8e.

2.6 References

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586.

Bartolo, R., & Averbeck, B. B. (2020). Prefrontal cortex predicts state switches during reversal learning. *Neuron.*

Berry, D. A., & Fristedt, B. (1985). Bandit problems: Sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, *5*, 71–87.

Bertsekas, D. P. (2000). *Dynamic programming and optimal control* (2nd edition). Athena Scientific.

Blanchard, T. C., & Gershman, S. J. (2018). Pure correlates of exploration and exploitation in the human brain. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(1), 117–126.

Bubeck, S., Munos, R., & Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, 23–37.

Byrd, R. H., Gilbert, J. C., & Nocedal, J. (2000). A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, *89*(1), 149–185.

Byrd, R. H., Hribar, M. E., & Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, *9*(4), 877–900.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). 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.

Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A., & Averbeck, B. B. (2016). Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron*, *92*(2), 505–517.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Dezza, I. C., Angela, J. Y., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific Reports*, *7*(1), 1–13.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277.

Gijsen, S., Grundei, M., Lange, R. T., Ostwald, D., & Blankenburg, F. (2020). Neural surprise in somatosensory bayesian learning. *bioRxiv*.

Gläscher, J., Hampton, A. N., & O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, *19*(2), 483–495.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 1–9.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360–8367.

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, *84*, 159–168.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

McFadden, D. (1973). *Conditional Logit Analysis of Qualitative Choice Behavior*. Institute of Urban and Regional Development, University of California.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

Nooner, K. B., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., Panek, L., Brown, S., Zavitz, S., Li, Q., et al. (2012). The NKI-

Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, *6*, 152.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., et al. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Ostwald, D., Kirilina, E., Starke, L., & Blankenburg, F. (2014). A tutorial on variational bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, *60*, 1–19.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177–188.

Ostwald, D., Starke, L., & Hertwig, R. (2015). A normative inference approach for optimal sample sizes in decisions from experience. *Frontiers in Psychology*, *6*, 1342.

Reverdy, P., & Leonard, N. E. (2015). Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, *13*(1), 54–67.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage*, *84*, 971–985.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *8*, e41703.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*(2), 351–367.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011). Planning to be surprised: Optimal bayesian exploration in dynamic environments. In J. Schmidhuber, K. R. Thórisson, & M. Looks (Eds.), *Artificial general intelligence* (pp. 41–51). Springer Berlin Heidelberg.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680.

Waltz, R. A., Morales, J. L., Nocedal, J., & Orban, D. (2006). An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical Programming*, *107*(3), 391–408.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

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), 915–924.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin*, *144*(2), 140.

Zhang, S., & Yu, A. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in Neural Information Processing Systems*, *26*.

3 | The neurocomputational mechanisms of sequential decision making in a multistep task with partially observable states

3.1 Introduction

In a class of sequential decision-making tasks, actions of the deciding agent affect next states of the environment and thereby future rewards (e.g., Bertsekas (2000), Puterman (2014), and Sutton and Barto (2018)). A central question in decision neuroscience research is whether humans and other biological deciding agents plan ahead or resort to model-free decision-making strategies in such tasks, usually discussed under the name multistep tasks. Both class of strategies assume that the agent tries to maximize its cumulative reward. However, to choose between the actions in a given state in light of this goal, planning strategies mentally simulate the consequences of actions, whereas model-free decision-making strategies simply rely on some instantaneously available reward-related information or the reward history (e.g., Collins and Cockburn (2020), Daw, Niv, and Dayan (2005), Dayan (2012, 2014), Dayan and Daw (2008), Dickinson and Balleine (2002), Dolan and Dayan (2013), Kaplan, Schuck, and Doeller (2017), Solway and Botvinick (2012), and Tolman (1948)).

While many studies have demonstrated that humans engage in planning, these studies examined sequential decision making in multistep tasks where the environmental states are fully observable (Cushman & Morris, 2015; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Doll, Duncan, Simon, Shohamy, & Daw, 2015; Gläscher, Daw, Dayan, & O'Doherty, 2010; Huys et al., 2012; Korn & Bach, 2018; Momennejad et al., 2017; Simon & Daw, 2011; Wunderlich, Smittenaar, & Dolan, 2012). Yet, in many real-life multistep tasks certain components of the states remain latent and the available observations only convey ambiguous information about these. To illustrate, consider someone trying to find their lost ring at the beach. The ring is buried deep in the sand but a metal detector can help to locate it by signalling at each step the direction towards it with some associated noise. Similar to this example, humans have to navigate various abstract representational spaces, for example, career paths, financial investment schemes, or maintaining a healthy lifestyle. Across all these domains, states are often only partially observable and therefore the decisions are imbued with state uncertainty (Bach & Dolan, 2012; Dayan & Daw, 2008; Ma & Jazayeri, 2014).

The Bayesian brain hypothesis generally posits that under uncertainty, the brain uses Bayes rule to probabilistically infer the hidden cause underlying sensory data. Abundant evidence supports the Bayesian brain hypothesis across various cognitive domains, such as perception (Harrison, Stephan, Rees, & Friston, 2007; Ostwald et al., 2012) and reasoning (Konovalov & Krajbich, 2018; Schwartenbeck, FitzGerald, & Dolan, 2016; Tenenbaum, Griffiths, & Kemp, 2006). Applied to sequential decision making under state uncertainty, the brain is postulated to maintain a Bayes optimal (or near-optimal) belief state representing its subjective uncertainty about the state given the history of observations and actions, on the basis of which actions can be evaluated (Ma, 2019; Rao, 2010). Crucially, the combination of this notion with the aforementioned findings on sequential decision making in multistep tasks suggests that if the states comprise latent components, humans plan ahead on the basis of belief states. However, as human sequential decision making in multistep tasks with partially observable states has so far received little attention in decision neuroscience research, evidence supporting this hypothesis is elusive.

State uncertainty introduces an additional critical computational dimension to planning (e.g., Dayan and Daw, 2008): A belief state-based planning agent may try to maximize its cumulative reward by exploiting its accumulated knowledge about the state. Alternatively, a belief state-based planning agent may explore first. That is, it may try to resolve its state uncertainty by choosing the action with the highest associated information gain to enable that the subsequent exploitative choices yield the maximum attainable cumulative reward. If and how biological agents alternate between exploration and exploitation has traditionally been studied using bandit tasks. Bandit tasks are sequential decision-making tasks with latent reward structures, where actions of an agent affect immediate rewards but do not have deferred consequences (Berry & Fristedt, 1985; Robbins, 1952; Sutton & Barto, 2018). Many studies, including ours as outlined in Chapter 2, have reported that humans combine the objectives of exploration and exploitation when performing bandit tasks (Gershman, 2018; Wilson, Geana, White, Ludvig, & Cohen, 2014; Zhang &

Yu, 2013), while other studies have only found support for exploitation (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Payzan-LeNestour & Bossaerts, 2011). Given the lack of studies investigating sequential decision making in multistep tasks with partially observable states, it is unclear if in these tasks humans adopt an exploitative or a hybrid explorative-exploitative objective.

In this work, we sought to computationally characterize the behavioral and neural correlates of human sequential decision making in multistep tasks with partially observable states. Specifically, we addressed the following questions: (1) Do humans engage in belief state-based planning or do they resort to decision-making strategies that eschew prospective computations? (2) If humans engage in belief state-based planning, do they adopt the objective of exploitation or do they combine exploitation with exploration? (3) Which brain regions implement the putative cognitive processes underlying behavior? To answer these questions, we designed a multistep spatial search task similar to the ring search example introduced above and analyzed 19 participants' behavioral and functional magnetic resonance imaging (fMRI) data. Specifically, we first evaluated a comprehensive set of agent-based computational models in light of participants' behavioral data (Russell & Norvig, 2010). This set included belief state-free agents that implement simple model-free decision-making strategies as well as belief state-based exploitative, explorative and hybrid explorative-exploitative agents that implement various belief state-based planning strategies. We subsequently assessed the neural correlates of the cognitive processes captured by the behaviourally most plausible agent using a model-based general linear model (GLM) approach (Friston & Dolan, 2010).

Our behavioral analyses revealed that participants' actions were best accounted for by a belief state-based agent adopting a purely exploitative objective. In terms of the neural representation of the constituent cognitive processes of this strategy, we hypothesized that a distributed network of cortical and subsortical areas would be involved. In particular, capitalizing on findings from recent neuroimaging studies (Fischer, Bourgeois-Gironde, & Ullsperger, 2017; O'Reilly, Jbabdi, Rushworth, & Behrens, 2013), we expected to observe activity in the frontal and posterior parietal cortices as well as the dorsal striatum related to belief state maintenance as indexed by the Bayesian surprise, a quantity commonly used in the Bayesian brain hypothesis framework (Itti & Baldi, 2009; Ostwald et al., 2012). We further expected that exploitative planning on the basis of belief states would engage parts of the ventromedial prefrontal and orbitofrontal cortices. These are key regions of the brain's reward-guided decision-making system (Rushworth, Noonan, Boorman, Walton, & Behrens, 2011) and have been implicated both in planning (in multistep tasks with fully observable states; Korn and Bach, 2018; Simon and Daw, 2011) and in exploitation (in bandit tasks; Chakroun, Mathar, Wiehler, Ganzer, and Peters, 2020; Daw et al., 2006). Our neural results largely corroborate these hypotheses. Consequently, in conjunction with our behavioral results, they provide evidence that in our multistep task with partially observable states, participants performed belief state-based exploitative planning.

3.2 General methods

3.2.1 Experimental methods

Participants and procedure The experimental data set was acquired at the Max Planck Institute for Human Development (Berlin, Germany). Behavioral and fMRI data was recorded from a group of 19 participants (10 female, 18 right-handed, 1 male participant both-handed with left hand preference, mean age: 27.11 years, standard deviation age: 2.92 years) with no reports of neurological or psychiatric disorders after providing written informed consent. The study was in line with the human subject guidelines of the Declaration of Helsinki and was approved by the ethics committee of the German Psychological Society (Deutsche Gesellschaft für Psychologie). Participants completed four consecutive runs of the experimental task in the Magnetic Resonance Imaging (MRI) scanner, after having received task instructions (Supplementary Material B.1) and having completed a training run on a desktop computer. Due to technical problems and the thus resulting time constraints, two participants performed only three complete runs in the scanner. Participants were reimbursed for their time with 10 € per hour and an additional 0.625 € for each task they solved.

Experimental design We framed a multistep task with partially observable states as a spatial search task. In the 'treasure hunt' task, participants were instructed to find two hidden 'treasures' in a 5×5 cell grid-world. Figure 3.1a shows the grid-world from above with an example task configuration and with a fictive participant's attempt decision sequence. A task corresponded to a specific location combination of the two treasures. Participants were given a maximum of three attempts to solve a task and secure a monetary reward.

Importantly, a task was considered to be solved only if both treasures were collected within a single attempt. In each attempt, participants had a limited number of steps at their disposal. If, in an attempt, participants failed to visit both treasure locations before the number of available steps was exhausted, their position was reset to the start position, which was the upper left cell of the grid in each attempt of each task. Note that participants were not presented with the bird's eye perspective of the grid-world when interacting with a task (Figure 3.1a) but on each trial were only given the information available from their current position (Figure 3.1b; please see below for details).

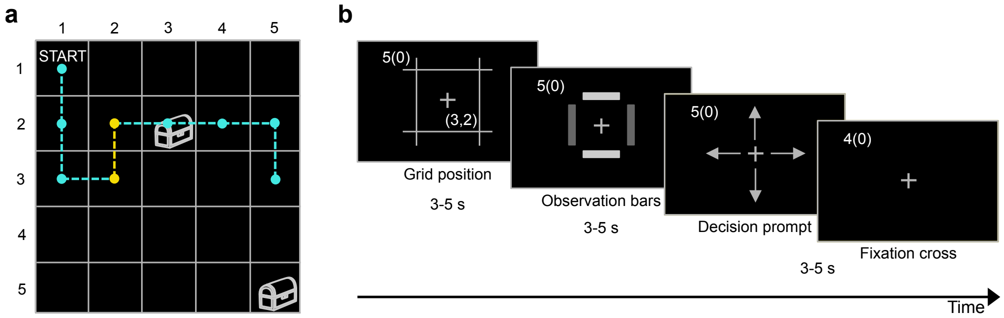

Figure 3.1. Treasure hunt task. a Bird's eye perspective of the 5×5 cell grid-world. In the example task shown, the treasures were hidden at cells (2,3) corresponding to the second row third column and (5,5) corresponding to the fifth row fifth column. The dashed line with full dots at the cells shows a fictive participant's decision sequence within an attempt of the task. Here, the participant discovered only the treasure at location $(2, 3)$ within the available number of steps. Therefore, to solve the task in a next attempt, the participant had to revisit this cell and collect the treasure at location (5,5). **b** Trial design. This example shows the stimuli sequence presented to the participant moving from cell (3,2) to (2,2) as shown by the yellow section of the dashed line in **a**. Before deciding on a northward movement, the participant received a signal from the treasure detector in the form of light and dark grey observation bars towards the neighbouring cells. The values in the upper left corner denote the number of remaining steps and the number of treasures collected (the value in parentheses) in the attempt.

For each task, the treasure locations were consecutively sampled online from a uniform distribution over the cells of the grid with two restrictions. First, no treasure was assigned to the cell in the upper left corner corresponding to the start position. Second, the treasures were assigned to different cells. Participants completed four tasks in each run, which yielded a total of 296 tasks presented to the participants. As shown in Figure 3.2a, the treasure location combinations were evenly represented across tasks and a certain combination occurred no more than five times.

In each attempt, the step limit was randomly determined in order to prevent participants from using this information to infer the treasure locations. Specifically, we first evaluated the number of optimal steps required to solve a task based on the l_1 distance between the start position and the two treasure locations. This subsequently served as the expectation of a discrete categorical distribution over ± 2 steps (with probabilities of $\pm 0 = 0.4$, $\pm 1 = 0.2$, $\pm 2 = 0.1$). As a consequence, the step limit sometimes sufficed and sometimes did not suffice to collect the two treasures within an attempt. Given the 5×5 grid-world layout, the maximal optimal step limit was 12 and accordingly, there were a maximum of 14 steps - and thus 14 trials - in an attempt. Figure 3.2b shows the number of tasks as a function of the optimal step limit. The majority of the tasks were solvable within 5-8 steps. Figure 3.2c shows the number of attempts as a function of the step limit deviation from optimum. In most attempts, the step limit corresponded to the optimal step limit.

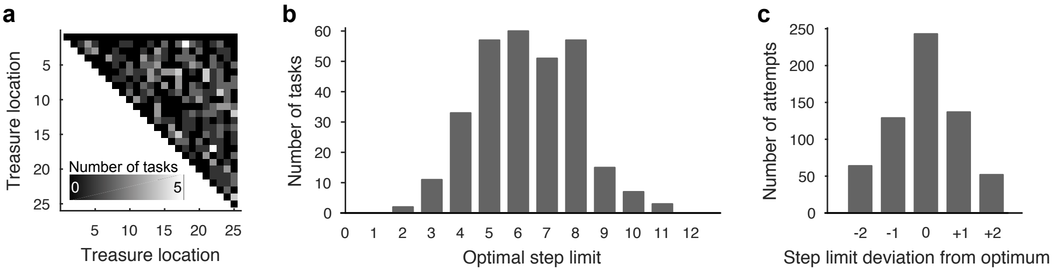

Figure 3.2. Task quality assurance. a Number of tasks as a function of treasure location combinations. In total, the group of participants completed 296 tasks. Of these, there were a maximum of 5 tasks with identical treasure location combinations. Note: The treasure locations are indicated with their corresponding node indices, for details please refer to Section 3.2.2. **b** Number of tasks as a function of optimal step limit. As given by the layout of the grid, most treasure location combination constituted problems that were solvable within 5-8 steps. **c** Number of attempts as a function of the step limit deviation from optimum. In most attempts, the step limit corresponded to the optimal step limit. Together, the results visualized in **a**, **b** and **c** validate the online sampling procedure used in the treasure hunt task by showing that the realized outcomes were consistent with the underlying distributions.

Every attempt ended with the presentation of the final grid position resulting from the decision on the last trial of the attempt. If participants, in a given attempt, solved a task, they were subsequently informed that both treasures were found during a single attempt and a new task was created. If participants did not solve the task in a given attempt, but one or two more task attempts remained, participants were informed that their step limit was reached and their position was being reset. Finally, if participants did not solve the task upon completing the third attempt, they were informed that both their step and attempt limits were reached and a new task was created. In each case, the visual information about the attempt and task outcomes was presented for a duration uniformly sampled online from an interval of 3 to 5 seconds. In addition, participants were given a short resting period after every fourth attempt. During these resting periods, participants were instructed to look at a fixation cross for a duration uniformly sampled online from an interval of 6 to 8 seconds.

Trial design Figure 3.1b depicts an exemplary trial as presented to a fictive participant. On every trial of the task, participants were first shown the cell of the grid they currently occupied, including its row and column indices. If participants entered a treasure location, they were additionally shown a treasure symbol. Next, a collection of light and dark grey 'observation bars' were presented towards the adjoining cells. These bars conveyed information about the treasure locations with certain accuracy and thus could be interpreted as noisy signals of a 'treasure detector'. Specifically, the detector always returned a dark grey bar for the direction leading away from the treasures while it displayed either a light or a dark grey bar in the direction of the treasures. The detector's accuracy of correctly returning a light grey bar towards the treasure locations depended on the participants' distance to the treasures: Its readout was completely unreliable at the most distant cell position and parametrically increased in accuracy as participants moved towards the treasures. It always signalled the true values in the immediate vicinity of the treasures. Following the presentation of the observation bars, participants were prompted to choose one of the available directions by means of pressing one of four buttons with the index to little finger of their right-hand. They could move to any of the neighbouring cells. Diagonal steps or steps off the grid were not allowed. The duration of the grid position, the observation bars and the response time window were uniformly sampled online from an interval of 3-5 seconds. If participants indicated their decision within the response time window, they were presented with a post-decision fixation cross until the response time window elapsed. Upon the post-decision fixation cross, the next trial commenced with the presentation of the new grid position. If, on the other hand, participants did not indicate their decision within the response time window, they maintained their current position on the next trial and lost one step. To help participants keep track of the attempt's evolution, throughout the trial they were visually informed about the number of remaining steps and the number of treasures collected.

3.2.2 Task model

To make the treasure hunt task amenable to agent-based computational modeling, we first represented the square grid-world of dimensionality $d = 5$ by a graph (N, E). Here, $N := \mathbb{N}_{d^2}$ denotes the set of nodes corresponding to the cells of the grid and E denotes a set of (unweighted) edges corresponding to the directions available from each cell (Figures 3.3a-b). Building on this graph representation, we specified central task components using concepts from the theory of partially observable Markov decision processes (Bertsekas, 2000). Specifically, the treasure hunt task on a given attempt can be conceived as a tuple (S, O, A, f, g), where for available trials $t = 1, 2, \ldots, T$

- $S := \mathbb{N}_{d^2} \times \mathbb{N}_2^0 \times \mathbb{N}_{d^2} \setminus \{1\} \times \mathbb{N}_{d^2} \setminus \{1, s^3\}$ denotes the set of states $s := (s^1, s^2, s^3, s^4)$. The first two state components comprise the directly observable part of the state, $y := (s^1, s^2)$. The first variable encodes the agent's current grid position. The second variable encodes the treasure discovery history and takes on the value 0 if no treasure, 1 if the treasure at the first location and 2 if the treasure at the second location was collected up to trial t. The second two components encode the location of the two treasures: s^3 encodes the location of the first treasure, i.e. the treasure with the the smaller node index and s^4 encodes the location of the second treasure, i.e. the treasure with the the larger node index. These state components comprise the latent part of the state, $x := (s^3, s^4)$. Note that we use the terms first and second treasure only to distinguish between them.

- $O := \{0, 1\}^4$ denotes the observation set. The components of an observation $o := (o^1, o^2, o^3, o^4) \in O$ encode the observation bars presented to the agent in the northern, eastern, southern and western directions, respectively. The value 0 represents a dark grey bar and 1 represents a light grey bar. The available observations depend on the current position of the agent. For example, for $s^1 := 1$, the available observation set $O_{s^1} \subseteq O$ is given by $O_1 = \{0, 1\}^2$ and elements of O_1 are of the form $o = (o^2, o^3)$.

- $A := \{-5, +1, +5, -1\}$ denotes the action set. The values encode the northward, eastward, southward and westward movements on the nodes of the graph representation of the grid, respectively. Like the set of available observations, the set of available actions $A_{s^1} \subseteq A$ depends also on the first state component s^1. For example, for $s^1 := 21$, $A_{21} = \{-5, +1\} \subseteq A$.

- f denotes the state transition function, which specifies the probability that the state takes on the value s_* at trial $t+1$ given state s and agent action a at trial t. In the treasure hunt task, actions have deterministic outcomes, i.e., deciding for an available action results in the indicated position change with certainty. Formally, assuming $a \in A_{s^1}$,

$$f\left(s, a, s_*\right) := p\left(s_{t+1} = s_* | s_t = s, a_t = a\right) := \begin{cases} 1, \text{ if } s_* = \left(y_*, x_*\right) \\ 0, \text{ else} \end{cases}, \tag{3.1}$$

where

$$y_* = \left(s^1 + a, 1 \cdot \lfloor s_*^1 = s^3 \rfloor + 2 \cdot \lfloor s_*^1 = s^4 \rfloor\right) \text{ and } x_* = \left(s^3, s^4\right) \tag{3.2}$$

with the initial state corresponding to

$$p\left(s_1 = \left(1, 0, s^3, s^4\right)\right) = 1 \tag{3.3}$$

in each attempt. Upon every action, the state changes according to Eq. 3.1 until there are steps left or until both treasures are collected within the attempt.

- Finally, g denotes the observation function, which specifies the probability of observation o given state s at trial t. This probability implements the distance-dependent accuracy of the observations and thus takes a central stance in the specification of the treasure hunt task. Omitting the restrictions to available observations for ease of exposition, this probability takes the form

$$g\left(s, o\right) := p\left(o_t = o | s_t = s\right) := \prod_{i=1}^{4} p\left(o^i | s\right) := \prod_{i=1}^{4} Bern\left(o^i; \pi_i(s)\right). \tag{3.4}$$

That is, the individual observations o^i, $i \in \mathbb{N}_4$ are distributed independently according to Bernoulli distributions with state-dependent parameters $\pi_i\left(s\right) \in \left[0, 1\right]$, $i \in \mathbb{N}_4$. These parameters are defined as follows. First, the probability of observing a light grey bar in the direction of i is expressed as an affine function of the l_1 distance between the agent and a treasure. The induced probability parameter set corresponds to

$$p_{j,k} := \beta_0 + \beta_1 l_1\left(j, k\right), \tag{3.5}$$

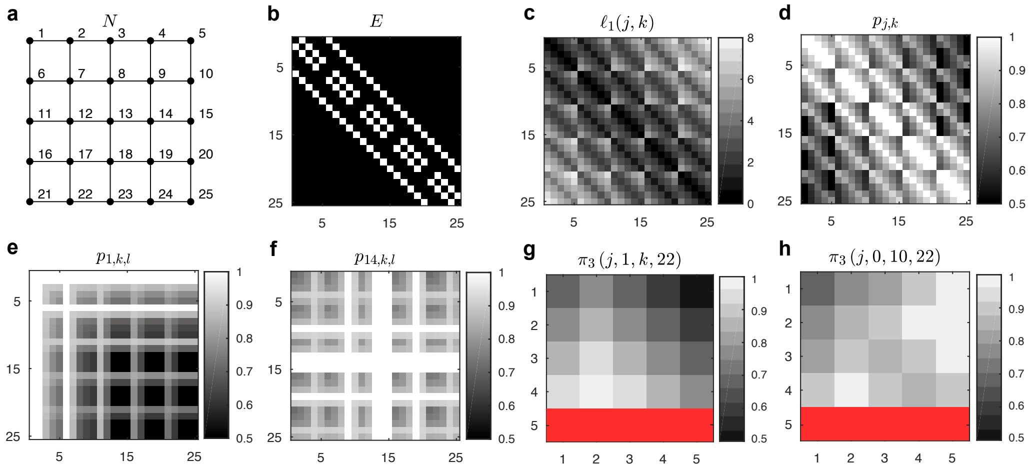

Figure 3.3. Task model. a and **b** Graph representation of the grid-world. **a** Nodes (circles) and edges (lines) of the graph representation of the 5×5 grid-world, numbers label the nodes. **b** Adjacency matrix representation of the set of edges visualized in **a**. **c** l_1 distance set between agent position j and treasure location k. **d** Observation probability parameter set as a function of the agent position and treasure location for single treasures. The probability was set to $p_{j,k} = 0.5$ at the maximum l_1 distance (yielding a maximally inaccurate signal) and to $p_{j,k} = 1$ at the minimum l_1 distance (yielding a maximally accurate signal). **e** and **f** Combined probability parameter set as function of two treasure locations for agent positions $j = 1$ and $j = 14$, respectively. **g** and **h** Bernoulli observation parameter $\pi_3(s)$ for southern bar observations for a single remaining treasure located at node 22 $(5, 2)$ and for two treasures located at nodes 10 $(2, 5)$ and 22 $(5, 2)$, respectively, as a function of the agent position. Note: The bottom rows are marked with red because at the corresponding nodes southern bars are not available.

where $l_1(j, k)$ denotes the l_1 distance between the agent's position node $j \in \mathbb{N}_{d^2}$ and the treasure location node $l \in \mathbb{N}_{d^2}$. The complete l_1 distance set is visualized in Figure 3.3c. β_0 and β_1 are real-valued parameters, which are determined according to the following constraints: First, the probability of observing a light grey bar is set to certainty ($p_{j,k} = 1$) if the agent is a single step away from the treasure, and second, it is set to full uncertainty ($p_{j,k} = 0.5$) if the agent is at its maximum possible distance from the treasure, i.e., $l_1(j, k) = \max_{j \in \mathbb{N}_{d^2}} l_1(j, k)$. These constraints are satisfied for

$$\beta_0 := 1 + \frac{0.5}{\max\limits_{j \in \mathbb{N}_{d^2}} l_1(j,k) - 1} \text{ and } \beta_1 := -\frac{0.5}{\max\limits_{j \in \mathbb{N}_{d^2}} l_1(j,k) - 1}. \tag{3.6}$$

The complete ensuing probability parameter set $P = \{p_{j,k}\}_{1 \le j,k \le d^2}$ for all combinations of agent and treasure nodes is visualized in Figure 3.3d. If $s^2 = 1$ or $s^2 = 2$, the parameter set visualized in Figure 3.3d forms the basis of $\pi_i(s)$. However, if $s^2 = 0$, i.e., none of the treasures have yet been collected, the parameter sets are combined into a single parameter set according to

$$p_{j,k,l} = \begin{cases} p_{j,l} + (p_{j,k} - p_{j,l})\, p_{j,k}, & \text{if } p_{j,k} \geq p_{j,l} \\ p_{j,k} + (p_{j,l} - p_{j,k})\, p_{j,l}, & \text{if } p_{j,k} < p_{j,l} \end{cases}, \tag{3.7}$$

where j encodes the agent position, and k and l encode the treasure locations ($j, k, l \in \mathbb{N}_{d^2}$). Figures 3.3e and f show the ensuing parameter set P for the agent positions $j = 1$ and $j = 14$, respectively. Based on the above, the Bernoulli distribution parameters $\pi_i(s)$ are then defined as

$$\pi_i(s) := \begin{cases} p_{j,k,l}, & \text{if } s = (j, 0, k, l) \\ p_{j,l}, & \text{if } s = (j, 1, k, l) \\ p_{j,k}, & \text{if } s = (j, 2, k, l) \end{cases} \tag{3.8}$$

if moving in the direction of i brings the agent closer to at least one of the treasures (in the case of $s^2 = 0$) or to the remaining treasure (in the case of $s^2 \in \{1, 2\}$),

$$l_1\left(j + a_i, k\right) < l_1\left(j, k\right) \text{ and or } l_1\left(j + a_i, l\right) < l_1\left(j, l\right). \tag{3.9}$$

In any other case,

$$\pi_i(s) = 0. \tag{3.10}$$

Figures 3.3g and h visualize the ensuing parameters $\pi_3(s)$ as a function of s^1, for $s^2 = 1$, an arbitrary s^3 and the remaining treasure location $s^4 = 22$, and for $s^2 = 0$, $s^3 = 10$ and $s^4 = 22$, respectively.

3.2.3 Agent models

We designed a set of nine agent-based computational models (Russell & Norvig, 2010) to formally capture different sequential decision-making strategies in the treasure hunt task. The main differences between the agents can be described along two dimensions. First, the agents differ in whether or not they make decisions on the basis of belief states. In general, as introduced above (see Section 3.1), the term belief state refers to a probabilistic representation of the state. For simplicity, we here consider the belief state with respect to the latent state components only. Thus, the belief state corresponds to a probabilistic map, which for each cell of the grid encompasses the agent's subjective uncertainty that a treasure is hidden there. Consequently, the belief state-free agents do not assume such a probabilistic map. Instead, to make decisions, they rely on information immediately available to them, such as the grid position or the observation bars. In contrast, the belief state-based agents assume a probabilistic map, which they update in a normative Bayesian fashion upon moving to a new grid position and being presented with a collection of observation bars. These agents plan ahead by relying on their evolving belief state. Second, the agents have different objectives: they are either purely exploitative, purely explorative or hybrid explorative-exploitative. The purely exploitative agents try to solve the task, i.e., collect the treasures, by harnessing their knowledge about the environment. This can be done both in a belief state-free and a belief state-based way. In contrast, the purely explorative agent tries to improve its knowledge about the environment, i.e., know where the treasures are hidden. This requires a belief state that stores the accumulating knowledge about the treasure locations. The hybrid explorative-exploitative agents combine these two objectives and try to acquire an accurate knowledge about the environment based on which the task can be solved, i.e., know where the treasures are hidden and then collect them. As the hybrid agents assume exploration, they require a belief state. In addition, we designed a cognitive null agent, which does not have any objective and thus makes decisions at random. This agent does not require a belief state. For an overview of the agent model space, please see the schematic in Figure 3.4.

In what follows, we introduce each agent in detail: First, we describe the scheme the agent uses to evaluate the actions' valence, i.e., the actions' desirability in light of the agent's objective. We then present a worked example of this scheme on a trial. To initialize the trial used in the worked example, consider the following situation: The agent is in the first attempt of a task and the treasures are hidden at nodes 10 and 16. The agent has not collected any of the treasures up to trial t. On trial t, the agent's grid position is $s_t^1 = 11$, where it makes the observation $o_t = (o^1 = 1, o^2 = 0, o^3 = 0)$. The available action set is $\{-5, +1, +5\}$ and the reachable nodes are $6, 12$ and 16. In addition, let us assume that the belief state-based agents have the non-zero trial-posterior belief state components $\mu_{t+1}^{(6)} = 0.25$, $\mu_{t+1}^{(13)} = 0.15$, $\mu_{t+1}^{(14)} = 0.2$, $\mu_{t+1}^{(15)} = 0.1$ and $\mu_{t+1}^{(21)} = 0.3$ (the formal notion of belief state is introduced below in Paragraph Belief state-based agents). We conclude by describing a simulated action sequence of the agent in the treasure hunt task. In the simulations, the treasures are also hidden at nodes 10 and 16. The agent is given the standard

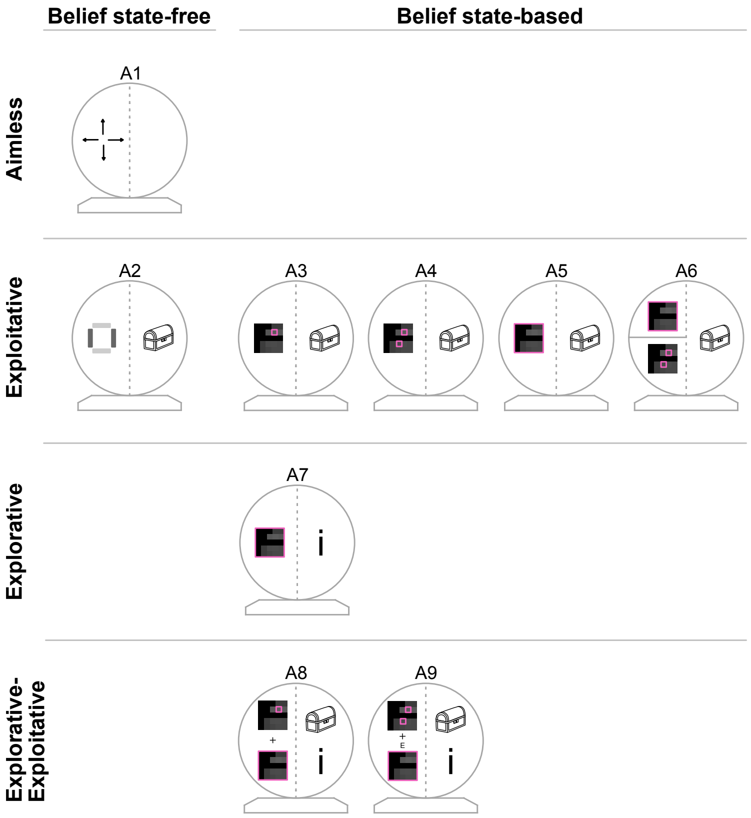

Figure 3.4. Agent models. The belief state-free agents A1 and A2 make decisions using merely the information available to them on the current trial. Specifically, the belief state-free cognitive null agent A1 implements a random choice strategy. The belief state-free exploitative agent A2 relies on the task instructions about the treasure detector and follows the light grey bar. The belief state-based agents A3, A4, A5, A6, A7, A8 and A9 make decisions using a dynamically evolving probabilistic representation of the treasure locations. Specifically, the belief state-based exploitative agent A3 moves towards the most probable treasure location. The belief state-based exploitative agent A4 identifies the two most probable treasure locations and moves towards the closer one. The belief state-based exploitative agent A5 takes its entire belief state into account and moves towards that part of the grid where it expects to find a treasure within the fewest steps. The belief state-based exploitative agent A6 behaves like A5, but, if it encounters an ambiguous decision situation (i.e., where there is more than one best action), it re-evaluates the actions using the strategy of A4. The belief state-based explorative agent A7 selects actions as to resolve its uncertainty about the treasure locations. The belief state-based hybrid explorative-exploitative agent A8 uses a constant combination of the strategies of A3 and A7. The belief state-based hybrid explorative-exploitative agent A9 uses a state uncertainty-dependent combination of the strategies of A4 and A7.

configurations, meaning that it has a maximum of three attempts to solve the task and in every attempt it has nine steps, which is the optimal step limit in the task. It is important to point out that while we conceive the belief state-based agents deterministic by assuming valence maximizing action selections, in these simulations the actions for all agents are sampled probabilistically. Specifically, for each agent, the actions are generated with the corresponding behavioral data analysis model detailed below in Paragraph Behavioral data analysis models. To ensure that the action with the highest valence is sampled with a high probability and all other actions are sampled with low probabilities, for the behavioral data analysis models that include a softmax operation (Reverdy & Leonard, 2015), the inverse temperature parameter is set to an arbitrary large value ($\beta = 100$).

Belief state-free agents The belief state-free agents denoted by A1 and A2 allocate action valences based on instantaneously available information. They do not maintain a probabilistic map encoding the subjective uncertainty about the treasure locations and are therefore belief state-free. To realize an action, both agents use a probabilistic decision rule by directly translating action valences into choice probabilities.

The belief state-free cognitive null agent A1 The belief state-free cognitive null agent A1 does not have any objective and therefore deems all actions equally desirable. Formally, on each trial A1 identifies the actions available at its current grid position and allocates equal valences to these according to

$$v_{\mathrm{A1}}(a) := \frac{1}{|A_{s^1}|}. \tag{3.11}$$

Worked trial example. On the example trial, A1 allocates the following action valences:

$$\begin{aligned} v_{\mathrm{A1}}(-5) &= \frac{1}{3} \\ v_{\mathrm{A1}}(+1) &= \frac{1}{3} \\ v_{\mathrm{A1}}(+5) &= \frac{1}{3}. \end{aligned}$$

Thus, all available actions are equally desirable for the agent.

Simulated task behavior. Figure 3.5 shows an exemplary behavior of A1. In this simulation, the agent does not solve the task: It collects one of the treasures in the first attempt but does not revisit this location in the next attempts nor collects the other treasure.

The belief state-free exploitative agent A2 The belief state-free exploitative agent A2 tries to collect the treasures by leveraging the task instructions about the treasure detector (cf. Supplementary Material B.1). Specifically, as A2 is informed that a light grey bar necessarily indicates a direction towards the treasures, on each trial the agent allocates high valence to the corresponding

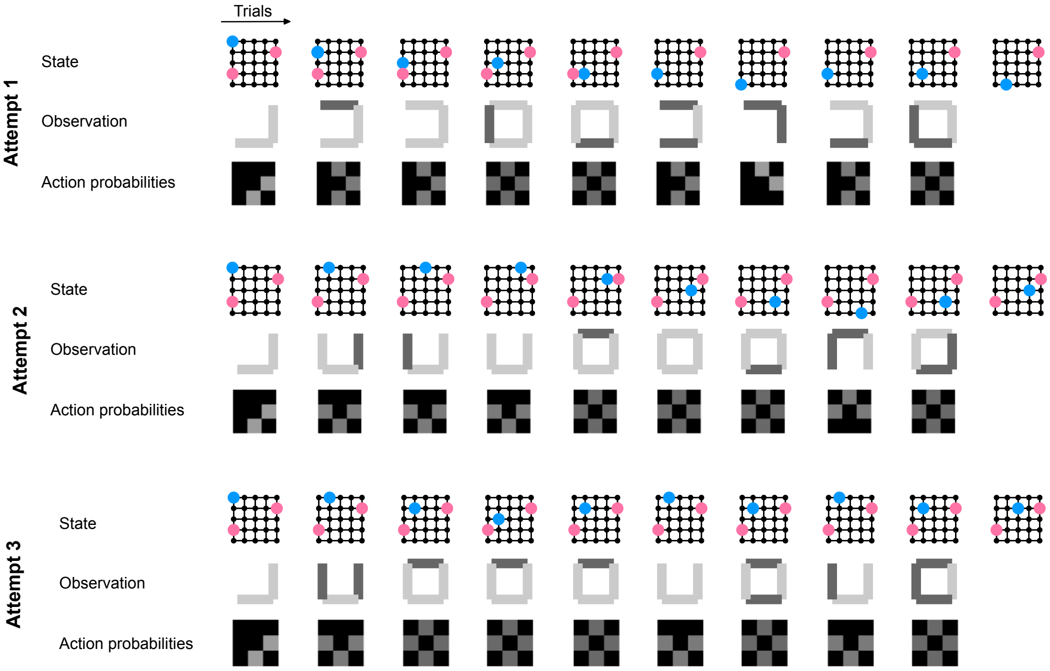

Figure 3.5. Agent A1 interacting with the treasure hunt task. A1 is a belief state-free cognitive null agent. This agent does not aim to optimize any quantity and on each trial allocates equal valences to the actions available from its position. In this simulation, the agent collects one of the treasures in the first attempt but does not solve the task. Note: Clusters of three rows show the attempts. The trial evolution is shown column-wise. The first row of each attempt shows the state with pink dots marking the treasure locations and the blue dot marking the agent position. The second row shows the collection of observation bars presented to the agent. The third row shows the action probabilities, which are allocated based on the agent-specific action valences. A lighter color marks a higher probability.

action and low valence to all other actions. Formally,

$$v_{\mathrm{A2}}\left(a_i\right) := \begin{cases} \frac{1}{|\mathcal{O}|}, & \text{if } i \in \mathcal{O} \\ 0, & \text{else} \end{cases}, \tag{3.12}$$

where, given observation o at the agent's position s^1 on trial t, $\mathcal{O} := \arg\max_i \left((o)^{i \in O_{s^1}}\right)$ denotes the set of directions marked with a light grey bar. On those few trials where no light grey bar is presented, A2 uses the action valence allocation scheme of agent A1 (eq. 3.11).

Worked trial example. On the example trial, A2 allocates the following action valences:

$$v_{\mathrm{A2}}(-5) = 1$$
$$v_{\mathrm{A2}}(+1) = 0$$
$$v_{\mathrm{A2}}(+5) = 0.$$

Thus, for agent A2, the most desirable action is the northward movement.

Simulated task behavior. Figure 3.6 shows an exemplary behavior of A2. In this simulation, the agent does not solve the task: Although A2 follows the light grey bar, it only collects one of the treasures in the first attempt, akin to A1.

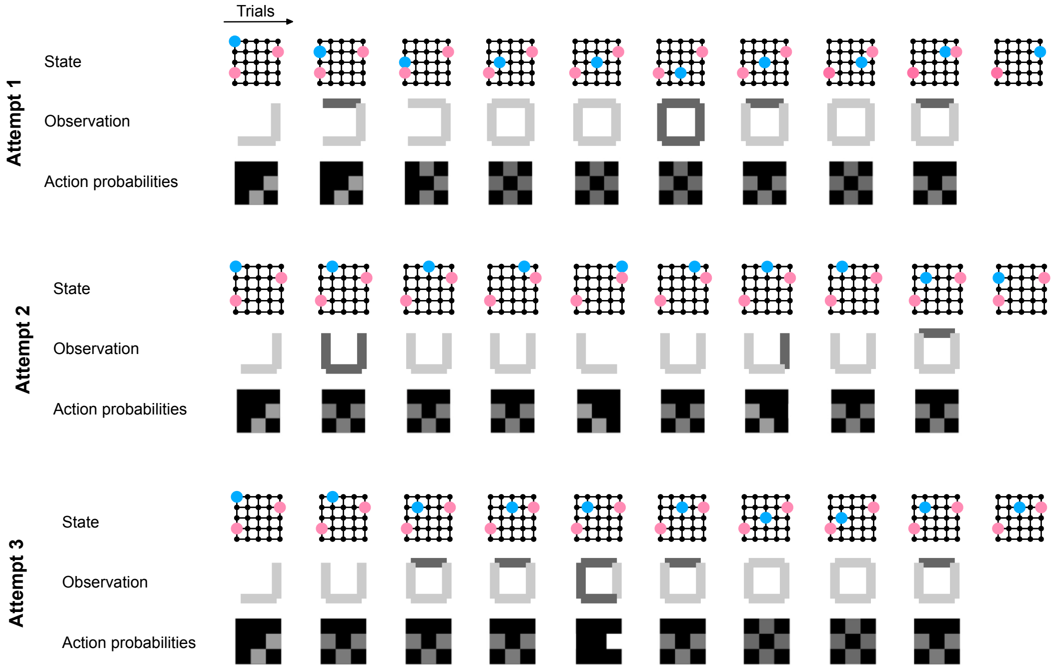

Figure 3.6. Agent A2 interacting with the treasure hunt task. A2 is a belief state-free exploitative agent. To try to collect the treasures, on each trial A2 allocates action valences based on the available observations: It allocates high valences to moving towards light grey bars and low valences to moving towards dark grey bars. In this simulation, the agent collects the treasure located at node 10 in the first attempt, but it does not revisit this location nor collects the other treasure in the second and third attempts. Note: For layout conventions, please see the legend of Figure 3.5.

Belief state-based agents The belief state-based agents denoted by A3 to A9 compute action valences based on their belief states, i.e., a probability mass function, which for every node of the grid specifies the agent's subjective uncertainty that a treasure is located there. The belief state is formed by assuming a uniform distribution over all possible treasure locations at the task's outset and employing a trial-by-trial two step belief state update. The first update happens after the agent enters a new grid position, where it either finds a treasure or not. The second update happens after the agent is presented with the observation bars. On each trial, the belief state is used to plan ahead following an exploitative (agents A3-A6), explorative (agent A7), or hybrid explorative-exploitative (agents A8-A9) objective. To realize an action, all belief state-based agents assume a maximizing decision rule, i.e., the action to issue is the action with the highest valence. Before we introduce these agents in detail, we next proceed with a formal description of the belief state.

Formally, we denote the agent's prior belief state on trial t based on the history of directly observable state components $y_{1:t-1}$ and observations $o_{1:t-1}$ by

$$\mu_t := \left(\mu_t^{(1)}, \ldots, \mu_t^{(d^2)}\right)^T \in \mathbb{R}^{d^2}. \tag{3.13}$$

The ith component of μ_t represents the agent's belief that a treasure is located at node $i \in \mathbb{N}_{d^2}$. At the beginning of the task, the components of μ_t are initialized as

$$\mu_0^{(1)} = 0 \text{ and } \mu_0^{(2)} = \mu_0^{(3)} = \ldots \mu_0^{(d^2)} = 1/\left(d^2 - 1\right) \tag{3.14}$$

representing the fact that a treasure is never encountered at the start node, and equal beliefs over all remaining nodes. The subsequent belief state updates depend on the state and take slightly different forms, for the case of two remaining treasures ($s_t^2 = 0$) or a single remaining treasure ($s_t^2 \in \{1, 2\}$). Specifically, in the former case, the belief state is first projected into a two-dimensional matrix form

$$M_t := \left(M_t^{(ij)}\right)_{1 \le i,j \le d^2} = \left(\mu_t^{(i)} \cdot \mu_t^{(j)}\right)_{1 \le i,j \le d^2} \in \mathbb{R}^{d^2 \times d^2} \tag{3.15}$$

encoding the current belief for all treasure location combinations. This two-dimensional belief state projection is necessitated to reflect that the agent does not discriminate between the first and second treasure, and thus allocates equal probabilities to the location combinations $s^3 = i, s^4 = j$ and $s^3 = j, s^4 = i$. To account for the fact that the treasures are located at different nodes, the diagonal of this belief state matrix is then set to zero

$$M_t^{(ij)} := 0, \text{ if } i = j \text{ for } i, j \in \mathbb{N}_{d^2}. \tag{3.16}$$

The two-dimensional belief state M_t is subsequently updated in two steps. First, the belief state of the agent is updated as it enters the new grid position s^1. If node k is visited on trial t and no treasure is found at this node ($y_t = (k, 0)$), the probabilities allocated to the treasure location combinations including node k are set to zero, i.e.,

$${M'_{t+1}}^{(ik)} := 0 \text{ and } {M'_{t+1}}^{(kj)} := 0, \text{ for all } i, j \in \mathbb{N}_{d^2}. \tag{3.17}$$

This reflects the agent's subjective certainty that node k is not a treasure location. The second update follows after the agent is presented with the observation bars and has the form

$$M''^{(ij)}_{t+1} := M'^{(ij)}_{t+1} \cdot p\left(o_t = o | s_t = (k, 0, i, j)\right). \tag{3.18}$$

This corresponds to a Bayesian update with prior probability $M'^{(ij)}_{t+1}$ and likelihood $p\left(o_t = o | s_t = (k, 0, i, j)\right)$. Finally, the belief state is normalized by setting

$$M_{t+1} := \frac{1}{\sum_{i=1}^{d^2} \sum_{j=1}^{d^2} M''^{(ij)}_{t+1}} M''_{t+1}. \tag{3.19}$$

Upon evaluation of this two-dimensional symmetric belief state projection, the trial posterior belief state is evaluated by marginalization

$$\mu_{t+1} := \sum_{i=1}^{d^2} M^{(ij)}_{t+1}. \tag{3.20}$$

If on trial t the agent enters node k and finds a treasure there ($y_t = (k, s_t^2 \in \mathbb{N}_2)$), a two-dimensional projection is not necessary anymore. Therefore, the position-updated belief state is given by the vector

$$\mu'_{t+1} := M_t^{(1:d^2, k)}, \tag{3.21}$$

which specifies the agent's subjective uncertainty about the remaining treasure location. This belief state is updated following the observation according to

$$\mu''^{(i)}_{t+1} := \mu'^{(i)}_{t+1} \cdot p(o_t = o | s_t = s). \tag{3.22}$$

The normalized posterior belief state

$$\mu_{t+1} := \frac{1}{\sum_{i=1}^{d^2} \mu''^{(i)}_{t+1}} \mu''_{t+1} \tag{3.23}$$

serves as the prior belief state on the subsequent trial, where the position-dependent belief state update takes the form

$$\mu'^{(s_t^1)}_{t+1} := \begin{cases} 1, \text{ if } s_t^2 \notin \left\{0, s_{t-1}^2\right\} \\ 0, \text{ else} \end{cases}. \tag{3.24}$$

If the agent does not solve the task within the attempt, the final belief state

μ_T in the attempt serves as the prior belief state in the next attempt, i.e.,

$$\mu_0 := \mu_T. \tag{3.25}$$

This holds if no treasure location was recovered in a previous attempt. Otherwise, the prior belief state in the new attempt is defined by setting the belief at the recovered treasure location to the maximum value. This modification is necessary, because, to solve the task, the agent has to collect both treasures within a single attempt. Consequently, in the new attempt, the agent has to revisit the previously recovered treasure location. Formally, the prior belief state in the new attempt is first defined as the final belief state in the previous attempt (eq. 3.25). Then, the prior belief state is modified and normalized according to

$$\mu_0^{(k)} := 1,\ \mu_0 := \frac{1}{\sum_{i=1}^{d^2} \mu_0^{(i)}} \mu_0, \tag{3.26}$$

where k denotes the previously recovered treasure location. The subsequent belief state updates are evaluated as described above with a minor modification: The two-dimensional belief state projection is given by assigning the prior belief state vector and its transpose to the kth column and kth row of a null matrix of the size $d^2 \times d^2$, respectively. Formally, the two-dimensional belief state is first defined as

$$M_t := 0^{d^2 \times d^2} \tag{3.27}$$

and then its entries are modified according to

$$M_t^{\left(1:d^2,k\right)} := \mu_t \text{ and } M_t^{\left(k,1:d^2\right)} := \mu_t^T. \tag{3.28}$$

Note that we throughout assume that the agent is endowed with the true probability of making observation o given state s as defined by the observation function g of the task model. As detailed above, these probabilities formalize the notions that (1) a direction leading away from the treasures is always marked with a dark grey bar and (2) the accuracy of the treasure detector correctly returning a light grey bar in the direction towards the treasures increases as the agent moves closer to the treasures. To motivate our assumption, recall the experimental procedure participants underwent prior to the runs in the scanner. In the task instructions, participants were truthfully informed about the first notion (Supplementary Material B.1). Although participants were not further informed about the second notion, before the runs in the scanner, they completed a training run by the end of which they arguably became familiar with the distance-dependent accuracy of the treasure detector.

Belief state-based exploitative agents Agents A3, A4, A5 and A6 implement variations of a belief state-based exploitative planning strategy: On each trial, these agents harness their accumulated knowledge about the location of the treasures to try to move closer to, and eventually collect them. Formally, each belief state-based exploitative agent allocates action valences by employing a heuristic real-time dynamic programming approach originally proposed by Korf (1990) and elaborated on by Geffner and Bonet (1998). The characteristic differentiating these agents is whether this approach is based on unitary state estimates or the entire belief state.

The belief state-based exploitative agent A3 The belief state-based exploitative agent A3 identifies the node with the highest subjective uncertainty that a treasure is located there and allocates the highest valence to the action which leads closest to this node. Formally, on every trial $t \in \mathbb{N}_T$, A3 first identifies the maximum-a-posteriori (MAP) belief state node

$$i_t = \arg\max_{i \in \mathbb{N}_{d^2}} \mu_{t+1}^{(i)}. \tag{3.29}$$

Then, adopting the heuristic real-time dynamic programming approach (Geffner & Bonet, 1998; Korf, 1990), the agent substitutes node i_t in the l_1 distance-based heuristic function

$$v_{\mathrm{A3}}(a) := -l_1\left(s_t^1 + a, i_t\right) \tag{3.30}$$

to allocate action valences. If there is more than one MAP belief state node, A3 evaluates the actions with respect to the node with the smallest index. Note that in eq. 3.30 the negative of the l_1 distances is taken. This is to ensure consistency across agents, in that a higher valence indicates a more desirable action.

Worked trial example. On the example trial, the MAP belief state node corresponds to

$$i_t = 21.$$

Consequently, A3 allocates the following action valences:

$$\begin{aligned}
v_{\mathrm{A3}}(-5) &= -l_1(6, 21) = -3 \\
v_{\mathrm{A3}}(+1) &= -l_1(12, 21) = -3 \\
v_{\mathrm{A3}}(+5) &= -l_1(16, 21) = -1.
\end{aligned}$$

Thus, for agent A3, the most desirable action is the southward movement.

Simulated task behavior. Figure 3.7 shows an exemplary behavior of A3. In this simulation, the agent does not solve the task: A3 recovers the first treasure location already in the first attempt but does not manage to recover the second treasure location. The reason for this is that the strategy of A3 has limited flexibility. Specifically, after the agent finds a treasure at node 10 in the first attempt, it becomes certain that this node is a treasure location. Consequently, at the outset of the second and third attempts, A3 allocates the highest belief to this node and goes directly to it. The problem is that from this node, the agent can not reach the other treasure located at node 16 within the step limit - even if the agent allocates the second highest belief to node 16. As a result, A3 closely approaches the second treasure location but does solve the task.

The belief state-based exploitative agent A4 The belief state-based exploitative agent A4 uses the same strategy as A3 with one important difference: Until there is only one treasure left, the agent identifies the most probable location of both treasures and evaluates the actions with respect to the location that is closer to its current position. Formally, while $s_t^2 = 0$, A4 first identifies the two nodes with the highest associated beliefs,

$$i_t^1 = \arg\max_{k \in \mathbb{N}_{d^2}} \mu_{t+1}^{(k)} \text{ and } i_t^2 = \arg\max_{l \in \left\{\mathbb{N}_{d^2} \setminus i_t^1\right\}} \mu_{t+1}^{(l)}. \tag{3.31}$$

Then, the agent identifies the node reachable within fewer steps

$$j_t = \arg\min_{m \in \left\{i_t^1, i_t^2\right\}} l_1\left(s_t^1, m\right) \tag{3.32}$$

and substitutes this node in the l_1 distance-based heuristic function to allocate action valences. After the recovery of a treasure location ($s_t^2 \in \{1, 2\}$), A4 uses the valence allocation scheme of A3 to collect the remaining treasure, i.e.,

$$v_{\text{A4}}(a) := \begin{cases} -l_1\left(s_t^1 + a, j_t\right), & \text{if } s_t^2 = 0 \\ v_{\text{A3}}(a), & \text{else} \end{cases}. \tag{3.33}$$

Worked trial example. On the example trial, the nodes with the highest associated beliefs are

$$i_t^1 = 21 \text{ and } i_t^2 = 6.$$

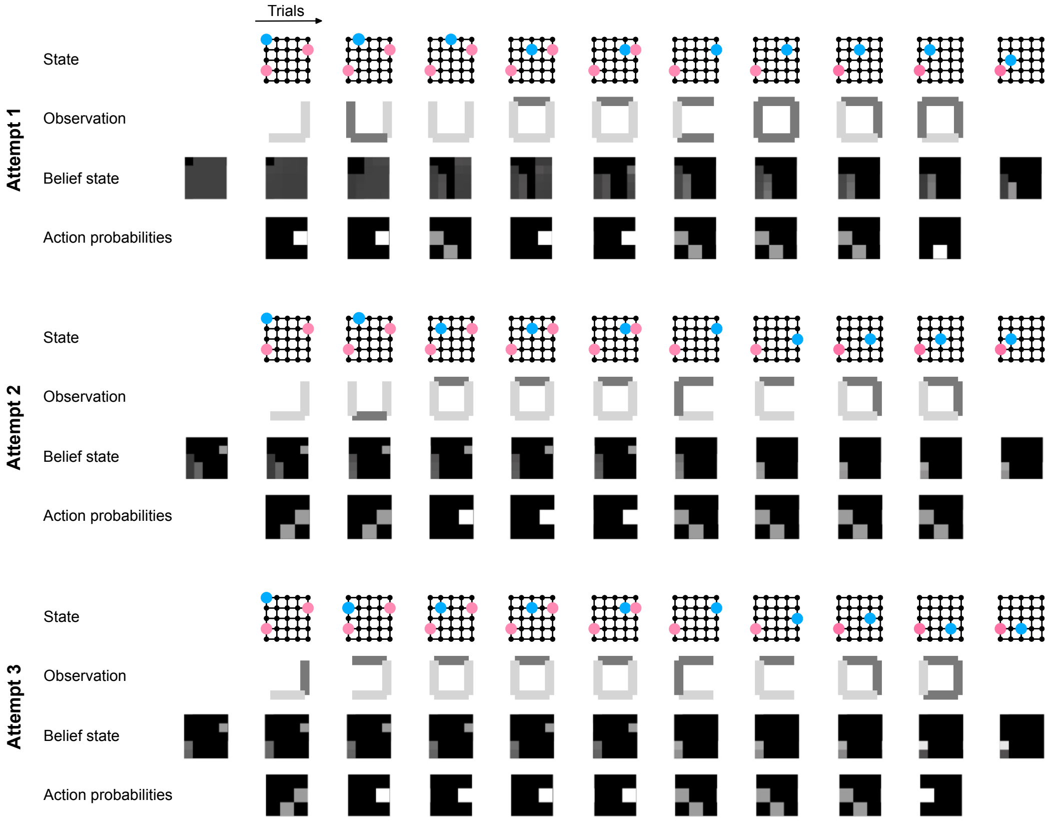

Figure 3.7. Agent A3 interacting with the treasure hunt task. A3 is a belief state-based exploitative agent. This agent tries to collect the treasures by relying on its belief state, a dynamically evolving probabilistic representation of the treasure locations. In particular, on each trial, this agent identifies the most likely to treasure location and allocates action valences depending on how close the new position would be to this location. Formally, this scheme corresponds to evaluating the MAP belief state node and computing action valences using a real time dynamic approach with an l_1 distance-based heuristic function. In this simulation, A3 collects the treasure located at node 10 in the first attempt. Although the agent revisits this node in the second and third attempts, it fails to collect the other treasure at node 16. Note: The trial-by-trial evolution of the agent's belief state is shown in the third row of each attempt, which complements the trial-by-trial states, observations, and action probabilities as shown for the belief state-free agents (cf. Figures 3.5 and 3.6). As in the case of the action probabilities, a lighter color indicates a higher probability.

Of these,

$$j_t = 6$$

is closer to the agent's position. Consequently, A4 allocates the following valences:

$$v_{\mathrm{A4}}(-5) = -l_1(6, 6) = 0$$
$$v_{\mathrm{A4}}(+1) = -l_1(12, 6) = -2$$
$$v_{\mathrm{A4}}(+5) = -l_1(16, 6) = -2.$$

Thus, for agent A4, the most desirable action is the northward movement.

Simulated task behavior. Figure 3.8 shows an exemplary behavior of A4. In this simulation, the agent solves the task: The agent recovers the first treasure location in the first attempt, the second treasure location in the second attempt and, by combining its knowledge about these locations, consecutively collects both treasures in the third attempt. This simulation demonstrates that by taking into account the potential location of both treasures A4 overcomes the limitation of A3.

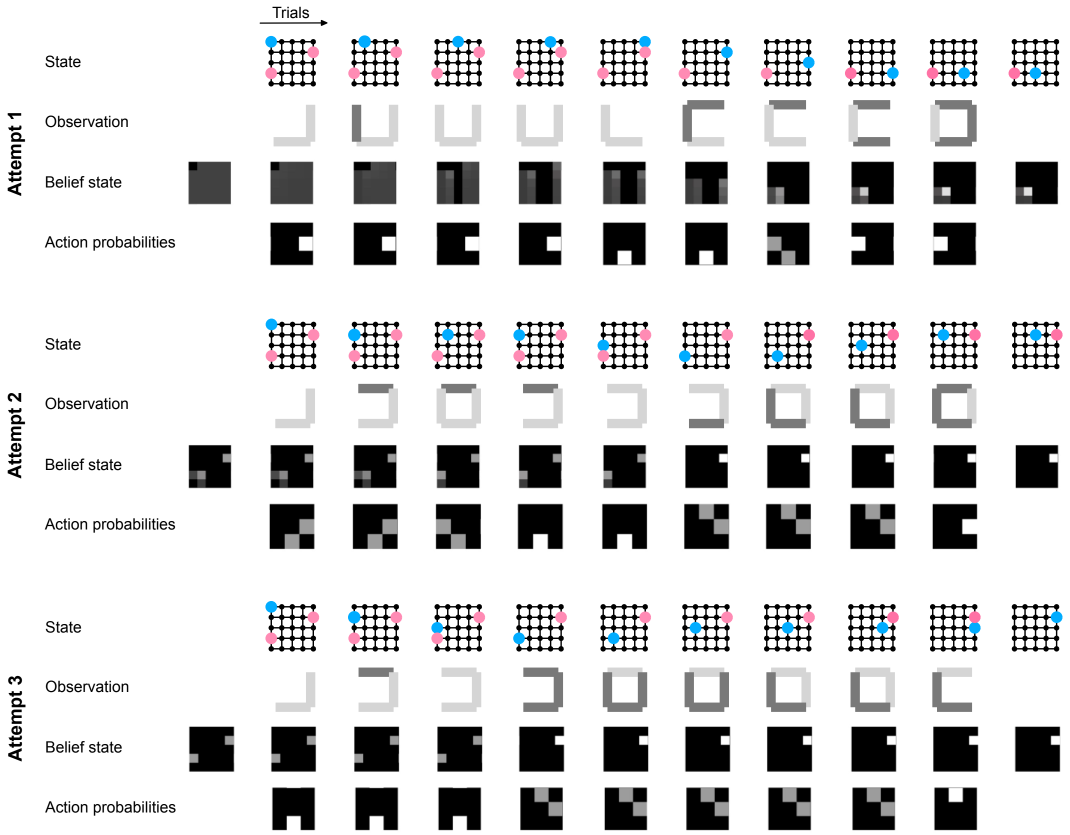

Figure 3.8. Agent A4 interacting with the treasure hunt task. A4 is a belief state-based exploitative agent with a similar strategy to that of A3. The difference is that A4 takes into account the potential location of both treasures. Specifically, until a treasure is collected in the attempt, A4 first identities the two most probable treasure locations, then the one closer to its current position and subsequently evaluates the actions based on the l_1 distance between this location and the new position. After a treasure is collected, A4 uses the same valence allocation scheme as A3. In this simulation, A4 recovers the treasure at node 10 in the first attempt, the treasure at node 16 in the second attempt and solves the task in the third attempt. Note: For layout conventions, please see the legend of Figure 3.7.

The belief state-based exploitative agent A5 The belief state-based exploitative agent A5 does not identify single most probable treasure locations. Instead, this agent takes into account its entire belief state when evaluating the actions. Specifically, for each available action $a \in A_{s^1}$, A5 evaluates the negative sum of the belief-weighted l_1 distances between the new position that would result upon the action and each node of the grid,

$$v_{\mathrm{A5}}(a) := -\sum\nolimits_{i=1}^{d^2} \mu_{t+1}^{(i)} l_1\left(s_t^1 + a, i\right). \tag{3.34}$$

Intuitively, agent A5 favors the action that leads towards the part of the grid, where a treasure is likely located and which is in close reach.

Worked trial example. On the example trial, agent A5 allocates the following action valences:

$$\begin{aligned}
v_{\mathrm{A5}}(-5) = -(&0.25 \cdot l_1(6,6) + 0.15 \cdot l_1(6,13)\\
&+ 0.2 \cdot l_1(6,14) + 0.1 \cdot l_1(6,15) + 0.3 \cdot l_1(6,21)) = -2.65\\
v_{\mathrm{A5}}(+1) = -(&0.25 \cdot l_1(12,6) + 0.15 \cdot l_1(12,13)\\
&+ 0.2 \cdot l_1(12,14) + 0.1 \cdot l_1(12,15) + 0.3 \cdot l_1(12,21)) = -2.25\\
v_{\mathrm{A5}}(+5) = -(&0.25 \cdot l_1(16,6) + 0.15 \cdot l_1(16,13)\\
&+ 0.2 \cdot l_1(16,14) + 0.1 \cdot l_1(16,15) + 0.3 \cdot l_1(16,21)) = -2.55.
\end{aligned}$$

Thus, for agent A5, the most desirable action is the eastward movement.

Simulated task behavior. Figure 3.9 shows an exemplary behavior of A5. In this simulation, the agent solves the task: A5 does not collect any of the treasures in the first attempt, but it collects the treasure at node 16 in the second attempt and both treasures in the third attempt. As this agent takes into account its entire belief state when evaluating the actions' desirability, tasks where the treasures are located at opposite ends of the grid can lead to to ambiguous decision situations, i.e., trials with multiple best actions leading in opposite directions. For example, in the exemplary simulation, the agent faces a series of ambiguous decision situations in the first attempt. This happens because the agent correctly believes that the treasures are located at opposite ends and therefore, midway between the treasures, it allocates equally high valences to converse actions. As a result, the agent moves back-and-forth in the first attempt but it manages to break this pattern in the second attempt and solve the task in the third attempt.

The belief state-based exploitative agent A6 The belief state-based exploitative agent A6 uses the same strategy as A5, but in ambiguous decision situations it switches to the strategy of A4. Formally, if there is more than one action maximizing the valence function of A5 (eq. 3.34), the agent re-evaluates

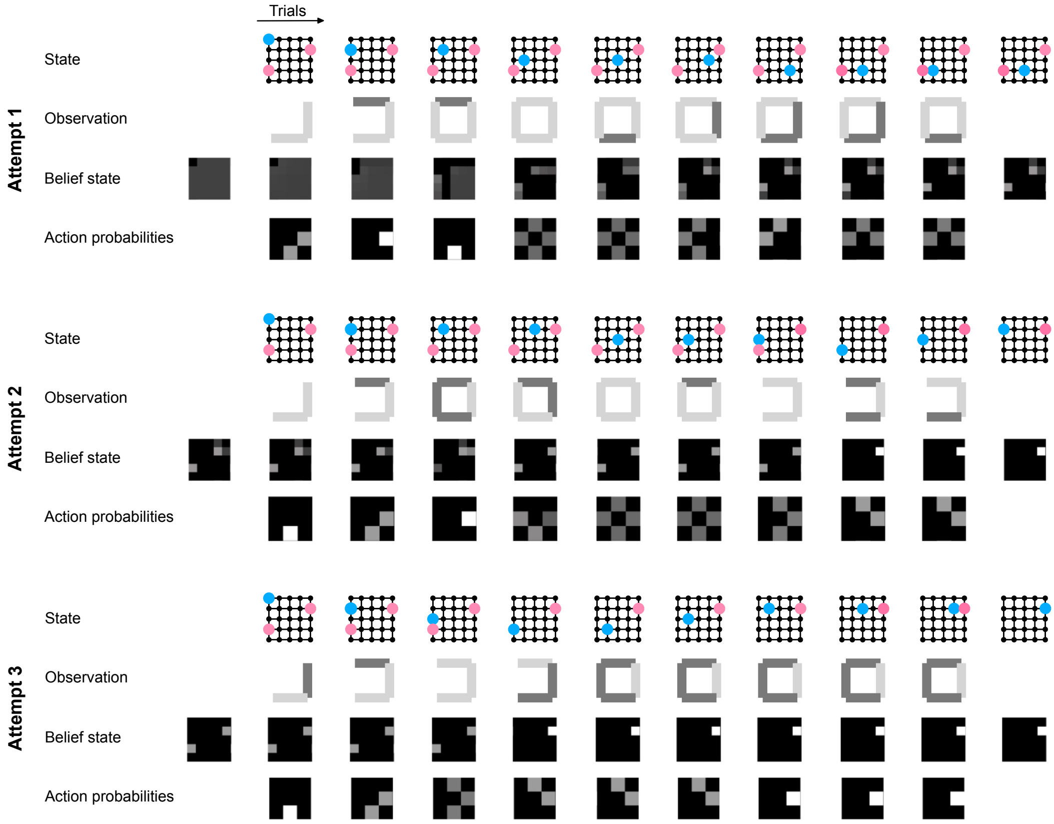

Figure 3.9. Agent A5 interacting with the treasure hunt task. A5 is a belief state-based exploitative agent. To evaluate the valence of an action, A5 does not identify single potential treasure locations - like agents A3 and A4 do - but uses its knowledge about all nodes and computes the sum of belief-weighted l_1 distances. In this simulation, the agent collects the treasure at node 16 in the second attempt and solves the task in the third attempt. Note: For layout conventions, please see the legend of Figure 3.7.

all available actions based the valence allocation scheme of A4 (eqs. 3.31-3.33),

$$v_{\text{A6}}(a) := \begin{cases} v_{\text{A5}}(a), & \text{if } \left| \underset{a \in A_{s^1}}{\arg\max}\, v_{\text{A5}}(a) \right| = 1 \\ v_{\text{A4}}(a), & \text{else} \end{cases}. \tag{3.35}$$

Worked trial example. On the example trial, there is only one action maximizing eq. 3.34,

$$\left| \underset{a \in A_{s^1}}{\arg\max}\, v_{\text{A5}}(a) \right| = 1$$

and thus

$$v_{\text{A6}}(a) = v_{\text{A5}}(a).$$

To illustrate an ambiguous decision situation, assume that the agent has the non-zero trial posterior belief state components $\mu_{t+1}^{(6)} = 0.3$, $\mu_{t+1}^{(13)} = 0.15$, $\mu_{t+1}^{(14)} = 0.2$

and $\mu_{t+1}^{(21)} = 0.35$. Then, the action valences based on the scheme of A5 evaluate to

$$\begin{aligned} v_{\mathrm{A5}}(-5) &= -(0.3 \cdot l_1(6,6) + 0.15 \cdot l_1(6,13) \\ &\qquad + 0.2 \cdot l_1(6,14) + 0.35 \cdot l_1(6,21)) = -2.3 \\ v_{\mathrm{A5}}(+1) &= -(0.3 \cdot l_1(12,6) + 0.15 \cdot l_1(12,13) \\ &\qquad + 0.2 \cdot l_1(12,14) + 0.35 \cdot l_1(12,21)) = -2.2 \\ v_{\mathrm{A5}}(+5) &= -(0.3 \cdot l_1(16,6) + 0.15 \cdot l_1(16,13) \\ &\qquad + 0.2 \cdot l_1(16,14) + 0.35 \cdot l_1(16,21)) = -2.2. \end{aligned}$$

As the eastward and southward movements are equally desirable, i.e.,

$$\left| \arg\max_{a \in A_{s^1}} v_{\mathrm{A5}}(a) \right| = 2,$$

A6 re-evaluates the actions using the A4 scheme:

$$i_t^1 = 21 \text{ and } i_t^2 = 6.$$

Consequently,

$$j_t = 6$$

and therefore,

$$\begin{aligned} v_{\mathrm{A4}}(-5) &= -l_1(6,6) = 0 \\ v_{\mathrm{A4}}(+1) &= -l_1(12,6) = -2 \\ v_{\mathrm{A4}}(+5) &= -l_1(16,6) = -2. \end{aligned}$$

Given that

$$v_{\mathrm{A6}}(a) = v_{\mathrm{A4}}(a),$$

the most desirable action for A6 is the northward movement.

Simulated task behavior. Figure 3.10 shows an exemplary behavior of A6. In this simulation, the agent solves the task: By switching between the strategies of A5 and A4, the agent recovers the first treasure location in the first attempt and manages to solve the task already in the second attempt.

The belief state-based explorative agent A7 The belief state-based explorative agent A7 tries to resolve its uncertainty about the treasure locations as quickly as possible. To accomplish this, A7 uses its belief state to evaluate the amount of information it would gain by choosing a certain action. Specifically, to allocate action valences, A7 adopts a one-step look-ahead Bayesian exploration

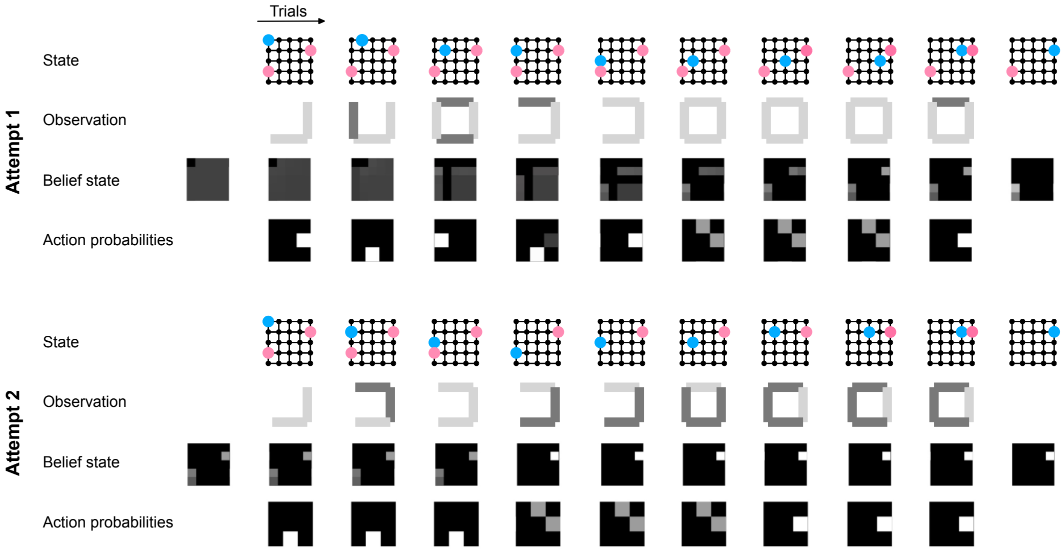

Figure 3.10. Agent A6 interacting with the treasure hunt task. A6 is a belief state-based exploitative agent that uses the same strategy as A5. If, however, the agent faces an ambiguous decision situation, it switches strategy and allocates action valences using the scheme of A4. In this simulation, A6 collects the first treasure in the first attempt and both treasures in the second attempt. Note: For layout conventions, please see the legend of Figure 3.7.

approach (e.g., Sun, Gomez, and Schmidhuber (2011)): For each available action $a \in A_{s^1}$, the agent evaluates the expected Bayesian surprise, which corresponds to the expected shift in the belief state that would result upon entering the new grid position and being presented with the observation bars. This shift is evaluated in two steps. First, A7 computes the Kullback-Leibler (KL) divergence between the current trial posterior belief state μ_{t+1} (which serves as the prior belief state on the next trial) and the simulated position-updated belief state μ'_{t+2} as

$$KL\left(\mu'_{t+2}||\mu_{t+1}\right) := \sum\nolimits_{i=1}^{d^2} {\mu'_{t+2}}^{(i)} \cdot \ln\left(\frac{{\mu'_{t+2}}^{(i)}}{\mu_{t+1}^{(i)}}\right). \tag{3.36}$$

Then, by taking into account all possible observations at the simulated new position s_*^1, A7 computes the expected KL divergence between the simulated position-updated belief state μ'_{t+2} and the simulated position- and observation-updated belief state μ''_{t+2},

$$E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right) = \sum\nolimits_{o \in O_{s_*^1}} p\left(o\right) KL\left(\mu''_{t+2}||\mu'_{t+2}\right). \tag{3.37}$$

Importantly, as eqs. 3.36 and 3.37 depend on whether there is a treasure discovery on the next trial or not, the overall expected shift in the belief state for an action $a \in A_{s^1}$ is given by

$$\begin{aligned} v_{\mathrm{A7}}(a) :=& z\mu_{t+1}^{(s_*^1)} \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 \neq s_t^2} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 \neq s_t^2} \right) \\ &+ \left(1 - z\mu_{t+1}^{(s_*^1)}\right) \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 = s_t^2} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 = s_t^2} \right). \end{aligned} \tag{3.38}$$

The first term in eq. 3.38 captures the expected shift in the belief state if there is a treasure discovery on trial $t+1$. The second term captures the expected shift in the belief state if there is no treasure discovery on trial $t+1$. The product $z\mu_{t+1}^{(s_*^1)}$ captures the agent's subjective uncertainty that a treasure is located at the new grid position s_*^1. Here, z is set to 2 if no treasure was discovered in the attempt up to trial t (i.e., $s_t^2 = 0$), and to 1 otherwise. This is to account for the fact that until a treasure is discovered in the attempt, μ represents the belief over both treasure locations. Therefore, maximal certainty about a treasure location is represented as a value of 0.5 at the corresponding component of μ.

Worked trial example. On the example trial, A7 allocates the valence

$$\begin{aligned} v_{\mathrm{A7}}(-5) &= z\mu_{t+1}^{(6)} \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 \in \{1,2\}} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 \in \{1,2\}} \right) \\ &\quad + \left(1 - z\mu_{t+1}^{(6)}\right) \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 = 0} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 = 0} \right) \\ &= 2 \cdot 0.25 \cdot (0.29 + 0.29) + (1 - 2 \cdot 0.25) \cdot (0.3 + 0) \\ &= 0.44 \end{aligned}$$

to the northward movement, the valence

$$\begin{aligned} v_{\mathrm{A7}}(+1) &= z\mu_{t+1}^{(12)} \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 \{1,2\}} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 \{1,2\}} \right) \\ &\quad + \left(1 - z\mu_{t+1}^{(12)}\right) \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 = 0} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 = 0} \right) \\ &= 2 \cdot 0 \cdot (0 + 0) + (1 - 2 \cdot 0) \cdot (0 + 0.45) \\ &= 0.45 \end{aligned}$$

to the eastward movement, and the valence

$$\begin{aligned} v_{\mathrm{A7}}(+5) &= z\mu_{t+1}^{(16)} \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 \{1,2\}} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 \{1,2\}} \right) \\ &\quad + \left(1 - z\mu_{t+1}^{(16)}\right) \cdot \left(KL\left(\mu'_{t+2}||\mu_{t+1}\right)_{s_{t+1}^2 = 0} + E\left(KL\left(\mu''_{t+2}||\mu'_{t+2}\right)\right)_{s_{t+1}^2 = 0} \right) \\ &= 2 \cdot 0 \cdot (0 + 0) + (1 - 2 \cdot 0) \cdot (0 + 0.4) \\ &= 0.4. \end{aligned}$$

to the southward movement. Thus, for agent A7, the most desirable action is the eastward movement.

Simulated task behavior. Figure 3.11 shows an exemplary behavior of agent A7. In this simulation, the agent does not solve the task: Although A7 recovers the first treasure location in the first attempt and the second treasure location in the second attempt, it fails to revisit these locations in the third attempt. It is notable that until the agent recovers the second treasure location, it behaves as if it tried to collect the treasures. The reason for this is that in order to reduce uncertainty about the treasure locations, A7 first visits parts of the grid that likely house a treasure. Then, once the agent recovers both locations, its uncertainty is resolved. From this point on, the agent allocates equal valences to all available actions because there is no further information to be gained. Consequently, after the recovery of the second treasure, A7 behaves like the belief state-free cognitive null agent A1 and does not solve the task.

Belief state-based hybrid explorative-exploitative agents We designed two belief state-based hybrid agents, A8 and A9, that combine the objectives of the explorative and exploitative agents. Specifically, these agents explore their belief states first to reduce their uncertainty about the treasure locations and then exploit their accumulated knowledge to collect the treasures. To this end, both agents allocate action valences by evaluating the convex combination of the valences of the belief state-based explorative agent and a belief state-based exploitative agent. However, there are two main differences between agents A8 and A9. First, they employ the valences of different belief state-based exploitative agents. Second, they differ with respect to the functional form of their weighting parameter.

The belief state-based hybrid agent A8 Agent A8 is the simpler belief state-based hybrid explorative-exploitative agent. To allocate action valences, this agent computes the convex combination of the normalized action valences of the belief state-based exploitative agent A3 and the belief state-based explorative agent A7. For this agent, the weighting parameter is constant throughout. Formally,

$$v_{\mathrm{A8}}(a) := \lambda \cdot - \frac{v_{\mathrm{A3}}(a)}{\sum_{a_* \in A_{s1}} v_{\mathrm{A3}}(a_*)} + (1 - \lambda) \cdot \frac{v_{\mathrm{A7}}(a)}{\sum_{a_* \in A_{s1}} v_{\mathrm{A7}}(a_*)}, \tag{3.39}$$

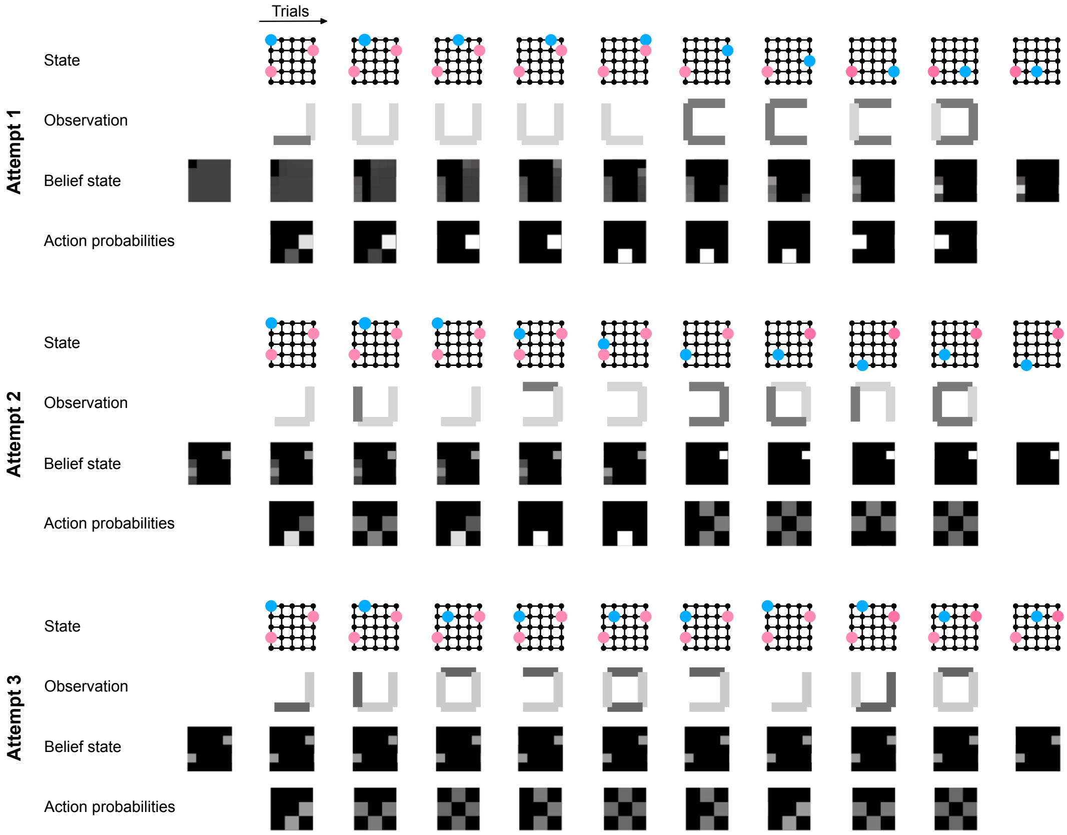

Figure 3.11. Agent A7 interacting with the treasure hunt task. A7 is a belief state-based explorative agent that tries to resolve its uncertainty about the treasure locations. To this end, on each trial, this agent allocates action valences based on the expected shift in the belief state that would result upon choosing the action. In this simulation, A7 collects the treasure at node 10 in the first attempt and the treasure at node 16 in the second attempt. After the recovery of both treasure locations, there is no further information to be gained and the agent allocates equal valences to the available actions. By doing so, A7 fails to solve the task in the third attempt. Note: For layout conventions, please see the legend of Figure 3.7.

where $\lambda \in [0, 1]$ is the weighting parameter. The reason we chose to adopt the action valences of A3 for this agent is that A3 arguably implements the purest form of a belief state-based exploitative strategy: A3 always favors the action that is the safest bet - i.e., the move towards the node which has the highest associated belief - and does not take into account other aspects of the task environment, such as the presence of two treasures. Therefore, for the simpler belief state-based hybrid explorative-exploitative agent the valences of A3 naturally contrast with the valences of the belief state-based exploitative agent A7.

Worked trial example. On the example trial, agent A8 allocates the following action valences using a weighting parameter of $\lambda = 0.7$:

$$
\begin{aligned}
v_{\mathrm{A8}}(-5) &= 0.7 \cdot -\frac{-3}{-3-3-1} + (1-0.7) \cdot \frac{0.44}{0.44+0.45+0.4} = -0.198 \\
v_{\mathrm{A8}}(+1) &= 0.7 \cdot -\frac{-3}{-3-3-1} + (1-0.7) \cdot \frac{0.45}{0.44+0.45+0.4} = -0.195 \\
v_{\mathrm{A8}}(+5) &= 0.7 \cdot -\frac{-1}{-3-3-1} + (1-0.7) \cdot \frac{0.4}{0.44+0.45+0.4} = -0.007.
\end{aligned}
$$

Thus, for agent A8, the most desirable action is the southward movement.

Simulated task behavior. Figure 3.12 shows an exemplary behavior of agent A8. In this simulation with the weighting parameter set to $\lambda = 0.7$, the agent does not solve the task: A8 recovers the treasure location at node 16 in the first attempt and the treasure location at node 10 in the second attempt. Although the agent revisits and thereby collects the treasure at node 10 in the third attempt, it does not manage to collect the other treasure before the step limit is reached. Notably, while a weighting parameter value of $\lambda = 0.7$ grants a stronger control to the belief state-based exploitative agent A3, the influence of the belief state-based explorative agent A7 is evident. In particular, in the second attempt agent A8 does not get trapped in revisiting the treasure location discovered in the first attempt, as A3 would do, but seeks for and recovers the other treasure location. After recovering both treasure locations, A7 does not differentiate between the available actions and therefore action selection on the remaining trials is driven by the valences of A3. Consequently, in the third attempt, agent A8 behaves similar to agent A3 and does not solve the task.

The belief state-based hybrid agent A9 Agent A9 is the more advanced belief state-based hybrid explorative-exploitative agent. This agent allocates action valences based on the convex combination of the valences of the belief state-based exploitative agent A4 and the belief state-based explorative agent A7. In contrast to the constant weighting parameter of agent A8, for this agent the value of the weighting parameter depends on its current state uncertainty. Formally,

$$
v_{\mathrm{A9}}(a) := \lambda_t \cdot -\frac{v_{\mathrm{A4}}(a)}{\sum_{a_* \in A_{s1}} v_{\mathrm{A4}}(a_*)} + (1-\lambda_t) \cdot \frac{v_{\mathrm{A7}}(a)}{\sum_{a_* \in A_{s1}} v_{\mathrm{A7}}(a_*)}, \qquad (3.40)
$$

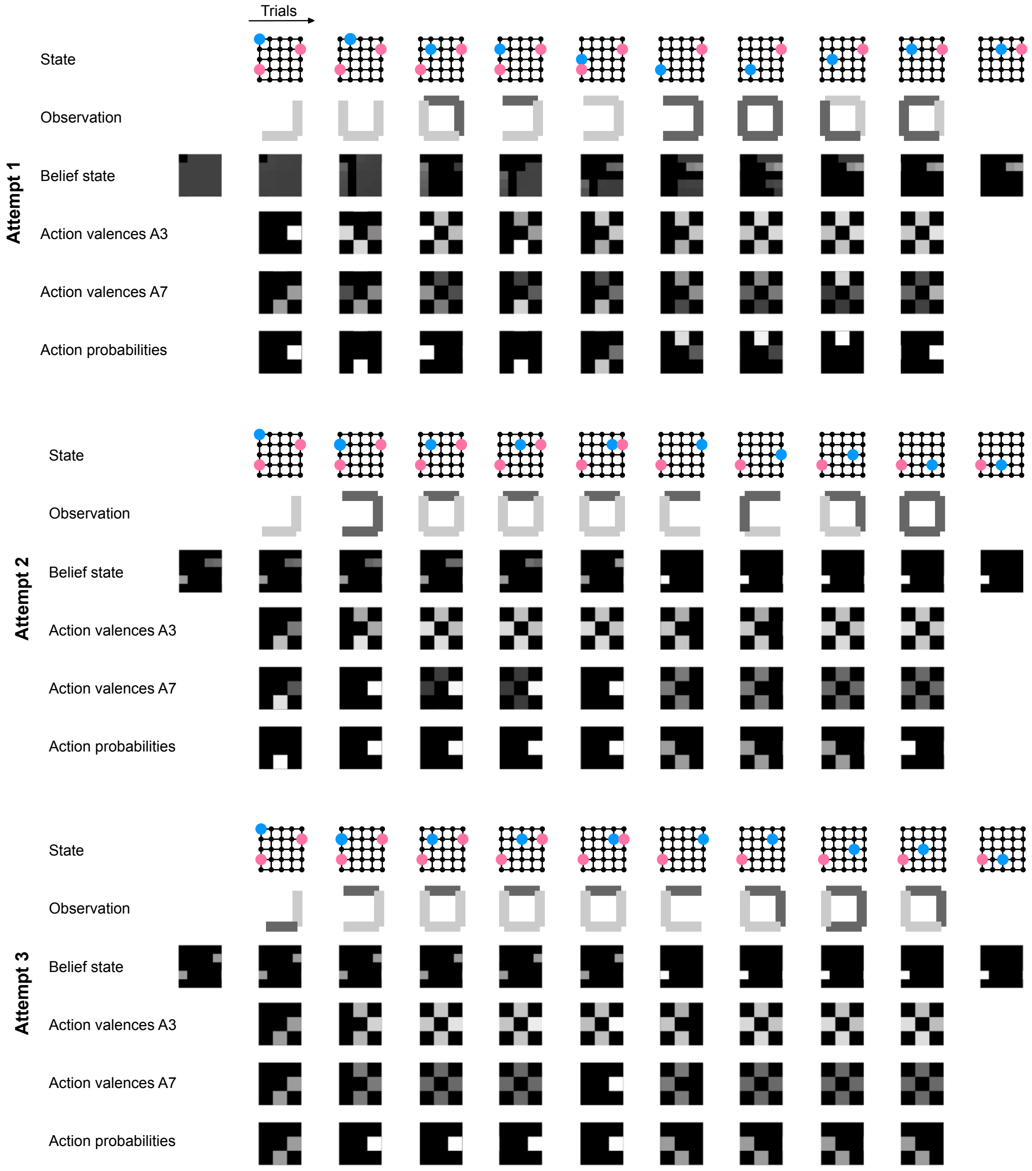

Figure 3.12. Agent A8 interacting with the treasure hunt task. A8 is a belief state-based hybrid explorative-exploitative agent that tries to resolve its uncertainty about the location of the treasures first and then collect them. To this end, agent A8 allocates action valences by evaluating the convex combination of the action valences of agents A3 and A7. In this simulation, the agent collects the first treasure in the first attempt and the second treasure in the second attempt. In the third attempt, the agent collects the first treasure but it does not manage to reach the second treasure location before the step limit is exhausted. Note: The trial-by-trial normalized action valences of agents A3 and A7 are shown in the fourth and fifth row of each attempt, respectively, with a lighter color indicating a higher value. These complement the trial-by-trial states, observations, belief states, and action probabilities as also shown for the other belief state-based agents (cf. Figure 3.7). The weighting parameter λ was set to 0.7 in this simulation.

where $\lambda_t \in [0, 1]$ is the weighting parameter on trial t and is evaluated according to

$$\lambda_t = \eta_0 \cdot \exp\left(-\eta_1 \cdot H\left(\mu_{t+1}\right)\right). \tag{3.41}$$

Here, $\eta_0 \in [0,1]$ and $\eta_1 \in \mathbb{R}_{\geq 0}$ are the offset and slope parameters of the exponential decay function, respectively, and

$$H\left(\mu_{t+1}\right) := -\sum\nolimits_{i=1}^{d^2} \mu_{t+1}^{(i)} \cdot \ln\left(\mu_{t+1}^{(i)}\right) \tag{3.42}$$

is the entropy of the agent's trial posterior belief state. Of note, our choice to express the dynamic weighting parameter with an exponential function is in line with similar previous work (e.g., Camerer and Ho, 1998; Gläscher et al., 2010). This agent combines the valences of agents A4 and A7. As discussed above for agent A4, by taking into consideration the possible location of both treasures, the strategy of A4 allows for more flexibility compared to the strategy of A3. Consequently, we reasoned that for the more advanced belief state-based hybrid explorative-exploitative agent the valences of A4 constitute a suitable contrast to the valences of the belief state-based exploitative agent A7.

Worked trial example. On the example trial, state uncertainty evaluates to

$$H\left(\mu_{t+1}\right) = 1.54.$$

By setting the offset and slope parameters to $\eta_0 = 1$ and $\eta_0 = 0.4$, respectively, the weighting parameter evaluates to

$$\lambda_t = 1 \cdot \exp(-0.4 \cdot 1.54) = 0.54.$$

Consequently, A9 allocates the following action valences:

$$\begin{aligned}
v_{\text{A9}}(-5) &= 0.54 \cdot -\frac{0}{0-2-2} + (1-0.54) \cdot \frac{0.44}{0.44+0.45+0.4} = 0.16 \\
v_{\text{A9}}(+1) &= 0.54 \cdot -\frac{-2}{0-2-2} + (1-0.54) \cdot \frac{0.45}{0.44+0.45+0.4} = -0.11 \\
v_{\text{A9}}(+5) &= 0.54 \cdot -\frac{-2}{0-2-2} + (1-0.54) \cdot \frac{0.4}{0.44+0.45+0.4} = -0.13\ .
\end{aligned}$$

Thus, for agent A9, the most desirable action is the northward movement.

Simulated task behavior. Figure 3.13 shows an exemplary behavior of agent A9. In this simulation with an offset of $\eta_0 = 1$ and a slope of $\eta_1 = 0.4$, the agent solves the task: A9 finds the treasure at node 16 in the first attempt, the treasure at node 10 in the second attempt and collects both treasures in the third attempt. For the η_0 and η_1 values used in this simulation, the weighting parameter λ_t grants dominant control to the belief state-based explorative agent A7 if the agent is maximally uncertain about the state (this is the case in the first trial of the first attempt, where the entropy of the trial posterior belief state is $H(\mu_2) \approx 3$ and thus $\lambda_1 = 0.3$). The weighting parameter λ_t grants full control to the belief state-based exploitative agent A4 if the agent's uncertainty is resolved ($\lambda_t = 1$; this is the case if both treasure locations are recovered and thus the entropy of the belief state is $H(\mu_{t+1}) \approx 0$). To demonstrate the changing contribution of agents A7 and A4 in the action valences of A9, consider the second and third attempts. In the beginning of the second attempt, the agent is still uncertain about the location of the first treasure. Consequently, the preference of A7 prevails over the preference of A4 and the agent searches for and finds the treasure. Then, in the third attempt, the valences of agent A4 prevail and, like agent A4, agent A9 solves the task (cf. Figure 3.8).

3.3 Behavioral methods

3.3.1 Descriptive behavioral data analyses

We evaluated six descriptive statistics of participants' decision-making behavior in the treasure hunt task. First, we evaluated the average performance by dividing the sum of solved tasks across participants by the number of participants. Second, we evaluated the average number of solvable tasks by dividing the sum of solvable tasks across participants by the number of participants. We evaluated this statistic because some tasks were not solvable due to the randomly allocated step limit. This was the case if the step limit in all three attempts of a task was smaller than the optimal step limit. For the last four statistics, we evaluated the average performance per run, per attempt, per optimal step limit and per treasure location combination. To evaluate the average performance per run, for each participant and run, we first divided the number of solved tasks by the number of solvable tasks and subsequently computed the mean and the standard error of the mean (SEM) across participants. To evaluate the average performance per attempt, for each participant, we first divided the number of tasks solved within one, two and three attempts by the number of all tasks that were solvable for that participant and subsequently computed the mean and the SEM across participants. To evaluate the average performance per optimal step limit, for each participant and optimal step limit, we first divided the number of solved tasks by the number of solvable tasks and then averaged across participants. Finally, to evaluate the average performance per treasure location combination, for each participant and treasure location

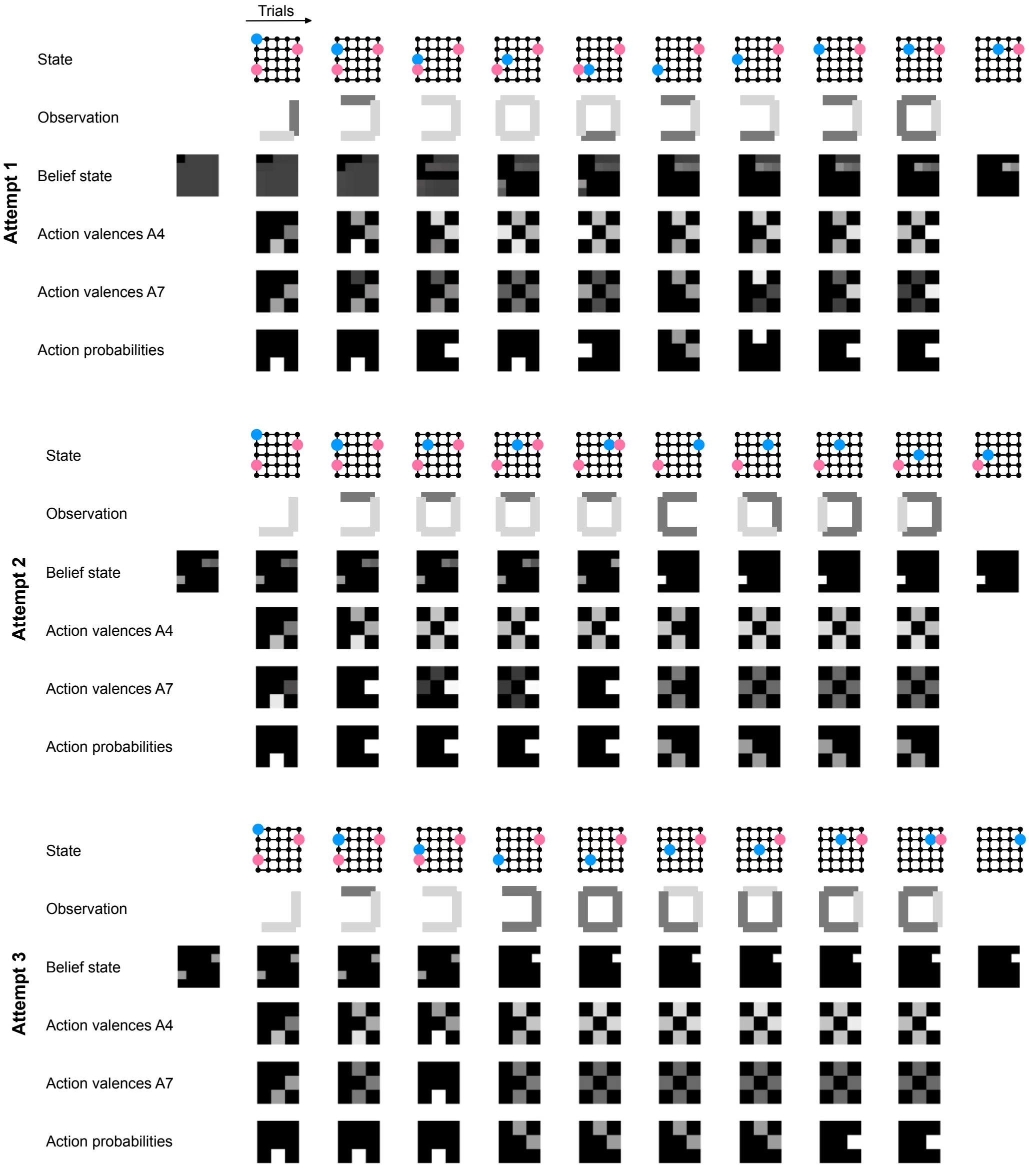

Figure 3.13. Agent A9 interacting with the treasure hunt task. A9 is a belief state-based hybrid explorative-exploitative agent, that, similar to A8, allocates action valences by combining the action valences of a belief state-based exploitative and explorative agent. Specifically, A9 evaluates the convex combination of the valences of A4 and A7 using a state uncertainty-dependent weighting parameter. In this simulation, the agent recovers the second treasure location in the first attempt. It subsequently searches for and recovers the first treasure location in the second attempt, and solves the task in the third attempt. Note: For layout conventions, please see the legend of Figure 3.12. In this simulation, the offset and slope parameters were set to $\eta_0 = 1$ and $\eta_1 = 0.4$.

combination, we first divided the number of solved tasks by the number of solvable tasks and then averaged across participants.

3.3.2 Model-based behavioral data analyses

Behavioral data analysis models To evaluate the agent models based on the experimentally acquired behavioral data, for each agent we first formulated a corresponding behavioral data analysis model that specifies the probability over action a_t given the history of directly observable state components $y_{1:t}$ and observations $o_{1:t}$. For the belief state-free cognitive null agent A1, the behavioral data analysis model is given by mapping the action valences into probabilities according to

$$p\left(a_t = a | y_{1:t}, o_{1:t}\right) = p\left(a_t = a | s_t^1\right) := v(a). \tag{3.43}$$

For all other agents, the behavioral data analysis models are given by nesting the agent-specific action valence functions in a softmax operation according to

$$p^{\beta}\left(a_t = a | y_{1:t}, o_{1:t}\right) := \frac{\exp\left(\beta\left(v\left(a\right)\right)\right)}{\sum_{a_* \in A_{s^1}} \exp\left(\beta\left(v\left(a_*\right)\right)\right)}, \tag{3.44}$$

where the inverse temperature parameter $\beta \in \mathbb{R}_{\geq 0}$ denotes the post-decision noise (Reverdy & Leonard, 2015). The larger the value of β, the higher the probability that the action with the highest valence is realized, i.e., the smaller the post-decision noise. Correspondingly, the smaller the value of β the less the probabilities reflect the action valence differences, i.e., the higher the post-decision noise.

Model simulations We ran two sets of simulations using each agent's behavioral data analysis model. First, to compare participants' performance with the performance of the agents, we presented the models with task configurations identical to those presented to the participants. In particular, in the first set of simulations, the models were presented with tasks that had the same treasure location combinations as the tasks presented to the participants. The attempt limit on each task was determined by the number of attempts the participant used and the step limit on each attempt was identical to the step limit given to the participant. Second, to obtain a global marker of the locally defined agent strategies and to complement the results of the first set of simulations, we presented the models with standard task configurations. In particular, in the second set of simulations, the models were presented with all possible treasure location combinations. On each task, the attempt limit was set to the maximal of three and the step limit corresponded to the optimal step limit.

Consistent with the simulations we conducted to show exemplary behaviors of the agents (cf. Figures 3.5-3.13), in both sets of simulations we set the value of the inverse temperature parameter β to an arbitrary large value ($\beta = 100$) to ensure minimal post-decision noise in the action generation. Similarly, we set the free parameters of the belief state-based hybrid explorative-exploitative agents to the values used in the exemplary simulations ($\lambda = 0.7$, $\eta_0 = 1$ and $\eta_0 = 0.4$). In both sets of simulations, for each task we generated one data set with each model and subsequently computed average model performances. Specifically, in the first set of simulations, for a given set of tasks presented to a single participant we evaluated the number of solved tasks for each model and subsequently computed the mean and SEM across participants. In the second set of simulations, for each model, we divided the number of solved tasks by the number of all tasks to obtain the mean performance and subsequently computed the SEM.

Model evaluation We employed a maximum likelihood (ML) approach to evaluate the agent-specific behavioral data analysis models in light of the participants' data. To this end, for each model and participant, we first evaluated the sum of the summed log probability of all actions on each task over all tasks, conditional on the participant-specific history of the directly observable state components and observations. Formally, let m denote the total number of tasks, n_T denote the number of valid trials (i.e., the number of trials with a valid action) on task $T = 1, ..., m$, and $K = \sum_{T=1}^{m} n_T$ denote the total number of valid trials across the experiment. Further, let y_t^T, o_t^T, a_t^T denote the directly observable state components, observation and action on trial $t = 1, ..., n_T$ of task T. Finally, let $y_{1:K} := \left(y_1^1, y_2^1, ..., y_{n_T}^T\right)$, $o_{1:K} := \left(o_1^1, o_2^1, ..., o_{n_T}^T\right)$ and $a_{1:K} := \left(a_1^1, a_2^1, ..., a_{n_T}^T\right)$ denote the set of directly observable state components, observations, and actions across the experiment. Then, for each model the directly observable state components-observation-action data log likelihood is given by

$$\ln p\left(a_{1:K}|y_{1:K}, o_{1:K}\right) = \sum_{T=1}^{m} \sum_{t=1}^{n_T} \ln p\left(a_t^T|y_{1:t}^T, o_{1:t}^T\right). \tag{3.45}$$

$p\left(a_t^T|y_{1:t}^T, o_{1:t}^T\right)$ denotes the model-specific probability of action a_t^T given the history of directly observable state components $y_{1:t}^T$ and observations $o_{1:t}^T$ on task T; intuitively, it evaluates the probability of the participant action a_t^T under the assumption that a given agent had experienced the identical directly observable state components and observations as the participant did up to trial t on task

T. For agent A1, this probability is evaluated directly. For all other agents, this probability depends on the post-decision noise parameter β, and, for agents A8 and A9, additionally on the weighting parameter λ and on the offset and slope parameters η_0 and η_1, respectively. To estimate these parameters, we maximized a model's log probability $\ln p\,(a_{1:K}|y_{1:K}, o_{1:K})$ as a function of its free parameters using Matlab's constrained nonlinear optimization function *fmincon* (Byrd, Gilbert, & Nocedal, 2000; Byrd, Hribar, & Nocedal, 1999; Waltz, Morales, Nocedal, & Orban, 2006). For β and η_1, the optimization was performed on the interval $[10^{-5}, 10]$ and the initial value was set to 5. For λ and η_0, the optimization was performed on the interval $[10^{-5}, 1]$ and the initial value was set to 0.5.

Model comparison To compare the models' relative plausibilities given the participants' data, we first computed the Bayesian Information Criterion (BIC; Schwarz, 1978) for each agent and participant as

$$\text{BIC} = -2\ln p^{\hat{\theta}}(a_{1:K}|y_{1:K}, o_{1:K}) + j\ln K \tag{3.46}$$

to account for the accuracy-complexity trade-off (Farrell & Lewandowsky, 2018). On the right hand-side of eq. 3.46, the first term denotes the model's maximized log probability multiplied by the factor -2, j denotes the model's number of free parameters and K denotes the number of data points (i.e., the number of valid trials). We then subjected the negative BIC scores of all agents and participants to a random-effects Bayesian model selection as implemented in the *spm_BMS* function in SPM12 (www.fil.ion.ucl.ac.uk/spm/; Rigoux, Stephan, Friston, and Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, and Friston, 2009). We report the ensuing protected model exceedance probabilities, which indicate the group-level probability that the particular model is more likely than any other models of the model space. We additionally evaluated a pseudo-r^2 statistic ρ according to McFadden (1973). This statistic offers a standardized measure of how well a model of an optimizing agent fits the data by quantifying the variance of a participant's actions explained by the respective model compared to a cognitive null model. In our model space, the cognitive null model corresponds to the model of agent A1 and consequently, for each participant we report ρ for all other agents. Formally, ρ is evaluated according to

$$\rho = 1 - \frac{\ln p^{\hat{\theta}}_{\text{Agent}\neq\text{A1}}(a_{1:K}|y_{1:K}, o_{1:K})}{\ln p_{\text{A1}}(a_{1:K}|y_{1:K}, o_{1:K})}. \tag{3.47}$$

Model recovery analysis To validate our agent-based behavioral modeling approach, we conducted a model recovery analysis to test if we can reliably arbitrate between the agent-specific behavioral data analysis models. To this end, we evaluated each model based on the synthetic data sets generated in the second set of simulations by means of the ML and BIC approaches discussed above. Specifically, for each model we evaluated the maximum log likelihood of the synthetic actions given the synthetic directly observable state components and observations generated with a given model on a given task. Then, we computed the BIC score for each model and synthetic data set and subsequently averaged across the synthetic data sets of a given data generating model. This way, we obtained an average BIC score for each model, for data generated with each model.

3.4 Behavioral results

3.4.1 Descriptive behavioral results

On average, the performance was high: Participants solved 11.26 ± 0.67 of 14.95 solvable tasks (Figure 3.14a). There was no difference in the performance between runs (one-way analysis of variance (ANOVA), $F(3, 70) = 0.72, p = 0.55$; Figure 3.14b), suggesting that participants performed at a constant level throughout the experiment. Of all solvable tasks, $59.12\% \pm 3.54$ were solved within one or two attempts and significantly less within three attempts ($16.09\% \pm 2.19$; one-way ANOVA, $F(2, 54) = 10.71, p < 0.001$; Figure 3.14c). This result indicates that participants were able to solve most task using only two attempts and it therefore reassures our decision to limit the number of attempts to three. We additionally evaluated the average performance per optimal step limit and per treasure location combination. Overall, the performance decreased with increasing optimal step limit (Figure 3.14d) and, on average, participants solved more tasks if at least one of the treasures was close to the start position (compare the second row with the last column in Figure 3.14e). These results are consistent with the notion that tasks with larger optimal step limits and with more distant treasure locations were harder. This is because in these tasks the solution required more steps and the initial accuracy of the treasure detector was lower.

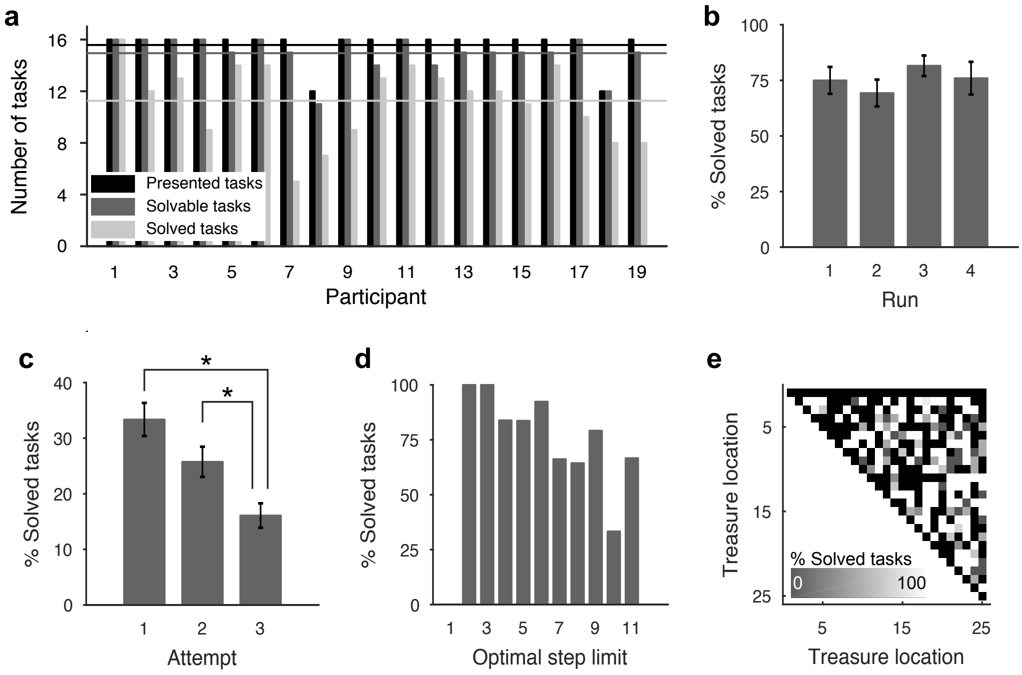

Figure 3.14. Descriptive behavioral results. **a** Number of presented, solvable and solved tasks for each participant. Horizontal lines show the participant-averaged values. The performance was high with approximately 11 solved tasks on average. **b** Average performance per run. The performance was stable throughout all runs. **c** Average performance per attempt. The majority of the tasks were solved in the first or the second attempt. **d** Average performance per optimal step limit. The performance showed a decreasing trend with increasing optimal step limit. **e** Average performance per treasure location combination. The performance was higher if at least one of the treasures was located at a cell close to the initial grid position. Black color marks those treasure location combinations that were not presented to the participants (cf. Figure 3.2a). Note: The error bars in **b** and **c** denote the SEM and $*$ in **c** denotes $p < 0.001$.

3.4.2 Model-based behavioral results

Model recovery results We validated our agent-based behavioral modeling approach by conducting a model recovery analysis of the agent-specific behavioral data analysis models. As shown in Figure 3.15, for data generated with a given model, the BIC score was minimal for the same recovering model except for the belief state-based hybrid explorative-exploitative agent A9. This indicates that all models except the model of A9 are reliably identifiable. For data generated with the model of A9, the BIC score was minimal for the model of the belief state-based exploitative agent A4. This is possibly due to the fact that with the offset and slope parameter values used in the simulations ($\eta_0 = 1$ and $\eta_0 = 0.4$), the behavior of agent A9 mimics the behavior of agent A4. As the model of agent A4 has two parameters less than that of A9, in the evaluation of the BIC score the penalization term is smaller for A4, which in turn leads to a smaller BIC score for this agent. Although the model of A9 is not recoverable for data generated with the aforementioned parameter values as evaluated based on the BIC scores, it is noteworthy that even with these parameter values there are certain characteristics of the behavior of A9 that differ from the behavior of A4, as discussed above in Paragraph Simulated task behavior of agent A9 (see also Figure 3.13).

Model simulation results To obtain global markers of the locally defined agent strategies and compare the agents' behavior to that of the participants, we evaluated synthetic performances based on data simulated with the behavioral data analysis model of each agent. The results from both sets of simulations show a very similar pattern (Figure 3.16a): The lowest performance was exhibited by agent A1, followed by a markedly higher performance of agents A7 and A2. On average, the latter two agents successfully solved 5-7 tasks with participants configurations and 49-59% of the tasks with standard configurations. Arguably, these values already suggest a decent performance given the complex nature of the treasure hunt task. Agents A3, A5, A8 and A9 solved approximately 11 tasks with participants configurations on average, which is comparable to the participants' performance (cf. Figure 3.14a). These agents solved 88-96% of the tasks with standard configurations. The best performing agents were A4 and A6 on both sets of simulations, with A6 solving almost all tasks (99.6%) with standard configurations.

The results of the simulations and their implications can be summarized as follows. First, a high performance as exhibited by the participants was only achieved by agents A3, A4, A5, A6, A8 and A9. This suggests that they are plausible models of the participants' decision-making behavior in the treasure hunt task, lending high face validity to the belief state-based exploitative or hybrid explorative-exploitative planning strategies. Second, the finding that these agents outperformed the belief state-free cognitive null agent A1 and the belief state-based explorative agent A7 is not surprising, because neither A1 nor A7 try to collect the treasures and thereby solve the task. However, their outperforming the belief state-free exploitative agent A2 is more surprising. While A2 does not maintain a belief state, it tries to collect the treasures by relying on the available signal of the treasure detector. As a light grey bar necessarily indicates a direction towards the treasures, this strategy could have turned out to be good enough to succeed on most tasks. Yet, as our finding highlights, a performance comparable to that of the participants can not be achieved merely based on instantaneously available information. Third, we found that of the belief state-based agents, the purely exploitative strategy of A6 prevailed. This agent exhibited a nearly perfect performance on the set of simulations with standard configurations. This suggests that optimal behavior in the treasure hunt task - by means of task solution given the standard configurations - requires belief state-based exploitative planning from the outset. It is notable that agent A6 implements the most advanced belief state-based exploitative planning strategy by dynamically switching between the strategies of agents A4 and A5.

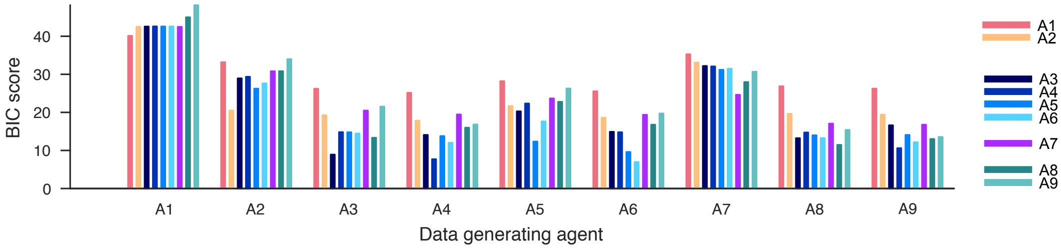

Figure 3.15. Model recovery results. Y-axis shows the average BIC score of each agent-specific behavioral data analysis model evaluated on the synthetic data sets of each data generating model shown on the x-axis. For data generated with a given model, the BIC score was minimal for the corresponding model indicating that the models are reliably recoverable. The only exception is the model of agent A9. For data generated with the model of agent A9, the BIC score was minimal for the model of its constituent agent A4.

Model comparison results We evaluated and compared the agent-specific behavioral data analysis models based on the participants' data. As shown in Figure 3.16b, for 15 of the 19 participants the BIC score was minimal under the model of agent A5. Accordingly, the group cumulative BIC score was minimal for this agent showing that among the set of models assessed the belief state-based exploitative planning strategy of A5 explained participants' behavior the best (left panel of Figure 3.16c). This conclusion is further supported by the result of the random-effects Bayesian model selection: The group-level protected exceedance probability of the model of A5 was larger than 0.99 indicating that within the group of participants the most frequently applied strategy resembled that of agent A5 (right panel of Figure 3.16c). On average, the winning A5 model explained 35.65% ± 2.06 of the participants actions as evaluated by the pseudo-r^2 statistic. Given the complexity of the treasure hunt task, this ρ value suggests a considerably good model fit.

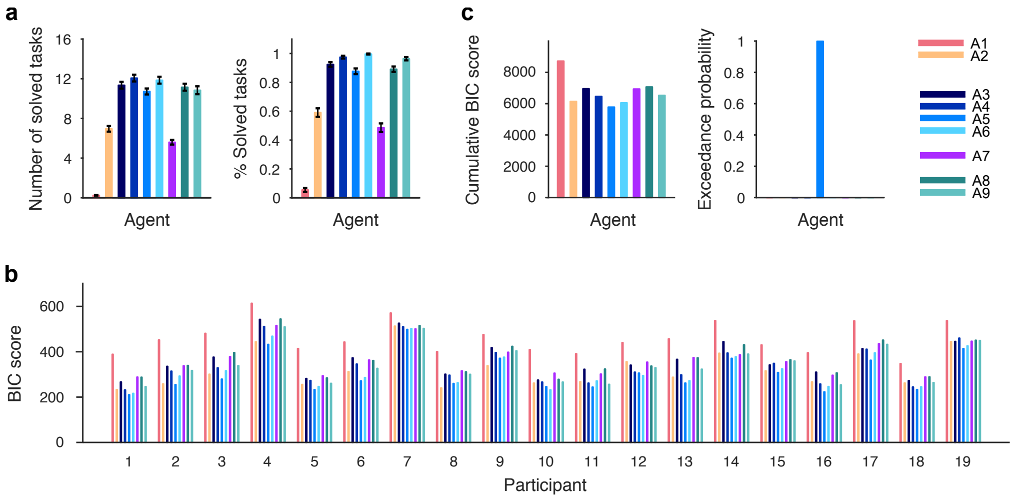

Figure 3.16. Model simulation and comparison results. **a** Synthetic agent performances evaluated based on the set of simulations with participants (left panel) and standard (right panel) configurations. The results from both analyses are similar: The highest performance was achieved by the belief state-based exploitative and hybrid exploitative-explorative agents, with a level comparable to that of the participants (cf. Figure 3.14a). **b** Participant-level BIC scores. For the majority of the participants (15 of 19), the BIC score was minimal for the model of agent A5. **c** Group-level cumulative BIC scores (left panel) and protected exceedance probabilities (right panel). The cumulative BIC score was minimal and the protected exceedance probability was maximal for the model of A5, providing evidence in favor of the belief state-based exploitative planning strategy of this agent.

3.5 FMRI methods

3.5.1 FMRI data acquisition and preprocessing

Functional imaging was performed on a 3T Siemens Magnetom Tim Trio MRI scanner (Siemens, Erlangen, Germany) with a 12-channel head coil. During task completion, 36 interleaved axial slices (flip angle: 80°, slice thickness: 3 mm, voxel size: $3{\times}3{\times}3$ mm^3, distance factor: 20%) of T2*-weighted echo-planar images (EPIs) (field of view: 216 mm) were acquired each 2000 ms. On each run, the data acquisition was terminated manually when the participant completed the tasks. Therefore, the number of volumes varied across runs, with a maximum of 600 volumes per run. To allow for the signal to saturate, the first 3 images of each run were discarded. Before the first run, T1-weighted anatomical images (voxel size: $1{\times}1{\times}1$ mm^3, field of view: 256 mm) were additionally acquired. The fMRI data were preprocessed and analyzed using SPM12. To prepare the participants' EPIs for analysis, each data set was motion-corrected by realigning the images to the first scan of the first run, normalized to the Montreal Neurological Institute (MNI) EPI reference template, and re-sampled to 2 mm isotropic voxel size. The images were subsequently smoothed using an

8 mm full-width half-maximum isotropic Gaussian kernel.

3.5.2 Model-based fMRI data analysis

Participant-level model formulation We analyzed the preprocessed fMRI data using a model-based GLM approach (Friston & Dolan, 2010). Specifically, our primary aim was to identify the functional anatomy of the putative cognitive processes underlying participants' behavior, as formally captured by the group-favored model of agent A5. To this end, we formulated the first-level design matrix with the following regressors: The first regressor served as a basis regressor and modeled the valid trials in a boxcar fashion with onsets corresponding to the time of the grid position presentation and with participant response time-dependent durations. To account for the maintenance of a belief state, the second regressor constituted a parametric modulation of the first regressor by the trial-by-trial Bayesian surprise (Itti & Baldi, 2009). The Bayesian surprise on a given trial corresponds to the KL divergence between the task- and trial-specific prior and posterior belief states (μ_t^T and μ_{t+1}^T, respectively) of a participant, i.e.,

$$KL\left(\mu_{t+1}^T || \mu_t^T\right) = \sum\nolimits_{i=1}^{d^2} {\mu_{t+1}^T}^{(i)} \cdot \ln\left(\frac{{\mu_{t+1}^T}^{(i)}}{\mu_t^{T(i)}}\right). \tag{3.48}$$

As the evaluation of the belief state is not dependent on any free parameters, the Bayesian surprise on trial t of task T depends only on the participant-specific history of the directly observable state components $y_{1:t}^T$ and observations $o_{1:t}^T$. For a consistent scaling across tasks, the Bayesian surprise was task-wise normalized by projecting the trial-by-trial values onto the interval $[0, 1]$ according to

$$KL\left(\mu_{t+1}^T || \mu_t^T\right)' = \frac{KL\left(\mu_{t+1}^T || \mu_t^T\right) - \min\limits_{t_* \in \mathbb{N}_{n_T}} KL\left(\mu_{t_*+1}^T || \mu_{t_*}^T\right)}{\max\limits_{t_* \in \mathbb{N}_{n_T}} KL\left(\mu_{t_*+1}^T || \mu_{t_*}^T\right) - \min\limits_{t_* \in \mathbb{N}_{n_T}} KL\left(\mu_{t_*+1}^T || \mu_{t_*}^T\right)}. \tag{3.49}$$

Here, $KL\left(\mu_{t+1}^T || \mu_t^T\right)'$ denotes the task-wise re-scaled Bayesian surprise and n_T denotes the number of valid trials on task T. The third regressor constituted a parametric modulation of the first regressor by the trial-by-trial chosen action valences as evaluated by agent A5. We included this regressor to account for action evaluation reflecting exploitative planning on the basis of belief states by means of the heuristic real-time dynamic programming approach of the winning agent. In particular, the chosen action valence on trial t of task T is given by the negative sum of the belief-weighted l_1 distances between the new position resulting from the participant's action a_t^T and each node of the grid, i.e.,

$$v_{\mathrm{A5}}(a_t^T) = -\sum\nolimits_{i=1}^{d^2} {\mu_{t+1}^T}^{(i)} \cdot l_1 \left({s_t^T}^{(1)} + a_t^T, i \right) \tag{3.50}$$

(cf. eq. 3.34). As in the case of the Bayesian surprise, the chosen action valence does not depend on any free parameters and for action a_t^T is evaluated directly based on the participant-specific history of the directly observable state components $y_{1:t}^T$ and observations $o_{1:t}^T$. Before combining the first regressor with the chosen action valences, they were also task-wise normalized on the interval $[0, 1]$ as

$$v_{\mathrm{A5}}(a_t^T)' = \frac{v_{\mathrm{A5}}(a_t^T) - \min\limits_{t_* \in \mathbb{N}_{n_T}} v_{\mathrm{A5}}(a_{t_*}^T)}{\max\limits_{t_* \in \mathbb{N}_{n_T}} v_{\mathrm{A5}}(a_{t_*}^T) - \min\limits_{t_* \in \mathbb{N}_{n_T}} v_{\mathrm{A5}}(a_{t_*}^T)} \tag{3.51}$$

for a consistent scaling across tasks. In addition to the model-based parametric regressors, as the fourth regressor we included a parametric modulation of the first regressor by the trial-by-trial average luminance to account for the confounding effect of varying light intensity across trials. To obtain trial-by-trial average luminance values, we first calculated the luminance of a three-dimensional RGB image $I \in \mathbb{R}^{J \times K \times 3}$ according to

$$l(I) = \frac{1}{JK} \sum\nolimits_{j=1}^{J} \sum\nolimits_{k=1}^{K} 0.299 I^{(jk1)} + 0.587 I^{(jk2)} + 0.114 I^{(jk3)} \tag{3.52}$$

and subsequently calculated the average luminance across the images presented to a participant on trial t of task T (Parekh, 2006). Finally, we also included three regressors modeling additional task events with 0-duration pulses at the event onsets. Specifically, the fifth regressor modeled the treasure discovery (i.e., grid positions with a treasure), the sixth regressor modeled the information display presented after each attempt and task, and the seventh regressor modeled the fixation crosses presented after every fourth attempt. Of note, the parametric regressors were only weakly correlated (the run- and subject-averaged correlation coefficient was 0.33 ± 0.02 between Bayesian surprise and chosen action valence, 0.25 ± 0.01 between Bayesian surprise and luminance, and 0.16 ± 0.03 between chosen action valence and luminance) and all seven regressors, referred to as regressors of interest, were entered in the participant-level design matrix without serial orthogonalization.

The regressors of interest were convolved with the canonical hemodynamic response function (HRF) and, as per SPM default, low frequency components were removed using a high-pass filter with a 128s cutoff. The residual error correlations were accounted for by SPM's standard AR(1) model (Friston, Glaser, et al., 2002). In addition, a constant run offset (as per SPM default) and six spatial realignment parameters estimated during preprocessing were entered in the participant-level design matrix as nuisance regressors of no interest. For an exemplary time course of the regressors of interest and the participant-level design matrix of a complete data set with four runs, please see Figure 3.17a and b, respectively.

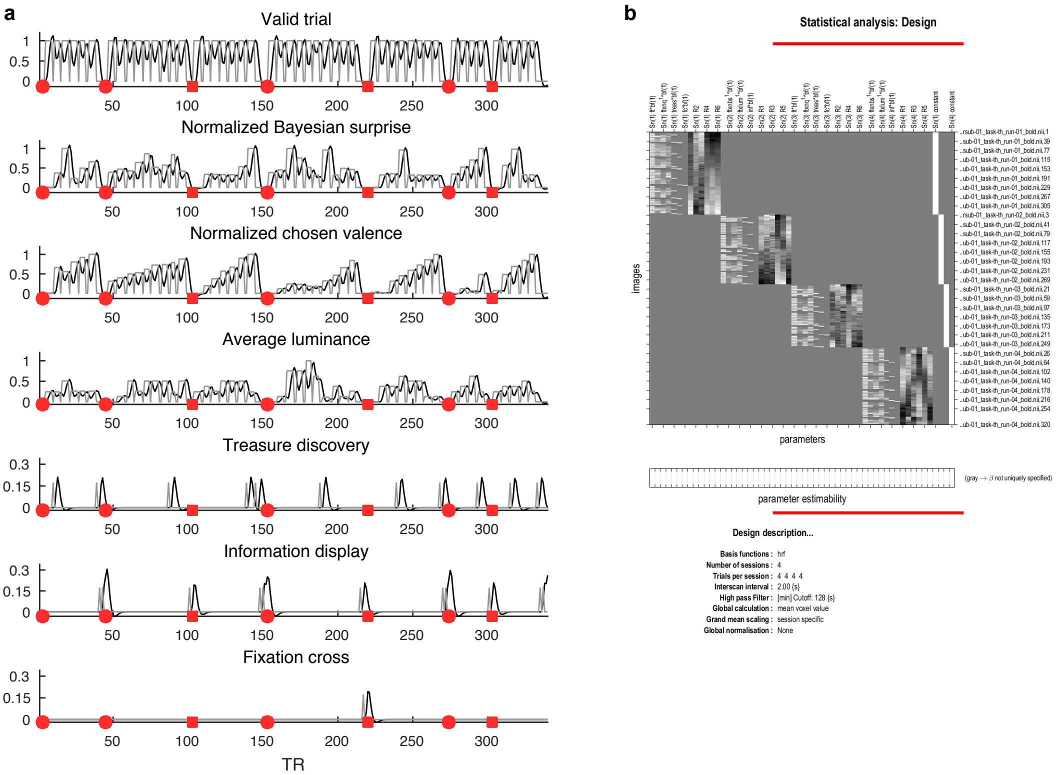

Figure 3.17. Participant-level GLM design matrix. a. Exemplary time course of the regressors of interest over a single run. From top to bottom: Valid trials modeled with a boxcar function with onsets corresponding to the onset of the grid position and offsets corresponding to the participant's button press. Normalized Bayesian surprise quantifying the trial-by-trial shift in the belief state. Normalized chosen action valence reflecting the trial-by-trial action evaluation process as assumed by the group-favored agent A5. Average luminance quantifying the light intensity of the images presented on a trial. Treasure discovery modeled with 0-duration pulses at the onsets of grid positions with a treasure. Information display modeled with 0-duration pulses at the respective stimulus onsets (presented after each attempt and task). Fixation cross modeled with 0-duration pulses at the respective stimulus onsets (presented after every fourth attempt). Note: Black lines depict the HRF-convolved regressors, grey lines depict the unconvolved regressors, orange dots denote the task onsets (or, equivalently, the onsets of the first task attempts), and small orange squares denote the attempt onsets. **b** Exemplary participant-level design matrix of a complete data set with four runs as output by SPM12. In addition to the regressors of interest, the participant-level design matrix comprised constant run offsets and motion parameters as regressors of no interest.

3.5.3 Participant- and group-level model estimation and evaluation

We estimated the parameters of the participant-level GLM using SPM's restricted maximum likelihood scheme (Friston, Penny, et al., 2002). We subsequently evaluated contrasted parameter estimates, first for the positive main effects of valid trials, luminance, treasure discovery and information display to validate the fMRI data quality, and then for the positive and negative main effects of the Bayesian surprise and chosen action valence to identify brain regions encoding model-based quantities. For example, we used the contrast vector $c = (0, 1, 0, 0, 0, 0, 0)^T$ to evaluate the positive main effect of Bayesian surprise. The resulting eight contrast images were entered for group-level voxel-wise one-sample t-tests. We applied a cluster-forming threshold of $p < 0.001$ (uncorrected) to the t-statistic maps and report the clusters with a family-wise error (FWE) corrected p-value smaller than 0.05. Anatomical labels were obtained using the third version of the Automated Anatomical Atlas (AAL3; Rolls, Huang, Lin, Feng, and Joliot, 2020).

3.6 FMRI results

3.6.1 FMRI data validation

Before interrogating the fMRI data for model-based effects, we conducted a set of analyses to investigate the neural underpinnings of cognitive processes involved in the treasure hunt task, which are not specifically assumed by the group-favored agent model A5. In particular, we evaluated the positive main effects of valid trials, luminance, treasure discovery, and information display. We reasoned that these regressors would capture variability in the fMRI data primarily related to movement planning and execution, visual processing, reward processing, and reading, respectively. Importantly, the goal of these analyses was to test if we can detect activity clusters in brain regions commonly implicated in the respective putative cognitive processes and to thereby validate the quality of our fMRI data. Therefore, we do not present these results in detail but show that they largely corroborate the validity of the fMRI data quality.

As shown in Figure 3.18, testing for the positive main effects of valid trials revealed a large cluster extending through mainly left-hemispheric areas of the pre- and postcentral gyri, supplementary motor area, middle cingulate gyrus, inferior parietal cortex, insula, rolandic operculum, putamen, pallidum and thalamus, consistent with participants using their right hand for button press (e.g., Ebbesen and Brecht, 2017; Gazzaniga, Ivry, and Mangun, 2013). For a list of all clusters, please refer to Supplementary Table B.1. As expected, activation in the bilateral visual cortex comprising parts of the superior and middle occipital gyri, calcarine sulcus and surrounding cortex, cuneus, as well as the lingual and fusiform gyri scaled positively with the trial-by-trial stimuli luminance, among other regions (eg., Gazzaniga et al., 2013; Grill-Spector and Malach, 2004; Supplementary Table B.2). For the positive main effects of treasure discovery and information display, we obtained single activity clusters spanning the majority of the brain. Specifically, both treasure discovery and information display engaged a similar extended bilateral network of cortical and subcortical areas, indicating the involvement of additional and partly shared cognitive processes at these events, over and above those hypothesized. For example, both analyses revealed increased activity in the hippocampus and the parahippocampus, which are typically not associated with reward processing or reading but have established roles in memory formation and recall (eg., Eichenbaum, 2017; Gazzaniga et al., 2013). Nevertheless, in keeping with our expectations, for the positive main effects of treasure discovery, the obtained large activity cluster also comprised classic reward regions, such as the bilateral dorsal and ventral striatum, insula, ventromedial prefrontal cortex, and ventral tagmental area (Bartra, McGuire, and Kable, 2013; Morales and Margolis, 2017; Supplementary Table B.3). While the activity pattern was similar for the positive main effects of information display, the obtained large activity cluster did not comprise the ventral striatum and the ventral tagmental area. It however comprised all regions commonly implicated in reading, including the bilateral lateral frontal cortex, supplementary motor areas as well as the temporal and occipital lobes, with slightly more extended activity in the left hemisphere (Martin, Schurz, Kronbichler, and Richlan, 2015; Supplementary Table B.4).

3.6.2 Model-based fMRI results

After validating the fMRI data quality, we commenced with investigating the neural substrates of the trial-by-trial cognitive processes belief state maintenance and action evaluation as assumed by the behaviorally most plausible agent model A5. To this end, we evaluated the main effects of the model-based

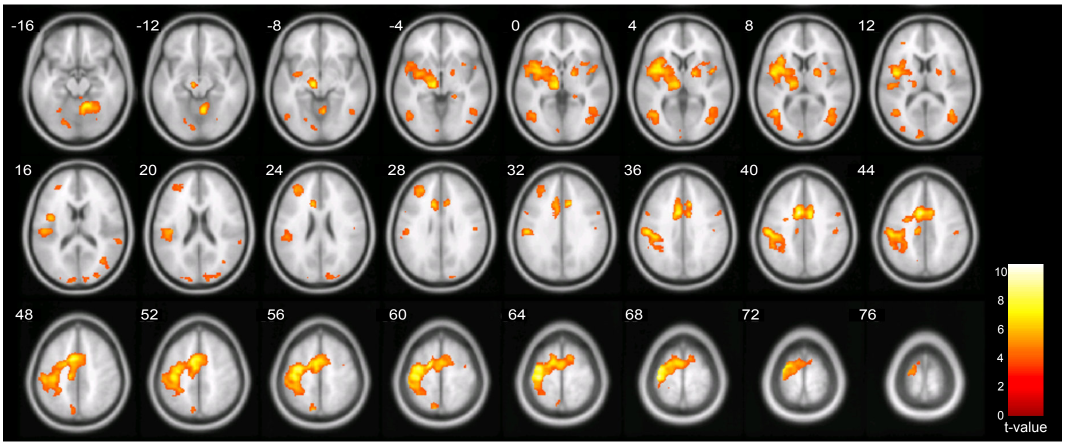

Figure 3.18. Positive main effects of valid trials. Group-level voxel-wise one-sample t-tests revealed increased activation in areas involved in motor planning and execution during valid trials. These include the left pre- and postcentral gyri, inferior parietal cortex, thalamus as well as the bilateral supplementary motor areas, middle cingulate gyri, insula, rolandic operculum, putamen, and pallidum, among other regions. Note: The t-value map is thresholded at a cluster-defining threshold corresponding to $p < 0.001$ (uncorrected) and overlaid on the SPM average T1 image. For each slice, the value in the upper left corner denotes the corresponding MNI z-coordinate. The results are visualized with the xjView toolbox available at https://www.alivelearn.net/xjview.

parametric regressors Bayesian surprise and chosen action valence.

Bayesian surprise We first examined the neural signals encoding the trial-by-trial Bayesian surprise quantifying belief state maintenance in terms of trial-by-trial belief state update. Drawing on previous work, we hypothesized that if participants maintain a belief state in the treasure hunt task, a distributed network of cortical and subcortical areas including the frontal and posterior parietal cortices, as well as the dorsal striatum would be engaged (Fischer et al., 2017; O'Reilly et al., 2013). Consistent with this hypothesis, we found increased activation in response to the Bayesian surprise in the bilateral inferior, middle, superior and medial frontal gyri. We also detected increased activation in more posterior cortical areas including the bilateral inferior and superior parietal lobules, precuneus and lingual gyri, and, in addition, in the right dorsal striatum and the bilateral cerebellum. The positive main effects of Bayesian surprise are visualized in Figure 3.19 and summarized in Supplementary Table B.5. Testing for the negative main effects of Bayesian surprise revealed a small cluster in the left precuneus and calcarine sulcus as listed in Supplementary Table B.6.

Chosen action valence We next sought for neural signals representing the trial-by-trial chosen action valences as evaluated by the belief state-based

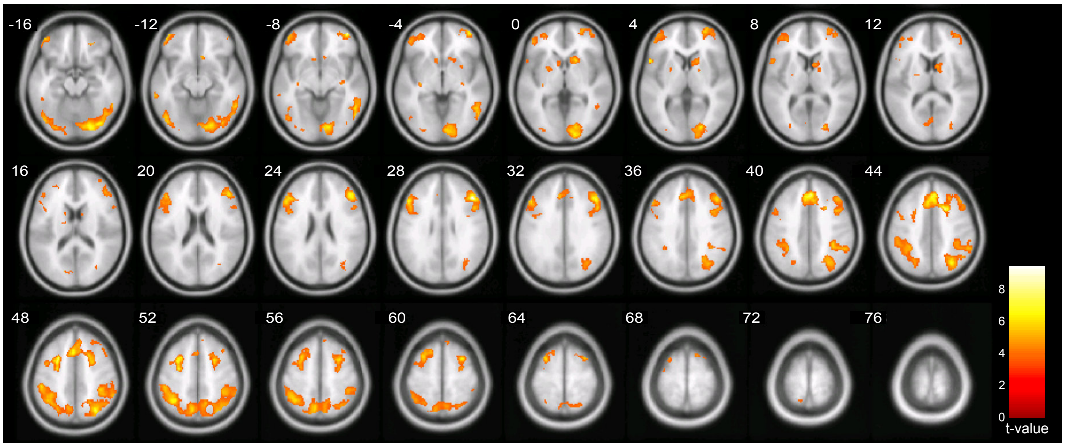

Figure 3.19. Positive main effects of Bayesian surprise. We detected clusters positively encoding the Bayesian surprise in the bilateral prefrontal cortices (inferior, middle, superior and superior medial frontal gyri), the bilateral inferior parietal cortices dorsally extending into the superior parietal gyri and medially extending into the precuneus, the bilateral lingual gyri, the right dorsal striatum and the bilateral cerebellum. Note: For the applied visualization conventions, please see the legend of Figure 3.18.

exploitative agent A5 adopting a heuristic real-time dynamic programming approach. Although the functional anatomy of belief state-based exploitative planning is understudied, fMRI studies on planning (Korn & Bach, 2018; Simon & Daw, 2011) and on exploitation (Chakroun et al., 2020; Daw et al., 2006) have implicated the orbitofrontal and ventromedial prefrontal cortices in encoding the subjective desirability of the chosen actions. As shown in Figure 3.20 and listed in Supplementary Table B.7, we observed a similar activation pattern. Specifically, activity in the right insula, posterior orbitofrontal gyrus and inferior frontal gyrus, as well as in the bilateral anterior cingulate cortices reaching into the medial frontal gyri scaled positively with the valences of the chosen actions. We additionally obtained positively activated clusters in the bilateral middle and superior temporal gyri (larger in the right hemisphere), and, more posteriorly and ventrally, in the bilateral inferior occipital gyri, also comprising parts of the fusiform and lingual gyri and the cerebellum. Seeking for areas negatively encoding the chosen action valence returned no activation clusters.

3.7 Discussion

In real-life, multistep tasks are often complicated by the states of the environment not being fully observable. Despite their relevance, how humans make sequential decisions in such tasks remains a largely open question. In the current work we addressed this issue and assessed if in multistep tasks with partially

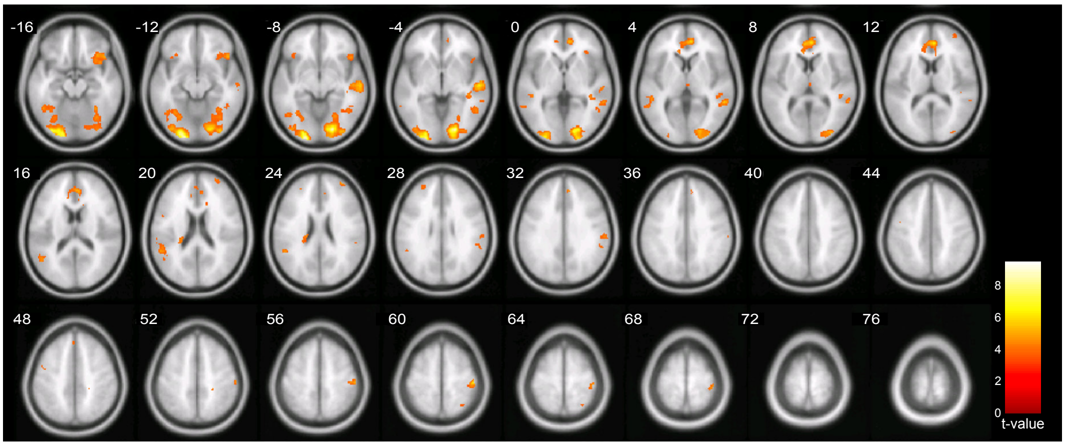

Figure 3.20. Positive main effects of chosen action valence. We observed clusters with increasing activation in response to increasing chosen action valences in the right insula extending into the orbitofrontal and inferior frontal cortices, the bilateral anterior cingulate cortices extending towards the medial frontal gyri, the bilateral middle and superior temporal gyri, and the bilateral inferior occipital gyri extending into the fusiform and lingual gyri. Note: For the applied visualization conventions, please see the legend of Figure 3.18.

observable states humans (1) rely on belief state-based planning strategies or use computationally simpler belief state-free strategies, and whether they (2) adopt an exploitative or a combined explorative-exploitative objective. We furthermore investigated (3) the network of brain regions supporting sequential decision making in these tasks. To this end, we developed a multistep task with partially observable states framed as a spatial search task and analyzed behavioral and fMRI data acquired from human participants using a combination of agent-based computational behavioral modeling and model-based GLM for fMRI. By doing so, we provide evidence for a belief state-based exploitative planning strategy engaging a distributed network of cortical and subcortical brain regions. In the following, the task design, computational modeling framework, and findings are discussed in turn.

Task design

As mentioned in Section 3.1, the defining characteristic of multistep tasks is that actions of the deciding agent affect next states and thereby future rewards (e.g., Bertsekas, 2000; Puterman, 2014; Sutton and Barto, 2018). Beside the fact that states can be fully or partially observable, the task environment can have various configurations. For example, the state transition and reward structures specifying how state-action pairs lead to next states and rewards, respectively, can be deterministic or probabilistic, and reward emissions can be restricted to terminal state-action pairs. Furthermore, the step horizon can, in principle, range from two to infinity. In the treasure hunt task, to focus on partially observable states, we opted for simple deterministic state transition and reward structures, with reward emissions restricted to terminal state-action pairs: participants' position changed according to their directional decision and they were only rewarded if they found both treasures. The step horizon depended on the combination of the treasure locations, with most combinations requiring participants to make decisions over six steps.

Tasks employed in two recent studies investigating model-free decision making versus planning in multistep tasks with fully observable states share key features with the treasure hunt task: The reward structures in the tasks both by Korn and Bach (2018) and Simon and Daw (2011) are also deterministic, and reward emissions in the task by Korn and Bach (2018) are restricted to terminal state-action pairs as in our task. However, different from our task, state-transition structures of both tasks are probabilistic. With respect to the step horizon, participants interact with the task by Korn and Bach (2018) over five steps, whereas they interact with the task by Simon and Daw (2011) over a total of one thousand steps. Although the step horizon of the task by Korn and Bach (2018) is more similar to ours, the spatial embedding of our task resembles the task employed by Simon and Daw (2011), while the task by Korn and Bach (2018) is framed as foraging.

Our task is also closely related to the spatial search task introduced by Yoshida and Ishii (2006) to test the Bayesian brain hypothesis in the context of sequential decision making under state uncertainty. Specifically, similar to the treasure hunt task, in the task by Yoshida and Ishii (2006) participants are instructed to navigate to a goal location in a grid-world by making decisions over a minimum of ten steps under state uncertainty. Moreover, Yoshida and Ishii (2006) also used deterministic state transition and reward structures with rewards emitted only for terminal state-action pairs. However, while their task constitutes an example of a multistep task with partially observable states similar to the treasure hunt task, the focus of the work by Yoshida and Ishii (2006) was not to examine various sequential decision-making strategies in such tasks but to investigate the neurocomputational mechanisms of state estimation. To this end, they induced state uncertainty by providing participants with ambiguous observations about their current position. This differs from our task, where state uncertainty results from ambiguous observations about the goal locations delivered by the treasure detector, which always shows a dark grey bar in the directions leading away from the treasures and either a light or a dark grey bar in the directions leading to the treasures. Crucially, introducing state uncertainty this way allowed us to contrast various belief state-based planning strategies in the treasure hunt task with belief state-free strategies, such as the simple but considerably effective strategy of following the light grey bar as further discussed below.

Computational modeling framework

Behavioral modeling To computationally characterize sequential decision making in the treasure hunt task on a behavioral level, we designed a set of agent-based behavioral models that can interact with the task (cf. Russell and Norvig (2010)). In our model set, the belief state-free agents A1 and A2 simply rely on immediately available information to evaluate the actions. Specifically, by merely acknowledging the available actions without differentiating between them, agent A1 corresponds to a generic cognitive null model. By contrast, agent A2 adopts a relatively refined, yet computationally inexpensive strategy: This agent takes advantage of the fact that the light grey bars necessarily indicate directions towards the treasures and always follows these. Model-free strategies that are specific to the task at hand, such as the strategy of A2, are often contrasted with planning strategies. For example, the aforementioned work by Korn and Bach (2018) applied a comparable approach, whereas Simon and Daw (2011) formally captured model-free decision making with a temporal-difference learning algorithm (e.g., Daw et al. (2005) and Sutton and Barto (2018)). In their standard form, temporal-difference learning algorithms evaluate the actions in a given state based on the associated reward history. Previous theoretical (Rao, 2010) and rodent work (Babayan, Uchida, & Gershman, 2018; Starkweather, Babayan, Uchida, & Gershman, 2017) has extended this framework by substituting states with belief states to accommodate decision making under state uncertainty. However, our task design is not suited to test decision making as accounted for by belief state-based temporal difference learning. This is because even if the agent's memory capacity allowed for storing a value estimate for each belief state-action pair, in order for such a strategy to be effective, the agent would need to re-encounter a belief state-action pair that has previously been associated with reward. Since the treasure locations change upon reward emission, this is unlikely to happen beyond the first few steps in a new task.

The belief state-based agents of our model space perform recursive Bayesian updating to infer the latent components of the state, i.e., the treasure locations.

All belief state-based agents plan ahead on the basis of their thereby ensuing belief states. However, they either pursue a purely exploitative, a purely explorative, or a hybrid explorative-exploitative objective. More specifically, to try to collect the treasures, the belief state-based exploitative agents A3-A6 harness their accumulated knowledge about the treasure locations and evaluate the actions using heuristic real-time dynamic programming (Geffner & Bonet, 1998; Korf, 1990). The belief state-based explorative agent A7 tries to resolve its uncertainty about the treasure locations and for each action quantifies the one-step look-ahead information gain as the expected Bayesian surprise (Itti & Baldi, 2009). The belief state-based explorative-exploitative agents A8-A9 try to resolve their uncertainty about the treasure locations and then collect these. To this end, they combine the heuristic action values of agents A3-A4 and the expected Bayesian surprise of agent A7. The algorithmic architectures of the belief state-based agents relate to earlier work central to our study as follows:

First, our approach to formalize belief state-based exploitative planning in the treasure hunt task is conceptually linked to the partial tree search algorithm used by Simon and Daw (2011) (Sutton & Barto, 2018). Although their algorithm operates on the basis of states and not belief states given the lack of state uncertainty in their task, it also accounts for planning by evaluating the reward-related consequences of the available actions in an approximate fashion. In contrast, Korn and Bach (2018) modelled planning by means of optimal action value computations. Even more similar to our formalization of belief state-based exploitative planning than the partial tree search algorithm of Simon and Daw (2011) is the model applied by Yoshida and Ishii (2006). As alluded to above, these authors did not compare the behavioral plausibility of various sequential decision-making strategies in terms of the dichotomies belief state-free versus belief state-based, and exploration versus exploitation. However, to uncover the neurocomputational mechanisms of state estimation in their spatial search task, they combined Bayesian state inference with a form of action evaluation, which, in essence, corresponds to the heuristic real-time dynamic programming algorithm employed by our belief state-based exploitative agents. Notably, the model of Yoshida and Ishii (2006) assumes action evaluation on the basis of unitary state estimates, which is consistent with early theoretical work (Ishii, Yoshida, & Yoshimoto, 2002). Newer theories, on the other hand, suggest action evaluation on the basis of the entire belief state (Rao, 2010). Therefore, to further refine belief state-based exploitative planning, agents A3-A4 and A5 implement heuristic real-time dynamic programming with unitary state estimates or with the entire belief state, respectively, while agent A6 implements heuristic real-time dynamic programming both with unitary state estimates and the entire belief state.

Second, our approach to cast belief state-based exploration as expected Bayesian surprise maximization rests on Bayesian accounts of exploration (Sun et al., 2011). By evaluating actions on the basis of the expected shifts in the belief state, our approach more broadly links to other (Bayesian or frequentist) methods quantifying the information gain associated with an action (Schwartenbeck et al., 2019).

Third, as mentioned in Section 3.1, whether humans adopt a purely exploitative or an explorative-exploitative objective has been primarily investigated in bandit tasks, i.e. sequential decision-making tasks where the reward structure is latent and the actions only affect immediate rewards (e.g., Berry and Fristedt (1985), Robbins (1952), and Sutton and Barto (2018)). To account for exploration-exploitation choice strategies in bandit tasks, quantities that capture information gain with respect to the reward structure are usually combined with action value estimates computed based on the current knowledge about the reward structure (e.g., Daw et al. (2006), Gershman (2018), and Wilson et al. (2014); see also Chapter 2). In our approach to model belief state-based explorative-exploitative planning in the treasure hunt task we followed a similar logic: Agents A8 and A9 combine the expected Bayesian surprise evaluated by agent A7 with the heuristic action values evaluated by agents A3 and A4, respectively. In these combinations we opted for the heuristic action values of agents A3 and A4 as opposed to agents A5 or A6 because, by relying on unitary state estimates, they arguably represent the simplest and strongest contrasts to belief state-based exploration.

fMRI modeling Given our behavioral model set, we found that the most plausible explanation of participants' actions in the treasure hunt task is provided by the belief state-based exploitative agent A5. To identify the underlying network of brain regions, we applied a model-based GLM approach (Friston & Dolan, 2010). Specifically, we evaluated the neural correlates of two key latent quantities derived from agent A5 on a trial-by-trial basis: the Bayesian surprise and the chosen action valence.

The Bayesian surprise offers a readout of the maintenance of the belief state by quantifying the extent to which the belief state changes following a new action and observation (Itti & Baldi, 2009). Of note, the Bayesian surprise as entered for fMRI data analysis is computed with respect to the current trial, which is different from our formalization of belief state-based exploration, where the expected Bayesian surprise is computed with respect to the next trial. In the context of the Bayesian brain hypothesis, Bayesian surprise has been extensively applied to model neural signals reflecting belief state maintenance; we therefore opted for this quantity in our fMRI modeling approach (e.g., Fischer et al. (2017), Gijsen, Grundei, Lange, Ostwald, and Blankenburg (2020), O'Reilly et al. (2013), Ostwald et al. (2012), and Schwartenbeck et al. (2016)). However, other closely related information theoretic measures have also been employed. For example, in their effort to map the functional anatomy of hidden state estimation, Yoshida and Ishii (2006) used the belief state entropy, which, akin to the Bayesian surprise, also declines with decreasing uncertainty.

The second key latent quantity whose neural correlates we analyzed was the chosen action valence. It corresponds to the desirability of participants' actions as evaluated by the heuristic real-time dynamic programming algorithm of the behaviorally most plausible agent A5. This quantity thus offers a readout of exploitative planning on the basis of belief states. As the desirability of an action is the primary variable underlying decision making, our approach is consistent with previous work aiming to uncover the neurocomputational mechanisms of various strategies such as exploitation and exploration (e.g., Chakroun et al. (2020) and Daw et al. (2006)), or model-free decision making and planning (e.g., Korn and Bach (2018) and Simon and Daw (2011)). We here modelled the fMRI data in terms of the absolute value of the trial-by-trial chosen action valence. An alternative would be to quantify the valence of the chosen action relative to the other actions available on a given trial. However, this quantity would indicate the easiness of a decision and not the desirability of the chosen action *per se*.

Findings

Behavioral findings On a behavioral level, we found evidence for a belief state-based exploitative planning strategy as formally captured by agent A5. The key implications of this result are twofold. First, our result implies that instead of resorting to belief state-free strategies, humans plan ahead on the basis of their belief states in the treasure hunt task. Second, our result further implies that in our task humans adopt a purely exploitative objective. In the following, these two key implications are discussed.

As alluded to in Section 3.1, there is a wealth of studies examining the dichotomy model-free decision making versus planning in multistep tasks with fully observable states and the Bayesian brain hypothesis. The first key implication of our behavioral result is consistent with findings from both lines of research, and, in particular, from the studies discussed in detail above with respect to their task design and computational modeling framework. Specifically, by investigating sequential decision making in a multistep task with fully observable states framed as spatial search, Simon and Daw (2011) demonstrated that participants planned ahead. Similarly, Korn and Bach (2018) also found that participants' actions reflected planning in their virtual foraging task, in addition to model-free strategies eschewing prospective computations. Our finding is perhaps most closely foreshadowed by Yoshida and Ishii (2006). These authors examined the Bayesian brain hypothesis in a multistep spatial search task with partially observable states and found striking correspondence between participants' actions and those predicted by their model that implemented a belief state-based planning strategy.

The second key implication of our behavioral result is also in line with the work by Yoshida and Ishii (2006), as their model assumed a purely exploitative objective. It is however more puzzling with respect to research on sequential decision making in bandit tasks. As mentioned in Section 3.1, there is growing evidence that in bandit tasks humans balance between exploration and exploitation depending on their current level of uncertainty about the environmental reward structure (e.g., Gershman (2018, 2019), Wilson et al. (2014), and Zhang and Yu (2013); see also Chapter 2). However, some studies did not find evidence for the combination of these two objectives (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011). These controversial findings suggest that certain configurations of the bandit task at hand, such as partial feedback (Dezza, Angela, Cleeremans, & Alexander, 2017; Gershman, 2018; Wilson et al., 2014), can veil exploration. Likewise, in the treasure hunt tasks we might have promoted exploitation, for example, by providing information about the treasure locations on each step or limiting the step horizon to approximately match the shortest path to the treasures. Despite these design-related considerations, the behavioral findings from the present work and the work by Yoshida and Ishii (2006) raise the intriguing possibility that in multistep tasks with partially observable states, people's tendency to explore is generally less pronounced as compared to bandit tasks. This might be due to the different sources of uncertainty in these two classes of sequential decision-making tasks. Alternatively, one might speculate that this is because multistep tasks are considerably more demanding and therefore, in these tasks purely exploitative strategies - which are simpler than explorative-exploitative strategies - prevail. Testing for these possibilities using carefully designed experimental tasks remains an interesting avenue for future research.

In addition to the two key implications of our behavioral result, it is noteworthy that the belief state-based exploitative planning strategy of the group-favored agent model assumes action evaluation that is reflective of the current level of subjective uncertainty about all possible treasure locations. This is in accordance with the theoretical work by Rao (2010) who postulated that, to make sequential decisions, biological agents rely on their entire belief state and not on unitary state estimates as suggested by Ishii et al. (2002). Given that Yoshida and Ishii (2006) only tested the behavioral plausibility of belief state-based exploitative planning on the basis of unitary state estimates, our behavioral result as afforded by our extensive model set offers a more fine-grained resolution to sequential decision-making behavior on multistep tasks with partially observable states.

FMRI findings On a neural level, we found that a distributed network of cortical and subcortical regions is involved in belief state-based exploitative planning. Specifically, belief state maintenance as indexed by the trial-by-trial Bayesian surprise engaged the bilateral prefrontal, parietal and medial occipitotemporal cortices as well as the right dorsal striatum. The trial-by-trial chosen action valence reflecting exploitative planning on the basis of belief states was neurally encoded in the right insula, orbitofrontal and inferior frontal cortices, as well as the bilateral anterior cingulate, medial frontal and lateral occipitotemporal cortices.

With respect to the neural correlates of belief state maintenance, our findings replicate previous human fMRI studies employing the Bayesian brain hypothesis framework: O'Reilly et al. (2013) and Fischer et al. (2017) sought to characterize the functional anatomy of belief state maintenance in a visual attention and bandit task, respectively, as captured by the Bayesian surprise. In striking correspondence with our findings, both studies reported activation related to the Bayesian surprise in the dorsal striatum, and the posterior parietal and frontal cortices. Furthermore, although medial occipitotemporal areas are primarily implicated in vision (e.g., Gazzaniga et al. (2013)), Fischer et al. (2017) also detected activity associated with Bayesian surprise in these areas. Notably, the frontal activation in our study was more focused to prefrontal areas than observed by O'Reilly et al. (2013) and Fischer et al. (2017). This is, however, in accordance with the study by Yoshida and Ishii (2006), who found that in their spatial search task the trial-by-trial belief state entropy reflecting belief state maintenance was represented in the anterior prefrontal cortex. In sum, our work provides further neural evidence for the Bayesian brain hypothesis in the context of sequential decision making in multistep tasks. In light of previous research, the brain regions identified here to be working together to integrate information about the state in a Bayesian fashion appear to be independent of the cognitive task and the nature of the latent task-relevant quantity.

The neural correlates of the trial-by-trial chosen action valence are largely consistent with prior human fMRI work on planning and exploitation: Akin to our findings, in their multistep task with fully observable states, Simon and Daw (2011) detected activation related to the planning-based desirability of the chosen action in the insular, orbitofrontal and inferior frontal cortices. Similarly, Korn and Bach (2018) also found the involvement of inferior frontal regions, and, in addition, the anterior cingulate and medial frontal cortices. These two latter areas are usually considered to be part of the ventromedial prefrontal cortex, which, together with the orbitofrontal cortex, is thought to play a central role in reward-guided decision making (e.g., Levy and Glimcher (2012) and Rushworth et al. (2011)). Accordingly, these regions have been implicated in exploitation in bandit tasks as captured by the value estimate of the chosen action, or a relative measure thereof (Boorman, Behrens, Woolrich, & Rushworth, 2009; Chakroun et al., 2020; Daw et al., 2006). Finally, the exploitative planning-related activity along the lateral occipitotemporal cortex as obtained in our work is perhaps surprising. However, of the aforementioned bandit studies, Chakroun et al. (2020) also identified a similar neural pattern, suggesting that beyond its established role in visual and auditory processing (e.g., Gazzaniga et al. (2013)), the occipitotemporal cortex may support higher cognitive processes, such as exploitation. Taken together, the regions uncovered here to be participating in exploitative planning likely ubiquitously contribute to decision making by evaluating the actions' reward-related consequences based on input from regions representing the state and other task-relevant quantities.

Conclusion

To conclude, we here computationally characterized 19 human participants' sequential decision-making behavior and underlying neural activity in a novel multistep spatial search task where the environmental states are only partially observable. On a behavioral level, we showed that instead of relying on simple belief state-free strategies that forgo prospective computations, participants engaged in belief state-based planning and adopted a purely exploitative objective. On a neural level, we showed that prefrontal, parietal, medial occipitotemporal and dorsal striatal areas subserve the maintenance of belief states, while the insular, orbitofrontal, vetromedial and lateral occipitotemporal cortices enable action evaluation according to exploitative planning. By accounting for state uncertainty in multistep tasks, the present work expands on previous research and contributes to a better understanding of the neurocomputational mechanisms of sequential decision making in such tasks.

3.8 Data and code availability

Data formatted according to the Brain Imaging Data Structure (Gorgolewski et al., 2016) and code implementing all analyses are hosted on Open Science Framework (Nosek et al., 2015) and are available at https://osf.io/pmnd6/?view_only=fa824ad40aad4be18b8b3ab0b9478c28 and https://osf.io/jrpg3/?view_only=934412d65a8e4460a19bea10eeb84ed3, respectively.

3.9 References

Babayan, B. M., Uchida, N., & Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature Communications*, *9*(1), 1–10.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of bold fmri experiments examining neural correlates of subjective value. *NeuroImage*, *76*, 412–427.

Berry, D. A., & Fristedt, B. (1985). Bandit problems: Sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, *5*, 71–87.

Bertsekas, D. P. (2000). *Dynamic programming and optimal control* (2nd edition). Athena Scientific.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, *62*(5), 733–743.

Byrd, R. H., Gilbert, J. C., & Nocedal, J. (2000). A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, *89*(1), 149–185.

Byrd, R. H., Hribar, M. E., & Nocedal, J. (1999). An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, *9*(4), 877–900.

Camerer, C., & Ho, T.-H. (1998). Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation. *Journal of Mathematical Psychology*, *42*(2-3), 305–326.

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, *9*, e51260.

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, 1–11.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, *112*(45), 13817–13822.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology*, *22*(6), 1068–1074.

Dayan, P. (2014). Rationalizable irrationalities of choice. *Topics in Cognitive Science*, *6*(2), 204–228.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Dezza, I. C., Angela, J. Y., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific Reports*, *7*(1), 1–13.

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Stevens' handbook of experimental psychology vol. 3: Learning, motivation and emotion* (pp. 497–533). Wiley.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767.

Ebbesen, C. L., & Brecht, M. (2017). Motor cortex—to act or not to act? *Nature Reviews Neuroscience*, *18*(11), 694.

Eichenbaum, H. (2017). On the integration of space, time, and memory. *Neuron*, *95*(5), 1007–1018.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fischer, A. G., Bourgeois-Gironde, S., & Ullsperger, M. (2017). Short-term reward experience biases inference despite dissociable neural correlates. *Nature Communications*, *8*(1), 1–14.

Friston, K., & Dolan, R. J. (2010). Computational and dynamic models in neuroimaging. *NeuroImage*, *52*(3), 752–765.

Friston, K., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: Applications. *NeuroImage*, *16*(2), 484–512.

Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., & Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: Theory. *NeuroImage*, *16*(2), 465–483.

Gazzaniga, M., Ivry, R., & Mangun, G. (2013). *Cognitive Neuroscience: The Biology of the Mind (Fourth Edition)*. W. W. Norton.

Geffner, H., & Bonet, B. (1998). Solving large POMDPs using real time dynamic programming. *Proc. AAAI fall symp. on POMDPs.*

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277.

Gijsen, S., Grundei, M., Lange, R. T., Ostwald, D., & Blankenburg, F. (2020). Neural surprise in somatosensory bayesian learning. *bioRxiv*.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 1–9.

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.*, *27*, 649–677.

Harrison, L. M., Stephan, K. E., Rees, G., & Friston, K. J. (2007). Extra-classical receptive field effects measured in striate cortex with fmri. *NeuroImage*, *34*(3), 1199–1208.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol*, *8*(3), e1002410.

Ishii, S., Yoshida, W., & Yoshimoto, J. (2002). Control of exploitation–exploration meta-parameter in reinforcement learning. *Neural Networks*, *15*(4-6), 665–687.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Kaplan, R., Schuck, N. W., & Doeller, C. F. (2017). The role of mental maps in decision-making. *Trends in Neurosciences*, *40*(5), 256–259.

Konovalov, A., & Krajbich, I. (2018). Neurocomputational dynamics of sequence learning. *Neuron*, *98*(6), 1282–1293.

Korf, R. E. (1990). Real-time heuristic search. *Artificial Intelligence*, *42*(2-3), 189–211.

Korn, C. W., & Bach, D. R. (2018). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications*, *9*(1), 1–15.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038.

Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron*, *104*(1), 164–175.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220.

Martin, A., Schurz, M., Kronbichler, M., & Richlan, F. (2015). Reading in the brain of children and adults: A meta-analysis of 40 functional magnetic resonance imaging studies. *Human Brain Mapping*, *36*(5), 1963–1981.

McFadden, D. (1973). *Conditional Logit Analysis of Qualitative Choice Behavior*. Institute of Urban and Regional Development, University of California.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692.

Morales, M., & Margolis, E. B. (2017). Ventral tegmental area: Cellular heterogeneity, connectivity and behaviour. *Nature Reviews Neuroscience*, *18*(2), 73.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., et al. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

O'Reilly, J. X., Jbabdi, S., Rushworth, M. F., & Behrens, T. E. (2013). Brain systems for probabilistic and dynamic prediction: Computational specificity and integration. *PLoS Biol*, *11*(9), e1001662.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177–188.

Parekh, R. (2006). *Principles of multimedia.* Tata McGraw-Hill Education.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput Biol*, *7*(1), e1001048.

Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming.* John Wiley & Sons.

Rao, R. P. (2010). Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, *4*, 146.

Reverdy, P., & Leonard, N. E. (2015). Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, *13*(1), 54–67.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage*, *84*, 971–985.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, *206*, 116189.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, *70*(6), 1054–1069.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Schwartenbeck, P., FitzGerald, T. H., & Dolan, R. (2016). Neural signals encoding shifts in beliefs. *NeuroImage*, *125*, 578–586.

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *8*, e41703.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*, *31*(14), 5526–5539.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120.

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, *20*(4), 581.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011). Planning to be surprised: Optimal bayesian exploration in dynamic environments. In J. Schmidhuber, K. R. Thórisson, & M. Looks (Eds.), *Artificial general intelligence* (pp. 41–51). Springer Berlin Heidelberg.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Waltz, R. A., Morales, J. L., Nocedal, J., & Orban, D. (2006). An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical Programming*, *107*(3), 391–408.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron*, *75*(3), 418–424.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, *50*(5), 781–789.

Zhang, S., & Yu, A. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in Neural Information Processing Systems*, *26*.

4 | General discussion

In this dissertation, I computationally modeled human sequential decision-making strategies under uncertainty with respect to the dichotomies model-free versus model-based and exploitative versus explorative. Chapter 1 conceptually situated the empirical work presented in Chapters 2 and 3 within relevant theories. To conclude the dissertation, in what follows, I first synthesize and discuss in a broader context the main findings from the empirical work. I then outline some interesting outstanding questions relating to these findings and the agent-based modeling framework applied throughout this dissertation.

4.1 Synthesis and discussion of the main findings

How humans make sequential decisions under uncertainty was investigated in two everyday choice environments previously neglected in decision neuroscience research. In the first choice environment, actions do not have effects on states - and it therefore belongs to the family of bandit tasks. Novel to the bandit task introduced in Chapter 2 is that an action is always associated with (positive or negative) reward gain, but not necessarily with information gain. In the second choice environment, state transitions depend on actions - and it therefore belongs to the family of multistep tasks. The distinctive characteristic of the multistep task introduced in Chapter 3 is that states are only partially observable. These novel tasks are referred to as information-selective reversal bandit task and treasure hunt task, respectively.

Using extensive sets of agent models to account for human participants' choice behavior in both tasks, I showed that the employed strategies reflect model-based processes: The behaviorally most plausible agent models maintained belief states and evaluated the desirability of actions by looking one step ahead. Yet, the nature of the belief state-based look-ahead differed in the two tasks. In the information-selective reversal bandit task, the look-ahead combined exploitation with exploration, whereas in the treasure hunt task, it conformed to pure exploitation. In this latter task, I additionally provided neural evidence for belief state-based exploitation, with a distributed network of cortical and subcortical brain regions playing an essential role in the realization of such a strategy. The remainder of this section contextualizes these key findings on a global level from a decision neuroscience standpoint.

4.1.1 Model-free versus model-based strategies

Decision neuroscientists have long been puzzled by the question whether biological agents apply model-free or model-based strategies. This question emerges from a dual-system view of behavior and is also discussed under the related terms habitual versus goal-directed (Dickinson, 1985; Dickinson & Balleine, 2002; Dolan & Dayan, 2013), associative versus rule-based (Sloman, 1996) and retrospective versus prospective (Dolan & Dayan, 2013; Economides, Kurth-Nelson, Lübbert, Guitart-Masip, & Dolan, 2015). According to this view, the first system governs behavior in a reactive fashion, which is resource-efficient albeit inflexible. By contrast, the second system governs behavior in a deliberate fashion, which is flexible albeit resource-intense (Collins & Cockburn, 2020; Evans, 2008; Gilovich, Griffin, & Kahneman, 2002).[1]

Early experimental work seeking for evidence for model-free strategies as foreseen by the first system was carried out by Thorndike (1911). In this work, cats were repeatedly locked in a box equipped with levers (or similar constructs) outside of which fish were placed visible from inside the box. To escape the box and reach the fish, the cats had to solve a puzzle by performing, for example, a certain sequence of lever-presses. Thorndike measured the amount of time it took for the cats to accomplish this and found that it decreased over several repeats. On the basis of his findings, Thorndike put forward the theory *Law of Effect*, which states that if, in a given situation, an action leads to a rewarding outcome, then it is likely to be repeated when the situation is encountered again. This theory later inspired the decision neuroscientists Rescorla and Wagner (1972), whose research, in turn, was fundamental to the development of temporal difference learning methods within the field of reinforcement learning

[1] Notably, while the terms model-free versus model-based are commonly used to differentiate strategies in bandit tasks (e.g., Fischer, Bourgeois-Gironde, and Ullsperger, 2017; Speekenbrink and Konstantinidis, 2015), the related terms listed here are usually applied in connection with multistep tasks only. This is because in bandit tasks, model-based strategies usually do not explicitly define a dynamic programming component accounting for (myopic or extended time horizon) look-ahead, just a Bayesian inference component accounting for belief formation (but see, e.g., Zhang and Yu, 2013). Nevertheless, given the flexible albeit resource-intense nature of Bayesian inference alone (e.g., Tavoni, Doi, Pizzica, Balasubramanian, and Gold, 2019), the dual-system view as described here naturally extends to bandit tasks (cf. Knox, Otto, Stone, and Love, 2012).

(Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 2018; Wiering & van Otterlo, 2014). Moreover, research on heuristics that rely on some readily available reward-related information - such as the win-stay-lose-shift heuristic - also owes its root to Thorndike's work (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Gigerenzer & Gaissmaier, 2011; Robbins, 1952; Tversky & Kahneman, 1974).

Shortly after the results presented by Thorndike, his contemporary, Edward Tolman, delivered evidence for the opposite model-based strategies as postulated by the second system (Tolman, 1948). In his experiments, rats were assessed based on their ability to navigate a maze. For example, when comparing the path undertaken by rats to a location novelly baited with food, Tolman found that those who previously freely ran the maze were more efficient than those who were previously not exposed to the maze. Similarly, after pre-exposure on a maze, rats were able to seek out the next shortest path if the shortest path to a rewarding location had been closed (Tolman & Honzik, 1930). To explain these findings, Tolman reasoned that biological agents plan ahead based on a *cognitive map* - which, in essence, corresponds to a representation of the environment entailing its statistical regularities.

Although the work of Tolman has since shaped the field of decision neuroscience by prompting it to turn to methods of dynamic programming to computationally characterize model-based strategies, it did not explicitly address uncertainty. This issue has, however, been the focus of Bayesian theories of decision making. Similar to Tolman, these theories also furnish biological agents with a representation of the statistical regularities of the environment, which they refer to as the *generative model*. The generative model underlies belief state formation and, consequently, serves as basis for action selection (Ma, 2019). Empirical testing of Bayesian theories of decision making traces back to the 19th century. Building on observations about the physiology of vision, Helmholtz (1866) argued that the brain turns sensory data into their hidden cause by means of *unconscious inference*. This notion was later linked with Bayesian inference and has been applied to various cognitive functions, including decision making, culminating in a line of research considering the brain to be inherently Bayesian (Doya, Ishii, Pouget, & Rao, 2007; Friston, 2010; Knill & Pouget, 2004).

When pitting model-free and model-based strategies against each other, a growing body of literature suggests the superiority of the latter in explaining human behavioral and neural data (Collins & Cockburn, 2020; Doll, Simon, & Daw, 2012; Doya et al., 2007). This holds true both for bandit tasks (e.g., Chakroun, Mathar, Wiehler, Ganzer, and Peters, 2020; Fischer et al., 2017; Hampton, Bossaerts, and O'Doherty, 2006; Knox et al., 2012; Stojić, Orquin, Dayan, Dolan, and Speekenbrink, 2020; Zhang and Yu, 2013) and multistep tasks (e.g., da Silva and Hare, 2020; Deserno et al., 2015; Doll, Duncan, Simon, Shohamy, and Daw, 2015; D. A. Simon and Daw, 2011; Yoshida and Ishii, 2006). The results presented in Chapters 2 and 3 lend further support to a predominantly model-based nature of sequential decision making under uncertainty. Yet, it is important to highlight that if resources are limited, model-free strategies can prevail (e.g., Dasgupta, Schulz, Hamrick, and Tenenbaum, 2019; Keramati, Smittenaar, Dolan, and Dayan, 2016; Otto, Gershman, Markman, and Daw, 2013). Together with studies providing evidence for decisions reliant on both model-based and model-free processes (e.g., Babayan, Uchida, and Gershman, 2018; Gläscher, Daw, Dayan, and O'Doherty, 2010; Korn and Bach, 2018; Kuperwajs, Van Opheusden, and Ma, 2019; Momennejad et al., 2017), this implies that instances of the two types of strategies - and combinations thereof - are flexibly realized to accommodate the ever-changing demands biological agents face (cf. Griffiths, Lieder, and Goodman, 2015; Lieder and Griffiths, 2020). Thus, viewing model-free and model-based strategies as cooperative, rather than competitive, is called for to refocus future research on deciphering how their cohabitation is organized. In Section 4.2.1, I review a theoretical advance in this regard and delineate related outstanding questions.

4.1.2 Exploitative versus explorative strategies

Besides scrutinizing strategies through the lens of the model-free versus model-based dichotomy, decision neuroscience has also striven to decipher whether biological agents engage in exploration beyond exploitation. Crucially, this question arises in the face of uncertainty: If the environment is not fully observable, humans and other animals can aim for a balance between maximizing information gain to improve their knowledge about the environment and maximizing reward gain in light of their accumulated knowledge. Alternatively, instead of combining exploration and exploitation, which is necessary to maximize reward gain *in the long run*, they might behave as if the environment was fully observable and settle for exploitation (Cohen, McClure, & Yu, 2007; Dayan & Daw, 2008; Schwartenbeck et al., 2019).

Experimental work on correlates of exploration and exploitation dates back to at least Tversky and Edwards (1966), who analyzed human participants' trial-by-trial choices on the observe-or-bet task, an extended variant of the pure exploration paradigm. As detailed in Chapter 2, in this task, the observe action only confers information but no reward, whereas the bet action only confers reward but no information. The authors found that while on most trials participants chose a bet action, pertaining to exploitation, they also explored as manifested by choices of the observe action. The observe actions were, however, spread across all trials, despite the fact that the bet actions' expected reward values were stationary. Seemingly, this behavioral pattern contradicted the optimal strategy, but only if participants trusted the instructions with respect to the stationary nature of the task, which possibly was not the case (Navarro, Newell, & Schulze, 2016). In accordance with this reasoning, Tversky and Edwards (1966) reported that this pattern was more pronounced in those participants who were explicitly malinformed and told that the task was non-stationary. In a subsequent study, Krebs, Kacelnik, and Taylor (1978) presented birds with a classical two-armed bandit task, wherein the arms corresponded to feeding places. Upon hopping on its feeding place of choice, the bird was rewarded with a mealworm emitted with a stationary probability. The results show that the birds first exhibited an explorative strategy by switching back and forth between the feeding places and eventually committed to the more rewarding feeding place, reflecting exploitation. Moreover, the time point of commitment was strikingly close to the optimum.

Research following up on these early studies providing evidence for exploration beyond exploitation has gained momentum in the last two decades. A particular focus has been placed on computationally assessing if biological agents *directly* seek out actions that maximize information gain as qualitatively motivated by an optimal balance - and, as such, as foreshadowed by the studies introduced above (Dayan & Daw, 2008; Gershman, 2018, 2019; Schwartenbeck et al., 2019; Wilson, Geana, White, Ludvig, & Cohen, 2014). Work in this vein has behaviorally and neurally modeled strategies using some combination of reward value estimates and measures of information gain, derived either in a model-free or model-based fashion. In doing so, mixed results have been reported. Some studies - including the study reported in Chapter 2 - found evidence for directed exploration additionally to exploitation (e.g., Chakroun et al., 2020; Navarro et al., 2016; Wu, Schulz, Speekenbrink, Nelson, and Meder, 2018; Zhang and Yu, 2013), whereas others - including the study reported in Chapter 3 - only found evidence for exploitation (e.g., Daw, O'Doherty, Dayan, Seymour, and Dolan, 2006; Payzan-LeNestour and Bossaerts, 2011; Speekenbrink and Konstantinidis, 2015).

On the one hand, these mixed results have been attributed to specifics of the applied task, such as the reward–information confound in classical bandit tasks, which has the potential to mask directed exploration (Dezza, Angela, Cleeremans, & Alexander, 2017; Gershman, 2018; Wilson et al., 2014). Consistent with this line of reasoning, the continuous availability of information and the probabilistic allocation of step limits in the treasure hunt task might have promoted exploitation, as also noted in Section 3.7 of Chapter 3. On the other hand, recent work emphasized the attenuating effect of increasing time pressure (Wu, Schulz, Gerbaulet, Pleskac, & Speekenbrink, 2019) and cognitive load (Cogliati Dezza, Cleeremans, & Alexander, 2019) on directed exploration, suggesting a resource-dependent account of the mixed results (cf. Griffiths et al., 2015; Lieder and Griffiths, 2020). Given that in the treasure hunt task, as compared to the information-selective reversal bandit task, the interdependency between actions together with the large state space arguably results in an increased cognitive load, the differing strategies used in these tasks might in fact support such a resource-dependent account.

The notion of resource-dependent arbitration between exploitative and explorative-exploitative strategies fits well with the cooperative view of model-free and model-based strategies as described above. Before delving into a corresponding integrative theory proposed by Griffiths et al. (2015), an issue with respect to exploration deserves mentioning. Besides computationally conceptualizing exploration in a directed sense, exploration has also been computationally conceptualized in a *random* sense. The virtue of random exploration as introduced in reinforcement learning research is that information can also be acquired by injecting some noise into the action selection process (Kaelbling, 1993; Sutton & Barto, 2018; Wiering & van Otterlo, 2014). Inspired by this idea, decision neuroscience research commonly interprets a biological agent's *observable* stochastic deviation from the action with the highest reward value estimate as random exploration (e.g., Chakroun et al., 2020; Daw et al., 2006; Dezza et al., 2017; Speekenbrink and Konstantinidis, 2015). However, as already argued in earlier chapters of this dissertation and recapitulated in some more detail in Section 4.2.2, this interpretation is problematic, because it conflates different sources of noise.

4.2 Outstanding questions

By embedding the main findings of Chapters 2 and 3 in the broader discussion they contribute to, I highlighted the remarkable capability of biological agents to employ a rich repertoire of strategies. This, in turn, raises the intriguing question of how the cohabitation of different strategies is organized. In this last section of the discussion, I first present an integrative theory arguing for *resource rationality* as the organizing principle and outline related open endeavors (Griffiths et al., 2015; Lieder & Griffiths, 2020). I then proceed by delineating ways to further refine how biological agents make sequential decisions under uncertainty, from the perspective of the agent-based modeling framework adopted in this dissertation.

4.2.1 How do different strategies cohabitate?

As reviewed above, while ample evidence has been provided in favor of model-based strategies, they are often complemented with or, if resources are cut short, may even be replaced by model-free strategies (e.g., Babayan et al., 2018; Gläscher et al., 2010; Huys et al., 2015; Keramati et al., 2016; Otto, Gershman, et al., 2013). Similarly, studies have shown that resource shortage dampens directed exploration, resulting in purely exploitative strategies to prevail (e.g., Cogliati Dezza et al., 2019; Wu et al., 2019). A common thread across these findings is that the available resources limit the complexity of the applied strategies - a notion readily accommodated by the theory of resource rationality (Griffiths et al., 2015; Lieder & Griffiths, 2020). According to this theory, biological agents make rational use of their limited resources. More specifically, drawing inspiration from Herbert Simon's work on bounded rationality (H. A. Simon, 1956; H. A. Simon, 1997), resource rationality posits that evolution has rendered biological agents' strategies optimal with respect to the trade-off between benefits in terms of accrued reward and costs in terms of processing time and related cognitive demands.

Although a substantial body of indirect evidence supports the theory of resource rationality, its formal testing is still in an early phase. To aid research in this regard, Lieder and Griffiths (2020) laid out an analysis scheme leveraging the constrained optimization problem biological agents, in particular humans, supposedly solve in an implicit fashion. Accordingly, the centerpiece of this scheme is that a cost-benefit trade-off is computed for each plausible strategy and the predictions of the strategy with the optimal trade-off is evaluated against empirical data. Using an early version of this analysis scheme, Callaway et al. (2018) demonstrated that human participants' choices in a multistep task were best explained by the constrained-optimal strategy, which rapidly adjusted to changes of the environment. While this finding constitutes an encouraging first step, much work remains to be done to conclude whether resource rationality truly is the principle according to which the cohabitation of pure and combined forms of model-free, model-based, explorative and exploitative strategies is organized.

A particular challenge for future research on resource rationality in humans concerns the quantification of costs. One option is to take on a biological perspective and consider general principles of neural activity, as Lieder and Griffiths (2020) point out. In my view, an important undertaking for work along this line will be to also carefully account for the possibility that the brain is better adapted to perform certain computations than others - even if, from an artificial agent perspective, the complexity of the operations is similar. For example, previous neuroimaging work, including the work presented in Chapter 3, revealed that Bayesian inference is enabled by a large network encompassing parts of the frontal, posterior parietal and occipital cortices as well as the dorsal striatum (e.g., Fischer et al., 2017; O'Reilly, Jbabdi, Rushworth, and Behrens, 2013; Yoshida and Ishii, 2006). This network is largely overlapping with a set of *task-positive* regions thought to routinely work together during performance of cognitive tasks (Fox et al., 2005; Raichle, 2011). In contrast, neural correlates of dynamic programming appear to be more localized to ventromedial prefrontal and orbitofrontal areas (e.g., Korn and Bach, 2018; D. A. Simon and Daw, 2011; see also Chapter 3). These findings might imply that Bayesian inference is more robustly represented than dynamic programming, which might, in turn, ease processing time and thereby diminish the associated costs.[2] Although the issue of cost quantification renders research on resource rationality a challenging endeavor, it could shed light on how humans arbitrate between sequential decision-making strategies to tackle the diversity of daily life and is therefore an important path forward.

[2] The hypothesis that the brain is better adapted to perform Bayesian inference than dynamic programming is also consistent with the observation that while Bayesian inference is resistant to stress (Trapp & Vilares, 2020), the reliance of humans on dynamic programming is negatively impacted by it (Otto, Raio, Chiang, Phelps, & Daw, 2013).

4.2.2 How can strategies be further refined?

In the agent-based modeling framework used in this dissertation, an agent's choice environment is formalized within the task model and alternative strategies are formally captured by means of the agent models. These agent models are then nested in statistical inference frameworks such as the softmax operation and model-based GLM for fMRI to evaluate their behavioral and neural plausibility, respectively, based on experimentally acquired data. To obtain a refined picture of human sequential decision-making strategies under uncertainty, setting up an extensive agent model space is thus essential. Consequently, while in Chapters 2 and 3 a variety of agent models implementing model-free, model-based, explorative and exploitative strategies were considered, future work should make expanding the agent model space an imperative.

One direction for expanding the agent model space is to also include random exploratory agents. In Chapters 2 and 3, exploration was only accounted for in a directed sense, which assumes an explicit evaluation of the attainable information. However, it is conceivable that humans rely on random exploratory strategies (Cohen et al., 2007; Dayan & Daw, 2008; Schwartenbeck et al., 2019; Wilson, Bonawitz, Costa, & Ebitz, 2021). As already mentioned in the introductory Chapter 1 and in Section 4.1.2 of the current chapter, the central idea underlying such strategies is that information can also be solicited by randomly deviating from the action with the highest reward value estimate. Instead of quantifying information gain as part of a directed exploratory agent's valence function, this idea can be implemented simply by defining stochastic agent decision rules (Bertsekas & Tsitsiklis, 1996; Kaelbling, 1993; Puterman, 2014; Sutton & Barto, 2018; Wiering & van Otterlo, 2014). Random exploratory strategies are thus comparatively computationally inexpensive and, in line with the resource rationality principle introduced above (Griffiths et al., 2015; Lieder & Griffiths, 2020), present themselves as viable alternatives to directed exploratory strategies.

In the last years, many studies have sought to decompose behavioral and neural correlates of directed and random exploration (e.g., Chakroun et al., 2020; Daw et al., 2006; Dezza et al., 2017; Speekenbrink and Konstantinidis, 2015; Tomov, Truong, Hundia, and Gershman, 2020; Wilson et al., 2014; Zajkowski, Kossut, and Wilson, 2017). Yet, a central issue tends to fall short in these investigations: Stochastic decision rules are turned into statistical inference frameworks to model data, thereby ascribing noise stemming from unrelated sources, such as lapses in attention or changes in motivation, to random exploration. This issue, once again, demonstrates the importance of explicitly separating agent models formally capturing strategies and their statistical embedding. Given that random exploratory agents yield latent random variables, future work expanding the agent model space in this regard will need to replace the maximum likelihood approach employed in Chapters 2 and 3 by a filtering approach. Performing model estimation and evaluation by means of filtering will enable partitioning of noise with respect to random exploration and unrelated sources (cf. Findling, Skvortsova, Dromnelle, Palminteri, and Wyart, 2019; Ostwald, Kirilina, Starke, and Blankenburg, 2014).

Another interesting direction that goes beyond the dichotomies model-free versus model-based and exploitation versus exploration is to expand the agent model space by formalizing possible ways in which humans construct a task model. Consistent with the experimental procedure adopted in Chapters 2 and 3, research on decision-making strategies in humans typically provides detailed task instructions prior to the experiment. Thus, it is assumed that participants can, in principle, enter the experiment with a complete copy of the task model. In life, however, exact instructions are rarely available, raising the question of how humans construct a task model in the first place. While this is still a largely open question, advances from machine learning could serve as insightful starting points. For example, previous theoretical work suggested that the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) could potentially capture how humans form representations of the statistical regularities of the environment, i.e., observation and reward emissions as well as state transitions, in the face of partially observable states (Dayan & Daw, 2008; Rao, 2010). Furthermore, it has recently been highlighted that deep neural networks could be leveraged to better understand humans' environmental representations in general, that is, also with respect to states, observations, actions and rewards (Botvinick, Wang, Dabney, Miller, & Kurth-Nelson, 2020; Ma & Peters, 2020). Testing these methods against empirical data would benefit decision neuroscience research. In addition, it might help elucidate why in certain aspects - such as generalization - biological intelligence still surpasses artificial intelligence and thereby also benefit research on the latter.

4.3 Conclusion

In this dissertation, I presented work on the computational characterization of human sequential decision making under uncertainty, with a focus on model-free, model-based, exploitative and explorative strategies. Using an agent-based modeling framework (Chapter 1), I demonstrated that humans rely on model-based strategies in an information-selective reversal bandit task (Chapter 2) and in a spatial multistep task with partially observable states (Chapter 3). In the former task, model-based strategies were deployed in an explorative-exploitative fashion. In the latter task, model-based strategies were deployed in a purely exploitative fashion, which was supported by the orchestrated activity of cortical and subcortical brain regions. By contextualizing these findings within a broader decision neuroscience discourse, I outlined key undertakings for future research, such as studying the arbitration between strategies from a resource-rational standpoint or adopting measures to capture random exploration and task model construction (Chapter 4).

Together, on a theoretical level, this dissertation offers a computational framework to decompose the behavioral and neural correlates of sequential decision making under uncertainty. On an empirical level, this dissertation contributes to a more fine-grained resolution of the strategies the human mind employs to master the challenges of everyday choice environments.

4.4 References

Babayan, B. M., Uchida, N., & Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature Communications*, *9*(1), 1–10.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming.* (Vol. 3). Athena Scientific.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron.*

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. *CogSci.*

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, *9*, e51260.

Cogliati Dezza, I., Cleeremans, A., & Alexander, W. (2019). Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General*, *148*(6), 977.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). 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.

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, 1–11.

Dasgupta, I., Schulz, E., Hamrick, J. B., & Tenenbaum, J. (2019). Heuristics, hacks, and habits: Boundedly optimal approaches to learning, reasoning and decision making. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1–2.

da Silva, C. F., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053–1066.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Deserno, L., Huys, Q. J., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., Dolan, R. J., Heinz, A., & Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, *112*(5), 1595–1600.

Dezza, I. C., Angela, J. Y., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific Reports*, *7*(1), 1–13.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67–78.

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Stevens' handbook of experimental psychology vol. 3: Learning, motivation and emotion* (pp. 497–533). Wiley.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*(5), 767.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., & Dolan, R. J. (2015). Model-based reasoning in humans becomes automatic with training. *PLoS Comput Biol*, *11*(9), e1004463.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature neuroscience*, *22*(12), 2066–2077.

Fischer, A. G., Bourgeois-Gironde, S., & Ullsperger, M. (2017). Short-term reward experience biases inference despite dissociable neural correlates. *Nature Communications*, *8*(1), 1–14.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

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

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360–8367.

Helmholtz, H. v. (1866). Concerning the perceptions in general (J. P. C. Southall, Trans.). In, *Treatise on physiological optics*. Dover.

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

Kaelbling, L. P. (1993). *Learning in Embedded Systems.* The MIT Press.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: The role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2*, 398.

Korn, C. W., & Bach, D. R. (2018). Heuristic and optimal policy computations in the human brain during sequential decision-making. *Nature Communications*, *9*(1), 1–15.

Krebs, J. R., Kacelnik, A., & Taylor, P. (1978). Test of optimal sampling by foraging great tits. *Nature*, *275*(5675), 27–31.

Kuperwajs, I., Van Opheusden, B., & Ma, W. J. (2019). Prospective planning and retrospective learning in a large-scale combinatorial game. *2019 Conference on Cognitive Computational Neuroscience*, 1091–1094.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron*, *104*(1), 164–175.

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: Towards using deep nets as models for human behavior. *arXiv*.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692.

Navarro, D. J., Newell, B. R., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology*, *85*, 43–77.

O'Reilly, J. X., Jbabdi, S., Rushworth, M. F., & Behrens, T. E. (2013). Brain systems for probabilistic and dynamic prediction: Computational specificity and integration. *PLoS Biol*, *11*(9), e1001662.

Ostwald, D., Kirilina, E., Starke, L., & Blankenburg, F. (2014). A tutorial on variational bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, *60*, 1–19.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*, *24*(5), 751–761.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences*, *110*(52), 20941–20946.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput Biol*, *7*(1), e1001048.

Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.

Raichle, M. E. (2011). The restless brain. *Brain Connectivity*, *1*(1), 3–12.

Rao, R. P. (2010). Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, *4*, 146.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasky (Eds.), *Classical conditioning II* (pp. 64–99). Appleton-Century-Crofts.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *8*, e41703.

Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*, *31*(14), 5526–5539.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129.

Simon, H. A. (1997). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). MIT press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*(2), 351–367.

Stojić, H., Orquin, J. L., Dayan, P., Dolan, R. J., & Speekenbrink, M. (2020). Uncertainty in learning, choice, and visual fixation. *Proceedings of the National Academy of Sciences*, *117*(6), 3291–3300.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Tavoni, G., Doi, T., Pizzica, C., Balasubramanian, V., & Gold, J. I. (2019). The complexity dividend: When sophisticated inference matters. *bioRxiv.*

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies.* MacMillan.

Tolman, E. C., & Honzik, C. H. (1930). "Insight" in rats. *University of California Publications in Psychology*, *4*, 215–232.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications*, *11*(1), 1–12.

Trapp, S., & Vilares, I. (2020). Bayesian decision-making under stress-preserved weighting of prior and likelihood information. *Scientific Reports*, *10*(1), 1–11.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Wiering, M., & van Otterlo, M. (2014). *Reinforcement learning: State-of-the-art.* Springer Publishing Company, Incorporated.

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, *38*, 49–56.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Wu, C. M., Schulz, E., Gerbaulet, K., Pleskac, T. J., & Speekenbrink, M. (2019). Under pressure: The influence of time limits on human exploration. *PsyArXiv.*

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), 915–924.

Yoshida, W., & Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*, *50*(5), 781–789.

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife*, *6*, e27430.

Zhang, S., & Yu, A. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in Neural Information Processing Systems*, *26*.

A | Supplementary material to Chapter 2

A.1 Sample characteristics

To characterize the group of participants, we measured symptoms of attention deficit hyperactivity disorder (ADHD), anxiety, depression and impulsivity. To this end, we used the questionnaires Conners Adult ADHD Rating Scale – Self Report, Short Version (CAARS-S:S; Conners, Erhardt, and Sparrow, 1999), State and Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, Lushene, Vagg, and Jacobs, 1983), Beck Depression Inventory II (BDI-II; Beck, Steer, and Brown, 1996) and UPPS-P Impulsive Behavior Scale (Lynam, Smith, Whiteside, & Cyders, 2006), respectively. As shown in Table A.1, the sample varied only moderately with respect to these symptoms. For example, on CAARS-S:S, our main questionnaire of interest, participants scored within ±2 standard deviations of the mean of their age- and gender-matched norm groups of the general population. We therefore argued that the sample represents the healthy population and did not relate individual variability in terms of ADHD or other clinical symptoms to behavioral strategies. Note that in Table A.1 we also report the IQ score, which was obtained by administering the Wechsler Abbreviated Scale of Intelligence (WASI-II; Wechsler, 1999) at the time of the Nathan Kline Institute Rockland Sample study (Nooner et al., 2012).

Measurement	Range	Median	Mean ± SD
Age (years)	18 - 35	23.5	24.5 ± 5.52
WASI-II (total score)	84 - 122	101.5	102.38 ± 9.14
CAARS-S:S (total T-score)	32 - 65	48	47.63 ± 9.15
BDI-II (total)	0 - 20	4	6.67 ± 6.28
STAI STATE (total T-score)	34 - 63	43	45.58 ± 8.52
STAI TRAIT (total T-score)	34 - 71	49.5	49.92 ± 10.29
UPPS-P (total)	73 - 190	118.5	124.04 ± 25.66

Table A.1. Sample characteristics. There was a moderate variability in terms of cognitive abilities (measured with Wechsler Abbreviated Scale of Intelligence (WASI-II)), attention deficit hyperactivity disorder (measured with Conners Adult ADHD Rating Scale – Self Report, Short Version (CAARS-S:S)), anxiety (measured with State and Trait Anxiety Inventory (STAI)), depression (measured with Beck Depression Inventory II (BDI-II)) and impulsivity (measured with UPPS-P Impulsive Behavior Scale). As the scores remained largely in the normal range we treated the group of participants as a healthy sample.

A.2 Task instructions

Participants were provided with the following instructions about the reversal bandit task:

Welcome to the main part of today's experiment! In the following we will introduce to you the decision making task that you will complete in the scanner. Please read the instructions carefully. If you have any questions, feel free to ask at any time. Once you read the instructions, you will complete a test run with the task to make sure you feel comfortable with it before going in the scanner. On every trial we will present to you two objects, an orange square and a blue triangle on either side of a black and grey screen and ask you to choose between them. One of these objects is profitable, meaning that it is going to give you a win most of the time, while the other object in not profitable meaning that it is going to give you a loss most of the time. Once you choose one of the objects, the outcome (win: $+1$ *or loss:* -1*) will be registered to your account. You will have 2.5 seconds to indicate your choice. If you do not respond within this time window, the message 'Too slow' will appear on the screen and you automatically lose* 1 *point. Here you see an example for a trial (Supplementary Figure A.1).*

You will start the experiment with a balance of 0 points and any wins or losses will be registered to your account. After the experiment, in addition to your standard payment for participation, you will receive up to $30 depending on your final account. Note that your balance cannot get below 0 and if you do not earn additional money on the task, you will not be penalized and you will

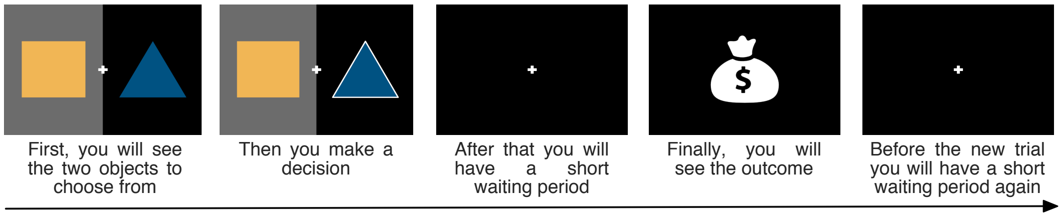
First, you will see the two objects to choose from
Then you make a decision
After that you will have a short waiting period
Finally, you will see the outcome
Before the new trial you will have a short waiting period again
$

Figure A.1. Reversal bandit task instructions 1. The figure shows the sequence of events within a trial as presented to the participants in the instructions.

still receive the standard payment for your participation. We would however encourage you to try to earn as much as possible on the task. After each run we will show you your balance. A run consists of 80 trials, which takes about 20 minutes to complete. You will have two runs in the scanner.

As mentioned above, one of the objects is profitable and it will bring you a win most of the times and every now and then it will bring you loss. At the same time, the other object is not profitable and it will bring you a loss most of the time and a win every now and then. You won't explicitly know which object is the profitable one and which is the non-profitable and you will need to conclude it from the outcomes. But be aware! These roles can switch, which means that the previously profitable object becomes non-profitable and the previously non-profitable object becomes profitable. Such a switch will happen only 1-4 times in the entire run and you will have enough trials without a switch to conclude which object is the profitable one.

Keep in mind that even the currently profitable object can from time to time deliver a loss and a couple of negative outcomes does not necessarily mean that a switch occurred. Similarly, even the non-profitable object can from time to time deliver a win and a couple of positive outcomes does not necessarily mean that a switch occurred. You can however assume that a switch has happened if you feel the previously rewarding object started to give you more losses than wins and the previously non-rewarding object started to bring you more wins than losses.

Before you do the test run, there is one more important aspect to the task: On each trial, one of the objects will be presented to you in front of a black background while the other object will be in front of a grey background. If you choose the object on the black-side, you will see the outcome of your choice. However, if you choose the grey-side object, the outcome will remain hidden from you but it will be registered to your account (Supplementary Figure A.2).

You will now complete a test run, which will be just like the ones you will complete in the scanner. We will discuss all your questions to make sure you

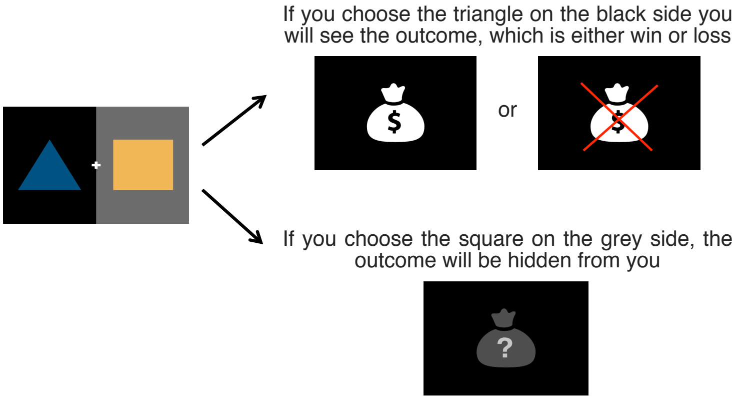
If you choose the triangle on the black side you will see the outcome, which is either win or loss
$
or
$
If you choose the square on the grey side, the outcome will be hidden from you
?

Figure A.2. Reversal bandit task instructions 2. The figure depicts the lucrativeness and informativeness associated with the actions as presented to the participants in the instructions.

feel comfortable with the task before going in the scanner.

A.3 Trial sequence

Participants were presented with the identical sequence of trials specifying the evolution of the first state component, s^1, which encodes the lucrative shape and the second state component, s^2, which encodes the choice options. The trial sequence of both runs is shown in Table A.2.

t	1	2	3	4	5	6	7	8	9	10
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$

t	11	12	13	14	15	16	17	18	19	20
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(2,1)$

t	21	22	23	24	25	26	27	28	29	30
s_t	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$

t	31	32	33	34	35	36	37	38	39	40
s_t	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$

t	41	42	43	44	45	46	47	48	49	50
s_t	$(2,2)$	$(2,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$

t	51	52	53	54	55	56	57	58	59	60
s_t	$(1,1)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$

t	61	62	63	64	65	66	67	68	69	70
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,1)$

t	71	72	73	74	75	76	77	78	79	80
s_t	$(2,2)$	$(2,1)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$

t	1	2	3	4	5	6	7	8	9	10
s_t	$(2,1)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,1)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,2)$

t	11	12	13	14	15	16	17	18	19	20
s_t	$(2,1)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$

t	21	22	23	24	25	26	27	28	29	30
s_t	$(2,1)$	$(2,1)$	$(2,2)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,1)$

t	31	32	33	34	35	36	37	38	39	40
s_t	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,2)$	$(1,2)$

t	41	42	43	44	45	46	47	48	49	50
s_t	$(1,2)$	$(1,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,2)$	$(2,2)$	$(2,1)$

t	51	52	53	54	55	56	57	58	59	60
s_t	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,1)$	$(2,2)$	$(2,1)$	$(2,2)$	$(2,1)$	$(1,2)$

t	61	62	63	64	65	66	67	68	69	70
s_t	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,1)$

t	71	72	73	74	75	76	77	78	79	80
s_t	$(1,2)$	$(1,1)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,2)$	$(1,1)$	$(1,1)$	$(1,1)$	$(1,2)$

Table A.2. State evolution function f. Upper table shows the trial sequence of the first run and lower table shows the trial sequence of the second run. t encodes the trial in the run and s_t encodes the state on the trial.

A.4 Belief state and posterior predictive distribution evaluation

Belief state Capitalizing on the probability distributions $p\left(s_1^1\right)$, $p\left(s_{t+1}^1|s_t^1\right)$ and $p^{a_t}\left(o_t|s_t^1\right)$ of the agent model M_{Agent} (eq. 2.4), the belief state at trial $t = 2, ..., T$ can be recursively evaluated according to eq. 2.11. To show that this equation holds, we first express the belief state as

$$p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) = \frac{p^{a_{1:t-1}}\left(s_t^1, o_{1:t-1}\right)}{p^{a_{1:t-1}}\left(o_{1:t-1}\right)}. \tag{A.1}$$

The numerator of eq. A.1 can be rewritten as

$$\begin{aligned}
p^{a_{1:t-1}}\left(s_t^1, o_{1:t-1}\right) &= \sum_{s_{t-1}^1} p^{a_{1:t-1}}\left(s_t^1, s_{t-1}^1, o_{1:t-1}\right) \\
&= \sum_{s_{t-1}^1} p^{a_{1:t-1}}\left(s_t^1|s_{t-1}^1, o_{1:t-1}\right) p^{a_{1:t-1}}\left(s_{t-1}^1, o_{1:t-2}, o_{t-1}\right) \\
&= \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-2:t-1}}\left(o_{t-1}|s_{t-1}^1, o_{1:t-2}\right) p^{a_{1:t-2}}\left(s_{t-1}^1, o_{1:t-2}\right) \\
&= \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right) p^{a_{1:t-2}}\left(o_{1:t-2}\right) \\
&= p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right).
\end{aligned} \tag{A.2}$$

Similarly, we can rewrite the denominator of eq. A.1 as

$$\begin{aligned}
p^{a_{1:t-1}}\left(o_{1:t-1}\right) &= \sum_{s_t^1} \sum_{s_{t-1}^1} p^{a_{1:t-1}}\left(s_t^1, s_{t-1}^1, o_{1:t-1}\right) \\
&= p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_t^1} \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right),
\end{aligned} \tag{A.3}$$

where in the last equality we used the numerator's derivation from eq. A.2. By substituting the derived expressions A.2 and A.3 in eq. A.1, we obtain

$$\begin{aligned}
p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) &= \frac{p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}{p^{a_{1:t-2}}\left(o_{1:t-2}\right) \sum_{s_t^1} \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)} \\
&= \frac{\sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}{\sum_{s_t^1} \sum_{s_{t-1}^1} p\left(s_t^1|s_{t-1}^1\right) p^{a_{t-1}}\left(o_{t-1}|s_{t-1}^1\right) p^{a_{1:t-2}}\left(s_{t-1}^1|o_{1:t-2}\right)}
\end{aligned} \tag{A.4}$$

for the belief state in trial t.

Posterior predictive distribution Given the agent's belief state $p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right)$ and the action-dependent state-conditional observation distribution $p^{a_t}\left(o_t|s_t^1\right)$, the posterior predictive distribution can be evaluated according to eq. 2.19. A proof of this equation is as follows:

$$\begin{aligned}
p^{a_{1:t}}\left(o_t|o_{1:t-1}\right) &= \frac{p^{a_{1:t}}\left(o_t, o_{1:t-1}\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t}}\left(o_t, o_{1:t-1}, s_t^1\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t}}\left(s_t^1|o_t, o_{1:t-1}\right) p^{a_{1:t}}\left(o_t, o_{1:t-1}\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) \sum_{s_t^1} p^{a_{1:t}}\left(o_t, o_{1:t-1}, s_t^1\right)}{p^{a_{1:t}}\left(o_{1:t-1}\right)} \\
&= \frac{\sum_{s_t^1} p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) \sum_{s_t^1} p^{a_{1:t}}\left(o_t|o_{1:t-1}, s_t^1\right) p^{a_{1:t}}\left(o_{1:t-1}, s_t^1\right)}{\sum_{s_t^1} p^{a_{1:t}}\left(o_{1:t-1}, s_t^1\right)} \\
&= \sum_{s_t^1} p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right) p^{a_t}\left(o_t|s_t^1\right) \\
&= b_t p^{a_t}\left(o_t|s_t^1 = 1\right) + \left(1 - b_t\right) p^{a_t}\left(o_t|s_t^1 = 2\right), \qquad \text{(A.5)}
\end{aligned}$$

where in the last equality we substituted the belief state with its scalar representation to complete the derivation of eq. 2.19.

A.5 Belief state and posterior predictive distribution implementation

For a concise implementation of the belief state and the posterior predictive distribution, we represented the probability distributions the agent model M_{Agent} (eq. 2.4) by stochastic vectors and stochastic matrices. Specifically, in our implementation

- $\mu_1 \in \mathbb{R}_{\geq 0}^{|S^1|}$ represents the initial belief state $p\left(s_1^1\right)$. The ith entry of μ_1 corresponds to the agent's subjective uncertainty that the non-observable state component takes on value $s_1^1 = i$ at trial $t = 1$. Formally,

$$\mu_1 := \begin{pmatrix} p\left(s_1^1 = 1\right) \\ p\left(s_1^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5. \end{pmatrix} \tag{A.6}$$

- $\mu_t \in \mathbb{R}_{\geq 0}^{|S^1|}$ represents the belief state $p^{a_{1:t-1}}\left(s_t^1|o_{1:t-1}\right)$ at trial t. The ith entry of μ_t corresponds to the agent's subjective uncertainty that the non-observable state component takes on value $s_t^1 = i$ at trial t given the history of observations $o_{1:t-1}$ and actions $a_{1:t-1}$. Formally,

$$\mu_t := \begin{pmatrix} p^{a_{1:t-1}}\left(s_t^1 = 1|o_{1:t-1}\right) \\ p^{a_{1:t-1}}\left(s_t^1 = 2|o_{1:t-1}\right) \end{pmatrix}. \tag{A.7}$$

- $\Phi \in \mathbb{R}_{\geq 0}^{|S^1|\times|S^1|}$ represents the state-state transition distribution $p\left(s_{t+1}^1|s_t^1\right)$. The jth entry of the ith row of Φ corresponds to the agent's subjective uncertainty that the non-observable state component takes on the value $s_{t+1}^1 = j$ in trial $t+1$ given that $s_t^1 = i$ in trial t. Formally,

$$\Phi := \begin{pmatrix} p\left(s_{t+1}^1 = 1|s_t^1 = 1\right) & p\left(s_{t+1}^1 = 2|s_t^1 = 1\right) \\ p\left(s_{t+1}^1 = 1|s_t^1 = 2\right) & p\left(s_{t+1}^1 = 2|s_t^1 = 2\right) \end{pmatrix} = \begin{pmatrix} 0.9625 & 0.0375 \\ 0.0375 & 0.9625 \end{pmatrix}. \tag{A.8}$$

- $\Omega^{a_t} \in \mathbb{R}_{\geq 0}^{|S^1|\times|O|}$ represents the action-dependent state-conditional observation distribution $p^{a_t}\left(o_t|s_t^1\right)$ for action $a \in A$. The kth entry of the ith row of $\Omega^{a_t=a}$ corresponds to the agent's subjective uncertainty that the observation takes on the value $o_t = k$ given that the non-observable state component takes on the value $s_t^1 = i$ and action the value $a_t = a$. Formally,

$$\begin{aligned}\Omega^{a_t=1} :&= \begin{pmatrix} p^1\left(o_t = 1|s_t^1 = 1\right) & p^1\left(o_t = 2|s_t^1 = 1\right) & p^1\left(o_t = 3|s_t^1 = 1\right) \\ p^1\left(o_t = 1|s_t^1 = 2\right) & p^1\left(o_t = 2|s_t^1 = 2\right) & p^1\left(o_t = 3|s_t^1 = 2\right)\end{pmatrix} \\ &= \begin{pmatrix} 0.15 & 0.85 & 0 \\ 0.85 & 0.15 & 0 \end{pmatrix} \end{aligned} \tag{A.9}$$

and

$$\begin{aligned}\Omega^{a_t=3} :&= \begin{pmatrix} p^3\left(o_t = 1|s_t^1 = 1\right) & p^3\left(o_t = 2|s_t^1 = 1\right) & p^3\left(o_t = 3|s_t^1 = 1\right) \\ p^3\left(o_t = 1|s_t^1 = 2\right) & p^3\left(o_t = 2|s_t^1 = 2\right) & p^3\left(o_t = 3|s_t^1 = 2\right)\end{pmatrix} \\ &= \begin{pmatrix} 0.85 & 0.15 & 0 \\ 0.15 & 0.85 & 0 \end{pmatrix} \end{aligned} \tag{A.10}$$

represent the action-dependent state-conditional observation distribution for the informative actions, and

$$\begin{aligned}\Omega^{a_t\in\{2,4\}} :&= \begin{pmatrix} p^{a_t}\left(o_t = 1|s_t^1 = 1\right) & p^{a_t}\left(o_t = 2|s_t^1 = 1\right) & p^{a_t}\left(o_t = 3|s_t^1 = 1\right) \\ p^{a_t}\left(o_t = 1|s_t^1 = 2\right) & p^{a_t}\left(o_t = 2|s_t^1 = 2\right) & p^{a_t}\left(o_t = 3|s_t^1 = 2\right)\end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned} \tag{A.11}$$

represent the action-dependent state-conditional observation distribution for the non-informative actions.

- $\Psi^{a_t} \in \mathbb{R}_{\geq 0}^{|S^1|\times|R|}$ represents the action-dependent state-conditional reward distribution $p^{a_t}\left(r_t|s_t^1\right)$ for action $a \in A$. The lth entry of the ith row of $\Psi^{a_t=a}$ corresponds to the agent's subjective uncertainty that the reward takes on the value $r_t = l - m$ given that the non-observable state component takes on the value $s_t^1 = i$ and action the value $a_t = a$. Note that m is introduced to convert the linear indices to reward values and takes on the value 2 if $l = 1$ and the value 1 if $l = 2$. Formally,

$$\Psi^{a_t\in\{1,2\}} := \begin{pmatrix} p^{a_t}\left(r_t = -1|s_t^1 = 1\right) & p^{a_t}\left(r_t = +1|s_t^1 = 1\right) \\ p^{a_t}\left(r_t = -1|s_t^1 = 2\right) & p^{a_t}\left(r_t = +1|s_t^1 = 2\right)\end{pmatrix} = \begin{pmatrix} 0.15 & 0.85 \\ 0.85 & 0.15 \end{pmatrix} \tag{A.12}$$

represent the action-dependent state-conditional reward distribution for the actions of choosing the square and,

$$\Psi^{a_t \in \{3,4\}} := \begin{pmatrix} p^{a_t}\left(r_t = -1 | s_t^1 = 1\right) & p^{a_t}\left(r_t = +1 | s_t^1 = 1\right) \\ p^{a_t}\left(r_t = -1 | s_t^1 = 2\right) & p^{a_t}\left(r_t = +1 | s_t^1 = 2\right) \end{pmatrix} := \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{pmatrix} \tag{A.13}$$

represent the action-dependent state-conditional reward distribution for the actions of choosing the triangle. Accordingly, $\sum_{l=1}^{|R|} \Psi_{i(l-m)}^{a_t} = 1$.

Belief state Using matrix multiplication (denoted as $\cdot$) as well as element-wise Hadamard matrix multiplication (denoted as $\circ$), the agent's prior belief state at trial t (cf. eq. 2.11) can be written as

$$\mu_t := \tilde{\mu}_t \cdot \left(\sum_{i=1}^{|S^1|} \tilde{\mu}_{t_i} \right)^{-1}, \tag{A.14}$$

where

$$\tilde{\mu}_t := \Phi \cdot \left(\Omega_k^a \circ \mu_{t-1} \right) \tag{A.15}$$

is the unnormalized belief state following action $a_{t-1} = a$ and observation $o_{t-1} = k$ and

$$\left(\sum_{i=1}^{|S^1|} \tilde{\mu}_{t_i} \right)^{-1} \tag{A.16}$$

is the normalization constant. In eq. A.15, Ω_k^a denotes the kth column of Ω^a and μ_{t-1} denotes the prior belief state on trial $t-1$, which corresponds to eq. A.7 if $t-1 > 1$ and to eq. A.6 if $t-1 = 1$.

Posterior predictive distribution Based on the stochastic matrix representation of the action-dependent state-conditional observation distribution Ω^{a_t} and the belief state μ_t, the posterior predictive distribution of A2 (cf. eq. 2.19) can be implemented using matrix multiplication as

$$\omega_t := \left(\Omega^a\right)^T \cdot \mu_t \tag{A.17}$$

for action $a_t = a$.

A.6 References

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II.* San Antonio, TX: Psychological Corporation.

Conners, C. K., Erhardt, D., & Sparrow, E. (1999). *CAARS. Adult ADHD Rating Scales.* Toronto, Ontario, Canada: Multi-Health Systems.

Lynam, D. R., Smith, G. T., Whiteside, S. P., & Cyders, M. A. (2006). *The UPPS-P: Assessing five personality pathways to impulsive behavior*. West Lafayette, IN: Purdue University.

Nooner, K. B., Colcombe, S., Tobe, R., Mennes, M., Benedict, M., Moreno, A., Panek, L., Brown, S., Zavitz, S., Li, Q., et al. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, *6*, 152.

Spielberger, C., Gorsuch, R., Lushene, R., Vagg, P., & Jacobs, G. (1983). *Manual for the state-trait anxiety inventory.* Palo Alto, CA: Consulting Psychologists Press.

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence.* The Psychological Corporation: Harcourt Brace & Company. New York, NY.

B | Supplementary material to Chapter 3

B.1 Task instructions

Participants were provided with the following instructions about the treasure hunt task (translated from the original German instructions):

Welcome to the treasure hunt task! The goal of this task is to find two hidden treasures in a 5-by-5 cell grid-world within a limited number of steps. You can navigate through the grid-world using the buttons. In each trial, you will first see the cell corresponding to your current position. To help you track your position, you will additionally be presented with the index of your current position. For example, (2,3) means that you are in the second row and in the third column. You start each attempt in (1,1), i.e., the first row and the first column, which corresponds to the upper left corner.

The right panel of Supplementary Figure B.1 displays the grid-world with two treasures from above. We are showing you this figure here to make you familiar with the layout. During the experiment, however, you will not see the grid-world from a bird's eye perspective. The left panel of Supplementary Figure B.1 displays how you will see your position in the grid-world during the experiment.

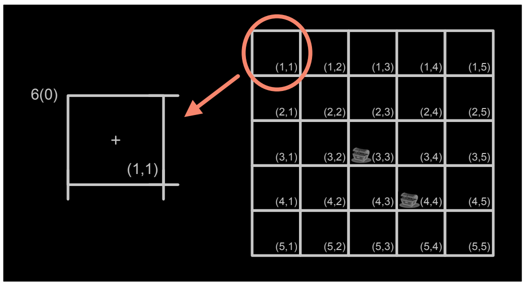

Figure B.1. Treasure hunt task instructions 1. The figure displays the grid-world from above (right panel) and a grid cell position (left panel) as presented to the participants in the instructions.

After you are presented with your position, you will see a light or a dark grey bar in each available direction (Supplementary Figure B.2). The bars convey *information about the location of the treasures and can be interpreted as noisy signals of a treasure detector. Light bars indicate directions that potentially lead you closer to a treasure. Dark bars indicate directions that potentially lead you away from the treasures. In the directions that lead you away from the treasures you will always receive the correct information. That is, in these directions you will always see a dark bar. In the directions that lead you closer to a treasure you will sometimes receive false information. That is, in these directions you will sometimes see a light bar and sometimes see a dark bar. Consequently, light bars always indicate a good direction. Dark bars, on the other hand, can indicate either a good or a bad direction.*

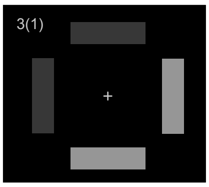

Figure B.2. Treasure hunt task instructions 2. The figure displays the light and dark grey bars of the "treasure detector" as presented to the participants in the instructions.

In the final stage, you can decide where to move in the grid-world. To this end, you will be presented with arrows pointing towards the available directions (Fig 1c). You will have approximately 4 seconds to indicate your decision using the buttons. If you do not make a decision, you will stay at your current position. Top left, you will see the number of remaining steps in your current attempt. In parentheses you will see the number of treasures you have found so far in your current attempt.

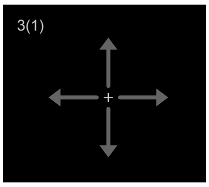

Figure B.3. Treasure hunt task instructions 3. The figure displays the arrows prompting for a decision as presented to the participants in the instructions.

A task is considered to be solved if you find both treasures within a single attempt. Each attempt will consist of a limited number of steps. The number of available steps will sometimes suffice to find the treasures, but sometimes it will not. You will have three attempts to solve a task. For each task, the treasure locations will be randomly assigned. The treasure locations will remain unchanged in all three attempts of a task. If you solve a task in fewer than three attempts, the next task begins immediately.

If you solve a task, you will see the following message printed on the screen: "Both targets found - Creating new task." (That is, you will be presented with a new task.) If in an attempt you exhaust the number of available steps but do not find the treasures you will see the following message: "Step limit reached - Resetting position" (That is, your position will be reset to the start position.) If you do not solve a task within three attempts, you will see the following message: "Step limit reached – Attempt limit reached – Creating new task" (That is, you will be presented with a new task.) A run consists of four tasks and takes approximately 10-15 minutes. After the fourth task, you will get a message informing you about the end of the run: "Task limit reached – Ending program"

Do you have any questions? Good luck and have fun!

B.2 FMRI result tables

We applied a model-based general linear model (GLM) approach to identify the neural substrates of the cognitive processes involved in the treasure hunt task (Friston & Dolan, 2010). Tables B.1-B.7 summarize the results of the group-level GLM analyses of the fMRI data obtained by applying a cluster forming threshold of $p < 0.001$ (uncorrected). Anatomical cluster labels are based on the Automated Anatomical Atlas (AAL3; Rolls, Huang, Lin, Feng, and Joliot, 2020).

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
R; cerebellum	2221	0.000	20	-52	-22	10.53
L; postcentral and precentral gyri, supplementary motor area (L/R), inferior parietal gyrus, supramarginal gyrus, insula, rolandic operculum, middle cingulate gyrus (L/R), putamen, pallidum, thalamus, superior temporal gyrus	11552	0.000	-6	6	42	9.98
R; putamen, pallidum, insula	395	0.000	16	-2	2	6.44
L; middle and inferior occipital gyri, middle temporal gyrus	694	0.000	-40	-62	8	6.34
R; insula, rolandic operculum	176	0.028	44	0	4	6.29
L; precuneus, superior parietal gyrus	179	0.026	-10	-72	58	6.23
R; middle temporal and middle occipital gyri	808	0.000	40	-62	10	5.98
L; cerebellum	377	0.000	-30	-52	-30	5.86
L; middle and superior frontal gyri	431	0.000	-28	40	30	5.77
L; superior occipital gyrus, cuneus, calcarine sulcus and surrounding cortex, lingual gyrus	315	0.002	-6	-100	10	5.55
R; superior occipital gyrus, cuneus	224	0.01	30	-90	18	5.11

Table B.1. Positive main effects of valid trials. Clusters with increased activity during valid trials are entered row-wise. Anatomical labels from AAL3 and other statistics are entered column-wise. Note: R denotes the right hemisphere, L denotes the left hemisphere, and L/R denotes both hemispheres. The leading annotation applies to all regions of a given cluster unless a region is individually annotated with the hemisphere location in parentheses.

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
L/R; superior and middle occipital gyri, cuneus, calcarine sulcus and surrounding cortex, lingual and fusiform gyri	4137	0.000	12	-92	4	12.51
R; superior and middle frontal gyri	870	0.000	22	16	56	9.96
R; inferior parietal and supramarginal gyri	448	0.000	38	-38	38	6.2
R; superior parietal gyrus, precuneus	245	0.004	24	-60	54	5.17
R; middle and inferior (triangular and opercular parts) frontal gyri	169	0.026	36	36	26	5.02

Table B.2. Positive main effects of luminance. Activity clusters positively relating to the trial-by-trial luminance value. Note: For table conventions, please see the legend of Table B.1.

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
L/R; lateral and medial surfaces of the occipital, temporal, parietal and frontal lobes, cingulate cortex, insula, rolandic operculum, amygdala, thalamus, basal ganglia, ventral tegmental area, cerebellum	84128	0.000	-30	20	-14	15.8

Table B.3. Positive main effects of treasure discovery. Using a cluster defining threshold of $p < 0.001$, we obtained a single large bilateral cluster with increased activity in response to treasure discovery as compared to implicit baseline. Note: For table conventions, please see the legend of Table B.1.

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
L/R; lateral and medial surfaces of the occipital, temporal, parietal and frontal lobes (except the right superior parietal gyrus), cingulate cortex, insula, rolandic operculum, amygdala, thalamus, basal ganglia (except ventral striatum), cerebellum	68694	0.000	0	-22	38	13.55

Table B.4. Positive main effects of information display. As in the case of the positive main effects of treasure discovery, applying a cluster defining threshold of $p < 0.001$ resulted in a single large bilateral cluster with increased activity in response to information display as compared to implicit baseline. Note: For table conventions, please see the legend of Table B.1.

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
R; middle and superior frontal gyri, superior medial frontal gyrus (L/R), inferior frontal gyrus triangular and opercular parts	3507	0.000	46	28	28	9.47
R; superior and inferior parietal gyri (L/R), angular gyrus, precuneus (L/R), lingual gyrus, inferior temporal gyrus	7032	0.000	30	-68	46	7.4
L; inferior (triangular, opercular and orbital parts), middle and superior frontal gyri, precentral gyrus	2425	0.000	-24	10	48	7.35
R; caudate, putamen, pallidum	291	0.001	18	16	0	6.05
L; cerebellum	154	0.023	-34	-54	-32	5.8
L; inferior occipital gyrus, fusiform and lingual gyri, cerebellum	575	0.000	-44	-76	-14	5.71
R; cerebellum	150	0.026	34	-56	-34	5.22

Table B.5. Positive main effects of Bayesian surprise. Activity clusters positively relating to the trial-by-trial Bayesian surprise. Note: For table conventions, please see the legend of Table B.1.

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
L; precuneus, calcarine sulcus and surrounding cortex	189	0.009	-12	-34	20	5.38

Table B.6. Negative main effects of Bayesian surprise. Activity clusters negatively relating to the trial-by-trial Bayesian surprise. Note: For table conventions, please see the legend of Table B.1.

Region	**Cluster size (number of voxels)**	**Cluster-level p-value (FWE-corrected)**	**Peak voxel MNI coordinates (mm)**			**Peak voxel t-value**
			x	**y**	**z**	
R; inferior occipital gyrus, lingual and fusiform gyri, calcarine sulcus and surrounding cortex, cerebellum, inferior temporal gyrus	1856	0.000	20	-84	-6	9.48
L; inferior occipital gyrus, lingual and fusiform gyri, middle occipital gyrus, inferior temporal gyrus, cerebellum	1240	0.000	-20	-96	-10	8.84
R; middle and superior temporal gyri	391	0.001	56	-20	-4	6.32
L/R; pregenual and subgenual anterior cingulate cortex, superior medial frontal gyrus	572	0.000	6	44	12	6.17
R; insula, posterior orbital gyrus, inferior frontal gyrus orbital part	373	0.001	38	14	-20	5.98
L; middle and superior temporal gyri, supramarginal gyrus	173	0.046	-52	-28	22	4.94

Table B.7. Positive main effects of chosen action valence. Activity clusters positively relating to the trial-by-trial chosen action valence. Note: For table conventions, please see the legend of Table B.1.

B.3 References

Friston, K., & Dolan, R. J. (2010). Computational and dynamic models in neuroimaging. *NeuroImage*, *52*(3), 752–765.

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., & Joliot, M. (2020). Automated anatomical labelling atlas 3. *NeuroImage*, *206*, 116189.

List of manuscripts

Chapters 2 and 3 of this dissertation served as the basis for manuscripts submitted to or prepared for peer-reviewed journals. These chapters are thus not identical with the manuscripts listed below.

- Horvath, L., Colcombe, S., Milham, M., Ray, S., Schwartenbeck, P., & Ostwald, D. (2020). Human belief state-based exploration and exploitation in an information-selective reversal bandit task. *bioRxiv*, https://doi.org/10.1101/2020.08.31.276139.

  Authors' contributions: L.H.: conceptualization, project administration, data curation, investigation, methodology, formal analysis, software, validation, visualization, writing – original draft, writing – review & editing; S.C.: project administration, data curation, investigation, experimental software, funding acquisition, resources, supervision, writing – review; M.M.: conceptualization, funding acquisition, resources, supervision, project administration, writing – review; S.R.: project administration, data curation, investigation, writing – review; P.S.: conceptualization, writing – review; D.O.: conceptualization, formal analysis, methodology, software, visualization, project administration, resources, validation, funding acquisition, supervision, writing – original draft, writing – review & editing

- Horvath, L., Tisdall, L., Mata, R., Hertwig, R., & Ostwald, D. (in preparation). The neurocomputational mechanisms of sequential decision making in a multistep task with partially observable states.

  Authors' contributions: L.H.: conceptualization, project administration, data curation, investigation, methodology, formal analysis, software, validation, visualization, writing – original draft, writing – review & editing, L.T.: investigation, R.M.: project administration, R.H.: funding acquisition, resources, D.O.: conceptualization, formal analysis, methodology, software, visualization, investigation, project administration, resources, validation, funding acquisition, supervision, writing – original draft, writing – review & editing

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt,

- dass ich die vorliegende Arbeit eigenständig und ohne unerlaubte Hilfe verfasst habe,
- dass Ideen und Gedanken aus Arbeiten anderer entsprechend gekennzeichnet wurden,
- dass ich mich nicht bereits anderweititg um einen Doktorgrad beworben habe und keinen Doktorgrad in dem Promotionsfach Psychologie besitze, sowie
- dass ich die zugrundeliegende Promotionsordnung vom 08.08.2016 anerkenne.

Lilla Horvath
Berlin, 9. April 2021