Mechanisms of reinforcement learning and decision making in different environments

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Hannah Kurtenbach
aus Bergisch Gladbach

Düsseldorf, Oktober 2024

aus dem Institut für Experimentelle Psychologie,
Biologische Psychologie des Entscheidungsverhaltens
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Gerhard Jocham

2. PD Dr. Jan Hirschmann

Tag der mündlichen Prüfung: 16. Dezember 2024

Preface

This dissertation comprises my work in the Institute of Experimental Psychology, Biological Psychology of Decision Making, at the Heinrich Heine University Düsseldorf from May 2019 to October 2024. The content of this dissertation is based on the following publications:

- **Study I**
  A role for acetylcholine in reinforcement learning and decision making under uncertainty
  Kurtenbach H, Froböse MI, Ort E, Bahners BH, Hirschmann J, Butz M, Schnitzler A, & Jocham G
  bioRxiv (2024)
  Digital Object Identifier (DOI): `https://doi.org/10.1101/2024.09.20.614105`
- **Study II**
  Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations
  Kurtenbach H, Ort E, Froböse MI, & Jocham G
  PLOS Computational Biology **18(12)**, e1010201 (2022)
  Digital Object Identifier (DOI): `https://doi.org/10.1371/journal.pcbi.1010201`

Kurzfassung

Seit Jahrhunderten beschäftigt Wissenschaftler die Frage, wie Menschen Entscheidungen treffen. Jedoch kann bei der Klärung der Frage die Entscheidung nicht in Isolation betrachtet werden; die Umwelt, in der die Entscheidungen getroffen werden, muss berücksichtigt werden. Je nach Umwelt, können verschiedene Entscheidungsstrategien verwendet werden und vorteilhaft sein. Wenn zusätzlich relevante Informationen fehlen, um eine Entscheidung zu treffen, spielt auch Lernen durch Ausprobieren eine tragende Rolle im Entscheidungsprozess, da Feedback in zukünftige Entscheidungen integriert werden muss. Der Entscheidungsprozess ist ein sensibles Konstrukt, das, wenn es aus der Balance gerät, häufig mit psychiatrischen Erkrankungen in Verbindung steht. Verantwortlich für die Aufrechterhaltung dieser Balance sind unter anderem Neurotransmitter und Neuromodulatoren. Welche Rolle sie genau in Lernen und Entscheidungsfindung spielen, ist jedoch nicht vollständig geklärt. Diese Dissertation befasst sich mit der Frage, wie Verhalten in Abhängigkeit von der Verlässlichkeit von Informationen und in Abhängigkeit von der Präsenz (oder Absenz) von Feedback angepasst wird. Neben den behavioralen Mechanismen haben wir zudem untersucht, welche Rolle der Neuromodulator Acetylcholin in den Entscheidungs- und Lernprozessen spielt. Hierzu wurden zwei Studien mit gesunden Probanden durchgeführt: In der ersten Studie führten die Probanden zwei Aufgaben nach der Verabreichung des muskarinergen Acetylcholin-Antagonisten Biperiden aus. Ihr Ziel war es, den Gewinn zu maximieren, indem sie sich für eine von zwei Optionen mit verschiedenen Gewinnhöhen und -wahrscheinlichkeiten entschieden. Allerdings waren nur in einer Aufgabe alle relevanten Informationen gegeben, während in der anderen Aufgabe Gewinnwahrscheinlichkeiten erlernt werden mussten und diese über den Verlauf der Aufgabe variierten. In der zweiten Studie führten Probanden eine weitere Aufgabe aus, in der sie die Assoziation zwischen visuellen Stimuli und entsprechenden Aktionen anhand von Feedback lernten, das während des Lernprozesses zeitweise entfernt wurde. Um die Daten

zu analysieren, verwendeten wir verschiedene komputationale Modellierungen mit dem Ziel die zugrundeliegenden Strategien und deren Veränderungen über die verschiedenen Bedingungen aufzudecken. Insgesamt zeigen unsere Ergebnisse, dass Entscheidungsstrategien entsprechend der Umwelt angepasst werden. Probanden verließen sich weniger auf Informationen, die mit Unsicherheit behaftet waren. Interessanterweise beeinträchtigte Biperiden die Schätzung unsicherer Optionsattribute, was auf eine maladaptiv erhöhte Lernrate zurückzuführen ist. Zudem fanden wir, dass die Performanz in der Abwesenheit von Feedback verbessert war, dieser Effekt allerdings nur zustande kam, weil Probanden vorsichtiger antworteten, was in diesem Kontext vorteilhaft war. Diese Dissertation trägt zum allgemeinen Verständnis von Entscheidungsstrategien sowie dem Wissensstand über den Effekt von Acetylcholin auf Verhalten bei.

Abstract

For centuries, the question of how humans make decisions has been a subject of scientific research. When addressing the question, however, a decision cannot be considered in isolation; the environment in which decisions are made must be taken into account. Depending on the environment, different decision strategies can be used and be advantageous. If, in addition, relevant information is lacking to form a decision, learning by trial and error also plays a key role in the decision-making process, as feedback must be integrated into future decisions. The decision-making process is a sensitive construct which, when out of balance, is often associated with psychiatric disorders. Neurotransmitters and neuromodulators are, among others, responsible for maintaining this balance. However, their exact role in learning and decision making is not fully understood. This dissertation focuses on the question of how behaviour is adapted depending on the reliability of information and the presence (or absence) of feedback. In addition to the behavioural mechanisms, we also investigated the role of the neuromodulator acetylcholine in decision-making and learning processes. To this end, two studies were conducted with healthy volunteers: In the first study, participants performed two tasks after administration of the muscarinic acetylcholine antagonist biperiden. Their goal was to maximise the gain by choosing one of two options with different reward magnitudes and probabilities. However, only in one task, all the relevant information was given, while in the other task, reward probabilities had to be learnt and these varied over the course of the task. In the second study, participants performed another task in which they learnt the association between visual stimuli and corresponding actions using feedback that was at times removed during the learning process. To analyse the data, we used different computational modelling approaches with the aim of uncovering the underlying strategies and their changes across conditions. Overall, our results show that decision strategies are adapted according to the environment. Participants relied less on information that was associated with

uncertainty. Interestingly, biperiden impaired the estimation of uncertain option attributes resulting from maladaptively increased learning rates. In addition, we found that performance improved in the absence of feedback, but this effect only emerged because participants responded more cautiously, which happened to be beneficial in this specific context. This dissertation contributes to the general understanding of decision strategies and the state of knowledge about the effect of acetylcholine on behaviour.

Contents

1 Introduction

Understanding the mind is a truly interdisciplinary endeavour. It ranges from studying fundamental cellular and network processes in biology and physiology to describing how emotions and personal preferences influence decisions e.g. in psychology, economy and philosophy. Remarkably, mathematics and physics significantly contribute to this field, helping to understand psychological concepts: For instance, the physical description of electrical activity is used for modelling physiological processes, neuronal activity, and network mechanisms within the brain (Soltani and Wang, 2006; Wang, 2002; Wong and Wang, 2006). This approach facilitates the understanding of how neurons communicate and form networks that underpin various cognitive functions. More abstractly, computational modelling based on mathematical concepts and models from thermodynamics and statistical mechanics are employed to comprehend complex behaviours and cognitive processes in humans and animals, such as decision making. For example, evidence accumulation in favour or against a choice option during decision making is often modelled as a drift-diffusion process, i.e., Brownian motion with drift - the phenomenon of a particle moving randomly in a medium (liquid or gas) under the additional influence of an external force (Ratcliff, 1978; Ratcliff and McKoon, 2008). Moreover, variability in human choices due to uncertainty or incomplete information is modelled using the softmax function, which is mathematically related to the Boltzmann distribution; the Boltzmann distribution describes the probability of a system being in a certain state, depending on the energy of this state and the temperature of the system (Luce, 1959; McFadden, 1974).

This dissertation aims to improve our understanding of how healthy adults learn and make decisions in different environments by combining computational models with experimental work. Specifically, in two projects, one of which involves pharmacological manipulation, we applied computational models that describe the dynamic interaction between an agent and its environment, providing a powerful

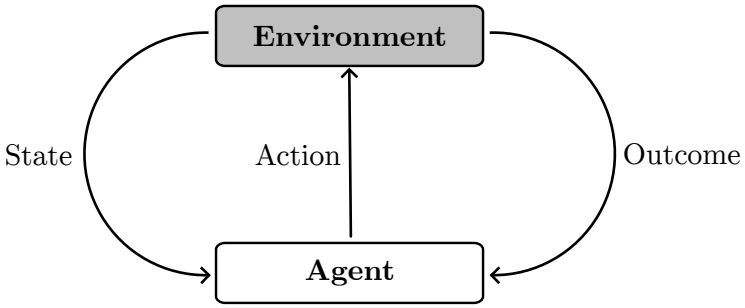

Figure 1.1: Schematic illustration of the interaction between an environment and an agent. The environment provides information for the agent through its current state. The agent observes this state and makes a decision. The selected action is then sent to the environment, which results in an outcome.

tool for understanding and predicting future decisions (Collins and Cockburn, 2020).

The interaction between an environment and an agent begins with a *state* (Figure 1.1), which provides environmental information to the agent. Usually, the agent can select between two or more *actions* to act on the state. For example, when deciding what to eat for dinner, you might consider cooking at home, ordering takeout, or dining at a restaurant. The choice depends on many factors, such as satisfaction, convenience, and cost. The question of how a human selects an action has been a topic of research for centuries. Modern decision theory traces back to 1654, when mathematicians Blaise Pascal and Pierre de Fermat elaborated the so-called problem of points (de Fermat et al., 1891). This classical problem in probability theory involves a gambling game played over several rounds, where two players have equal chances of winning money, awarded to the first player to win n times. The challenge arises when the game is interrupted prematurely: How should the prize money be fairly divided? This problem is directly applicable to decision-making scenarios where an agent must choose between two possible actions. Pascal and Fermat proposed a solution by evaluating the potential outcomes of each action and their associated probabilities to calculate an average expected outcome, called the *expected value*. The economically optimal choice is the action with the highest expected value of reward (Rachlin et al., 1983; Sutton and Barto, 2018). In several studies it has been found that the expected value of choice options drives reward-guided decision making (e.g. Dias Maile et al. (2024); Farashahi et

al. (2019); Jocham et al. (2012)). Neuroimaging studies complement behavioural findings by pointing towards neural representations of expected value in several brain regions during learning and reward-guided decision-making tasks, including the ventromedial prefrontal cortex (vmPFC), posterior parietal cortex, orbitofrontal cortex, and amygdala (Dorris and Glimcher, 2004; Gottfried et al., 2003; Jocham et al., 2014; Padoa-Schioppa and Assad, 2006). After the agent decided between options and performed an action, the environment provides feedback in the form of an *outcome*. In reinforcement learning (RL) frameworks, agents use the outcome to evaluate the consequences of their actions and gradually ascertain the probabilities of different possible actions leading to desired rewards, as described by *instrumental learning* (Skinner, 1938; Thorndike, 1927). Computationally, this learning process is implemented as *prediction error*, which represents the discrepancy between the expected value of an action and the actual reward received (Sutton and Barto, 2018). For instance, if you decided to order takeout and the food exceeds your expectations in taste and quality, the prediction error is positive, making it more likely that you order food from that restaurant again. Conversely, if the food falls short of your expectations, the negative prediction error will prompt you to reassess and possibly alter your future dining decisions. That the prediction error is not only a computational construct is supported by neuronal data. In 1994, a remarkable similarity between prediction errors and dopaminergic neural firing rates was discovered in primates (Mirenowicz and Schultz, 1994). Subsequent studies have confirmed the relationship between dopamine and error-driven learning mechanisms in humans with the computation of prediction errors being identified in the ventral striatum and ventral tegmental area (D'Ardenne et al., 2008; Pessiglione et al., 2006).

In summary, reinforcement learning and decision making are deeply interconnected, as learning influences decision making by providing a repository of past experiences and knowledge about the environment, which in turn guides future choices. Key aspects of computational modelling, such as RL algorithms, correspond well with behavioural and neuronal data, thereby helping to investigate the underlying cognitive processes during decision making across different environmental states.

1.1 Decision strategies

Goal-directed, or reward-guided, decision making is characterised by comparing the potential outcomes of each action and selecting the action which is most likely to generate a desired outcome (Daw and O'Doherty, 2013). In real life, the environment is often fraught with *risk* or *uncertainty*. Under risky conditions, although the probability distribution of a future outcome is known, the individual outcome is not. During decision making under uncertain conditions, the probability distribution is also not known (Knight, 1921). Different decision strategies are required for optimal performance under risk and uncertainty. Hence, computational models must incorporate different decision strategies based on the underlying environment to accurately reflect and explain human decision making.

1.1.1 Risk

A decision under risk typically involves deciding between options, each with known reward magnitude and associated probability. Roulette is a classical example: The stake can be placed on colours, number ranges, or specific numbers, with the selected bet determining both the reward magnitude and the reward probability. The expected value EV of each option's reward is calculated as product of reward magnitude M and probability P:

$$EV = MP \tag{1.1}$$

The EV varies significantly between different bets, indicating the economically best choice: For instance, imagine playing with a €10 stake and betting on red, 18 out of 37 fields potentially yield a reward, but the payout is only $1:1$. This results in a reward magnitude of €10 with reward probability of about 48.7 %. On the other hand, betting on number 26 means only one out of 37 fields offer a potential reward, but the payout is significantly higher at $35:1$. With the same stake of €10, this leads to a reward magnitude of €350 with a probability of around 2.7 %. Consequently, the EV of betting on a specific number is almost double that of betting on a colour.

But even if the bet on a specific number might be objectively the best, why take the risk if the probability of losing money is higher than for other bets? According

to Daniel Bernoulli, these decisions depend on the *utility* (Bernoulli, 1954): The less money a person has, the more utility they gain from winning more money. Thus, the value of each option is subjective and depends on the reference. How human decision making deviates from economically optimal choices is described in the *prospect theory*, one of the most influential theories for choice behaviour under risk. Among others, it describes that losses have a greater emotional impact than equivalent gains, such that humans avoid losses more vigorously than they seek rewards (Kahneman and Tversky, 1979). This results in risk-averse choices favouring known, lower rewards over unknown, higher rewards (Ellsberg, 1961; Tversky and Kahneman, 1992).

Expected value theory, prospect theory, and many other decision theories assume that the integration of option attributes into a subjective value is multiplicative (see Equation 1.1) (Bernoulli, 1954; Birnbaum, 2008; Busemeyer and Townsend, 1993; Kahneman and Tversky, 1979; Quiggin, 1982), allowing to fuse all attributes into one value without differential weighting (Farashahi et al., 2019). This is also supported by evidence from studies using decision making under risk (Farashahi et al., 2019; Molter et al., 2022; Tversky, 1967). Correlates of subjective values in risky reward-guided decision making tasks are also identified in the human brain. For example, the difference between subjective values of two options were found to correlate with activity in the vmPFC (Boorman et al., 2009; Hunt et al., 2012; Jocham et al., 2014). Additionally, vmPFC activity covaried positively with the value of the chosen option, but was independent of motor preparation (Wunderlich et al., 2012, 2009, 2010). Together this suggests, that the vmPFC plays a key role in valuation, independent of action selection.

1.1.2 Uncertainty

Uncertainty is characterised by incomplete knowledge about the environmental state (Scholz, 1983). However, uncertainty can have different forms. For example, consider ordering food: You might order from an unfamiliar restaurant and after five good experiences, it has become one of your favourite restaurants, as the food quality, although it varies slightly, is on average very high and gives you an estimate of what to expect when ordering again. The associated variability is classified as *expected uncertainty*. Expected uncertainty refers to situations

where the probability of outcomes can be anticipated based on past experiences or available information (Yu and Dayan, 2005). However, the sixth time you order food from this restaurant, the food is disgusting. This would be ascribed to *unexpected uncertainty*. Unexpected uncertainty arises from unforeseen events or changes that are difficult to predict or quantify, such as a new chef in your favourite restaurant, requiring adaptive responses and flexible strategies (Soltani and Izquierdo, 2019; Yu and Dayan, 2005). In another scenario a restaurant alternates between chefs, leading to inconsistent food quality over time. This type of frequent fluctuation is sometimes considered as unexpected uncertainty, but theoretical work suggests a distinct term, coined *volatility* (Bland and Schaefer, 2012). Each of the three scenarios demand different decision strategies (Farashahi et al., 2019).

In order to adapt choice behaviour, agents need to learn uncertain attributes. Beliefs about an option's value are updated in light of receiving evidence (Eckstein et al., 2022). In reinforcement learning models, the *prediction error PE* computationally implements this approach:

$$PE_t = r_t - Q_t \tag{1.2}$$

with the obtained outcome r_t (i.e. $r_t = 1$ for reward, $r_t = 0$ for no reward, $r_t = -1$ for punishment) and the estimated value Q_t of the unknown component at time point t (Sutton and Barto, 2018). The influence of the prediction error on the updated value Q_{t+1} is weighted by the *learning rate* λ in a Q-learning approach with a delta update rule (Rescorla and Wagner, 1972):

$$Q_{t+1} = Q_t + \lambda PE_t \tag{1.3}$$

The learning rate can be adjusted to suit the stochastics of the environment (Iglesias et al., 2021; Jocham et al., 2009; Soltani and Izquierdo, 2019): A low learning rate accumulates the estimate over a long run, while a high learning rate favours recent outcomes. Hence, a high learning rate enables the agent to quickly adjust to changes, being favourable in highly volatile environments. This adjustment of learning rates has been demonstrated in human behaviour (Blain and Rutledge, 2020; Browning et al., 2015) and in neuroimaging studies. Neural correlates of learning rates have been found in the anterior cingulate cortex, which reflects

volatility estimates and outcome predictions (Behrens et al., 2007; Rushworth et al., 2004).

Decisions under uncertainty often involve multi-attribute options consisting of both known and unknown attributes. For example, when deciding which restaurant to order from, not only is the (unknown or learned) quality of the food taken into account, but also the explicit price. Thus, next to learning unknown attributes for these decisions, it is required to integrate both the known and the unknown attributes (which are learnt via Equation 1.3) into a single value per option to facilitate comparison between options (Lee et al., 2012). Unlike decision making under risk, where a multiplicative integration of attributes is considered favourable (as explained in Subsection 1.1.1), a theoretical framework suggests that in uncertain environments, it can be beneficial to integrate attributes additively into a subjective value per option (Stewart, 2011). For options with reward magnitude and probability, the subjective value SV would then be calculated as follows:

$$SV = \omega_M M + \omega_P P \tag{1.4}$$

In contrast to multiplicative integration, additive integration allows for direct comparison between attributes and differential weighting of reward information via weighting parameters for magnitude ω_M and probabilities ω_P. When one attribute is unknown, additive integration offers greater flexibility by allowing the unknown attribute to be weighted less relative to the known attribute. A recent study found that both humans and non-human primates tend to adopt a more additive integration approach in environments with higher volatility (Farashahi et al., 2019). Therefore, adapting decision strategies in the context of volatility can occur either during learning or during the integration of attributes. In this thesis, we investigated how the explained decision strategies are adapted under risk and under different degrees of uncertainty.

Yet another type of uncertainty is given by the presence or absence of reinforcement. When learning is involved, feedback plays a crucial role in guiding the decision-making process. In learning experiments, feedback on a response provided by the environment is referred to as *external feedback* (Asher and Hibbard, 2020). When no explicit feedback is provided, it is also possible that agents evaluate their choices themselves, known as *internal feedback* (Ptasczynski et al., 2022). There is

evidence from the domain of perceptual learning that such an internal feedback signal does exist and is similar to the external one. Perceptual learning describes, for example, testing vision using the Landolt ring test, where patients are asked to determine where the opening of the ring is. In such tasks, performance improved over time, regardless of the presence or absence of external feedback (Asher and Hibbard, 2020; Haddara and Rahnev, 2022; Petrov et al., 2006). In some cases, performance without external feedback was even better than with external feedback (Herzog and Fahle, 1997). Imaging studies support the concept of internal feedback: In tasks without external feedback, brain activity in mesolimbic regions shows patterns similar to those seen with prediction errors following external feedback (Daniel and Pollmann, 2012; Guggenmos et al., 2016). Interestingly, when manipulating presence and absence of feedback to instrumental learning, where external feedback is required to learn the association between action and outcome, it has been observed that during initial learning feedback removal improved performance relative to phases where feedback was provided (Kuchibhotla et al., 2019). This would imply that animals use different strategies during task acquisition that do not fully reflect their latent knowledge of the task.

All in all, decision strategies depend heavily on the environmental uncertainty, such as the volatility of outcomes and the availability of reinforcement.

1.1.3 Biases

Humans and animals deviate from optimal decision behaviour, and these deviations are referred to as biases. There are various biases, that stem from factors such as attention, expectations, and reward (Cerracchio et al., 2023). Some cognitive biases can be explained by normative decision strategies, like the preference for known outcomes and the aversion to loss, and are incorporated into established decision theories (Ellsberg, 1961; Rahnev, 2021; Tversky and Kahneman, 1992). However, there are other biases that also need to be considered when developing comprehensive computational models of decision making.

The response bias reflects the systematic preference for one of the possible actions (Macmillan and Creelman, 2005). Potential outcomes can induce this bias: Prospect of reward leads to more engagement, while threat of punishment results in refraining actions (Dayan et al., 2006; Guitart-Masip et al., 2014; Swart et al.,

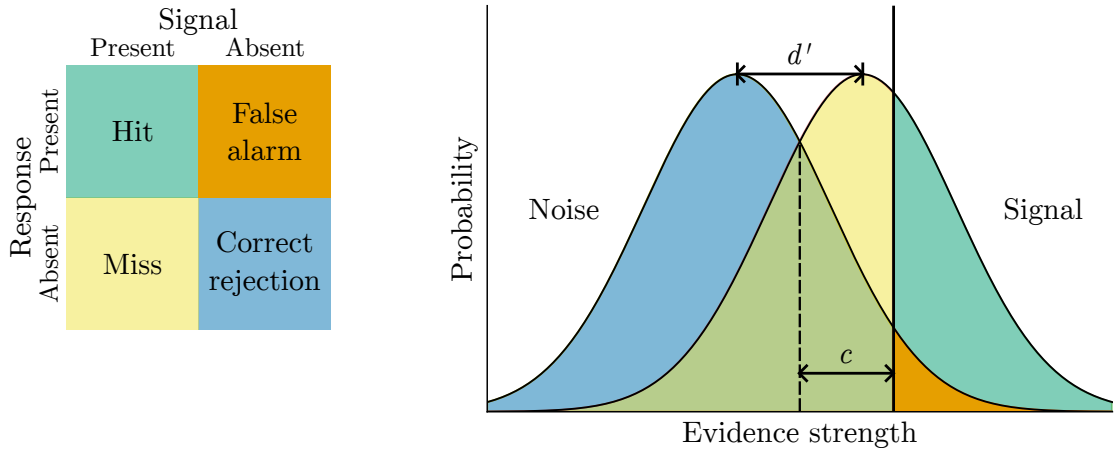

Figure 1.2: Graphical representation of the SDT. For signal and noise (absent signal), responses are categorised in hits, misses, false alarms and correct rejections (left). Evidence strength for signal and noise are given by the hit rate (reflecting the ratio between hits and misses) and false alarm rate (reflecting the ratio between false alarms and correct rejections), respectively (right). The sensitivity d' represents the difference between signal (right distribution) and noise (left distribution). Here, a positive criterion c is illustrated, leading to a conservative response strategy with less false alarms but also less hits.

2017). Moreover, the response bias is found to be more pronounced in the beginning of a task and decreases during learning (Jones et al., 2015). This suggests that the response bias may depend on the level of uncertainty associated with the available options.

In order to distinguish between informed choices and biased choices in the presence of uncertainty the *signal detection theory* (SDT) is typically applied. Originally developed to assess the performance of radar operators (Peterson et al., 1954), SDT has since been widely applied in decision making research (Green and Swets, 1966; Lynn and Barrett, 2014; Macmillan and Creelman, 2005). SDT is designed to accurately distinguish an actual signal from noise (i.e., absent signal), while also assessing potential biases towards one of the responses. These aspects are quantified using the sensitivity measure d' and the criterion c, which represents the response bias. To assess these measures, the signal and response are categorised as either present or absent (Figure 1.2). When the signal is present, the agent's response can lead to either a hit (if detected) or a miss (if not detected). Conversely, when the signal is absent (i.e., only noise is present), the agent's response can result

in either a false alarm (if incorrectly detected) or a correct rejection (if correctly not detected). Sensitivity and criterion are then calculated using the normalised hit rate $z(HR)$, which captures the ratio between hits and misses, and false alarm rate $z(FAR)$, which reflects the ratio between false alarms and correct rejections (Green and Swets, 1966):

$$d' = z(HR) - z(FAR) \tag{1.5}$$

$$c = -\frac{1}{2}(z(HR) + z(FAR)) \tag{1.6}$$

A criterion of 0 corresponds to a neutral criterion, negative values to a liberal and positive values to a conservative response strategy (Lynn and Barrett, 2014). A liberal response strategy, i.e., a negative criterion, implies a higher overall tendency to act, leading to increased detection of signals, but also more false alarms. In contrast, a conservative response strategy, i.e., a positive criterion, describes a low tendency to act implying more correct rejections, but also a higher number of misses.

1.2 The cholinergic system

Acetylcholine (ACh) was the first neurotransmitter to be discovered (Dale, 1914; Ewins, 1914), and it has since been recognised as a key neurotransmitter in both the peripheral nervous system (PNS) and central nervous system (CNS) of mammals. In the PNS, ACh serves as the primary neurotransmitter responsible for muscular movement (Katz and Miledi, 1965; Peper et al., 1982). In the CNS, ACh acts as a widely distributed neuromodulator, modulating the likelihood of synaptic release of neurotransmitters (Ananth et al., 2023). Cholinergic neurons are classified into motor neurons, interneurons, and projection neurons (Ananth et al., 2023). Cholinergic motor neurons are found in the hindbrain and spinal cord (Stifani, 2014). Cholinergic interneurons are located in the striatum, cortex, and hippocampus (Dudai et al., 2021; Higley et al., 2011; Kljakic et al., 2017). Cholinergic projection neurons are found in the brainstem and basal forebrain (Mesulam et al., 1983; Woolf, 1991). Due to its widespread distribution in the brain, ACh plays a crucial role in essential functions such as the regulation of sleep, attention, memory, and learning (Ananth et al., 2023; Everitt and Robbins, 1997). Particularly the basal

forebrain cholinergic neurons, which innervate the whole cortex and hippocampus, play a key role in behavioural functions (Ananth et al., 2023; Hasselmo and Sarter, 2011). Dysregulation of ACh transmission can have serious consequences; for instance, cholinergic dysfunction is associated with conditions like Alzheimer's disease, schizophrenia, and attention-deficit hyperactivity disorder (ADHD) (Chen et al., 2022; English et al., 2009; Higley and Picciotto, 2014). Hence, understanding cholinergic mechanisms and their causal role for cognitive processes is fundamental.

The diverse functioning of the cholinergic system is achieved by cholinergic neurons acting via two receptor classes: ionotropic nicotinic ACh receptors (nAChRs) and metabotropic muscarinic ACh receptors (mAChRs) (Ananth et al., 2023). The rapid acting nAChRs are cation channels consisting of several combinations of subunits (α2 - α10 and β2 - β4) (Dani and Bertrand, 2007; Jones et al., 1999; Role and Berg, 1996). Although nAChRs are mainly located in neuromuscular junctions due to their fast acting nature (Sine, 2012), they are also represented in the CNS with α7* and α4β2* being the most common subtypes (Dineley et al., 2015; McKay et al., 2007). Both receptor subtypes are found to be critical for memory, learning, and attention (Levin et al., 2006). The slower acting mAChRs use G proteins as signalling mechanism (Ballinger et al., 2016). There are five main receptor subtypes (M1 - M5) with M1, M3, and M5 being coupled with $G_{q/11}$ proteins and M2 and M4 being coupled with $G_{i/o}$ proteins (Thiele, 2013). The mAChRs predominate in the CNS (Carlson, 2010). M1, M2, and M4 receptors are widely distributed in the cortex and striatum and, additionally, M1-M4 receptors are prevalent in the hippocampus, while M5 receptors are mostly expressed on dopaminergic neurons in the substantia nigra and the ventral tegmental area (as reviewed in Thiele (2013)). Especially the M1 subtype is associated with behavioural functions, such as cognitive flexibility and working memory (Bradley et al., 2016; Galvin et al., 2020; Shirey et al., 2009).

In this thesis, I will focus on the muscarinic M1 receptor subtype in order to determine its role in uncertainty computations during decision making. Physiologically, the activation of muscarinic M1 receptors in the prefrontal cortex (PFC) has been shown to exert top-down control over sensory cortices by enhancing the activity of pyramidal cells in layer II/III of the visual cortex and, thereby, enhance the signal-to-noise ratio (Ballinger et al., 2016; Bentley et al., 2011; Eggermann and Feldmeyer, 2009). Behavioural studies further support this: For

example, cholinergic antagonism, which diminishes top-down control, was found to suppress post-error behavioural adjustments, while cholinergic enhancement improved stimulus detection amidst distractions (Danielmeier et al., 2015; Gratton et al., 2017). Additionally, theoretical frameworks propose that ACh modulates decision making under expected uncertainty (Avery et al., 2012; Yu and Dayan, 2005). This hypothesis is supported by a study in humans, where pharmacologically enhanced ACh levels led to faster updating of beliefs about cue validity in a spatial attention task (Vossel et al., 2014). Furthermore, pharmacological studies have demonstrated that cholinergic antagonism impairs environmental adaptation in humans and impairs reversal learning in mice (Cools and Arnsten, 2022; Marshall et al., 2016; Robbins and Roberts, 2007). Thus, ACh seems to play a critical role in uncertainty computations.

However, it remains elusive how decision making under risk, when all information is given, is affected by ACh. *N*-methyl-D-aspartate (NMDA) and γ-Aminobutyric acid (GABA) are two neurotransmitters which are typically considered to be relevant for decision making irrespective of learning. ACh has been found to modulate both of these neurotransmitters: NMDA and GABA receptor function are enhanced following activation of cholinergic M1 receptors (Bessie Aramakis et al., 1997; Kuchibhotla et al., 2017; Marino et al., 1998; Obermayer et al., 2017; Zwart et al., 2018). On a neural level, decision making is typically modelled using recurrent cortical circuit models, in which competition between options is governed via excitatory NMDA and inhibitory GABA receptor activity (Wang, 2002). In reward-guided decision-making tasks in humans, enhancement of NMDA led to more optimal decisions (Scholl et al., 2014), while higher concentrations of GABA relative to glutamate in the vmPFC led to higher decision accuracy (Jocham et al., 2012; Kaiser et al., 2021). Therefore, ACh could possibly also play a role in decision making irrespective of learning.

1.3 Hypotheses

This dissertation seeks to explain the nature of reward-guided learning and decision making in different environments in two studies.

Study I had two primary aims. The first aim was to investigate how muscarinic M1 receptor activity influences learning and decision making depending on the uncertainty of the environment. For this, we had the following hypotheses:

1. *Cholinergic antagonism of the muscarinic M1 receptor impairs learning under uncertainty.* Extensive research has shown that learning depends on ACh (Everitt and Robbins, 1997; Hasselmo and Sarter, 2011). Additionally, several studies suggest that ACh seems to be crucial for learning in uncertain environments (Avery et al., 2012; Marshall et al., 2016; Yu and Dayan, 2005).

2. *Cholinergic antagonism of the muscarinic M1 receptor leads to suboptimal information integration.* MAChR transmission potentiates GABA and NMDA receptor activity (Bessie Aramakis et al., 1997; Zwart et al., 2018). Both neurotransmitters are found to enhance information integration in risky decision-making tasks (Jocham et al., 2012; Kaiser et al., 2021; Scholl et al., 2014), thus, reduction of these neurotransmitters via blocking cholinergic receptors should result in less information integration.

The second aim was to examine the adjustment of decision strategies under risk and under different degrees of uncertainty and was accompanied by the following hypotheses:

3. *The more uncertain the environment, the more additive the information integration.* Research in both non-human primates and humans observed multiplicative information integration without uncertainty, namely under risk, while it became additive with increased volatility levels (Farashahi et al., 2019).

4. *The more uncertain the environment, the higher the learning rate.* Learning is involved in uncertain environments. Previous research suggests that the learning rate increases with the volatility level (Behrens et al., 2007; Browning et al., 2015).

Study II aimed to assess the impact of external reinforcement on learning performance and decision strategies. Concerning this aim, we had the following hypothesis:

5. *Instrumental performance during learning is increased when reinforcement is absent compared to present.* A recent study with animals showed improved instrumental performance during early learning in blocks without reinforcement compared to reinforced blocks (Kuchibhotla et al., 2019).

1.4 Overview

My dissertation is composed of two studies, of which the relevant methods are outlined in Chapter 2, key results are summarised in Chapter 3 and discussed in Chapter 4.

In the first study, we investigated the influence of muscarinic M1 receptor activity on learning and decision making and the used decision strategies under risk and different degrees of uncertainty. For this purpose, we used two reward-guided tasks, described in Subsections 2.1.1 and 2.1.2. The pharmacological intervention is specified in Section 2.2. Computational models, which are used for data analysis, are formulated in Section 2.3. The results of this project are summarised in Section 3.1.

In the second study, we investigated how learning behaviour depends on the presence versus absence of reinforcement. We used a go/no-go instrumental learning task described in Subsection 2.1.3. The computational modelling approach used for the analysis is specified in Subsection 2.3.2. The results are summarised in Section 3.2.

2 Methods

The following chapter provides an overview of key experimental and modelling methods used in the two studies. More detailed information can be found in the original work (Kurtenbach et al., 2024, 2022).

2.1 Experimental paradigms

This dissertation comprises experimental research conducted with healthy human participants, focusing on decision making and learning in variable environments. Two studies were carried out, each set up with specific experimental paradigms to address our research questions.

2.1.1 Reward-guided decision-making paradigm

In Study I, our objective was to investigate decision strategies in risky environments. To this end, we employed a reward-guided decision-making paradigm, called **gambling task** (Figure 2.1). In each trial, participants are presented with a choice between two options, each featuring two explicitly stated attributes: the reward magnitude and the probability of receiving the reward. Participants' goal is to maximise their total reward, with monetary compensation granted after the experiment based on the points they accumulated. In this task, outcomes of both options are independent of each other.

To make an informed decision, participants need to compare the two options by integrating the provided information about the reward magnitude and probability. Since both attributes are provided, participants could theoretically estimate the EV for each option and, based on the EV, make decisions that would maximise their potential rewards. The task setup allows to analyse how participants utilise the

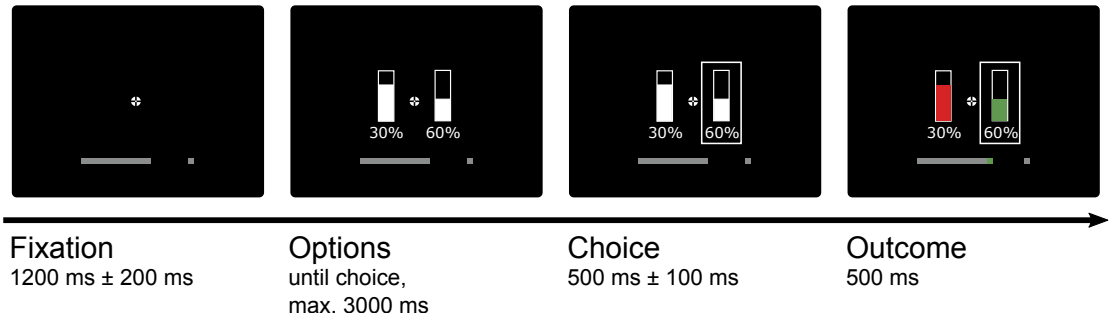

Figure 2.1: Example trial of the gambling task. The gambling task consisted of 500 trials. The height of each bar represents the reward magnitudes, while the numeric percentage below indicates the reward probabilities. The outcome of each trial, either a monetary win or no win, is shown by the fill colour — green for a win and red for no win. Adapted from (Kurtenbach et al., 2024). CC BY 4.0.

given information to make decisions under risk, revealing insights into the cognitive processes underlying risky decision making.

2.1.2 Reward-guided learning paradigm

In addition to decision making under risk, in Study I we investigated decision making under two levels of uncertainty. To accomplish this, we set up a **learning task**, similar to the gambling task described in Subsection 2.1.1 (Figure 2.2A). However, unlike the gambling task where both reward magnitude and probability are explicitly stated, in this learning task, while the reward magnitude is explicitly provided, the reward probability is not. Instead, participants have to learn the probability throughout the task, with the probability being implicitly signalled by colour: One colour represents a low reward probability of 30 %, and the other indicates a high reward probability of 70 %. In each trial, only one of the two options is guaranteed to yield a reward. In order to investigate decision making under different levels of uncertainty, the task is divided into two distinct phases: a stable phase characterised by low uncertainty, and a volatile phase characterised by high uncertainty. During the stable phase, the contingency between colour and outcome remains constant, allowing participants to form reliable expectations for each response option (i.e. colour). In contrast, during the volatile phase, reward contingencies reverse multiple times, introducing high uncertainty and requiring participants to continuously update their probability estimates based on

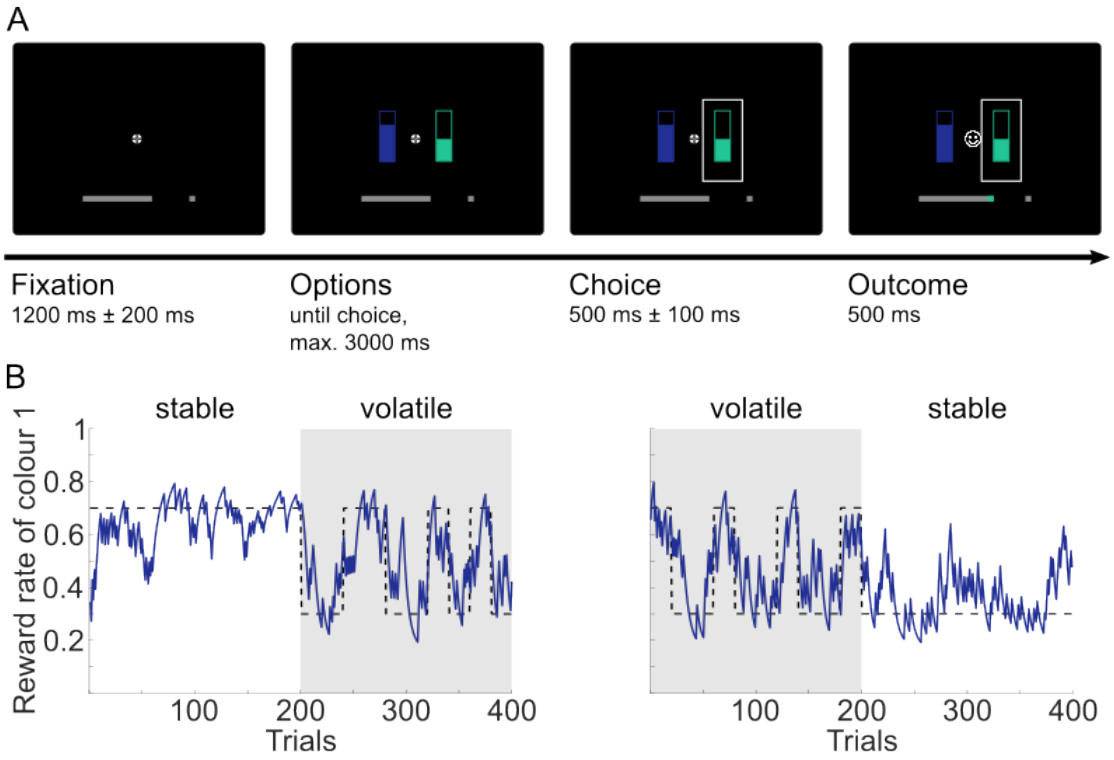

Figure 2.2: Example trial and time courses of the learning task. A The learning task consisted of 400 trials. In each trial, reward magnitudes are indicated by the height of the bar and reward probabilities by the colour of the bars and, thus, need to be learnt during the task. When the winning option is chosen, a smiley is presented, otherwise, a frowny is presented. **B** Two representative time courses of reward contingencies for the blue colour (colour 1). In the stable phase, reward contingencies remain stable and in the volatile phase, they switch several times. The true probabilities are represented as dashed black lines, probability estimates obtained from a statistically optimal Bayesian learner as blue line (Behrens et al., 2007). Adapted from Kurtenbach et al. (2024). CC BY 4.0.

the changing environment (Figure 2.2B).

Crucially, because the reward probability has to be learnt and is not explicitly given, participants cannot directly compute the EV of each option, like in the gambling task. Instead, they have to estimate the probabilities over time. Computationally, probabilities can be optimally estimated using a Bayesian learning model (Behrens et al., 2007). However, participants' behaviour can both deviate from these optimal probability estimations and the utilising of the computed EV. This allows to compare how decision strategies differ when participants are faced with known risks versus when they must learn and adapt to uncertainty.

2.1.3 Go/no-go learning paradigm

In Study II, we aimed to replicate improved performance in non-reinforced blocks observed in animals performing a go/no-go task (Kuchibhotla et al., 2019). To achieve this, we developed a visual go/no-go reinforcement learning task which induces a slow and incremental learning process (Figure 2.3). We utilised twelve abstract figurines, so-called greebles, as stimuli (Gauthier and Tarr, 2002). In each trial, one of the twelve stimuli is presented, with half of them designated as go options and the other half as no-go options. Participants have to learn, through trial and error, to press a button when presented with go stimuli and to refrain from responding when presented with no-go stimuli. Correct go responses are rewarded with both monetary gain and a positive visual cue, i.e., a smiley, while incorrect go responses are penalised with a monetary loss and a negative visual cue, i.e., a frowny. In contrast, no-go responses are neither rewarded nor punished, following the asymmetry described in the original animal study (Kuchibhotla et al., 2019).

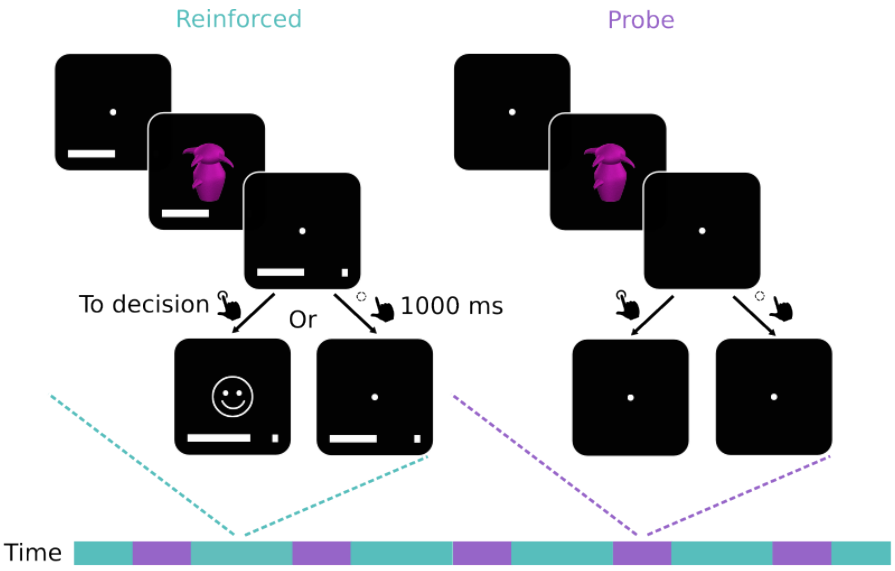

Figure 2.3: Task structure of the go/no-go learning task. Each trial starts with the presentation of a fixation cross, followed by the stimulus. Participants have to decide whether to press the button (go response) or not (no-go response). In reinforced trials (cyan), participants receive reinforcement: Correct go responses are followed by a smiley and monetary win, incorrect go responses by a frowny and monetary loss; no-go responses are not reinforced. Reinforced trials are interleaved by five non-reinforced blocks (purple), in which participants receive neither reward nor punishment for each action. Adapted from (Kurtenbach et al., 2022). CC BY 4.0.

To assess instrumental performance without reinforcement during the learning process, the reinforced trials are interspersed with several blocks of non-reinforced probe trials. These probe blocks are crucial for understanding the internalised decision strategies and the persistence of learned associations in the absence of immediate reinforcement.

2.2 Pharmacological intervention

In Study I, we applied a psychopharmacological approach: Our objective was to investigate how ACh, specifically muscarinic M1 receptor activity, affects decision making and learning within the same participants as a function of the environment. Therefore, we pharmacologically blocked ACh receptors while participants performed the two behavioural tasks. In order to reduce expectations, we set up a double-blind and randomised design. Additionally, all participants received both placebo and drug to reduce errors caused by individual differences. We used the ACh receptor antagonist biperiden, which primarily blocks muscarinic M1 receptors.

Biperiden is commonly used in the treatment of Parkinson's disease, a condition characterised by dopamine deficiency that leads to an overactive cholinergic system, resulting in excessive ACh release (Aosaki et al., 2010). This cholinergic overactivity contributes to movement disorders, such as tremors, which can be treated by pharmacologically reducing ACh levels (Brocks, 1999). Besides its clinical applications, biperiden is also suitable for pharmacological studies in healthy participants due to its relatively specific binding to M1 receptors, which reduces the risk of side effects compared to other muscarinic antagonists that also target M3 receptors. The M3 receptor subtype is found in the visual system, and its manipulation could impair stimulus perception (Bolden et al., 1992; Danielmeier et al., 2015). Additionally, biperiden effectively crosses the blood-brain barrier, allowing it to influence brain regions involved in learning and decision making (Yokagawa et al., 1992). Its pharmacokinetic properties, including a rapid peak in plasma concentration of 1 to 1.5 hours after oral administration and a short elimination half-life of about 18 to 24 hours, make it well-suited for psychopharmacological applications (Brocks, 1999; Grimaldi et al., 1986).

2.3 Computational models

The three behavioural paradigms, which are used in this dissertation, involve complex cognitive processes. In these experiments, only few behavioural variables are observable, namely choices and reaction times. While fundamental effects, like task effects, can be detected using straightforward, regression-based methods, these approaches are not suitable to capture the subtle, yet critical, components of decision-making strategies and learning mechanisms. In contrast, computational modelling offers a more sophisticated framework that generates possible strategies and mechanisms by linking observed behaviour to task-specific variables using mathematical equations (Wilson and Collins, 2019). Although computational models might not depict the ground truth, they are nevertheless useful to decipher the intricacies of cognitive processes (Eckstein et al., 2022).

There are several ways to utilise computational models. In this dissertation, we employed two approaches to analyse the behavioural data. The first involves fitting parameters to match a specific theory. This allows to reveal effects of variables, such as experimental conditions and pharmacological intervention. The second approach, the model comparison, assesses which theory best fits the observed behaviour (Wilson and Collins, 2019). In Study I we implemented the parameter fitting approach to compare behaviour across different tasks and conditions. We set up a valuation model (Subsection 2.3.1) and a reinforcement learning model (Subsection 2.3.2) and implemented these as Bayesian hierarchical model (explained in Subsection 2.3.3). In Study II, where we explored the behavioural mechanisms during a specific task, we made use of the model comparison approach. For this purpose, we introduced four different reinforcement learning models (Subsection 2.3.2) and compared model fits.

Thus, computational modelling allows us to gain a deeper understanding of decision dynamics, such as how participants weight probabilities and rewards, how they adapt their strategies over time, and how they update their beliefs in response to feedback.

2.3.1 Valuation models

Decisions under risk require a choice between two options with explicitly presented attributes. To computationally assess the process from option presentation to action, valuation models are employed, which consist of two components: the valuation process and the action selection (Figure 2.4). As outlined in Section 1.1, the valuation process involves calculating the subjective value for each option by integrating the available information. However, animals and humans do not always choose actions based solely on the calculated subjective value. Hence, valuation models need to account for the stochasticity during action selection.

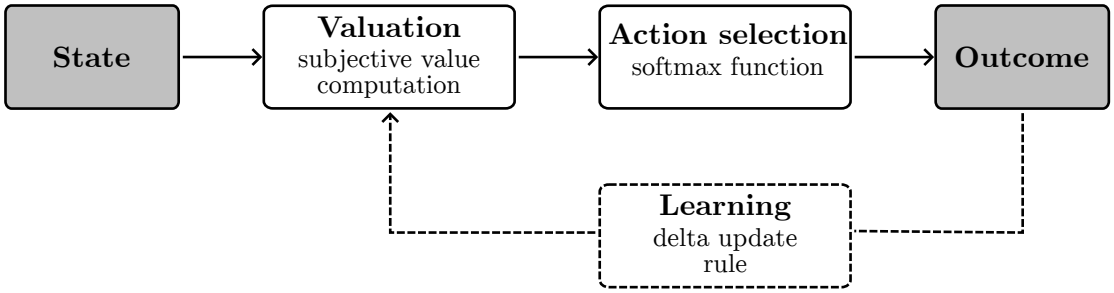

Figure 2.4: Structure of computational models. Computational models link environmental variables (grey) to learning and decision-making processes. After observing the state, the valuation process is modelled with the computation of the subjective value. Based on the subjective value, the action is selected using a softmax function. Following the choice, the environment provides an outcome. When learning is involved, as further described in Section 2.3.2, the agent learns from this information, computationally modelled using the delta update rule, which in turn adjusts the subjective value.

Reward-guided decision-making paradigm

We fitted parameters to participants' behaviour in the gambling task (Subsection 2.1.1) using a valuation model. The valuation model tests if the information integration is rather additive or multiplicative via a hybrid model, which captures both strategies. Thus, the subjective value $SV_{i,t}$ for option i in each trial t of each option is computed as:

$$SV_{i,t} = \omega_{mult} M_{i,t} P_{i,t} + (1 - \omega_{mult})((1 - \omega_P) M_{i,t} + \omega_P P_{i,t}) \tag{2.1}$$

where $M_{i,t}$ and $P_{i,t}$ are magnitude and probability of option i in trial t, ω_{mult} represents the weighting of multiplicative information integration and ω_P denotes the relative weighting of probability. A purely multiplicative integration, corresponding to optimal decision making, is represented by $\omega_{mult} = 1$, while a purely additive integration is indicated by $\omega_{mult} = 0$. Additionally, the model accounts for how participants prioritise different option attributes: If they select high-magnitude rewards only, ω_P would be 0, and if they exclusively opt for high-probability options, ω_P would be 1.

To convert the subjective value of each option into probabilities and account for participants' variability during action selection, a softmax function is implemented:

$$p_{l,t} = \frac{1}{1 + exp(-(SV_{l,t} - SV_{r,t})\zeta)} \tag{2.2}$$

where $p_{l,t}$ is the probability of choosing the left-side option in trial t, $SV_{l,t}$ and $SV_{r,t}$ represent the subjective value of the left-side and right-side option in trial t, respectively, and the inverse temperature parameter ζ reflects the level of stochasticity in choice behaviour. Higher values of inverse temperature represent less random behaviour and lower values more random behaviour.

2.3.2 Reinforcement learning models

To make informed decisions under uncertainty, the agent is required to learn from the outcome. Therefore, next to valuation and action selection, RL models consist of a third component, the learning process (Figure 2.4).

Reward-guided learning paradigm

In order to compare behaviour in both tasks of Study I, we used a computational model closely following the valuation model for the learning task (Section 2.1.2). In contrast to the gambling task, the reward probability needs to be learnt by trial and error in the learning task. To assess each option's value, a subjective probability estimate is therefore required. In the model, the subjective probability of chosen options $SP_{c,t}$ is adjusted in each trial t using a Q-learning approach with a delta update rule (Equation 1.3). Previous work suggests that humans learn differently for positive versus negative prediction errors (Eckstein et al., 2022; Gershman,

2015; Palminteri et al., 2016), thus, separate learning rates λ_r and λ_u for rewarded and unrewarded choices, respectively, are implemented. Additionally, since the outcomes of the two options are interdependent, the model assumes a dependent update mechanism, where the subjective probability of the unchosen option $SP_{u,t}$ is simply $1 - SP_{c,t}$. The subjective probability is then used to integrate information into the subjective value $SV_{i,t}$ of option i in each trial t:

$$SV_{i,t} = (1 - \omega_P)M_{i,t} + \omega_P SP_{i,t} \tag{2.3}$$

Note, that we used an additive model for the learning task due to convergence issues. Additionally, the softmax function is implemented to model action selection (Equation 2.2).

The learning task consists of two different phases - the stable phase and the volatile phase. Parameters are fitted separately for both phases to capture differences in behaviour between volatility levels.

Both the valuation model (Section 2.3.1) and this reinforcement learning model were fitted as Bayesian hierarchical models, which implementation and validation is described in Subsection 2.3.3.

Go/no-go learning paradigm

The go/no-go learning task (Section 3.2) consists of reinforced and non-reinforced blocks. We aimed to reveal the mechanism of adjusted behaviour in non-reinforced blocks, thus, we set up four different RL models: a baseline model, a temperature model, a bias model, and a full model. The learning process is equal in all models. In this paradigm, instead of making choices between two choice options, participants are presented with a single stimulus and have to choose between two actions - press a button or refrain from a button press. After performing a go response in reinforced trials, the value $Q_{i,t}$ of the presented stimulus i in trial t is updated according to the delta update rule (Equation 1.3). In non-reinforced probe trials, stimulus' values after performing a go response are not updated, such that $Q_{i,t+1} = Q_{i,t}$. However, participants learn to retain an active response and perform a no-go response after certain stimuli. The non-monotonic plasticity theory implies that synaptic connections are weakened for unchosen options (Ritvo et al., 2019), which is also supported by imaging studies (Luettgau et al., 2020). Therefore, a decay

parameter θ is implemented to enable passive forgetting when a no-go response is performed: $Q_{i,t+1} = \theta Q_{i,t}$. Since participants have no information about the presented stimuli, initial Q-values Q_0 when stimuli are presented for the first time are also treated as a free parameter. Note that in the go/no-go task, the learned value Q serves immediately as subjective value determining the following action. Based on the subjective value per stimulus $Q_{i,t}$, the probability of performing a go-action p_t in trial t is modelled using a softmax function:

$$p_t = \frac{1}{1 + exp(\frac{-(Q_{i,t}+b_k)}{\tau_k})} \tag{2.4}$$

where τ_k is the softmax temperature and b_k is the response bias, dependent on the environment k. The four models differ in the implementation of the bias and the temperature terms. The baseline model does not differentiate between reinforced (R) and probe (P) trials, thus, temperature and bias are equal in both contexts: $\tau_k = \tau_R = \tau_P$ and $b_k = b_R = b_P$. The temperature model enables different temperatures in reinforced and probe trials, such that $\tau_k = \tau_R$ in reinforced trials and $\tau_k = \tau_P$ in probe trials. In contrast, the bias model enables different biases in reinforced and probe trials with $b_k = b_R$ in reinforced trials and $b_k = b_P$ in probe trials. The full model combines the temperature and the bias model, as it enables separate temperatures and biases in reinforced and probe trials. Thus, the models differ in the number of free parameters n (baseline model: $n = 4$, temperature model: $n = 5$, bias model: $n = 5$, full model: $n = 6$) and the interpretation of which processes are affected by the task manipulation (i.e. reinforcement vs. non-reinforcement).

All four models were fitted to participants' behaviour using the maximum-likelihood approach. We then compared model fits using the Bayesian information criterion (BIC). While a greater number of free parameters can improve a model's ability to fit observed behaviour, it also increases the risk of overfitting. The BIC helps to prevent overfitting by penalising models with higher complexity (Stoica and Selen, 2004). As a result, more complex models should only outperform simpler ones if their additional parameters are truly necessary to accurately describe the observed behaviour.

Moreover, model validation and parameter recovery were conducted for the

best-fitting model. For the model validation, behaviour was simulated based on fitted parameters and compared to observed behaviour in order to test whether the model captures key behavioural effects. This step is crucial to ensure that the model makes sense of the data (Wilson and Collins, 2019). With the parameter recovery we ensured the reliability of fitted parameters. To this end, parameters were fitted to simulated behaviour and correlated with the original parameter fits (Wilson and Collins, 2019).

2.3.3 Bayesian hierarchical modelling

In Study I, the valuation model of the gambling task (Subsection 2.3.1) and the RL model of the learning task (Subsection 2.3.2) were implemented as Bayesian hierarchical modelling approach with Markov chain Monte Carlo sampling for parameter estimations. In contrast to the maximum-likelihood approach, Bayesian hierarchical modelling allows for the analysis of complex data structures by modelling relationships simultaneously at multiple levels (Lee, 2011). Parameters are estimated as probability distributions, namely posterior distributions, which are continuously updated (Lee and Wagenmakers, 2013). These posterior distributions provide a full probabilistic description of the parameters, which quantify the uncertainty of fitted parameters and shrink outliers (Baribault and Collins, 2023).

Bayesian hierarchical modelling is particularly useful when individual differences and group-level effects need to be considered simultaneously. In the models, prior information about the parameters are used as hyperprior for group-level parameters and group-level parameters serve as prior for the estimation of subject-level parameters (Figure 2.5). Thus, parameter estimates are iteratively updated with observed data and yield to posterior distributions, which in turn inform group-level parameter estimates.

In Study I, we applied pharmacological intervention in a within-subjects design. Participants performed the gambling and learning task twice, once under placebo and once under the influence of biperiden. Thus, Bayesian hierarchical models account for variability across participants while also identifying drug effects within the group (Lee and Wagenmakers, 2013). To achieve this, we implemented a drug-induced shift parameter s_x for all free parameters x with $x + \delta_{bip} s_x$, where δ_{bip} is 0 or 1 for placebo and biperiden sessions, respectively.

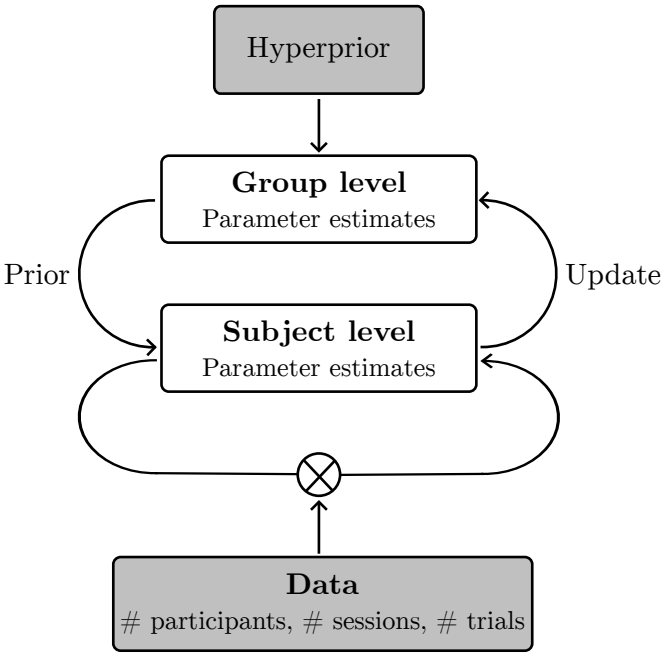

Figure 2.5: Schematic of the Bayesian hierarchical models. A hyperprior informs group-level parameters, which are used as prior for subject-level parameters. Subject-level parameters are updated with observed data and the resulting posterior serves as prior for group-level parameter estimates.

Since parameters are fitted as probability distributions, we cannot infer the significance. Instead, we inferred the credibility using the 95 % highest density interval (HDI) of the posterior predictive distribution: When the interval does not overlap with 0, it is credible, that the fitted parameter is not equal to 0 (Kruschke and Liddell, 2018). Moreover, we validated if the model captures participants' behaviour by conducting posterior predictive checks. Based on the posterior distributions of subject-level parameters, we simulated data and compared it to the observed data (Baribault and Collins, 2023).

3 Results

The following chapter provides an overview of key results of the two studies. More detailed information can be found in the original work (Kurtenbach et al., 2024, 2022).

3.1 Study I: A role for acetylcholine in reinforcement learning and decision making under uncertainty

The following section is based on the preprint available at bioRxiv (see Research articles):

Kurtenbach H, Froböse MI, Ort E, Bahners BH, Hirschmann J,
Butz M, Schnitzler A, & Jocham G
A role for acetylcholine in reinforcement learning and decision making under uncertainty
bioRxiv (2024)

The study aimed to investigate decision making and learning under risk and different levels of uncertainty. The results can be divided into two parts following from the two aims defined above. The first part (Subsection 3.1.1) set out to reveal strategy shifts dependent on the environment. Building on that, the cholinergic effect on decision behaviour in different environments is described in the second part (Subsection 3.1.2).

3.1.1 Decision strategies under risk versus uncertainty

We investigated learning and decision making in three different environments: under risk (gambling task, Subsection 2.1.1), and under uncertainty, with a stable and a volatile environment (learning task, Subsection 2.1.2). We assumed that decision strategies are adapted in two ways to suit the environment. First, we hypothesised that information integration of option attributes is more multiplicative under risk and increasingly additive with higher uncertainty, in order to weight certain option attributes more heavily (Farashahi et al., 2019; Stewart, 2011). The next hypothesis addresses decision strategies under uncertainty, when learning is involved. We assumed that the learning rate increases with increasing volatility in the reward-guided learning task, in order to favour more recent outcomes (Behrens et al., 2007; Browning et al., 2015).

We used Bayesian hierarchical modelling to identify varying decision strategies (see Section 2.3). For the gambling task, we fitted a valuation model with three free parameters: ω_{mult}, which determines the degree of multiplicative versus additive information integration, ω_P, which defines the degree of probability versus magnitude weighting, and the inverse softmax temperature ζ. For both the stable and the volatile phase of the learning task, we fitted four free parameters: the relative attribute weighting ω_P, a learning rate for rewarded and unrewarded choices, λ_r and λ_u, respectively, and the inverse softmax temperature ζ.

In line with our hypothesis, we found that in the gambling task, participants applied a hybrid strategy of information integration, comprising both additive and multiplicative integration (Figure 3.1A). Moreover, they weighted reward probability credibly stronger than in the learning task, where the probability was implicitly presented and, thus, needed to be learnt (Figure 3.1B). Lastly, in the gambling task, participants were credibly less stochastic than in the learning task, as reflected by a higher inverse softmax temperature (Figure 3.1E). This is also in line with our expectations, since option attributes were explicitly provided in the gambling task, hence, requiring less exploration.

However, we found no credible adjustments of decision strategies for the two phases within the learning task. There was no credible difference in relative attribute weighting between stable and volatile phase, albeit reward probability was associated with more uncertainty in the volatile phase (Figure 3.1B). Additionally, despite

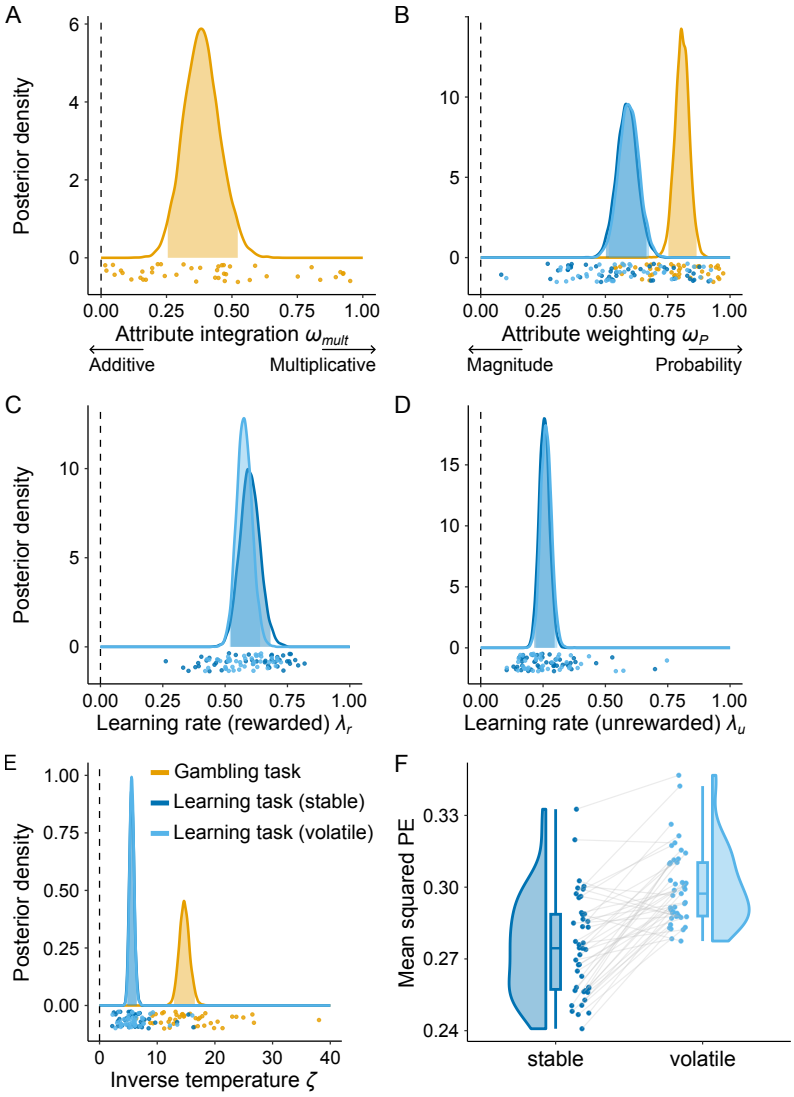

Figure 3.1: Parameter fits for choice behaviour under risk versus different degrees of uncertainty. Posterior distributions of parameter estimates of the gambling task (orange), stable phase of the learning task (dark blue), and volatile phase of the learning task (light blue). **A** The attribute integration ω_{mult} indicates whether the attributes were integrated multiplicatively or additively. **B** The attribute weighting ω_P represents how the reward probability is weighted relative to the magnitude. **C, D** For the learning task, learning rates for rewarded choices λ_r and for unrewarded choices λ_u were fitted. **E** The stochasticity of choice behaviour was fitted via the inverse softmax temperature ζ. Shaded areas represent the 95 %-HDI and points single-subject means. **F** Mean squared prediction error (PE) in the stable and volatile phase of the learning task. A higher PE corresponds to a greater surprise of outcomes. Reproduced from (Kurtenbach et al., 2024). CC BY 4.0.

the higher volatility in the volatile phase, learning rates did not credibly differ between both phases (Figure 3.1C,D). Even the response stochasticity remained equal in both phases (Figure 3.1E). Although no behavioural adjustments between the stable and volatile phases were found, follow-up analyses revealed a difference in mean squared prediction errors. In the volatile phase, mean squared prediction errors were significantly higher than in the stable phase (Figure 3.1F). This reflects that participants were more surprised by outcomes under higher volatility.

In sum, participants adapted decision strategies across tasks, where one task involved learning and the other not. However, within the learning task, participants did not adjust response behaviour according to the underlying volatility.

3.1.2 Cholinergic effect on decision strategies

The study set out to investigate the influence of ACh on decision making and learning under risk and different degrees of uncertainty. Therefore, participants performed the gambling task and the learning task twice, once after placebo and once after the administration of biperiden, a muscarinic M1 ACh receptor antagonist (see Section 2.2). As a first step, we used logistic mixed-effects regression (see Kurtenbach et al. (2024) for more information) to analyse the effects of biperiden on task parameters. While there were no significant behavioural modulations by biperiden in the gambling task or in the stable phase of the learning task, we found, specifically for the volatile phase of the learning task, a reduced use of (learned) reward probabilities under biperiden (Figure 3.2). Note that for the learning task, reward probabilities in the regression model reflect estimates derived from a Bayesian optimal learner (based on Behrens et al. (2007)), because participants did not know the underlying probabilities. Therefore, the observed biperiden effect could be caused by less use of probability information or by impaired tracking of probabilities or both.

In order to analyse the origin of the biperiden-induced effect, we used Bayesian hierarchical modelling and implemented shifts from placebo sessions on each fitted parameter in the biperiden session (see Subsection 2.3.3). A diminished use of probability information would be reflected in a reduced probability weighting parameter under biperiden, while a modulation of learning rates would correspond to impaired tracking of probabilities. We found that, in the volatile phase of

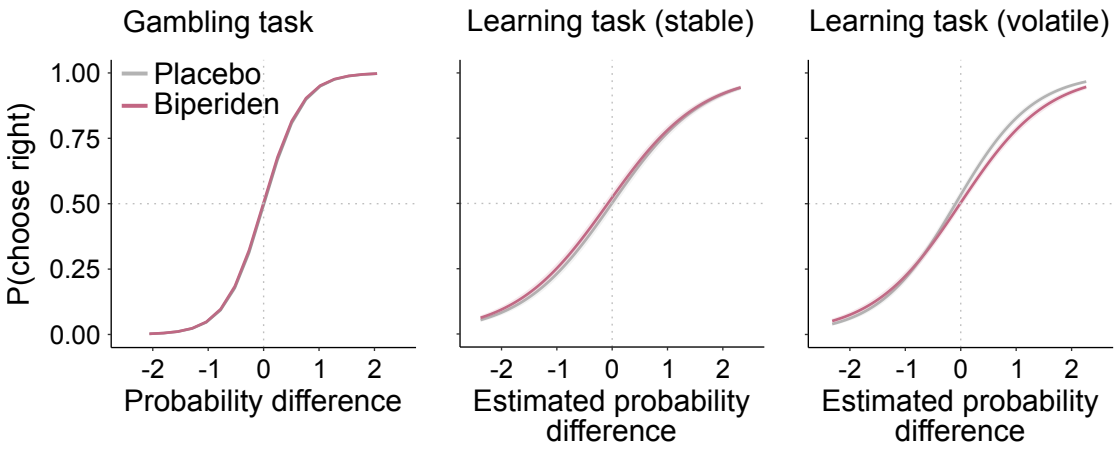

Figure 3.2: Probability for a right-side choice as a function of differences in reward probability. Interaction between reward probability and drug from logistic-mixed effects regression in the gambling task (left) and in the stable (middle) and the volatile phase (right) of the learning task. Sensitivity to reward probabilities was significantly reduced under biperiden (pink) compared to placebo (grey) in the volatile phase. Solid lines represent mean, shaded areas SEM across participants. Adapted from (Kurtenbach et al., 2024). CC BY 4.0.

the learning task, the learning rate for rewarded choices increased credibly after biperiden administration (Figure 3.3), indicating that participants adjusted more quickly to changes in contingencies. As there were no credible biperiden-induced effects on relative probability weighting, we conclude that biperiden influences specifically the learning process rather than the valuation process.

However, it is generally considered advantageous to increase the learning rate in volatile environments compared to stable ones (Behrens et al., 2007; Browning et al., 2015), while we hypothesised that biperiden would impair performance. To test this, we compared participants' estimated reward probabilities, based on their fitted learning rates, to those of a Bayesian optimal learner. We found that participants' estimates deviated significantly stronger from the optimal estimates during the biperiden session than during the placebo session (Figure 3.3E).

All in all, biperiden induces a maladaptive increase in learning rate in volatile environments, resulting in noisier value estimates.

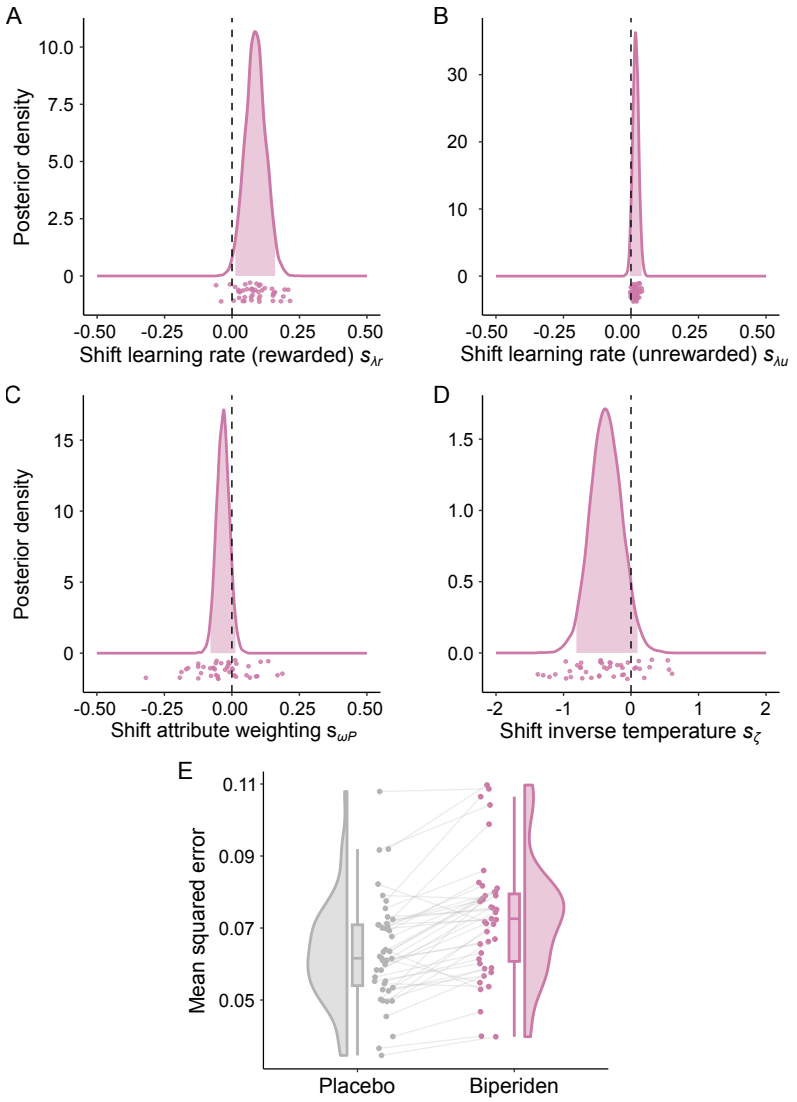

Figure 3.3: Biperiden effect in the volatile phase of the learning task. Posterior distributions of biperiden-induced shifts on **A**, **B** learning rate in rewarded choices $s_{\lambda r}$ and unrewarded choices $s_{\lambda u}$, respectively, **C** relative attribute weighting $s_{\omega P}$, and **D** inverse temperature s_{ζ}. A positive shift in parameters displays an increase under biperiden compared to placebo. Shaded areas represent the 95 %-HDI and points single-subject means. **E** Mean squared error (MSE) of participants' estimated reward probabilities relative to optimal estimates in the placebo session (grey) and the biperiden session (pink). Reproduced from Kurtenbach et al. (2024). CC BY 4.0.

3.2 Study II: Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations

The following section is based on the manuscript published in PLOS Computational Biology (see Research articles):

Kurtenbach H, Ort E, Froböse MI, & Jocham G
Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations
PLOS Computational Biology **18(12)**, e1010201 (2022)

In this study, we aimed to translate findings from rodent work to a human sample. The rodent study observed improved performance in non-reinforced compared to reinforced trials in an instrumental learning task (Kuchibhotla et al., 2019). To investigate this effect in humans, we set up a visual go/no-go learning task (Subsection 2.1.3). In line with Kuchibhotla et al. (2019), we conducted an SDT-based analysis and found increased sensitivity, as measured by the sensitivity index d', in non-reinforced probe compared to reinforced trials (Figure 3.4). For the computation of SDT measures, however, a window of several trials needs to be considered. This results in an artefact when assessing d' during a learning process: Before reaching stable levels, performance on later trials is naturally higher than in earlier trials. Thus, d' is always higher in later compared to earlier trials, irrespective of task manipulations. To differentiate whether our findings indeed reflect a performance increase or follows spuriously from the SDT analysis approach, we used RL models, which provide trialwise estimates of behaviour to decipher underlying mechanisms when reinforcement is removed.

We fitted four distinct models to account for different response strategies during probe trials (Subsection 2.3.2). The first, the baseline model, assumed identical choice behaviour across both reinforced and probe trials. The second, the temperature model, allowed for varying softmax temperatures between conditions, where lower temperatures in probe trials would indicate increased sensitivity to learned values, suggesting that the removal of reinforcement reveals latent task knowledge.

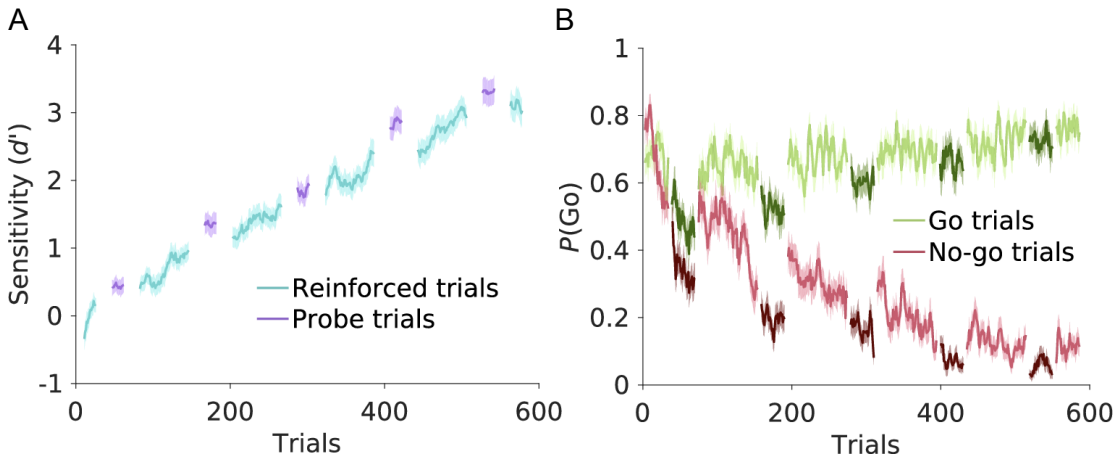

Figure 3.4: Time course of participants behaviour in the go/no-go learning task. A Sensitivity index d' in reinforced (cyan) and probe trials (purple). **B** Probability for go responses (P(Go)) for go trials (green) and no-go trials (red). Probe trials are indicated by darker shades of green and red. Solid lines represent mean, shaded areas SEM across participants. Adapted from (Kurtenbach et al., 2022). CC BY 4.0.

The third, the bias model, introduced a varying general bias between conditions, with a lower bias parameter in probe trials indicating a reduction in go responses across both go and no-go stimuli. Finally, the full model incorporated both varying softmax temperatures and bias parameters. If the performance increase as observed using SDT reflects improved latent task knowledge, we expected the temperature model to have the best fit. However, we found that the bias model provided the best fit (Figure 3.5A).

We observed that the response (i.e. go) bias was lower in probe trials compared to reinforced trials (Figure 3.5B). With the general bias being the only difference between reinforced and probe trials, simulated data based on the bias model could successfully reproduce participants' choice behaviour (Figure 3.5E). It seemed counterintuitive in the first place that the bias model, which captures a general reduction in overall button presses could account for the observed behaviour. In theory, less button presses should lead to a lower false alarm, but also lower hit rate and should therefore not lead to an increase in d'. Further analysis, however, shed light on this: Initial Q-values Q_0 were positive in the bias model, reflecting frequent go responses early in the task (Figure 3.5C). Thus, starting with already high values, values for go stimuli increase further during the task, while values for no-go stimuli decrease. Combined with the softmax function's sigmoid shape

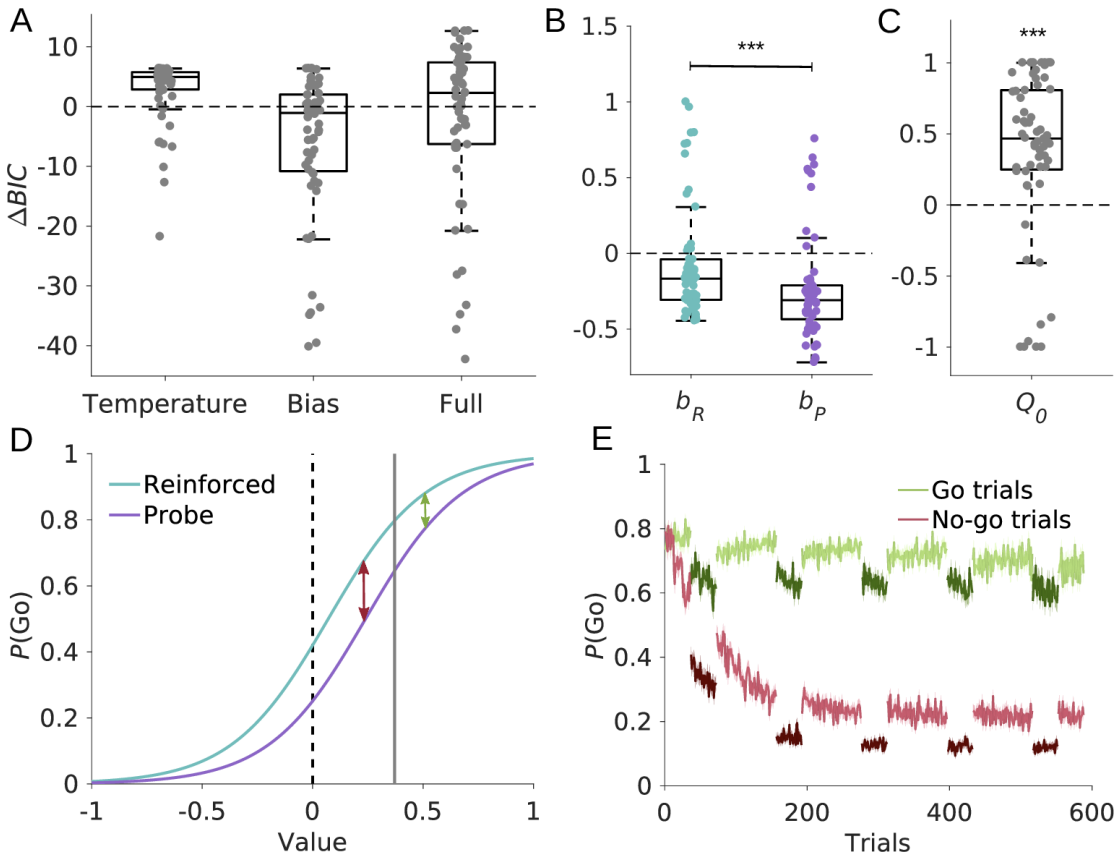

Figure 3.5: Computational modelling results of the bias model. **A** Model comparison of fitted models relative to the baseline model. The lower the BIC, the better the fit. The bias model provides the best fit and, thus, is further described in the following. **B** Comparison of the bias parameter in reinforced trials b_R and probe trials b_P. **C** Parameter fits for the initial estimate of option values Q_0. Points represent individual participants' fits. **D** Schematic illustration of the differential effect on false alarm rate and hit rate. Softmax go-response probabilities P(Go) are reduced for probe compared to reinforced trials. In combination with positive initial estimates Q_0 (solid grey line), the difference between P(Go) in reinforced and probe trials decreases over time for go stimuli, as the values increase (green arrow). Conversely, values for no-go stimuli decrease over time and the difference between P(Go) in reinforced and probe trials increases (red arrow). **E** Time course of simulated probabilities for go responses P(Go). Probe trials are indicated by darker shades of green and red. Solid lines represent mean, shaded areas SEM across simulations. Reproduced from Kurtenbach et al. (2022). CC BY 4.0.

(Figure 3.5D), a reduced go-bias in probe trials leads to a disproportional reduction of false alarms relative to the reduction in hits.

In summary, we could replicate the improved instrumental performance in non-

reinforced trials. However, this improvement resulted from reduced response probabilities in general rather than an increased sensitivity to learned values.

4 Discussion

This dissertation addresses the mechanisms of reinforcement learning and decision making depending on the reliability of information and the availability of feedback. In two studies, we aimed to

- investigate how muscarinic M1 receptor activity influences learning and decision making depending on the uncertainty of the environment (Study I).
- examine the adjustment of decision strategies under risk and under different degrees of uncertainty (Study I).
- assess the impact of external reinforcement on learning performance and decision strategies (Study II).

I will start by discussing the results of the two studies concerning our hypotheses. Following this, I will present an outline of future research based on the findings. Finally, I will draw a conclusion regarding the work presented in this dissertation.

4.1 Discussion of hypotheses

1. *Cholinergic antagonism of the muscarinic M1 receptor impairs learning under uncertainty.* In Study I, we observed that biperiden administration resulted in impaired learning in the volatile phase, as learning rates were maladaptively increased. Exploratory analyses additionally revealed that this increase in learning rate led to impaired estimates of reward probability and, thus, enhanced distractibility by recent outcomes. This supports the idea, suggested by physiological and behavioural studies, that ACh may be crucial for top-down control (Ballinger et al., 2016; Bentley et al., 2011; Danielmeier et al., 2015; Eggermann and Feldmeyer, 2009; Gratton et al., 2017). Additionally, the biperiden-induced increase in learning rate

was specific for rewarded choices. This aligns with an optogenetic study in mice, which demonstrates that basal forebrain cholinergic neurons responded to outcome surprise, with a stronger response for rewarded compared to punished outcomes (Hangya et al., 2015). Moreover, our findings aligns with studies which found that ACh is involved in shaping behaviour under uncertainty. For example, cholinergic manipulation led to impaired attentional set-shifting and increased distractibility in mice (Cools and Arnsten, 2022; Robbins and Roberts, 2007). However, our findings appear to contrast with theoretical predictions regarding the role of ACh. Biperiden specifically affected behaviour in the volatile phase, the only condition where both expected and unexpected uncertainty were present. Historically, ACh has been linked to expected uncertainty, while noradrenaline has been associated with unexpected uncertainty (Avery et al., 2012; Yu and Dayan, 2005). However, these models also suggest that ACh depletion can increase distractibility by causing an underestimation of environmental volatility. This aligns with our findings, as the biperiden effect was observed in the volatile phase, where outcome surprise was greater compared to the stable phase, indicated by a higher mean squared prediction error. Additionally, a recent study in humans demonstrated that biperiden increased the update rate of volatility estimates in a probabilistic serial response time task, impairing adaptation to environmental changes (Marshall et al., 2016). Thus, muscarinic M1 receptor activity might be involved in uncertainty processing, particularly in volatile environments where surprise levels are increased.

2. *Cholinergic antagonism of the muscarinic M1 receptor leads to suboptimal information integration.* In Study I, we found no effect of biperiden on information integration in both tasks. This may be due to suboptimal performance across tasks, even without cholinergic manipulation. Compared to other studies, our participants demonstrated less multiplicative attribute integration (Farashahi et al., 2019; Scholl et al., 2014), and their learning rates were notably high in the learning task. One possible explanation is that our tasks were more difficult. In particular, the learning task may have been more challenging; for example, some reversal learning tasks feature larger differences in reward probabilities between options, making reversals easier to detect (Behrens et al., 2007; Browning et al., 2015), while others explicitly signal volatile phases (Blain and Rutledge, 2020; Massi et al., 2018). The lack of involvement of muscarinic M1 receptor activity in information integration in our

study may have multiple causes, but our behavioural results for now suggest that M1 receptor activity does not affect reward-guided decision making in the absence of learning, but is specifically relevant to learning in volatile environments.

3. *The more uncertain the environment, the more additive the information integration.* In Study I, we found partial evidence for adaptive information integration under uncertainty. In the gambling task, participants employed a hybrid valuation strategy, combining both additive and multiplicative approaches, consistent with findings in humans and non-human primates (Farashahi et al., 2019). However, in the learning task, we could fit only an additive, but not a hybrid model due to convergence issues. While this suggests that participants may have relied on a purely additive integration strategy in the learning task, we cannot definitively conclude this. Nonetheless, we observed that participants weighted reward probabilities significantly more in the gambling task than in the learning task. This supports the hypothesis that participants adjust their valuation strategy by placing less emphasis on the unknown attribute (i.e., probability) under uncertainty (Stewart, 2011). However, within the learning task, we found no differences in information integration, which may be due to the difficulty of the task, as the different volatility phases were not easily distinguishable. Overall, we propose that agents adapt their valuation strategies based on environmental uncertainty, though this flexibility may be influenced by participants' awareness of the environment.

4. *The more uncertain the environment, the higher the learning rate.* In Study I, we did not find evidence that learning rates adapt based on environmental uncertainty. Numerous studies have reported that learning rates increase during volatile phases compared to stable ones (Behrens et al., 2007; Blain and Rutledge, 2020; Browning et al., 2015; Massi et al., 2018). We found no increase in learning rate under higher volatility, however, explorative analyses indicated that there is a difference between stable and volatile phases: The mean squared prediction error was higher in the volatile compared to the stable phase, revealing that participants were more surprised by outcomes in the volatile phase. Nevertheless, they did not adjust their learning rates accordingly. A recent study also failed to find increased learning rates in volatile environments (Cook et al., 2019). They suggested that the absence of this effect could be attributed to the complexity of the task. Similarly, the difficulty of our learning task may have prevented participants from distinguishing

between the two phases. All in all, this suggests that increased volatility does not lead to changes in learning rates in general.

5. *Instrumental performance during learning is increased when reinforcement is absent compared to present.* In Study II, our results confirmed that instrumental performance improved during blocks without reinforcement compared to reinforced blocks. However, computational modelling indicated that this improvement was not due to participants demonstrating latent task knowledge in the absence of feedback, as would be reflected by a reduced softmax temperature (i.e., increased sensitivity to learned values). Instead, participants adopted a more cautious response strategy. In non-reinforced blocks, there was a decrease in general response bias, meaning participants were less likely to act. In reinforced blocks, go responses provided feedback and thus an information bonus. It is assumed that humans and animals seek information to reduce uncertainty about the environment (Bromberg-Martin and Hikosaka, 2009; Stagner and Zentall, 2010; van Lieshout et al., 2021). Thus, when no information is available, such as in non-reinforced blocks, participants held back responses. Additionally, the models revealed that initial values were positively biased, reflecting a high propensity for go responses. This is driven by the asymmetric nature of our task structure where only go responses provided feedback. Thus, our findings suggest that the response strategy is modulated by the presence of reinforcement, and affects instrumental performance.

4.2 Future research

Our results displayed that humans adapt decision strategies to the underlying environment and that ACh modulates learning in environments with outcome surprise. These findings but also limitations of the presented work, certainly invite future research.

In Study I, we investigated how the neuromodulator ACh influences choice behaviour, but its impact on neural dynamics remains an open question. The next step would be to link ACh-driven impairments in learning under volatile conditions to neuronal data using magnetoencephalography (MEG). MEG not only identifies specific brain areas involved in task performance but also allows for the examination of cortical oscillations, which reflect communication within and

between brain regions (Florin and Baillet, 2015; Fujisawa and Buzsáki, 2011). ACh has been shown to modulate cortical oscillations: In visual attention tasks, ACh has been found to influence gamma oscillations (Howe et al., 2017; Rodriguez et al., 2004), while theta oscillations are thought to be modulated during memory tasks (Gedankien et al., 2023). However, there is limited knowledge about how ACh modulates cortical oscillations during learning. Our study links ACh modulation to learning rates, which are closely tied to reward processing. Reward processing is broadly distributed across several brain regions, including the PFC, orbitofrontal cortex, amygdala, striatum, and anterior cingulate cortex (Marco-Pallares et al., 2008). Typically, beta oscillations are associated with learning from gains, while both beta and theta oscillations are linked to learning from losses (Marco-Pallares et al., 2008; Van De Vijver et al., 2011). To further investigate the neural network and temporal dynamics biperiden acts on during learning, we plan to correlate biperiden-specific behavioural effects with neuronal data (in preparation).

Although we manipulated ACh in Study I, the role of ACh in decision-making processes under risk remains an open question. Decision making under risk and under uncertainty are considered distinct processes. For instance, patients with Parkinson's disease exhibit impairments in decision making under risk but not under uncertainty (Euteneuer et al., 2009). Our findings revealed that ACh did not affect valuation, a process involved in both the gambling and learning tasks, but it did influence learning, which was specific to the learning task. One possibility is that ACh does not play a significant role in decisions under risk. However, it is also plausible that ACh influences subtle decision-making dynamics that do not manifest directly in choice behaviour. For example, a study with nicotine-dependent humans found improvements in decision making in abstinent individuals using drift-diffusion models (DDMs), albeit task performance did not differ between groups (Biernacki et al., 2023). DDMs offer the possibility to explore decision dynamics, such as the speed of evidence accumulation. Thus, in future work, it would be useful to apply DDMs to our behavioural results.

Another way to get more insights into the cholinergic modulation is to apply biophysically plausible network models. These models simulate firing rates via competition by mutual inhibition (Soltani et al., 2016; Wang, 2002). This process is driven by excitatory NMDA and inhibitory GABA activity, however, ACh is known to modulate both neurotransmitter systems (Bessie Aramakis et al., 1997;

Kuchibhotla et al., 2017; Marino et al., 1998; Obermayer et al., 2017; Zwart et al., 2018). These models could not only clarify whether we should predict behavioural effects based on the cholinergic modulation of GABA and NMDA, but could also be used as predictors for neural signals (Hunt et al., 2012).

Study II suggests that performance is not only influenced by external factors such as the availability of feedback, but also interacts with response biases that may or may not be beneficial given a particular task structure. However, the dynamics of these biases are not fully understood, especially how they change behaviour depending on the availability of reinforcement. We assumed that the task structure plays a critical role in shaping behaviour, particularly in changing contexts like the removal of reinforcement, since behaviour might not solely be driven by maximising reward, but also by maximising information. When this latter possibility is removed, an overall reduction in button presses resulted in improved performance in this specific task structure. To follow up this point, we proposed that a task using different asymmetric feedback, such as rewarded no-go stimuli and unrewarded go stimuli, would actually impair instrumental performance when reinforcement is removed, due to an initial negative bias towards no-go responses. Conversely, a symmetrical task structure, where both go and no-go stimuli receive feedback, should result in no difference in behaviour after reinforcement is withdrawn. To test this, we conducted a follow-up study with modifications to the task structure. Interestingly, we found that instrumental performance improved in both the inverse asymmetrical task and the symmetrical task. This aligns with a recent study, which also reported improved performance in the absence of reinforcement using a symmetrical learning task (Vahedi et al., 2024). These findings suggest that the response bias might be more dynamic than we anticipated.

Further to this, a recent study found that the motivational bias (i.e. the tendency to respond more actively to reward-related go responses) originates in the PFC, which processes external signals, before being integrated with internal signals in the striatum (Algermissen et al., 2024). This suggests that the motivational bias may represent a more flexible, strategic mechanism, rather than a rigid response pattern. To better understand the neural mechanisms underlying our bias effect in Study II, future research could investigate performance in the three variations of our go/no-go task (asymmetric, inverse asymmetric, symmetric) using imaging techniques like MEG. This approach could provide valuable insights into how the

bias effect operates at the neural level.

Additionally, future research could investigate how a neuromodulator like ACh affects the bias parameter depending on the availability of reinforcement. We found that cholinergic antagonism impaired learning by causing an overshoot in the learning rate. Furthermore, research suggests that ACh regulates not only learning but also influences negative encoding bias (as reviewed in Mineur and Picciotto (2021)). Thus, a cholinergic antagonist such as biperiden could potentially lead to faster learning from punishment, i.e., false alarms in our task. The reduction in false alarm rate observed in the absence of reinforcement could be even more pronounced. However, it remains unclear what happens when reinforcement is absent and agents must rely solely on intrinsic reinforcement.

As a next step, resting state MEG activity could be used to predict behaviour. Previous research suggests that the ratio between resting-state theta and beta oscillatory activity is involved in reward-related processing (Massar et al., 2014). Participants with a higher theta/beta ratio distinguished less well between gains and losses and learnt more slowly, which was additionally accompanied by increased risk-taking behaviour (Massar et al., 2012; Schutter and Van Honk, 2005). Moreover, recent research found that participants with lower beta oscillatory activity during resting state had weaker frontoparietal connectivity, resulting in higher flexibility and, thus, improved accuracy in a visual attention task (Rogala et al., 2020). Thus, resting-state theta and beta oscillatory activity could also be predictive of the use of decision strategies in our tasks, which involve both risk and certain forms of flexibility.

The present work focused on investigating task-related behaviour in healthy participants. However, from a clinical perspective, resting-state oscillatory activity is also of interest, as neuropsychiatric diseases are associated not only with altered decision-making behaviour but also with abnormal neuronal synchronization (Uhlhaas and Singer, 2006). For instance, patients with Alzheimer's disease exhibit increased power in delta and theta oscillations and reduced power in alpha and beta oscillations compared to healthy controls (Kopčanová et al., 2024). Similarly, patients with Parkinson's disease and cognitive dysfunction show increased delta, alpha, beta, and gamma power, and decreased theta power (Anjum et al., 2024; Jaramillo-Jimenez et al., 2021). Hence, electroencephalography and MEG could serve as tools for identifying early-stage biomarkers of neuropsychiatric dis-

eases (Anjum et al., 2024). Oscillatory patterns, especially those observed during resting-state measurements, are of particular interest as they provide insight into network-level interactions independent of specific behavioural tasks. Understanding how specific neuromodulators affect these interactions may provide a foundation for future clinical applications. Therefore, as a next step, we plan to compare neuronal oscillations during resting-state measurements between the biperiden and placebo sessions (in preparation).

Next to these follow-up studies, future research could also address the limitations of our studies. We conducted the pharmacological study during the COVID-19 pandemic. A recent study found impaired reward learning in volatile environments caused by heightened expectations of environmental volatility in a dataset collected during the pandemic (Guitart-Masip et al., 2023). They linked the impaired performance to increased state anxiety. It is possible that these circumstances may also have an impact on our study, e.g. by leading to generally worse performance in the tasks. Even though we excluded participants with symptoms of depression, as assessed by Beck's Depression Inventory (Beck et al., 1996), and measured trait anxiety (Spielberger, 1983), we did not account for state anxiety. A further limitation is that participants were recruited in university settings and, thus, were mainly young and highly educated. Moreover, our pharmacological study included males only. Especially in pharmacological studies, females are still underrepresented. Although, a recent behavioural study using biperiden included both females and males and found no gender-specific drug effects on cost-benefit decision making (Erfanian Abdoust et al., 2024), future studies should include a broader range of participants to be more representative.

4.3 Conclusion

In this dissertation, we investigated the dynamics of reinforcement learning and decision making, and the role of ACh in these processes. Across two studies, we demonstrated that participants' choice strategies adapt to external environmental conditions and that ACh plays a crucial role in modulating behaviour, particularly in volatile environments. The right balance of flexibility is essential for successful performance in everyday tasks. Maladaptive adjustments to the environment are

associated with neuropsychiatric disorders; for example, patients with depression, anxiety, or obsessive-compulsive disorder (OCD) exhibit a maladaptive increase in learning rate under uncertainty (Aylward et al., 2019; Huang et al., 2017; Pulcu and Browning, 2019; Scholl and Rushworth, 2017). In order to understand these pathological conditions, it is crucial to, first, understand the response strategies in different environments, and, second, to understand how neuromodulators and neurotransmitters affect neuronal dynamics and ultimately behaviour. This dissertation contributes to the broader understanding of flexible decision making and the role of ACh in reinforcement learning, laying a foundation for potential therapeutic approaches to diseases associated with altered ACh levels, such as Alzheimer's disease.

In both studies, computational models were key to understanding the mechanisms underlying observed behaviours. These models uncovered hidden variables that conventional analyses may have missed. In Study I, Bayesian hierarchical models revealed that blocking muscarinic M1 receptors impairs learning in volatile environments, rather than valuation, leading to noisier estimates of the environment. In Study II, model comparisons showed that improved instrumental performance following the removal of reinforcement was not due to latent task knowledge, but rather to a more cautious response mode. These findings underscore the importance of computational modelling in avoiding misinterpretation of behavioural effects and in providing insights beyond what behavioural measures alone can reveal.

In conclusion, this thesis advances our understanding of decision strategies and reinforcement learning processes, particularly the role of muscarinic ACh in these functions. It also highlights the value of computational methods in gaining deeper insights into behavioural dynamics, offering a more nuanced approach than conventional analyses.

References

Algermissen J, Swart JC, Scheeringa R, Cools R, Den Ouden HEM (2024) Prefrontal signals precede striatal signals for biased credit assignment in motivational learning biases. *Nature Communications* 15:19.

Ananth MR, Rajebhosale P, Kim R, Talmage DA, Role LW (2023) Basal forebrain cholinergic signalling: development, connectivity and roles in cognition. *Nature Reviews Neuroscience* 24:233–251.

Anjum MF, Espinoza AI, Cole RC, Singh A, May P, Uc EY, Dasgupta S, Narayanan NS (2024) Resting-state EEG measures cognitive impairment in Parkinson's disease. *npj Parkinson's Disease* 10:6.

Aosaki T, Miura M, Suzuki T, Nishimura K, Masuda M (2010) Acetylcholine–dopamine balance hypothesis in the striatum: An update. *Geriatrics & Gerontology International* 10.

Asher JM, Hibbard PB (2020) No effect of feedback, level of processing or stimulus presentation protocol on perceptual learning when easy and difficult trials are interleaved. *Vision Research* 176:100–117.

Avery MC, Nitz DA, Chiba AA, Krichmar JL (2012) Simulation of cholinergic and noradrenergic modulation of behavior in uncertain environments. *Frontiers in Computational Neuroscience* 6.

Aylward J, Valton V, Ahn WY, Bond RL, Dayan P, Roiser JP, Robinson OJ (2019) Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature Human Behaviour* 3:1116–1123.

Ballinger E, Ananth M, Talmage D, Role L (2016) Basal Forebrain Cholinergic Circuits and Signaling in Cognition and Cognitive Decline. *Neuron* 91:1199–1218.

Baribault B, Collins AGE (2023) Troubleshooting Bayesian cognitive models. *Psychological Methods* .

Beck AT, Steer RA, Brown G (1996) Beck Depression Inventory–II Institution: American Psychological Association.

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. *Nature Neuroscience* 10:1214–1221.

Bentley P, Driver J, Dolan RJ (2011) Cholinergic modulation of cognition: Insights from human pharmacological functional neuroimaging. *Progress in Neurobiology* 94:360–388.

Bernoulli D (1954) Exposition of a New Theory on the Measurement of Risk. *Econometrica* 22:23.

Bessie Aramakis V, Bandrowski AE, Ashe JH (1997) Muscarinic reduction of GABAergic synaptic potentials results in disinhibition of the AMPA/kainate-mediated EPSP in auditory cortex. *Brain Research* 758:107–117.

Biernacki K, Molokotos E, Han C, Dillon DG, Leventhal AM, Janes AC (2023) Enhanced decision-making in nicotine dependent individuals who abstain: A computational analysis using Hierarchical Drift Diffusion Modeling. *Drug and Alcohol Dependence* 250:110890.

Birnbaum MH (2008) New paradoxes of risky decision making. *Psychological Review* 115:463–501.

Blain B, Rutledge RB (2020) Momentary subjective well-being depends on learning and not reward. *eLife* 9:e57977.

Bland AR, Schaefer A (2012) Different Varieties of Uncertainty in Human Decision-Making. *Frontiers in Neuroscience* 6.

Bolden C, Cusack B, Richelson E (1992) Antagonism by antimuscarinic and neuroleptic compunds at the five cloned human muscarinic cholinergic receptors expessed in Chinese hamster ovary cells. *Journal of Pharmacology and Experimental Therapeutics* 260:576–580.

Boorman ED, Behrens TE, Woolrich MW, Rushworth MF (2009) How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron* 62:733–743.

Bradley SJ, Bourgognon JM, Sanger HE, Verity N, Mogg AJ, White DJ, Butcher AJ, Moreno JA, Molloy C, Macedo-Hatch T, Edwards JM, Wess J, Pawlak R, Read DJ, Sexton PM, Broad LM, Steinert JR, Mallucci GR, Christopoulos A, Felder CC, Tobin AB (2016) M1 muscarinic allosteric modulators slow prion neurodegeneration and restore memory loss. *Journal of Clinical Investigation* 127:487–499.

Brocks DR (1999) Anticholinergic drugs used in Parkinson's disease: An overlooked class of drugs from a pharmacokinetic perspective. *Journal of Pharmacy & Pharmaceutical Sciences: A Publication of the Canadian Society for Pharmaceutical Sciences, Societe Canadienne Des Sciences Pharmaceutiques* 2:39–46.

Bromberg-Martin ES, Hikosaka O (2009) Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards. *Neuron* 63:119–126.

Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ (2015) Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience* 18:590–596.

Busemeyer JR, Townsend JT (1993) Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* 100:432–459.

Carlson N (2010) *Physiology of Behavior* MyPsychKit Series. Allyn & Bacon, 10th edition.

Cerracchio E, Miletić S, Forstmann BU (2023) Modelling decision-making biases. *Frontiers in Computational Neuroscience* 17:1222924.

Chen ZR, Huang JB, Yang SL, Hong FF (2022) Role of Cholinergic Signaling in Alzheimer's Disease. *Molecules* 27:1816.

Collins AGE, Cockburn J (2020) Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience* 21:576–586.

Cook JL, Swart JC, Froböse MI, Diaconescu AO, Geurts DE, Den Ouden HE, Cools R (2019) Catecholaminergic modulation of meta-learning. *eLife* 8:e51439.

Cools R, Arnsten AFT (2022) Neuromodulation of prefrontal cortex cognitive function in primates: the powerful roles of monoamines and acetylcholine. *Neuropsychopharmacology* 47:309–328.

Dale H (1914) The action of certain esters and ethers of choline, and their relation to muscarine. *Journal of Pharmacology and Experimental Therapeutics* 6:147–190.

Dani JA, Bertrand D (2007) Nicotinic Acetylcholine Receptors and Nicotinic Cholinergic Mechanisms of the Central Nervous System. *Annual Review of Pharmacology and Toxicology* 47:699–729.

Daniel R, Pollmann S (2012) Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage* 59:3457–3467.

Danielmeier C, Allen E, Jocham G, Onur O, Eichele T, Ullsperger M (2015) Acetylcholine Mediates Behavioral and Neural Post-Error Control. *Current Biology* 25:1461–1468.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area. *Science* 319:1264–1267.

Daw N, O'Doherty J (2013) *Multiple Systems for Value Learning*, pp. 393–410 Elsevier Inc.

Dayan P, Niv Y, Seymour B, D. Daw N (2006) The misbehavior of value and the discipline of the will. *Neural Networks* 19:1153–1160.

de Fermat P, Henry C, Tannery P (1891) *Oeuvres de Fermat*, Vol. 1 of *Oeuvres de Fermat* Gauthier-Villars, Paris.

Dias Maile AA, Gründler TO, Froböse MI, Kurtenbach H, Kaiser LF, Jocham G (2024) Bidirectional modulation of reward-guided decision making by dopamine `http://biorxiv.org/lookup/doi/10.1101/2024.03.27.586793`.

Dineley KT, Pandya AA, Yakel JL (2015) Nicotinic ACh receptors as therapeutic targets in CNS disorders. *Trends in Pharmacological Sciences* 36:96–108.

Dorris MC, Glimcher PW (2004) Activity in Posterior Parietal Cortex Is Correlated with the Relative Subjective Desirability of Action. *Neuron* 44:365–378.

Dudai A, Yayon N, Soreq H, London M (2021) Cortical VIP $^{+}$ /ChAT $^{+}$ interneurons: From genetics to function. *Journal of Neurochemistry* 158:1320–1333.

Eckstein MK, Master SL, Dahl RE, Wilbrecht L, Collins AG (2022) Reinforcement learning and Bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience* 55:101106.

Eggermann E, Feldmeyer D (2009) Cholinergic filtering in the recurrent excitatory microcircuit of cortical layer 4. *Proceedings of the National Academy of Sciences* 106:11753–11758.

Ellsberg D (1961) Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics* 75:643.

English BA, Hahn MK, Gizer IR, Mazei-Robison M, Steele A, Kurnik DM, Stein MA, Waldman ID, Blakely RD (2009) Choline transporter gene variation is associated with attention-deficit hyperactivity disorder. *Journal of Neurodevelopmental Disorders* 1:252–263.

Erfanian Abdoust M, Froböse MI, Schnitzler A, Schreivogel E, Jocham G (2024) Dopamine and acetylcholine have distinct roles in delay- and effort-based decision-making in humans. *PLOS Biology* 22:e3002714.

Euteneuer F, Schaefer F, Stuermer R, Boucsein W, Timmermann L, Barbe MT, Ebersbach G, Otto J, Kessler J, Kalbe E (2009) Dissociation of decision-making under ambiguity and decision-making under risk in patients with Parkinson's disease: A neuropsychological and psychophysiological study. *Neuropsychologia* 47:2882–2890.

Everitt BJ, Robbins TW (1997) Central cholinergic systems and cognition. *Annual Review of Psychology* 48:649–684.

Ewins AJ (1914) Acetylcholine, a New Active Principle of Ergot. *Biochemical Journal* 8:44–49.

Farashahi S, Donahue CH, Hayden BY, Lee D, Soltani A (2019) Flexible combination of reward information across primates. *Nature Human Behaviour* 3:1215–1224.

Florin E, Baillet S (2015) The brain's resting-state activity is shaped by synchronized cross-frequency coupling of neural oscillations. *NeuroImage* 111:26–35.

Fujisawa S, Buzsáki G (2011) A 4 Hz Oscillation Adaptively Synchronizes Prefrontal, VTA, and Hippocampal Activities. *Neuron* 72:153–165.

Galvin VC, Yang ST, Paspalas CD, Yang Y, Jin LE, Datta D, Morozov YM, Lightbourne TC, Lowet AS, Rakic P, Arnsten AF, Wang M (2020) Muscarinic M1 Receptors Modulate Working Memory Performance and Activity via KCNQ Potassium Channels in the Primate Prefrontal Cortex. *Neuron* 106:649–661.e4.

Gauthier I, Tarr MJ (2002) Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance* 28:431–446.

Gedankien T, Tan RJ, Qasim SE, Moore H, McDonagh D, Jacobs J, Lega B (2023) Acetylcholine modulates the temporal dynamics of human theta oscillations during memory. *Nature Communications* 14:5283.

Gershman SJ (2015) Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review* 22:1320–1327.

Gottfried JA, O'Doherty J, Dolan RJ (2003) Encoding Predictive Reward Value in Human Amygdala and Orbitofrontal Cortex. *Science* 301:1104–1107.

Gratton C, Yousef S, Aarts E, Wallace DL, D'Esposito M, Silver MA (2017) Cholinergic, But Not Dopaminergic or Noradrenergic, Enhancement Sharpens Visual Spatial Perception in Humans. *The Journal of Neuroscience* 37:4405–4415.

Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* Wiley, New York.

Grimaldi R, Perucca E, Ruberto G, Gelmi C, Trimarchi F, Hollmann M, Crema A (1986) Pharmacokinetic and pharmacodynamic studies following the intravenous and oral administration of the antiparkinsonian drug biperiden to normal subjects. *European Journal of Clinical Pharmacology* 29:735–737.

Guggenmos M, Wilbertz G, Hebart MN, Sterzer P (2016) Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* 5:e13388.

Guitart-Masip M, Duzel E, Dolan R, Dayan P (2014) Action versus valence in decision making. *Trends in Cognitive Sciences* 18:194–202.

Guitart-Masip M, Walsh A, Dayan P, Olsson A (2023) Anxiety associated with perceived uncontrollable stress enhances expectations of environmental volatility and impairs reward learning. *Scientific Reports* 13:18451.

Haddara N, Rahnev D (2022) The Impact of Feedback on Perceptual Decision-Making and Metacognition: Reduction in Bias but No Change in Sensitivity. *Psychological Science* 33:259–275.

Hangya B, Ranade S, Lorenc M, Kepecs A (2015) Central Cholinergic Neurons Are Rapidly Recruited by Reinforcement Feedback. *Cell* 162:1155–1168.

Hasselmo ME, Sarter M (2011) Modes and Models of Forebrain Cholinergic Neuromodulation of Cognition. *Neuropsychopharmacology* 36:52–73.

Herzog MH, Fahle M (1997) The role of feedback in learning a vernier discrimination task. *Vision Research* 37:2133–2141.

Higley MJ, Gittis AH, Oldenburg IA, Balthasar N, Seal RP, Edwards RH, Lowell BB, Kreitzer AC, Sabatini BL (2011) Cholinergic Interneurons Mediate Fast VGluT3-Dependent Glutamatergic Transmission in the Striatum. *PLoS ONE* 6:e19155.

Higley MJ, Picciotto MR (2014) Neuromodulation by acetylcholine: examples from schizophrenia and depression. *Current Opinion in Neurobiology* 29:88–95.

Howe WM, Gritton HJ, Lusk NA, Roberts EA, Hetrick VL, Berke JD, Sarter M (2017) Acetylcholine Release in Prefrontal Cortex Promotes Gamma Oscillations and Theta–Gamma Coupling during Cue Detection. *The Journal of Neuroscience* 37:3215–3230.

Huang H, Thompson W, Paulus MP (2017) Computational Dysfunctions in Anxiety: Failure to Differentiate Signal From Noise. *Biological Psychiatry* 82:440–446.

Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MFS, Behrens TEJ (2012) Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience* 15:470–476.

Iglesias S, Kasper L, Harrison SJ, Manka R, Mathys C, Stephan KE (2021) Cholinergic and dopaminergic effects on prediction error and uncertainty responses during sensory associative learning. *NeuroImage* 226:117590.

Jaramillo-Jimenez A, Suarez-Revelo JX, Ochoa-Gomez JF, Carmona Arroyave JA, Bocanegra Y, Lopera F, Buriticá O, Pineda-Salazar DA, Moreno Gómez L, Tobón Quintero CA, Borda MG, Bonanni L, Ffytche DH, Brønnick K, Aarsland D (2021) Resting-state EEG alpha/theta ratio related to neuropsychological test performance in Parkinson's Disease. *Clinical Neurophysiology* 132:756–764.

Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M (2009) Adaptive Coding of Action Values in the Human Rostral Cingulate Zone. *Journal of Neuroscience* 29:7489–7496.

Jocham G, Furlong PM, Kröger IL, Kahn MC, Hunt LT, Behrens TE (2014) Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. *NeuroImage* 100:498–506.

Jocham G, Hunt LT, Near J, Behrens TEJ (2012) A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nature Neuroscience* 15:960–961.

Jones PR, Moore DR, Shub DE, Amitay S (2015) The role of response bias in perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41:1456–1470.

Jones S, Sudweeks S, Yakel JL (1999) Nicotinic receptors in the brain: correlating physiology with function. *Trends in Neurosciences* 22:555–561.

Kahneman D, Tversky A (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47:263.

Kaiser LF, Gruendler TOJ, Speck O, Luettgau L, Jocham G (2021) Dissociable roles of cortical excitation-inhibition balance during patch-leaving versus value-guided decisions. *Nature Communications* 12:904.

Katz B, Miledi R (1965) The measurement of synaptic delay, and the time course of acetylcholine release at the neuromuscular junction. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 161:483–495.

Kljakic O, Janickova H, Prado VF, Prado MAM (2017) Cholinergic/glutamatergic cotransmission in striatal cholinergic interneurons: new mechanisms regulating striatal computation. *Journal of Neurochemistry* 142:90–102.

Knight F (1921) *Risk, Uncertainty, and Profit* Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Co.

Kopčanová M, Tait L, Donoghue T, Stothart G, Smith L, Flores-Sandoval AA, Davila-Perez P, Buss S, Shafi MM, Pascual-Leone A, Fried PJ, Benwell CS (2024) Resting-state EEG signatures of Alzheimer's disease are driven by periodic but not aperiodic changes. *Neurobiology of Disease* 190:106380.

Kruschke JK, Liddell TM (2018) Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review* 25:155–177.

Kuchibhotla KV, Gill JV, Lindsay GW, Papadoyannis ES, Field RE, Sten TAH, Miller KD, Froemke RC (2017) Parallel processing by cortical inhibition enables context-dependent behavior. *Nature Neuroscience* 20:62–71.

Kuchibhotla KV, Hindmarsh Sten T, Papadoyannis ES, Elnozahy S, Fogelson KA, Kumar R, Boubenec Y, Holland PC, Ostojic S, Froemke RC (2019) Dissociating task acquisition from expression during learning reveals latent knowledge. *Nature Communications* 10:2151.

Kurtenbach H, Froböse MI, Ort E, Bahners BH, Hirschmann J, Butz M, Schnitzler A, Jocham G (2024) A role for acetylcholine in reinforcement learning and decision making under uncertainty `http://biorxiv.org/lookup/doi/10.1101/2024.09.20.614105`.

Kurtenbach H, Ort E, Froböse MI, Jocham G (2022) Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations. *PLOS Computational Biology* 18:e1010201.

Lee D, Seo H, Jung MW (2012) Neural Basis of Reinforcement Learning and Decision Making. *Annual Review of Neuroscience* 35:287–308.

Lee MD (2011) How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology* 55:1–7.

Lee MD, Wagenmakers EJ (2013) *Bayesian cognitive modeling: a practical course* Cambridge University Press, Cambridge.

Levin ED, McClernon FJ, Rezvani AH (2006) Nicotinic effects on cognitive function: behavioral characterization, pharmacological specification, and anatomic localization. *Psychopharmacology* 184:523–539.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* Wiley, New York.

Luettgau L, Tempelmann C, Kaiser LF, Jocham G (2020) Decisions bias future choices by modifying hippocampal associative memories. *Nature Communications* 11:3318.

Lynn SK, Barrett LF (2014) "Utilizing" Signal Detection Theory. *Psychological Science* 25:1663–1673.

Macmillan N, Creelman C (2005) *Detection Theory: A User's Guide* Lawrence Erlbaum Associates.

Marco-Pallares J, Cucurell D, Cunillera T, García R, Andrés-Pueyo A, Münte TF, Rodríguez-Fornells A (2008) Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia* 46:241–248.

Marino MJ, Rouse ST, Levey AI, Potter LT, Conn PJ (1998) Activation of the genetically defined m1 muscarinic receptor potentiates *N*-methyl-d-aspartate (NMDA) receptor currents in hippocampal pyramidal cells. *Proceedings of the National Academy of Sciences* 95:11465–11470.

Marshall L, Mathys C, Ruge D, De Berker AO, Dayan P, Stephan KE, Bestmann S (2016) Pharmacological Fingerprints of Contextual Uncertainty. *PLOS Biology* 14:e1002575.

Massar S, Rossi V, Schutter D, Kenemans J (2012) Baseline eeg theta/beta ratio and punishment sensitivity as biomarkers for feedback-related negativity (frn) and risk-taking. *Clinical Neurophysiology* 123:1958–1965.

Massar SA, Kenemans JL, Schutter DJ (2014) Resting-state eeg theta activity and risk learning: sensitivity to reward or punishment? *International Journal of Psychophysiology* 91:172–177.

Massi B, Donahue CH, Lee D (2018) Volatility Facilitates Value Updating in the Prefrontal Cortex. *Neuron* 99:598–608.e4.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior In *Frontiers in Econometrics*, pp. 105–142. Academic press, New York.

McKay BE, Placzek AN, Dani JA (2007) Regulation of synaptic transmission and plasticity by neuronal nicotinic acetylcholine receptors. *Biochemical Pharmacology* 74:1120–1133.

Mesulam M, Mufson EJ, Levey AI, Wainer BH (1983) Cholinergic innervation of cortex by the basal forebrain: Cytochemistry and cortical connections of the septal area, diagonal band nuclei, nucleus basalis (Substantia innominata), and hypothalamus in the rhesus monkey. *Journal of Comparative Neurology* 214:170–197.

Mineur YS, Picciotto MR (2021) The role of acetylcholine in negative encoding bias: Too much of a good thing? *European Journal of Neuroscience* 53:114–125.

Mirenowicz J, Schultz W (1994) Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology* 72:1024–1027.

Molter F, Thomas AW, Huettel SA, Heekeren HR, Mohr PNC (2022) Gaze-dependent evidence accumulation predicts multi-alternative risky choice behaviour. *PLOS Computational Biology* 18:e1010283.

Obermayer J, Verhoog MB, Luchicchi A, Mansvelder HD (2017) Cholinergic Modulation of Cortical Microcircuits Is Layer-Specific: Evidence from Rodent, Monkey and Human Brain. *Frontiers in Neural Circuits* 11:100.

Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441:223–226.

Palminteri S, Kilford EJ, Coricelli G, Blakemore SJ (2016) The Computational Development of Reinforcement Learning during Adolescence. *PLOS Computational Biology* 12:e1004953.

Peper K, Bradley RJ, Dreyer F (1982) The acetylcholine receptor at the neuromuscular junction. *Physiological Reviews* 62:1271–1340.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442:1042–1045.

Peterson W, Birdsall T, Fox W (1954) The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory* 4:171–212.

Petrov AA, Dosher BA, Lu ZL (2006) Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research* 46:3177–3197.

Ptasczynski LE, Steinecker I, Sterzer P, Guggenmos M (2022) The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PLOS Computational Biology* 18:e1010580.

Pulcu E, Browning M (2019) The Misestimation of Uncertainty in Affective Disorders. *Trends in Cognitive Sciences* 23:865–875.

Quiggin J (1982) A theory of anticipated utility. *Journal of Economic Behavior & Organization* 3:323–343.

Rachlin H, Battalio R, Kage J, Green L (1983) The concept of leisure in maximization theory. *Behavioral and Brain Sciences* 6:330–333.

Rahnev D (2021) Response Bias Reflects Individual Differences in Sensory Encoding. *Psychological Science* 32:1157–1168.

Ratcliff R (1978) A theory of memory retrieval. *Psychological Review* 85:59–108.

Ratcliff R, McKoon G (2008) The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation* 20:873–922.

Rescorla RA, Wagner AR (1972) A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement In *Classical Conditioning II: Current Research and Theory*, pp. 64–99. Appleton-Century-Crofts, New York.

Ritvo VJ, Turk-Browne NB, Norman KA (2019) Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. *Trends in Cognitive Sciences* 23:726–742.

Robbins T, Roberts A (2007) Differential Regulation of Fronto-Executive Function by the Monoamines and Acetylcholine. *Cerebral Cortex* 17:i151–i160.

Rodriguez R, Kallenbach U, Singer W, Munk MHJ (2004) Short- and Long-Term Effects of Cholinergic Modulation on Gamma Oscillations and Response Synchronization in the Visual Cortex. *The Journal of Neuroscience* 24:10369–10378.

Rogala J, Kublik E, Krauz R, Wróbel A (2020) Resting-state eeg activity predicts frontoparietal network reconfiguration and improved attentional performance. *Scientific Reports* 10.

Role LW, Berg DK (1996) Nicotinic Receptors in the Development and Modulation of CNS Synapses. *Neuron* 16:1077–1085.

Rushworth M, Walton M, Kennerley S, Bannerman D (2004) Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences* 8:410–417.

Scholl J, Günthner J, Kolling N, Favaron E, Rushworth MF, Harmer CJ, Reinecke A (2014) A Role Beyond Learning for NMDA Receptors in Reward-Based Decision-Making—a Pharmacological Study Using d-Cycloserine. *Neuropsychopharmacology* 39:2900–2909.

Scholl J, Rushworth MF (2017) Obsessing about Uncertainty? *Neuron* 96:250–252.

Scholz RW (1983) Introduction to Decision Making Under Uncertainty: Biases, Fallacies, and the Development of Decision Making In *Advances in Psychology*, Vol. 16, pp. 3–18. Elsevier.

Schutter DJ, Van Honk J (2005) Electrophysiological ratio markers for the balance between reward and punishment. *Cognitive Brain Research* 24:685–690.

Shirey JK, Brady AE, Jones PJ, Davis AA, Bridges TM, Kennedy JP, Jadhav SB, Menon UN, Xiang Z, Watson ML, Christian EP, Doherty JJ, Quirk MC, Snyder DH, Lah JJ, Levey AI, Nicolle MM, Lindsley CW, Conn PJ (2009) A Selective Allosteric Potentiator of the M_1 Muscarinic Acetylcholine Receptor Increases Activity of Medial Prefrontal Cortical Neurons and Restores Impairments in Reversal Learning. *The Journal of Neuroscience* 29:14271–14286.

Sine SM (2012) End-Plate Acetylcholine Receptor: Structure, Mechanism, Pharmacology, and Disease. *Physiological Reviews* 92:1189–1234.

Skinner B (1938) *The behavior of organisms* Appleton-Century-Crofts, New York.

Soltani A, Izquierdo A (2019) Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience* 20:635–644.

Soltani A, Khorsand P, Guo C, Farashahi S, Liu J (2016) Neural substrates of cognitive biases during probabilistic inference. *Nature Communications* 7:11393.

Soltani A, Wang XJ (2006) A Biophysically Based Neural Model of Matching Law Behavior: Melioration by Stochastic Synapses. *The Journal of Neuroscience* 26:3731–3744.

Spielberger CD (1983) State-Trait Anxiety Inventory for Adults Institution: American Psychological Association.

Stagner JP, Zentall TR (2010) Suboptimal choice behavior by pigeons. *Psychonomic Bulletin & Review* 17:412–416.

Stewart N (2011) Information Integration in Risky Choice: Identification and Stability. *Frontiers in Psychology* 2.

Stifani N (2014) Motor neurons and the generation of spinal motor neuron diversity. *Frontiers in Cellular Neuroscience* 8.

Stoica P, Selen Y (2004) Model-order selection. *IEEE Signal Processing Magazine* 21:36–47.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* The MIT Press, second edition.

Swart JC, Froböse MI, Cook JL, Geurts DE, Frank MJ, Cools R, den Ouden HE (2017) Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *eLife* 6:e22169.

Thiele A (2013) Muscarinic signaling in the brain. *Annual Review of Neuroscience* 36:271–294.

Thorndike EL (1927) The Law of Effect. *The American Journal of Psychology* 39:212.

Tversky A (1967) Utility theory and additivity analysis of risky choices. *Journal of Experimental Psychology* 75:27–36.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5:297–323.

Uhlhaas PJ, Singer W (2006) Neural Synchrony in Brain Disorders: Relevance for Cognitive Dysfunctions and Pathophysiology. *Neuron* 52:155–168.

Vahedi J, Mundorf A, Bellebaum C, Peterburs J (2024) Emotional cues reduce Pavlovian interference in feedback-based go and nogo learning. *Psychological Research* 88:1212–1230.

Van De Vijver I, Ridderinkhof KR, Cohen MX (2011) Frontal Oscillatory Dynamics Predict Feedback Learning and Action Adjustment. *Journal of Cognitive Neuroscience* 23:4106–4121.

van Lieshout LLF, Traast IJ, de Lange FP, Cools R (2021) Curiosity or savouring? Information seeking is modulated by both uncertainty and valence. *PLOS ONE* 16:e0257011.

Vossel S, Bauer M, Mathys C, Adams RA, Dolan RJ, Stephan KE, Friston KJ (2014) Cholinergic Stimulation Enhances Bayesian Belief Updating in the Deployment of Spatial Attention. *The Journal of Neuroscience* 34:15735–15742.

Wang XJ (2002) Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron* 36:955–968.

Wilson RC, Collins AG (2019) Ten simple rules for the computational modeling of behavioral data. *eLife* 8:e49547.

Wong KF, Wang XJ (2006) A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *The Journal of Neuroscience* 26:1314–1328.

Woolf N (1991) Cholinergic systems in mammalian brain and spinal cord. *Progress in Neurobiology* 37:475–524.

Wunderlich K, Dayan P, Dolan RJ (2012) Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience* 15:786–791.

Wunderlich K, Rangel A, O'Doherty JP (2009) Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences* 106:17199–17204.

Wunderlich K, Rangel A, O'Doherty JP (2010) Economic choices can be made using only stimulus values. *Proceedings of the National Academy of Sciences* 107:15005–15010.

Yokagawa K, Nakashima E, Ishizaki J, Hasegawa M, Kido H, Ichimura F (1992) Brain regional pharmacokinetics of biperiden in rats. *Biopharmaceutics & Drug Disposition* 13:131–140.

Yu AJ, Dayan P (2005) Uncertainty, Neuromodulation, and Attention. *Neuron* 46:681–692.

Zwart R, Reed H, Sher E (2018) Oxotremorine-M potentiates NMDA receptors by muscarinic receptor dependent and independent mechanisms. *Biochemical and Biophysical Research Communications* 495:481–486.

List of abbreviations

ACh	acetylcholine
BIC	Bayesian information criterion
CNS	central nervous system
DDM	drift-diffusion model
EV	expected value
FAR	false alarm rate
GABA	γ-Aminobutyric acid
HDI	highest density interval
HR	hit rate
mAChR	muscarinic acetylcholine receptor
MEG	magnetoencephalography
MSE	mean squared error
nAChR	nicotinic acetylcholine receptor
NMDA	N-methyl-D-aspartate
PE	prediction error
PFC	prefrontal cortex
PNS	peripheral nervous system
RL	reinforcement learning
SDT	signal detection theory
vmPFC	ventromedial prefrontal cortex

List of publications

This dissertation comprises two publications:

- **Kurtenbach H**, Froböse MI, Ort E, Bahners BH, Hirschmann J, Butz M, Schnitzler A, & Jocham G. *A role for acetylcholine in reinforcement learning and decision making under uncertainty.* bioRxiv (2024). Digital Object Identifier (DOI): `https://doi.org/10.1101/2024.09.20.614105`.

  Statement of contribution
  Funding acquisition and conceptualisation of the study (GJ), study design (HK, MIF, EO, GJ), medical support (BHB, MB, AS), MEG support (JH, MB, AS), data acquisition (HK, MIF, EO), data analysis, computational modelling, visualisation of results, and writing of the original draft of the manuscript (HK), discussion of analyses (HK, MIF, EO, GJ), discussion of results, editing of the manuscript and approving of the final version of the manuscript (all authors)

- **Kurtenbach H**, Ort E, Froböse MI, & Jocham G. *Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations.* PLOS Computational Biology **18(12)**, e1010201 (2022). Digital Object Identifier (DOI): `https://doi.org/10.1371/journal.pcbi.1010201`.

  Statement of contribution
  Funding acquisition (GJ), conceptualisation of the study and study design (all authors), data analysis, computational modelling, visualisation of results, and writing the original draft of the manuscript (HK), discussion of analyses and results, editing of the manuscript and approving of the final version of the manuscript (all authors)

Publications produced during my PhD, but not included in this dissertation:

- Dias Maile AA, Gründler TOJ, Froböse MI, **Kurtenbach H**, Kaiser LF, & Jocham G. *Bidirectional modulation of reward-guided decision making by dopamine*. bioRxiv (under review at Psychopharmacology) (2024). Digital Object Identifier (DOI): `https://doi.org/10.1101/2024.03.27.586793`.

Conference contributions

- **Kurtenbach H**, Froböse MI, Ort E, Butz M, Schnitzler A, & Jocham G. *The Role of Acetylcholine in Reward-Guided Decision Making Under Different Degrees of Uncertainty*. 48. Jahrestagung Psychologie & Gehirn 2023 in Tübingen, Germany (poster).
- **Kurtenbach H**, Froböse MI, Ort E, Butz M, Schnitzler A, & Jocham G. *The Role of Acetylcholine in Reward-Guided Decision Making Under Different Degrees of Uncertainty*. Eleventh Symposium on Biology of Decision Making 2023 in Paris, France (poster).
- **Kurtenbach H**, Ort E, Froböse MI, & Jocham G. *Removal of reinforcement reduces propensity to respond in instrumental learning*. 47. Jahrestagung Psychologie & Gehirn 2022 in Freiburg, Germany (poster).
- **Kurtenbach H**, Ort E, Froböse MI, & Jocham G. *Unmasking instrumental associations by removing reinforcement*. International Conference of Cognitive Neuroscience 2022 in Helsinki, Finland (poster).
- **Kurtenbach H**, Ort E, Froböse MI, Rakhshan M, Soltani A, Butz M, Schnitzler A, & Jocham G. *From local computations to system-level dynamics: The role of neurotransmitters in neural oscillations and decision making*. 5th Symposium on cutting-edge methods for EEG research 2021 in Aix-en-Provence, France (poster).

Research articles

Study I: A role for acetylcholine in reinforcement learning and decision making under uncertainty

Reproduced from

Kurtenbach H, Froböse MI, Ort E, Bahners BH, Hirschmann J,
Butz M, Schnitzler A, & Jocham G
A role for acetylcholine in reinforcement learning and decision making under uncertainty
bioRxiv (2024)

Digital Object Identifier (DOI): `https://doi.org/10.1101/2024.09.20.614105`

A role for acetylcholine in reinforcement learning and decision making under uncertainty

Hannah Kurtenbach[1*], Monja Isabel Froböse[1], Eduard Ort[1], Bahne Hendrik Bahners[2,3], Jan Hirschmann[2], Markus Butz[2], Alfons Schnitzler[2,3], Gerhard Jocham[1]

[1] Biological Psychology of Decision Making, Institute of Experimental Psychology, Heinrich Heine University Düsseldorf, Germany

[2] Institute of Clinical Neuroscience and Medical Psychology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Germany

[3] Department of Neurology, Center for Movement Disorders and Neuromodulation, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Germany

*hannah.kurtenbach@hhu.de

Abstract

The neuromodulator acetylcholine has been suggested to govern learning under uncertainty. Here, we investigated the role of muscarinic receptors in reward-guided learning and decision making under different degrees of uncertainty. We administered the muscarinic M1 antagonist biperiden (4 mg) to healthy male participants (n = 43) in a within-subjects, placebo-controlled design. Participants performed two tasks that both involved choices between options characterized by two attributes, reward probability and magnitude. In the gambling task, both attributes were explicitly provided, whereas in the learning task, reward probabilities had to be inferred from past experience. In addition, uncertainty was manipulated within the learning task by inclusion of a stable phase with fixed reward contingencies, and a volatile phase with frequent contingency reversals. We show that biperiden did not affect decision making in the gambling task, where no learning was required. However, in the learning task, biperiden reduced the sensitivity to the learnt reward probabilities. Notably, this was primarily driven by choices under higher uncertainty in the volatile phase. Using reinforcement learning models, we reveal that the change in behaviour was caused by noisier estimates of probabilities resulting from maladaptively increased learning rates under biperiden. Together, these findings suggest that muscarinic acetylcholine transmission is involved in controlling learning in highly uncertain contexts, when the demand for carefully calibrated adjustments is highest.

Introduction

Decision making usually involves some degree of uncertainty. Often, information relevant for a choice is not known and needs to be inferred from experience. Alternatively, or in addition to this, the information to be learnt, i.e., choice-outcome contingencies, may be probabilistic and even change over time. When option attributes are not explicitly presented and have to be learnt from trial and error, choice behaviour has been formally described by relatively simple reinforcement learning algorithms (Daw et al., 2011; Lee et al., 2012; Sutton and Barto, 2014; Kurtenbach et al., 2022). Core to these algorithms is the updating of value estimates using the prediction error, i.e., the discrepancy between obtained and expected outcomes. The prediction error is scaled by a learning rate parameter that determines the degree to which the error is used to update value estimates. In real life, the link between choices and outcomes is often fraught with uncertainty, which requires agents to estimate this uncertainty for adaptive decision making (Behrens et al., 2007; Lee et al., 2012). It has been demonstrated that participants adjust learning rates to the statistics of the environment (Jocham et al., 2009; Soltani and Izquierdo, 2019; Iglesias et al., 2021). In particular, learning rates have been found to increase in volatile environments when contingencies change frequently (Behrens et al., 2007; Browning et al., 2015).

Instead of modifying the learning rate, another possibility to adjust choice behaviour to the environment is to change the strategy for value construction. In reward-guided tasks, options are often characterized by two attributes, a reward magnitude and a reward probability, which are used to construct the value of each option. These attributes can be integrated either additively or multiplicatively. Additive integration of these attributes offers more flexibility, which may be beneficial in volatile environments, as it allows for direct comparison between attributes and differential weighting of reward information (Stewart, 2011). Conversely, multiplicative integration by computing the expected value is statistically optimal. This however relies on relatively precise knowledge, or estimates, of decision attributes, which is more likely in stable environments. Unlike the additive strategy, it does not confer the opportunity to down-weight an attribute when it is very uncertainty-laden. Uncertainty-dependent shifts towards a more additive strategy have recently been described in both humans and monkeys (Farashahi et al., 2019).

Acetylcholine is a neurotransmitter that has been suggested to play an important role in learning under uncertainty. A long tradition of research, in particular using

pharmacological and lesion approaches, has firmly established a role for basal forebrain cholinergic neurons in learning and attention (reviewed in e.g. Everitt and Robbins, 1997; Hasselmo and Sarter, 2011). Beyond this, theoretical work proposes that acetylcholine governs decision making in environments with known uncertainty (Yu and Dayan, 2005; Avery et al., 2012). In addition to this presumed role in uncertainty-dependent learning, acetylcholine has been reported to modulate neural circuit dynamics supporting reward-guided choice, irrespective of learning: Activation of cholinergic receptors, especially the muscarinic M1 subtype, enhances the function of excitatory NMDA receptors and activates GABAergic inhibitory circuits (Bessie Aramakis et al., 1997; Marino et al., 1998; Kuchibhotla et al., 2017; Obermayer et al., 2017; Zwart et al., 2018). These two neurotransmitter systems are the key components in recurrent cortical circuit models of decision making, where competition between options is governed by slow excitation at NDMA receptors and GABAergic feedback inhibition (Wang, 2002; Wong and Wang, 2006). This suggests that acetylcholine may affect decision-making computations, beyond its role in learning. Nevertheless, there is a limited understanding of how acetylcholine influences human decision-making processes in the absence of learning and under different kinds of uncertainty within the same individuals.

The current study aimed to investigate the role of muscarinic acetylcholine receptors in both decision-making computations and in learning, using paradigms probing reward-guided decision making under risk and under varying degrees of uncertainty. For this purpose, participants completed two reward-guided choice tasks, once under placebo and once under the muscarinic M1 acetylcholine receptor antagonist biperiden (4 mg). In the first task, participants had to select between two options with explicitly provided attributes, involving risk (gambling task). In the second task, one attribute had to be estimated from experience, requiring participants to learn throughout the task (learning task). Furthermore, the learning task consisted of a stable and a volatile phase, involving varying degrees of uncertainty. We hypothesized that cholinergic M1 antagonism would impair decision making, leading to impaired choice performance in all tasks. Additionally, we expected biperiden to impair learning in both the stable and the volatile phases of the learning task. Notably, we found that biperiden had no effect on choice behaviour in the gambling task. Instead, it decreased the reliance on estimates of optimally learnt reward probability in the learning task.

Reinforcement learning models revealed that biperiden maladaptively increased the learning rate in the volatile phase of the learning task.

Results

We administered the muscarinic M1 antagonist biperiden (4 mg) to healthy male participants (n = 43) in a within-subjects, placebo-controlled design (Fig. 1A). 35 participants guessed correctly on which testing day they had received biperiden with a certainty of 77.5 ± 3.9 (mean ± SEM) on a scale of 1 to 100. Participants performed two reward-guided choice tasks with the goal of maximizing their reward. In the gambling task, participants selected between pairs of independent gambles each associated with a reward probability and reward magnitude (Fig. 1B). Both attributes were explicitly presented to participants: The reward magnitude was provided by the height of rectangular bars and the reward probability was presented numerically underneath each bar. Whether or not the reward was paid out depended on the explicit reward probability, therefore, the gambling task involved risk only. In the learning task, the reward probability had to be learnt from experience. While reward magnitudes were again explicitly expressed by the height of the rectangular bars, reward probabilities were now indicated by two colours, one of which represented a reward probability of 0.7, and the other of 0.3 (Fig. 1B). In the stable phase, the mapping of reward probabilities to the two colours was fixed, whereas in the volatile phase, the contingencies between reward probability and colour reversed several times over the course of the experiment, (Fig. 1C). As a result, the learning task involved different degrees of uncertainty in the stable and volatile phase, respectively.

Biperiden modulates sensitivity to learnt values

To quantify how task manipulations and cholinergic intervention affected choices in both tasks, we used logistic mixed-effects models (see supplementary Fig. S1 for raw behaviour). For these analyses, probability and magnitude were mean-centred and multiplied to calculate the expected values (EVs), which serve as measure for the multiplicative attribute integration, above and beyond the main effects of reward probability and magnitude. For the learning task, probabilities were estimated using a

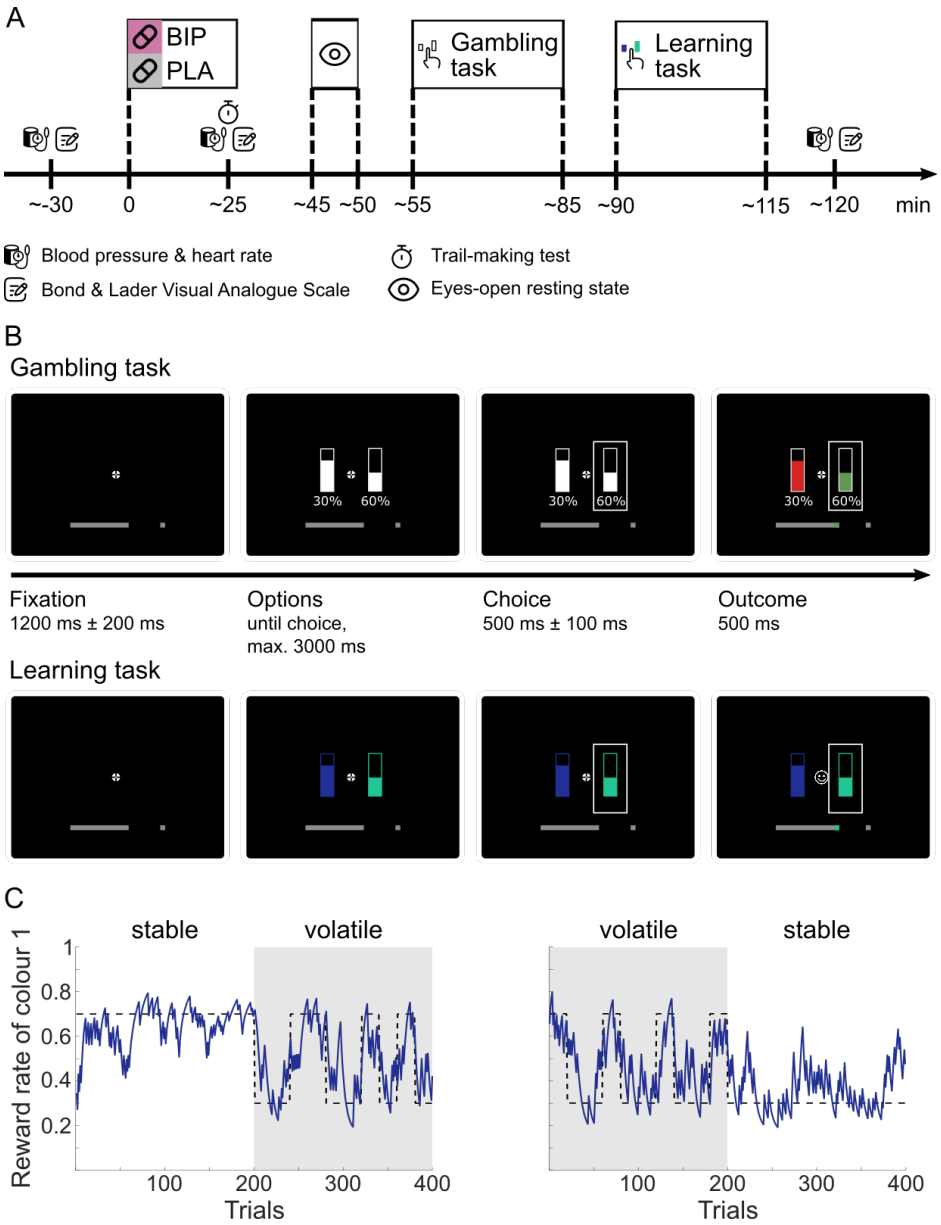

Figure 1. Schematic of study procedure and experimental tasks. A Study procedure. At the beginning of the testing day, blood pressure, heart rate, and mood were measured using the Bond and Lader Visual Analogue Scales. Approximately 30 minutes later, participants received either the drug or a placebo pill. 25 minutes after this, blood pressure, heart rate, and mood were acquired again. Additionally, a trail-making test was conducted. 45 minutes (45.7 min ± 1.6 min, mean ± SEM) after drug, a 5 minutes eyes-open resting state MEG scan was acquired. The gambling task began 55 minutes (55.9 min ± 1.7 min) and the learning task approximately 90 minutes (91.3 min ± 3.6 min) after drug intake. After completion of both tasks, approximately 120 minutes after drug intake, a final measurement of blood pressure, heart

rate, and mood was collected. **B** Example trial of the gambling and the learning task. In the gambling task, reward magnitudes were indicated by the height of the bar and reward probabilities by the numeric percentage below each bar. The outcome of each option was presented as fill colour in green or red to indicate win or no win, respectively. In the learning task, reward probabilities were indicated by the colour of the bars and needed to be learnt during the task. Additionally, outcome was presented as smiley when the winning option was chosen and as frowny otherwise. **C** Example time course of reward contingencies during the learning task. Reward contingencies either remained stable for 200 trials or switched every 20 to 40 trials (volatile phase). Dashed black lines represent true underlying probabilities, solid blue lines represent probability estimates derived from a statistically optimal Bayesian learner (Behrens et al., 2007).

Bayesian optimal learner (Behrens et al., 2007), because participants could not know the true reward probabilities at the outset of the task. In line with previous work (Jocham et al., 2012, 2014; Farashahi et al., 2019; Dias Maile et al., 2024), all task parameters had a significant main effect on choice in the gambling task (probability: β = 2.90, *SEM* = 0.03, $z(42303) = 95.78$, $p < .001$; magnitude: β = 1.58, *SEM* = 0.02, $z(42303) = 73.66$, $p < .001$; EV: β = 0.50, *SEM* = 0.02, $z(42303) = 30.05$, $p < .001$, Fig. 2A). In the learning task, choice behaviour was significantly driven by probability and magnitude, while participants did not use the EV (probability: β = 1.27, *SEM* = 0.02, $z(34175) = 80.28$, $p < .001$; magnitude: β = 0.75, *SEM* = 0.01, $z(34175) = 51.45$, $p < .001$, EV: β = -0.04, *SEM* = 0.02, $z(34175) = -2.30$, $p = .021$; Fig. 2A, see supplementary tables S1 and S2 for all results). Surprisingly, the EV effect was even negative, although very small.

Contrary to our expectation, biperiden diminished the effect of probability on choice only in the learning task, while leaving choices in the gambling task unaffected. Under biperiden, participants relied less on probability (β = -0.05, *SEM* = 0.02, $z(34175) = -3.07$, $p = .002$). This drug effect was dependent on phase (drug x probability x phase: β = 0.04, *SEM* = 0.02, $z(34175) = 2.29$, $p = .022$). Post-hoc tests indicated that there was no significant effect of drug on probability weighting in the stable phase, whereas in the volatile phase, reliance on learnt reward probability was significantly reduced under biperiden (β = -0.08, *SEM* = 0.03, $z(17102) = -3.36$, $p < .001$; Fig. 2B).

Control analyses revealed that, even though biperiden significantly reduced heart rate and subjective reports of alertness, calmness, and contentedness on the Bond and Lader Visual Analogue Scale (BL VAS; Bond and Lader, 1974, see supplementary tables S3-S8, Fig. S2-S4), the biperiden-induced reduction in sensitivity

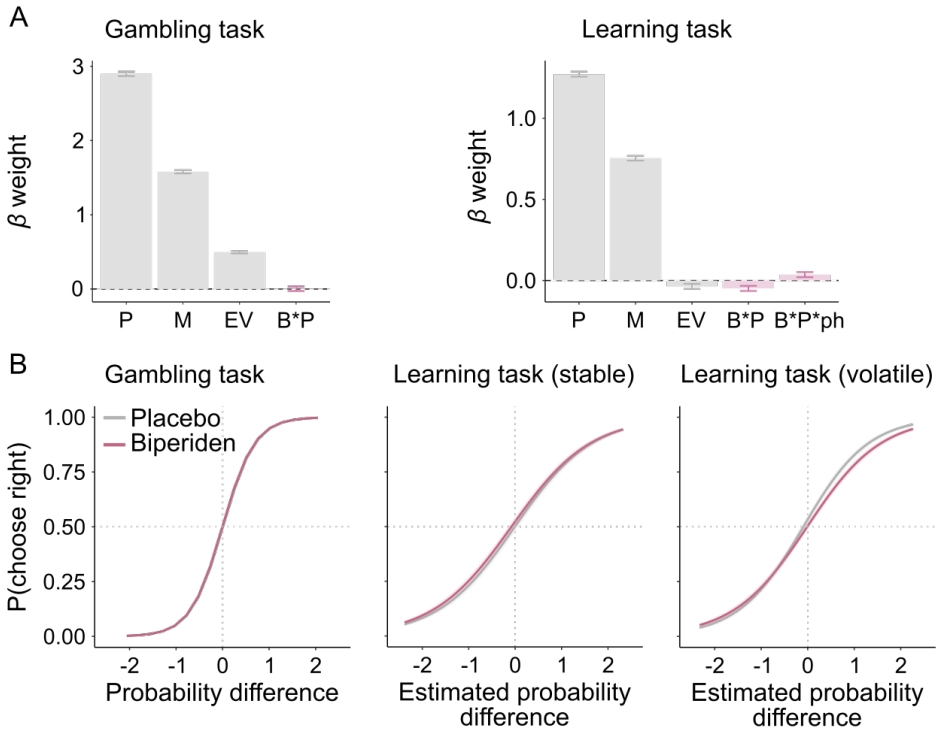

Figure 2. Results from logistic mixed-effects models. A Estimates of task (probability P, magnitude M, and expected value EV) and drug effects for the gambling task (left) and the learning task (right). All regressors are *z*-scored. The interaction between drug B and probability is illustrated for both tasks, the interaction between drug, probability and phase ph is shown for the learning task. **B** Interaction between reward probability and drug. Probability for a right-side choice as a function of differences in reward probability (right versus left option) for the gambling task, and for the stable (middle) and the volatile phase (right) of the learning task. Post-hoc tests following up a significant interaction indicate that in the volatile phase only, sensitivity to reward probabilities was significantly reduced under biperiden (pink) compared to placebo (grey). Solid lines represent mean, shaded areas SEM across participants.

to reward probability in the volatile phase was independent of these effects (see supplementary table S9). Similarly, the effect was independent of the order of stable versus volatile phases (which varied between participants), independent of session days, and independent of the order of drug (see supplementary table S9).

In sum, logistic mixed-effect models revealed that cholinergic modulation occurred only in the volatile phase of the learning task. We observed that, under biperiden, participants made less use of estimated reward probability in the volatile phase. However, note that this probability estimate is derived from a Bayesian optimal learner and thus reflects what participants could ideally know. It is not given however that participants tracked probabilities in such a statistically optimal way. Thus, rather

than reflecting a diminished use of probability information under biperiden, our result could also indicate a failure to optimally track probabilities. These two mechanisms are not mutually exclusive, they might jointly contribute to our behavioural findings. To adjudicate between these possible mechanisms, we used Bayesian hierarchical modelling. Specifically, we asked (i) whether participants adjusted strategies across tasks, irrespective of drug and (ii) whether any such adjustments were modulated by biperiden.

Choice strategies depend on whether learning is involved

Behaviour under risk and uncertainty can be adapted in different ways. One approach is to adjust information integration between tasks; under higher volatility it is assumed that additive rather than multiplicative integration of option attributes is more favourable, because reliable attributes can be weighted more strongly (Farashahi et al., 2019). In situations where one or more attributes have to be learnt from experience, another approach is to increase the learning rate when contingencies change frequently (Behrens et al., 2007). To capture both possible approaches, we set up Bayesian hierarchical models. To foreshadow the findings, using these models, we found adaptations in attribute integration between tasks (gambling *vs.* learning task), but no adaption within the learning task (stable *vs.* volatile).

In the models of both the gambling and the learning task, value can be constructed either in an additive or multiplicative fashion, or using a combination of both, which we call a hybrid strategy. In the additive strategy, values are constructed by simply adding probabilities and magnitudes for one option, after scaling both attributes by a parameter ω_P which indicates the degree to which participants rely more on probabilities relative to magnitudes, or vice versa. In the multiplicative strategy, values result from direct multiplication of probabilities and magnitudes, which corresponds to the economically optimal expected value. Finally, the hybrid strategy features a weighted combination of the additive and the multiplicative strategy, in which the relative dominance of the latter over the former is indicated by the weight ω_{mult} (see equation 1 for details). Note that the purely additive and multiplicative models are nested within this hybrid model - they are special cases with values of ω_{mult} of either 0 or 1, respectively. For the gambling task, we fitted a hybrid model, comprising both additive and multiplicative integration of option attributes, and softmax action selection. For the learning task, we used a similar model and added *Q*-learning with a delta

update rule with learning rates λ_r and λ_u, for rewarded and unrewarded choices, to capture how participants tracked probabilities. Because our logistic mixed-effects models (see above) indicated that, in the learning task, multiplicative integration did not play a role in guiding participants' choices (resulting in convergence issues in the hybrid models), we allowed only additive value construction here (see methods for details). Stable and volatile phases were fitted separately, because the ultimate aim was to detail the phase-specific biperiden effects observed in the logistic mixed-effects models.

In line with the results from logistic mixed-effects modelling, in the gambling task, where both option attributes were explicitly presented, participants used both additive and multiplicative integration. Within the hybrid model, the parameter ω_{mult} was larger than 0 but below 1, which indicates a mixture of both strategies (ω_{mult}: HDI_{mdn} = 0.38, $HDI_{.95}$ = [0.25, 0.52], Fig. 3A). Reward probabilities were weighted more strongly than magnitudes (ω_P: HDI_{mdn} = 0.81, $HDI_{.95}$ = [0.75, 0.87], Fig. 3B). Similarly, in the learning task, participants also focussed slightly more on (inferred) reward probabilities than on magnitude information (stable: ω_P: HDI_{mdn} = 0.59, $HDI_{.95}$ = [0.50, 0.66]; volatile: ω_P: HDI_{mdn} = 0.59, $HDI_{.95}$ = [0.51, 0.67], Fig. 3B), but relative probability weighting was much lower compared to the gambling task. There was no credible difference in relative attribute weighting between stable and volatile phase (Fig. 3B).

In the learning task, we observed that learning rates did not differ credibly between the stable and volatile phase, which was unexpected given earlier reports on learning rate adjustments (Behrens et al., 2007; Browning et al., 2015; Blain and Rutledge, 2020). In both phases, the learning rate for rewarded choices was higher than for unrewarded choices (stable: λ_r: HDI_{mdn} = 0.60, $HDI_{.95}$ = [0.52, 0.68]; λ_u = 0.25, $HDI_{.95}$ = [0.21, 0.30], Fig. 3C, volatile: λ_r: HDI_{mdn} = 0.58, $HDI_{.95}$ = [0.52, 0.64]; λ_u: HDI_{mdn} = 0.26, $HDI_{.95}$ = [0.22, 0.31], Fig. 3D).

Furthermore, as can be expected from the more uncertainty-laden value estimates in the learning task, choice stochasticity was increased compared to the gambling task. The softmax inverse temperature ζ was lower in the learning task compared to the gambling task, irrespective of stable and volatile phases (gambling task: ζ: HDI_{mdn} = 14.70, $HDI_{.95}$ = [12.92, 16.53]; learning task, stable: HDI_{mdn} = 5.58, $HDI_{.95}$ = [4.80, 6.47], learning task, volatile: ζ: HDI_{mdn} = 5.60, $HDI_{.95}$ = [4.83, 6.40], Fig.

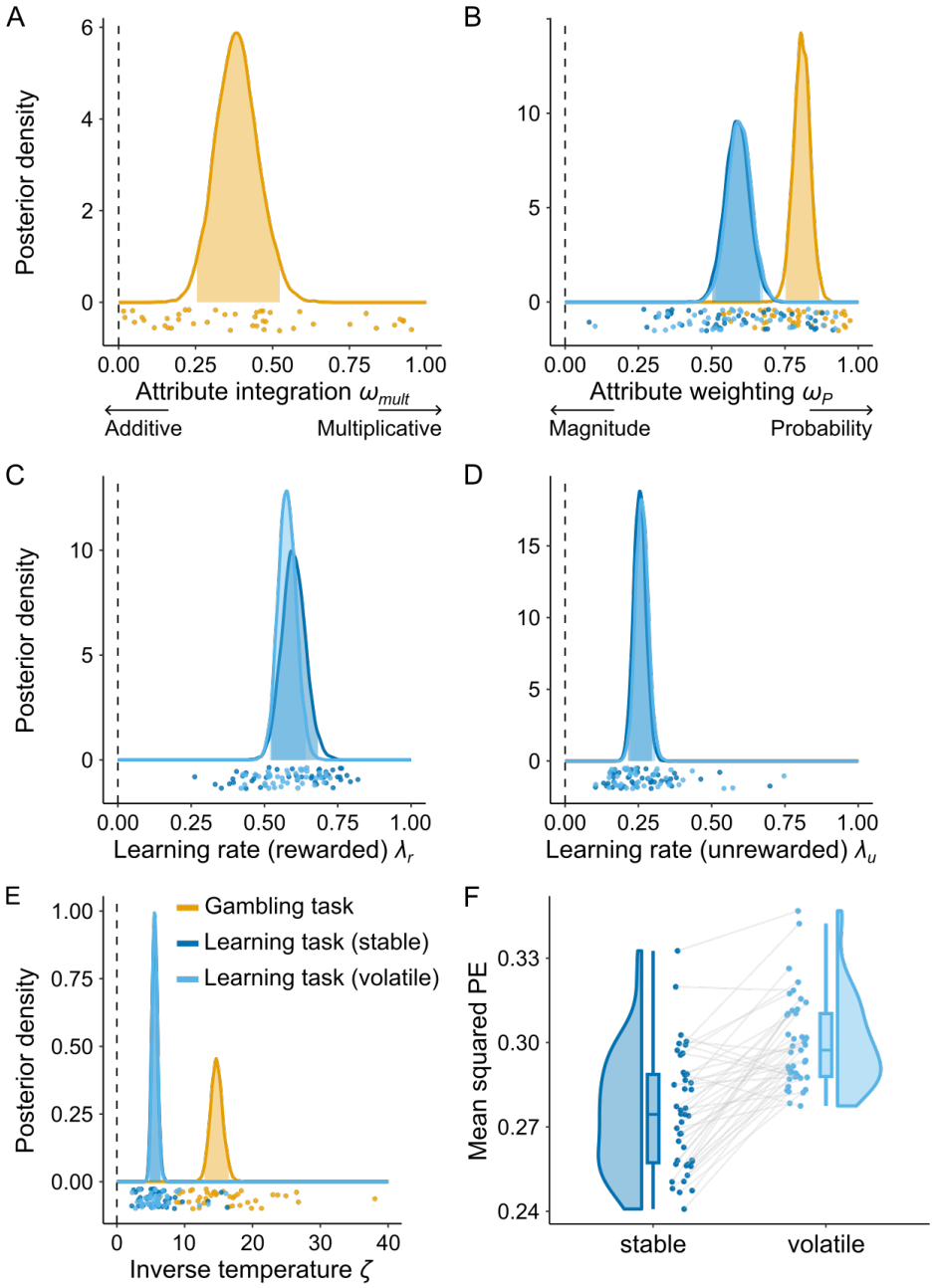

Figure 3. Parameter fits of the Bayesian hierarchical model. A-E Posterior densities for the gambling task (orange) and for the learning task (stable: dark blue, volatile: light blue), irrespective of drug effects. Shaded areas represent the 95 %-HDI of the posterior predictive distribution, points are single-subject means. **A** Strategy used for value construction ω_{mult}. A purely multiplicative integration corresponds to ω_{mult} of 1, whereas ω_{mult} of 0 reflects a purely additive integration of option attributes. For the learning task, ω_{mult} was fixed at 0 (additive

strategy). **B** Weighting of reward probability relative to magnitude ω_P, where decisions based either entirely on probabilities or on magnitudes are reflected by values for ω_P of 1 or 0, respectively. In the gambling task, participants weighted reward probabilities more strongly than in the learning task. **C, D** Learning rates λ_r for rewarded and λ_u for unrewarded choices in the learning task. **E** Softmax inverse temperature ζ. The higher the inverse temperature, the more deterministic the choice behaviour. Choice behaviour in the gambling task was more deterministic compared to the learning task. **F** Mean squared prediction error (PE) in the stable and volatile phase of the learning task. Points reflect individual participants' fit. In the volatile phase, the mean squared PE is significantly higher, indicating a higher degree of overall surprise.

3E). Thus, participants were more deterministic in the gambling task than in the learning task, indicating that they were more sensitive to the estimated value of the options.

Although the fitted model parameters did not differ between the stable and volatile phase, we observed that, overall, participants experienced a higher degree of outcome-related surprise in the volatile phase, evidenced by a significantly higher mean squared prediction error (MSPE) in the volatile compared to the stable phase ($\Delta(MSPE_{vol} - MSPE_{stab}) = 2.48 \cdot 10^{-2} \pm 0.05 \cdot 10^{-2}$, mean ± SEM, $t_{42} = 8.22$, $p < .001$, Cohen's $d = 1.25$, Fig. 3F).

In sum, when both option attributes were explicitly provided, participants used a hybrid attribute integration, consisting of mostly additive but also multiplicative integration. When learning was involved, participants showed no difference in decision strategies for stable versus volatile phases.

Biperiden increases learning rate in highly uncertain environments

Logistic mixed-effects models revealed that biperiden changed the influence of the implicit reward probability on choice behaviour in the volatile phase of the learning task. This effect could emerge from two different causes which can be identified by biperiden-induced shifts of the model parameters introduced above: Biperiden may have diminished the impact of the learnt (uncertainty-laden) attribute on choice, or alternatively, participants' probability estimate is less accurate in the first place. The former possibility should be reflected in the relative attribute weighting, whereas the latter would indicate an effect on learning mechanisms (rather than on choice) and should be reflected in changes in the learning rate. To test for cholinergic effects on these model parameters, we extended the Bayesian hierarchical models with a biperiden-specific shift in each fitted model parameter.

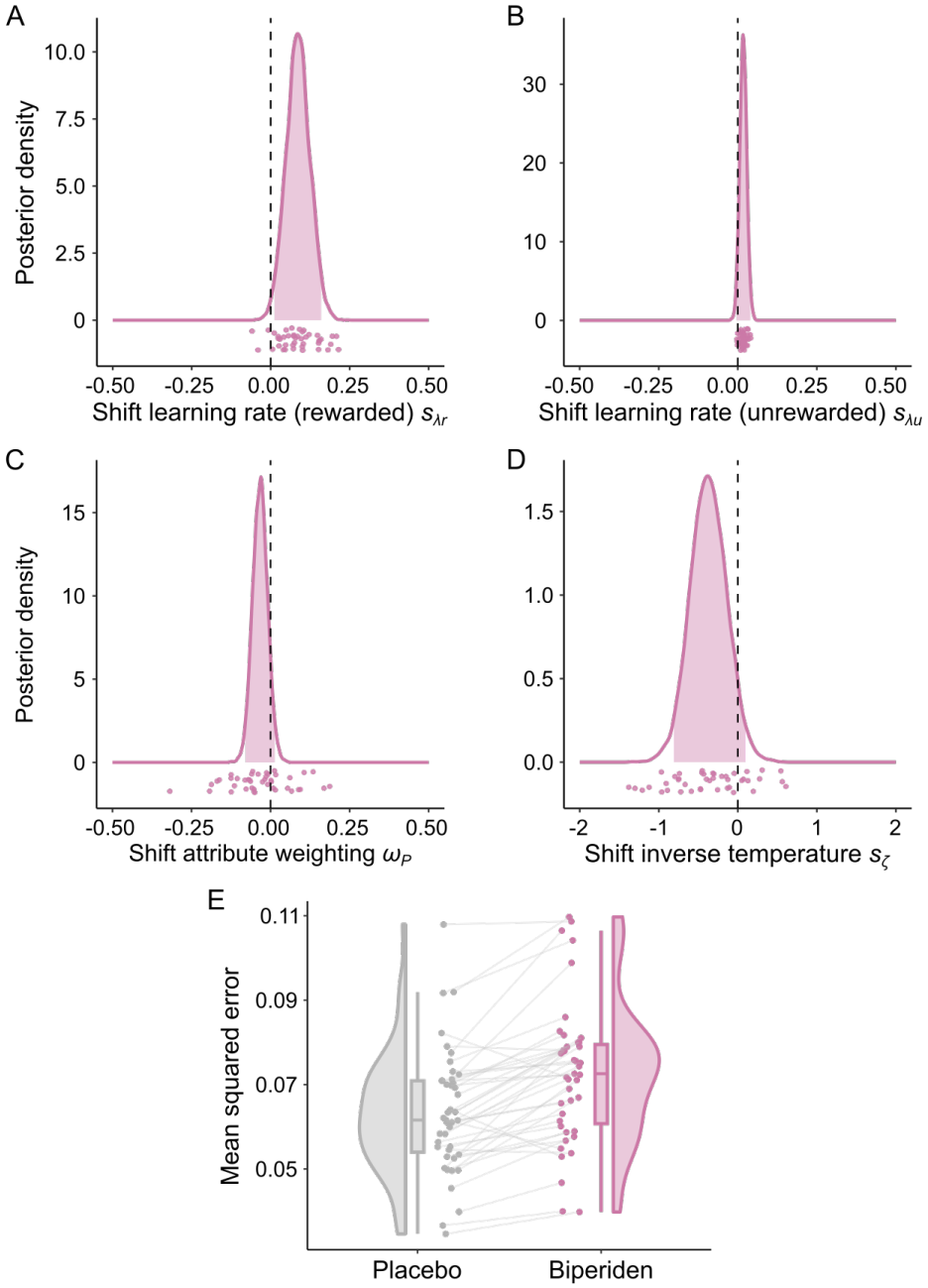

Figure 4. Biperiden-induced shifts of model parameters in the volatile phase of the learning task, derived from the Bayesian hierarchical model. Density of posterior predictive distributions of the biperiden-specific shift in **A**, **B** learning rate in rewarded $s_{\lambda r}$ and unrewarded trials $s_{\lambda u}$, **C** attribute weighting $s_{\omega P}$, and **D** inverse temperature s_ζ. Positive shifts represent an increase under biperiden relative to placebo. The learning rate for rewarded choices is credibly increased under biperiden. Shaded areas represent the 95 %-HDI of the

posterior predictive distribution, points are single-subject means. **E** Mean squared error (MSE) of the deviation of participants' estimated reward probabilities from Bayes optimal estimated probabilities in the placebo session (grey) and biperiden session (pink). Points reflect individual participants' fit. Under biperiden, the MSE is significantly higher, reflecting impaired estimation of the learnt attribute.

In line with the results from logistic mixed-effects models, there were neither significant biperiden-specific shifts in the gambling task nor in the stable phase of the learning task (see supplementary Fig. S5, S6, and tables S10, S11). In the volatile phase, however, there was a credible biperiden-induced increase of the learning rate for rewarded choices $s_{\lambda r}$ ($s_{\lambda r}$: $HDI_{mdn} = 0.09$, $HDI_{.95} = [0.01, 0.16]$, Fig. 4A), but not for other parameters, such as the learning rate for unrewarded choices $s_{\lambda u}$, attribute weighting $s_{\omega P}$, or inverse temperature s_{ζ} (see table 1, Fig. 4B-D). As increased learning rates are generally considered to be more optimal in volatile environments (Behrens et al., 2007; Browning et al., 2015), we explored to what extent the estimated probabilities using participants' fitted learning rates diverged from an optimal learner's estimate. To this end, we computed, for each participant, the mean squared errors (MSE) of these two (trial-wise) probability estimates and compared them between the placebo and biperiden sessions. Under biperiden, the MSE was significantly increased compared to placebo ($\Delta(MSE_{BIP} - MSE_{PLA}) = 9.25 \cdot 10^{-3} \pm 0.21 \cdot 10^{-3}$, $t_{42} = 6.61$, $p = 0.009$, Cohen's $d = 1.01$, Fig. 4E), indicating that the increase learning rate was indeed maladaptive. The learning rate for rewarded choices was already rather high under placebo. Thus, a further increase in learning rates, as observed under biperiden is suboptimal, as it caused noisier value estimates of implicit information.

Taken together, these results indicate that the diminished impact of probability on choice observed under biperiden does not result from participants using this information less to guide their choices. Instead, acetylcholine appears to be involved in appropriately setting the learning rate, particularly under conditions of high uncertainty.

Table 1. Group-level parameter estimates of the Bayesian hierarchical model for the volatile phase of the learning task. Median (Mdn), standard deviation (SD), and lower and upper bounds of the 95 %-HDI interval are presented. The model included parameter estimates for the learning rate of rewarded choices λ_r, the learning rate of unrewarded choices λ_u, the attribute weighting ω_P, the inverse temperature ζ, and the corresponding biperiden-specific shifts on these parameters $s_{\lambda r}$, $s_{\lambda u}$, $s_{\omega P}$, and s_{ζ}.

Parameter	Mdn	SD	2.5 %	97.5 %
λ_r	*0.58*	*0.03*	*0.52*	*0.64*
λ_u	*0.26*	*0.02*	*0.22*	*0.31*
ω_P	*0.59*	*0.04*	*0.51*	*0.67*
ζ	*5.60*	*0.40*	*4.83*	*6.40*
$s_{\lambda r}$	*0.09*	*0.04*	*0.01*	*0.16*
$s_{\lambda u}$	0.02	0.01	0.00	0.04
$s_{\omega P}$	-0.03	0.02	-0.08	0.01
s_{ζ}	-0.37	0.23	-0.81	0.10

Discussion

Decision making requires flexibility, especially, when options or outcomes are uncertain. One approach to adapt behaviour involves increasing the learning rate in volatile environments (Behrens et al., 2007). Alternatively, the choice strategy can be adapted to weight more reliable information more strongly, e.g., using additive rather than multiplicative integration of option attributes under higher volatility (Farashahi et al., 2019). Cholinergic transmission is crucial for learning and uncertainty processing (Everitt and Robbins, 1997; Hasselmo and Sarter, 2011), but its role in the adjustment of decision strategies to environmental risk and uncertainty remains elusive. Therefore, in the present study, we investigated the effects of biperiden, a cholinergic M1 receptor antagonist, on decision making in healthy (male) participants in three scenarios: (i) under risk, when all attributes were explicitly presented, (ii) under uncertainty, when one attribute needed to be learnt, and, on top of learning, (iii) under volatility.

Under the influence of biperiden, participants exhibited no change in choice behaviour under risk in general, but instead only when option attributes needed to be learnt. More specifically, in the learning task, biperiden reduced sensitivity to the learnt reward probabilities selectively in the volatile phase. However, using computational models we could demonstrate that this reduced sensitivity to Bayes-optimal reward probabilities was caused by impairments in the accurate estimation of implicit reward

probabilities rather than by reduced reliance on the learnt reward probability. Biperiden increased the learning rate for rewarded choices in the volatile phase, but in a maladaptive manner, leading to higher deviations of estimated from optimally tracked probabilities in the biperiden relative to placebo sessions. Together, this reveals that blocking cholinergic M1 receptor activity results in a maladaptive overshoot of the learning rate in volatile environments.

Role of acetylcholine in volatile environments

Acetylcholine is widely distributed in the brain and known to influence behaviour, such as learning, memory and attention (Ananth et al., 2023). Historically, it has been assumed that acetylcholine release amplifies bottom-up thalamocortical processing at the expense of intracortical processing (Hasselmo, 2006). However, more recent studies suggest a more complex role of acetylcholine. In contrast to the traditional view, cholinergic activity seems to be particularly crucial for top-down processes (Ballinger et al., 2016). Physiological studies have demonstrated that acetylcholine release in the medial prefrontal cortex enhanced pyramidal cell activity in layer 2/3 of the visual cortex, boosting top-down influences on sensory cortices (Eggermann and Feldmeyer, 2009). In addition, in line with a role in top-down control and response adaptation, cholinergic antagonism abolished post-error adjustments in both behaviour and sensory cortical areas (Danielmeier et al., 2015). We observed a maladaptive increase in learning rate in the volatile phase of the learning task caused by blocking cholinergic M1 receptors. This demonstrates that, under biperiden, participants put more weight on recent relative to more distant experience, pointing towards reduced integration over time, or even enhanced distractibility by current sensory (reward) information. This might support the notion that acetylcholine plays a critical role in coordinating the ratio of bottom-up to top-down processes, thereby balancing the signal-to-noise ratio in uncertain environments. Yet, it remains elusive whether top-down processes are down-regulated or bottom-up processes are up-regulated following cholinergic M1 antagonism. In any case, the effect reflects noisier, suboptimal estimation of the uncertain reward probability under cholinergic M1 antagonism. These findings do not stand in isolation, as for example Bucci et al. (1998) suggested that acetylcholine is involved in increased attentional processing, as presumably required during high volatility. This is in line with studies reporting that, in a set-shifting task, cholinergic manipulation impairs serial reversal learning of mice while leaving reversal-free

learning intact (Robbins and Roberts, 2007; Cools and Arnsten, 2022). Similarly, Marshall et al. (2016) also observed impaired adaptations to environmental changes in a probabilistic serial response time task under biperiden. They argued this occurred due to increased distractibility. In line with this, we found that the increase in learning rate was specific for the volatile phase which is also characterized by a higher degree of overall outcome surprise. This supports the hypothesis that participants tended to interpret probabilistic outcomes more readily as indicators for context changes, leading to noisier estimations of volatility (Yu and Dayan, 2005; Marshall et al., 2016). It is also noteworthy that basal forebrain cholinergic neurons have been shown to display asymmetrical responses to outcome surprise, where responses were more pronounced to appetitive as opposed to aversive outcomes (Hangya et al., 2015). This might provide some explanation as to why we observed effects of cholinergic blockade selectively on the learning rate for rewarded, but not unrewarded outcomes.

Furthermore, acetylcholine acting at M1 receptors enhances NMDA and GABA receptor function (Bessie Aramakis et al., 1997; Obermayer et al., 2017; Zwart et al., 2018). The balance between recurrent NMDA-mediated excitation and GABAergic feedback inhibition is a fundamental determinant in cortical circuit models of decision making (Wang, 2002; Wong and Wang, 2006). In line with this, administration of an NMDA-receptor agonist has been reported to lead to more optimal integration of reward information in healthy adults while it did not affect learning (Scholl et al., 2014). Moreover, human participants with higher concentrations of GABA relative to glutamate in the ventromedial prefrontal cortex were found to have a higher decision accuracy in a reward-guided choice task (Jocham et al., 2012; Kaiser et al., 2021). Therefore, we hypothesized that cholinergic M1 antagonism would lead to less optimal decisions in all tasks, irrespective of learning and volatility manipulations. The absence of any effect on choice behaviour in the gambling task and the stable phase of the learning task under biperiden is therefore unexpected. We can only speculate about the reason that biperiden had effects only in the volatile environment of the learning task and did not lead to suboptimal choices in the stable environment of the learning task nor in the gambling task. One possible reason is that our tasks were more difficult than the paradigms used in earlier studies. In support of this, it is noteworthy that, overall, our participants used less multiplicative attribute integration compared to other studies (Scholl et al., 2014; Farashahi et al., 2019), and reached remarkably high learning rates in the learning task, indicative of suboptimal behaviour. Rather than

observing any general attentional deficits, such as more stochastic choices across all tasks, as would be reflected in the softmax inverse temperature, our effects were very specific to the only task phase that included volatility.

At first glance, this volatility-specific biperiden effect appears surprising given that acetylcholine is proposed to play a critical role in environments with known unreliability, termed expected uncertainty, whereas uncertainty arising from unpredictable switches of context, such as reversal of cue-outcome-contingencies, termed unexpected uncertainty, has been ascribed to the neuromodulator noradrenaline (Yu and Dayan, 2005; Avery et al., 2012). By this definition, the gambling task and the stable phase of the learning task should be associated with expected uncertainty, whereas the volatile phase of the learning task involves elements of both expected and unexpected uncertainty. However, one notable prediction of this framework is that, under reduced cholinergic transmission, the degree of randomness in the environment is underestimated, which in turn should amplify the effect of unexpected outcomes - as these are then more likely to be taken as an indication that stimulus-outcome-contingencies have switched. Indeed, the authors refer to the acetylcholine-depleted state of their model as "hyper-distractible" (Yu and Dayan, 2005). In line with both this theoretical framework and our experimental results, Marshall et al. (Marshall et al., 2016) also observed more rapid updating of higher-order volatility estimates under biperiden. In this context, it is also worth noting that the mean squared prediction error was higher in the volatile compared to the stable environment. This might explain why biperiden specifically affected the volatile phase of the learning task, where the increased level of surprise in the environment under reduced cholinergic transmission is more readily interpreted as a change in outcome contingencies.

The observed effects of cholinergic antagonism are complemented by pharmacological studies increasing catecholaminergic transmission that found similar effects. In particular, the catecholamine reuptake inhibitor methylphenidate has been shown to increase learning rates in a volatile environment of a learning task similar to ours (Cook et al., 2019), suggestive of opposing actions of muscarinic acetylcholine and catecholaminergic transmission. For the dopaminergic system, there is evidence for a reciprocal antagonistic interaction between cholinergic M1 and dopaminergic D2 receptors at the cellular level in the striatum (Di Chiara et al., 1994), which is paralleled by opposing effects at the functional level (Brocks, 1999; Stanhope et al., 2001). In line

with this, we recently observed that biperiden had effects on effort-based decision making that were opposite to those of the D2 receptor antagonist haloperidol (Erfanian Abdoust et al., 2024).

Behavioural adaptions across tasks

While numerous studies have reported an increased learning rate during volatile relative to stable phases in humans (Behrens et al., 2007; Browning et al., 2015; Blain and Rutledge, 2020), we did not observe this pattern. Cook et al. (2019), who, as in our present results, found no evidence for learning rate adjustments either, hypothesized that this could be because their task consisted of two sources of information for learning. However, our task consisted of only one source of information for learning, and one explicit task parameter, aligning more closely with studies that have identified such adjustments (Behrens et al., 2007; Browning et al., 2015). Nevertheless, in studies that observed learning rate adjustments, volatility has typically been more discernible. In particular, the differences in reward probabilities between options were larger, which both makes probability estimates less uncertain, and, more importantly, contingency switches easier to detect (Behrens et al., 2007; Browning et al., 2015). In some studies, volatility levels were even explicitly signalled (Massi et al., 2018; Blain and Rutledge, 2020). Our data suggests that the effect is less robust when the difficulty of learning is high, such as when the value difference between options or the average reward rate in the environment are low.

In addition to adjusting the learning rate, there is also evidence that the valuation strategy, i.e., the way in which values are computed, may vary with different levels of uncertainty (Farashahi et al., 2019). Although most commonly used models, such as prospect theory, assume that reward probability and magnitude are combined multiplicatively to estimate the options' values (Bernoulli, 1954; Kahneman and Tversky, 1979), it has been suggested that agents also employ an additive strategy to combine information (Stewart, 2011). Findings in humans and non-human primates support this hypothesis, revealing that they rely on multiplicative integration more strongly when all choice information is provided, but switch towards mainly additive integration when reward probabilities need to be learnt (Farashahi et al., 2019). We found an adjustment of value construction between the gambling and the learning task. We observed that, in the gambling task, participants used a mixture of both multiplicative and additive value construction, the latter with a strong weighting of

reward probability. In contrast, in the learning task, there was no evidence for multiplicative integration. Indeed, we even observed a small, yet significant, negative effect of expected value difference. The reasons for this counterintuitive effect remain unknown, it should be noted though that it was only present when analysing data from the entire learning task, not when separately analysing the stable and volatile phases. Furthermore, in the learning task, participants' choices relied less on (estimated) reward probabilities relative to magnitudes compared to the gambling task. Thus, participants appear to adjust their decision strategy to be more flexible under conditions of higher uncertainty. Further to these between-task adaptations, we were surprised not to observe significant differences in valuation strategy between volatility levels within the learning task. Again, this could result from the difficulty of our learning task; Farashahi et al. (2019) used a higher reward probability and explicitly signalled volatility levels in their task.

Conclusion

Although blocking M1 muscarinic acetylcholine receptors had no effect on decision making under risk, it increased learning rates under uncertainty, which lead to suboptimal value estimates. This effect did not occur for learning in general, instead it was specific for highly volatile environments, characterized by frequent changes of stimulus-outcome contingencies. Altogether, this suggests that modulation of the signal-to-noise ratio in cortical circuits by M1 cholinergic receptors is particularly crucial in highly uncertain environments.

Materials and methods

Ethics statement

All procedures were approved by the Ethics Committee of the Medical Faculty of the Heinrich Heine University Düsseldorf (reference 2018-211_1). The study was performed in compliance with the Code of Ethics of the World Medical Association (Declaration of Helsinki, 1975).

Participants

Participants were recruited from the local student community of the Heinrich Heine University, Germany. Participants signed a written informed consent prior to participation and received monetary compensation for their participation. Each experiment was run with healthy males who reported normal or corrected-to-normal vision. Due to the pharmacological challenge, recruited participants were extensively screened for medical exclusion criteria (see supplementary list S1). Additionally, participants were only included if they succeeded at both tasks in the screening session. For that reason, we set a performance criterion of expected value choices > 60 % for the gambling task and > 55 % for the learning task. In total, 43 participants aged between 18 and 35 years (mean age = 23.7 ± 3.1 years) took part in the study.

General design

We pharmacologically manipulated the levels of acetylcholine in a double-blind, randomized, placebo-controlled, within-subjects design. Each participant completed four experimental sessions: one screening session, two experimental sessions, which took place in the MEG, consisting of drug or placebo intake, and one MRI session. The screening session took place prior to the pharmacological experimental sessions. After providing informed consent, we tested whether medical inclusion criteria were fulfilled and measured heart rate, blood pressure and the Beck's Depression Inventory score (Beck et al., 1996). In addition, participants conducted the State-Trait Anxiety Inventory (STAI; Spielberger, 1983) and a modified version of the Edinburgh Handedness Inventory (Oldfield, 1971). Finally, participants performed the two behavioural tasks to assure that they exceeded pre-defined performance thresholds (choice of high expected value option > 60 % and > 55 % for the gambling and learning task, respectively). The two experimental sessions took place at University Hospital Düsseldorf. After over-night fasting and a standardized breakfast, participants received

a single oral dose of biperiden on one day and a placebo on the other day 45 minutes prior to the MEG recording (Fig. 1A). Blood pressure, heart rate and mood (Bond and Lader, 1974; using the Bond and Lader Visual Analogue Scales) were measured after breakfast, before entering the MEG chamber, and at the end of the session. Before entering the MEG chamber, participants additionally conducted a modified trail-making test (Rodewald et al., 2012; part A). During MEG measurement, participants were seated on a chair inside a dimly lit, magnetically shielded MEG chamber. Each MEG measurement started with a 5 minutes eyes-open resting state task followed by the two behavioural paradigms with 500 trials of the gambling task (~30 min) and 400 trials of the learning task (~20 min). However, MEG results are not further addressed here. After the last experimental session, participants were asked to guess on which testing day they received biperiden and to indicate on a scale from 1 to 100 how certain they were. In a last and fourth session we recorded an anatomical scan in the MRI.

Pharmacological intervention

We administered the muscarinic M1 acetylcholine receptor antagonist biperiden (4 mg). Typically, peak plasma concentrations are reached between 1 and 1.5 hours after oral administration of biperiden and the elimination half-time is about 18 to 24 hours (Grimaldi et al., 1986; Brocks, 1999). In view of this pharmacokinetic profile, MEG measurements started 45 minutes after drug administration and lasted approximately 1.5 hours. In order to allow plasma concentration levels to return to baseline, the two experimental sessions were scheduled at least 6 days apart.

Behavioural tasks

The experiment was designed and presented using the PsychoPy software package (version 3.1.5; Peirce, 2007). Inside the MEG chamber, stimuli were presented on a projector (Panasonic PT-D7700E, screen dimensions: 43 cm x 31 cm) with a resolution of 1280 x 1024 pixels and a refresh rate of 60 Hz (viewing distance: 80 cm). For the behavioural tasks, participants responded bi-manually with their left and right index fingers on a custom-made button box with only two buttons available.

Gambling task

We implemented a gambling task in which participants had to decide between two options on each of 500 trials in order to maximize their reward. Each trial started with

the presentation of a fixation dot (radius = 0.4 dva), for a pseudo-randomly selected duration of 1000 ms to 1400 ms to keep the participants' attention to the centre of the screen. Afterwards, the two options (width = 0.6 dva, height = 1.36 dva) appeared on the left and right side (2.2 dva distance) of the screen until a response was made. The options were visualized as vertical bars, with the fill level indicating the reward magnitude and a numeric percentage below each bar indicating the reward probability (Fig. 1). The cumulative reward already earned was visualized by a progress bar at the bottom of the screen. Participants had 3000 ms to make a choice by pressing the left or right button, followed by the presentation of a frame around the chosen option for 400 ms to 600 ms. Next, the outcome of both options was presented for 500 ms by changing the colour of the bars to either green or red indicating whether the option was rewarded or not, respectively. The outcomes of both options were independent of each other. When no response was given in time, a warning was presented for 500 ms, urging participants to respond within the time frame. The task consisted of ten blocks of 50 trials. Participants were encouraged to rest between blocks as long as they needed, but needed to wait for at least 10 s before the next block could be started.

Learning task

In the learning task, participants again had to choose between two options to maximize their reward. However, in contrast to the gambling task, not all relevant information of the options was explicitly presented, requiring participants to learn an implicit choice attribute. Here, the reward probability was implicit, similar to Farashahi et al. (2019). The trial structure of the gambling and learning tasks was similar: After the presentation of the fixation dot for 1000 ms to 1400 ms (radius = 0.4 dva), both options were presented as vertical bars on the left and right side of the screen until response (distance = 2.2 dva, width = 0.6 dva, height = 1.36 dva). The reward magnitude was explicitly presented as fill level of both bars; however, the reward probability was linked to the colour of the bars, requiring participants to learn which of the two colours was associated with a high reward probability. The high-probability colour had a reward probability of 0.7, while the other had a reward probability of 0.3. Participants had 3000 ms to make a choice by pressing the left or right button which was followed by a frame around the chosen option for 400 ms to 600 ms. Whether the chosen option was rewarded or not was indicated by a smiley or frowny, which was presented for 500 ms in the centre of the screen (radius = 0.67 dva). Note that the outcomes of the two

options were dependent in the learning task. A rewarded choice led to an increase in the progress bar. When no response was given in time, a warning was presented for 500 ms. Volatility levels were manipulated during the task by changing reward contingencies over the course of the experiment, involving a stable and volatile phase, each lasting for 200 trials. While reward probabilities were fixed in the stable phase, probabilities reversed six times in the volatile phase: Successive reversals were separated by either 20 (three times) or 40 trials (four times). The order of the change points as well as the order of stable and volatile phase was counterbalanced across participants. Every 50 trials, participants had the possibility to rest as long as they needed.

For each participant, the task structure remained the same for all sessions. The colour pairs indicating the reward probabilities were different for each session.

Statistical analyses

We tested how the option attributes, and, for the learning task, the phase (stable *vs.* volatile) explained participants' choice (left *vs.* right choices). Specifically, we investigated how biperiden changed the influence of the attributes on choices. We used logistic mixed-level models using the lme4 package in R (R version 4.0.2 (2020-06-22); Bates et al., 2015). To account for within-subjects variability, we set the subjects' ID as random effect for both tasks. The fixed effects, which account for the between-subjects variability, were dependent on the task: For the gambling task, we used drug (biperiden *vs.* placebo), the demeaned difference of expected value (EV), reward magnitudes, and reward probabilities of both options, and the previous choice side as between-subjects factor. Note, that the demeaned difference of EV recapitulates the multiplicative integration of reward magnitude and probability only. For the learning task, we additionally used the phase (stable *vs.* volatile) as between-subjects factor. However, since participants did not know the objective probabilities in the learning task, we used estimated probabilities from a Bayesian optimal learner, based on Behrens et al. (2007). We applied sum-to-zero contrasts and z-scored all continuous predictors to achieve standardized estimates. The Bayesian optimal learner was set up in MATLAB (MATLAB Version R2016b, Massachusetts: The Mathworks Inc.) and run individually for each participant and session. In a model comparison approach, we tested whether the testing day improved the model fit. Additionally, we incorporated control measures (BL-VAS, TMT, heart rate and blood

pressure) as fixed effects only when we observed significant drug effects on these measures. All *p*-values were based on asymptotic Wald tests.

Computational modelling

Computational modelling was performed in R (R version 4.0.2 (2020-06-22)).

Gambling task

To determine how participants used option attributes to shape a decision, we used a hybrid model, incorporating both additive and multiplicative value integration (Scholl et al., 2014; Farashahi et al., 2019). The subjective value *SV* for each option *i* at trial *t* was computed as follows:

$$SV_{i,t} = \omega_{mult} M_{i,t} P_{i,t} + (1 - \omega_{mult}) \left((1 - \omega_p) M_{i,t} + \omega_p P_{i,t} \right) \quad [1]$$

Where $M_{i,t}$ is the reward magnitude, $P_{i,t}$ is the reward probability, ω_{mult} is the degree of multiplicative relative to additive integration and ω_P is the degree of probability relative to magnitude weighting within the additive component. ω_{mult} would be either 1 or 0 if only the multiplicative or additive integration was used, respectively. Similarly, if participants would only use the information of probability or magnitude, ω_P would be either 1 or 0, respectively. Magnitudes were scaled to values between 0.1 and 1.0 to allow for the comparison with reward probabilities. Based on the subjective values, the probability for choosing the left option $p_{l,t}$ at trial *t* was generated using a softmax choice rule:

$$p_{l,t} = \frac{1}{1 + e^{-(SV_{l,t} - SV_{r,t})\zeta}} \quad [2]$$

Where $SV_{l,t}$ and $SV_{r,t}$ are the subjective value of the left-side and right-side option, respectively, and ζ is the inverse temperature parameter capturing the stochasticity of action selection. We used hierarchical Bayesian estimation of group-level and subject-level parameters to incorporate the within-subjects design, similar to Swart et al. (2017). For individual-level parameters *x* group-level parameters *X* were used as priors $x \sim \mathcal{N}(X, \sigma)$ and a half-Cauchy with a scale of 2 served as hyperprior for σ (Gelman, 2006). Weakly informative distributions were used as hyperpriors for *X*: $X_{mult,P} \sim \mathcal{N}(0,2)$, $X_{\zeta} \sim \mathcal{N}(2,3)$. While ω_{mult} and ω_P were constrained between 0 and 1 using an inverse logit transform, ζ was positively bounded using an exponential transform. Initial

parameter estimates were determined using an independent training dataset. The model allowed for a biperiden-induced shift s_x in all fitted parameters x:

$$x = x + \delta_{BIP} s_x \quad [3]$$

With δ_{BIP} being 0 or 1 for the placebo and biperiden session, respectively. The parameter shifts were unconstrained and $\mathcal{N}(0,3)$ served as hyperprior. In total, six free parameters were fitted. We performed Markov chain Monte Carlo (MCMC) sampling in RStan (RStan version 2.21.8, Stan Development Team, 2016). We used four Markov chains for sampling with 2500 iterations, including 500 warm-up iterations, per chain. Models successfully converged for a maximal potential scale reduction factor $\hat{R} < 1.05$ and after verifying convergence and diagnostic criteria, as provided by RStan. To verify if the model captures participants' behaviour, we conducted posterior predictive checks by simulating 500 datasets based on the posterior distributions of subject-level parameters. We then compared and correlated simulated and real data (see supplementary Fig. S7-S9 and tables S12-S14).

Learning task

Computational modelling was similar for both tasks. Since our logistic mixed-effects models revealed that participants did not use the multiplicative strategy in the learning task, fitting of the hybrid model led to convergence issues. Thus, we fixed ω_{mult} at 0 for the computation of the subjective value, making the value integration additive only:

$$SV_{i,t} = (1 - \omega_p) M_{i,t} + \omega_p SP_{i,t} \quad [4]$$

With the subjective probability $SP_{i,t}$. The learning of reward probabilities was modelled using *Q*-learning. The probability estimate of the chosen colour $SP_{c,t}$ was updated on each trial t via two separate learning rates for rewarded and unrewarded choices, λ_r and λ_u, respectively:

$$SP_{c,t+1} = SP_{c,t} + \lambda_{r,u}(r_t - SP_{c,t}) \quad [5]$$

Where r_t reflects the outcome of the current trial and was either 1 or 0, depending on whether the choice was rewarded or not. The probability estimate of the unchosen colour $SP_{u,t}$ was dependent on the chosen colour:

$$SP_{u,t+1} = 1 - SP_{c,t} \quad [6]$$

Again, parameters were estimated using Bayesian hierarchical modelling with a shift on each parameter in the biperiden session. Thus, eight free parameters were fitted. Learning rates were constrained between 0 and 1 via an inverse logit transform and $\mathcal{N}(0,2)$ was used as hyperprior. All other priors, hyperpriors and transformations were defined in the same way as in the gambling task.

Acknowledgments

The work was supported by a European Research Council grant (ERC-CoG 771432) to GJ. The Bayesian optimal learner is a slightly modified version of Matlab code kindly provided by Tim Behrens and James Whittington. We thank Hanin Alejel, Ana Antonia Dias Maile, Helena El Kholy, Judith Geusen, Paul Höchter, Marlene Hüsken, Christina Kalinicenko, Joshua Saal, Kouta Sasaki, Georg Schäfer, and Helena Schmidt for their support during data acquisition. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

References

Ananth MR, Rajebhosale P, Kim R, Talmage DA, Role LW (2023) Basal forebrain cholinergic signalling: development, connectivity and roles in cognition. Nat Rev Neurosci 24:233–251.

Avery MC, Nitz DA, Chiba AA, Krichmar JL (2012) Simulation of cholinergic and noradrenergic modulation of behavior in uncertain environments. Front Comput Neurosci 6. DOI: http://journal.frontiersin.org/article/10.3389/fncom.2012.00005/abstract.

Ballinger EC, Ananth M, Talmage DA, Role LW (2016) Basal Forebrain Cholinergic Circuits and Signaling in Cognition and Cognitive Decline. Neuron 91:1199–1218.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using **lme4**. J Stat Soft 67. DOI: http://www.jstatsoft.org/v67/i01/.

Beck AT, Steer RA, Brown G (1996) Beck Depression Inventory–II. DOI: https://doi.apa.org/doi/10.1037/t00742-000.

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221.

Bernoulli D (1954) Exposition of a New Theory on the Measurement of Risk. Econometrica 22:23.

Bessie Aramakis V, Bandrowski AE, Ashe JH (1997) Muscarinic reduction of GABAergic synaptic potentials results in disinhibition of the AMPA/kainate-mediated EPSP in auditory cortex. Brain Research 758:107–117.

Blain B, Rutledge RB (2020) Momentary subjective well-being depends on learning and not reward. eLife 9:e57977.

Bond A, Lader M (1974) The use of analogue scales in rating subjective feelings. British Journal of Medical Psychology 47:211–218.

Brocks DR (1999) Anticholinergic drugs used in Parkinson's disease: An overlooked class of drugs from a pharmacokinetic perspective. J Pharm Pharm Sci 2:39–46.

Browning M, Behrens TE, Jocham G, O'Reilly JX, Bishop SJ (2015) Anxious individuals have difficulty learning the causal statistics of aversive environments. Nat Neurosci 18:590–596.

Bucci DJ, Holland PC, Gallagher M (1998) Removal of Cholinergic Input to Rat Posterior Parietal Cortex Disrupts Incremental Processing of Conditioned Stimuli. J Neurosci 18:8038–8046.

Cook JL, Swart JC, Froböse MI, Diaconescu AO, Geurts DE, Den Ouden HE, Cools R (2019) Catecholaminergic modulation of meta-learning. eLife 8:e51439.

Cools R, Arnsten AFT (2022) Neuromodulation of prefrontal cortex cognitive function

in primates: the powerful roles of monoamines and acetylcholine. Neuropsychopharmacol 47:309–328.

Danielmeier C, Allen EA, Jocham G, Onur OA, Eichele T, Ullsperger M (2015) Acetylcholine Mediates Behavioral and Neural Post-Error Control. Current Biology 25:1461–1468.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron 69:1204–1215.

Di Chiara G, Morelli M, Consolo S (1994) Modulatory functions of neurotransmitters in the striatum: ACh/dopamine/NMDA interactions. Trends in Neurosciences 17:228–233.

Dias Maile AA, Gründler TO, Froböse MI, Kurtenbach H, Kaiser LF, Jocham G (2024) Bidirectional modulation of reward-guided decision making by dopamine. DOI: http://biorxiv.org/lookup/doi/10.1101/2024.03.27.586793.

Eggermann E, Feldmeyer D (2009) Cholinergic filtering in the recurrent excitatory microcircuit of cortical layer 4. Proc Natl Acad Sci USA 106:11753–11758.

Erfanian Abdoust M, Froböse MI, Schnitzler A, Schreivogel E, Jocham G (2024) Dopamine and acetylcholine have distinct roles in delay- and effort-based decision-making in humans Kaplan. PLoS Biol 22:e3002714.

Everitt BJ, Robbins TW (1997) Central cholinergic systems and cognition. Annu Rev Psychol 48:649–684.

Farashahi S, Donahue CH, Hayden BY, Lee D, Soltani A (2019) Flexible combination of reward information across primates. Nat Hum Behav 3:1215–1224.

Gelman A (2006) Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal 1.

Grimaldi R, Perucca E, Ruberto G, Gelmi C, Trimarchi F, Hollmann M, Crema A (1986) Pharmacokinetic and pharmacodynamic studies following the intravenous and oral administration of the antiparkinsonian drug biperiden to normal subjects. Eur J Clin Pharmacol 29:735–737.

Hangya B, Ranade SP, Lorenc M, Kepecs A (2015) Central Cholinergic Neurons Are Rapidly Recruited by Reinforcement Feedback. Cell 162:1155–1168.

Hasselmo ME (2006) The role of acetylcholine in learning and memory. Current Opinion in Neurobiology 16:710–715.

Hasselmo ME, Sarter M (2011) Modes and Models of Forebrain Cholinergic Neuromodulation of Cognition. Neuropsychopharmacol 36:52–73.

Iglesias S, Kasper L, Harrison SJ, Manka R, Mathys C, Stephan KE (2021) Cholinergic and dopaminergic effects on prediction error and uncertainty responses during

sensory associative learning. NeuroImage 226:117590.

Jocham G, Furlong PM, Kröger IL, Kahn MC, Hunt LT, Behrens TEJ (2014) Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. NeuroImage 100:498–506.

Jocham G, Hunt LT, Near J, Behrens TEJ (2012) A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. Nat Neurosci 15:960–961.

Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M (2009) Adaptive Coding of Action Values in the Human Rostral Cingulate Zone. Journal of Neuroscience 29:7489–7496.

Kahneman D, Tversky A (1979) Prospect Theory: An Analysis of Decision under Risk. Econometrica 47:263.

Kaiser LF, Gruendler TOJ, Speck O, Luettgau L, Jocham G (2021) Dissociable roles of cortical excitation-inhibition balance during patch-leaving versus value-guided decisions. Nat Commun 12:904.

Kuchibhotla KV, Gill JV, Lindsay GW, Papadoyannis ES, Field RE, Sten TAH, Miller KD, Froemke RC (2017) Parallel processing by cortical inhibition enables context-dependent behavior. Nat Neurosci 20:62–71.

Kurtenbach H, Ort E, Froböse MI, Jocham G (2022) Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations Palminteri. PLoS Comput Biol 18:e1010201.

Lee D, Seo H, Jung MW (2012) Neural Basis of Reinforcement Learning and Decision Making. Annu Rev Neurosci 35:287–308.

Marino MJ, Rouse ST, Levey AI, Potter LT, Conn PJ (1998) Activation of the genetically defined m1 muscarinic receptor potentiates *N* -methyl- D -aspartate (NMDA) receptor currents in hippocampal pyramidal cells. Proc Natl Acad Sci USA 95:11465–11470.

Marshall L, Mathys C, Ruge D, De Berker AO, Dayan P, Stephan KE, Bestmann S (2016) Pharmacological Fingerprints of Contextual Uncertainty. PLoS Biol 14:e1002575.

Massi B, Donahue CH, Lee D (2018) Volatility Facilitates Value Updating in the Prefrontal Cortex. Neuron 99:598-608.e4.

Obermayer J, Verhoog MB, Luchicchi A, Mansvelder HD (2017) Cholinergic Modulation of Cortical Microcircuits Is Layer-Specific: Evidence from Rodent, Monkey and Human Brain. Front Neural Circuits 11:100.

Oldfield RC (1971) The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9:97–113.

Peirce JW (2007) PsychoPy—Psychophysics software in Python. Journal of Neuroscience Methods 162:8–13.

Robbins T, Roberts A (2007) Differential Regulation of Fronto-Executive Function by the Monoamines and Acetylcholine. Cerebral Cortex 17:i151–i160.

Rodewald K, Bartolovic M, Debelak R, Aschenbrenner S, Weisbrod M, Roesch-Ely D (2012) Eine Normierungsstudie eines modifizierten Trail Making Tests im deutschsprachigen Raum. Zeitschrift für Neuropsychologie 23:37–48.

Scholl J, Günthner J, Kolling N, Favaron E, Rushworth MF, Harmer CJ, Reinecke A (2014) A Role Beyond Learning for NMDA Receptors in Reward-Based Decision-Making—a Pharmacological Study Using d-Cycloserine. Neuropsychopharmacol 39:2900–2909.

Soltani A, Izquierdo A (2019) Adaptive learning under expected and unexpected uncertainty. Nat Rev Neurosci 20:635–644.

Spielberger CD (1983) State-Trait Anxiety Inventory for Adults. DOI: https://doi.apa.org/doi/10.1037/t06496-000.

Stanhope KJ, Mirza NR, Bickerdike MJ, Bright JL, Harrington NR, Hesselink MB, Kennett GA, Lightowler S, Sheardown MJ, Syed R, Upton RL, Wadsworth G, Weiss SM, Wyatt A (2001) The muscarinic receptor agonist xanomeline has an antipsychotic-like profile in the rat. J Pharmacol Exp Ther 299:782–792.

Stewart N (2011) Information Integration in Risky Choice: Identification and Stability. Front Psychology 2:301.

Sutton RS, Barto A (2014) Reinforcement learning: an introduction. Cambridge, Massachusetts: The MIT Press.

Swart JC, Froböse MI, Cook JL, Geurts DE, Frank MJ, Cools R, den Ouden HE (2017) Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. eLife 6:e22169.

Wang X-J (2002) Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. Neuron 36:955–968.

Wong K-F, Wang X-J (2006) A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. J Neurosci 26:1314–1328.

Yu AJ, Dayan P (2005) Uncertainty, Neuromodulation, and Attention. Neuron 46:681–692.

Zwart R, Reed H, Sher E (2018) Oxotremorine-M potentiates NMDA receptors by muscarinic receptor dependent and independent mechanisms. Biochemical and Biophysical Research Communications 495:481–486.

Supplementary information: A role for acetylcholine in reinforcement learning and decision making under uncertainty

Contents

1. Raw behaviour

Plotting the raw choice data as a function of the separate choice attributes illustrates that choice behaviour in both tasks was driven by the highest expected value (EV), which is the product of reward magnitude and probability, (figure S1, left side), by reward probabilities (figure S1, middle) and by reward magnitudes (figure S1, right side). Interestingly, in the gambling task, choices were only guided by magnitude information if the magnitude difference was large, for small magnitude differences participants opted for the other option (figure S1).

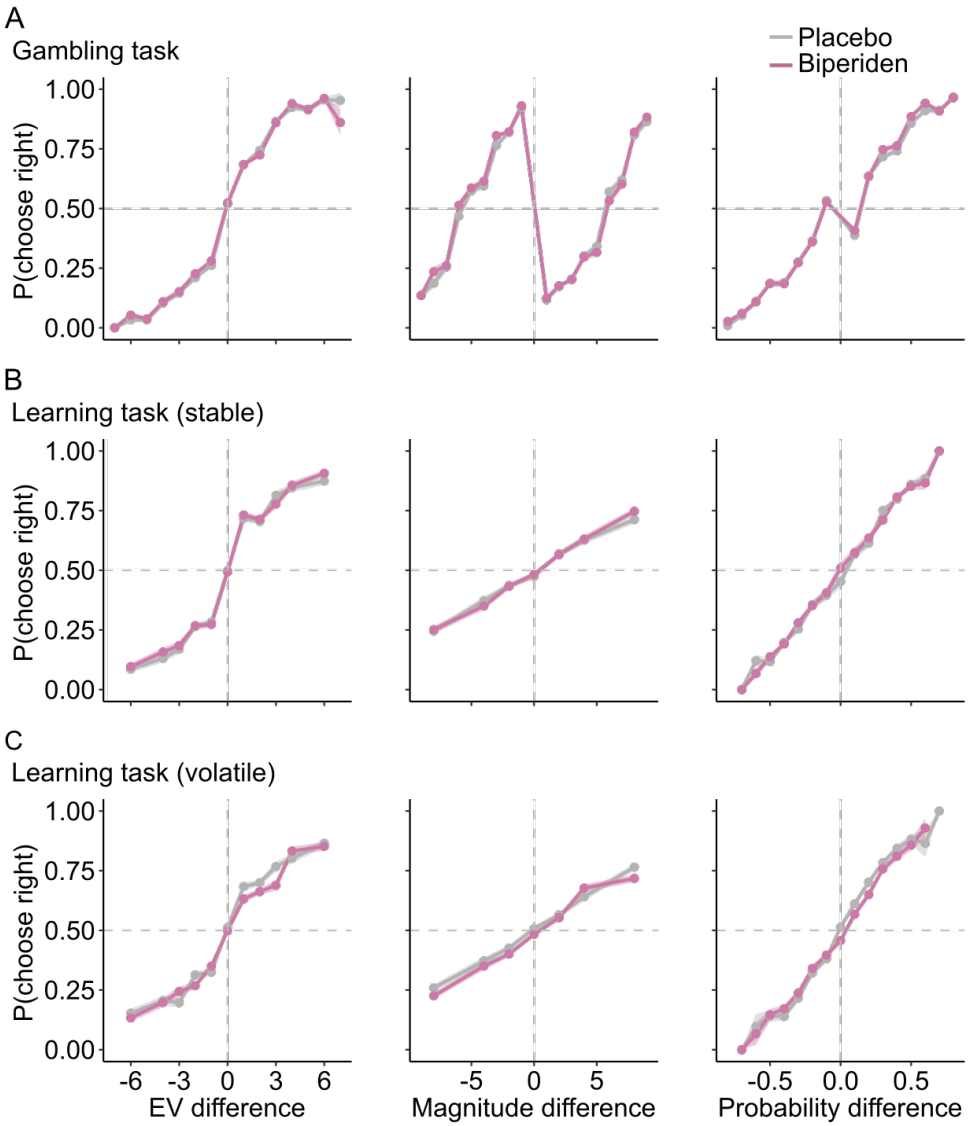

Figure S1. Participants' choice behaviour. Influence of task parameters on choice in the gambling task (**A**) and in both phases of the learning task (stable: **B**, volatile: **C**). Left: Probability of right-side choice as a function of difference in expected value (EV) between the two options. Note, that the shown EV is not mean-centred and, thus, is highly correlated with reward magnitude and probability. Middle: Probability of right-side choice as a function of difference in reward magnitude between the two options. Right: Probability of right-side choice as a function of difference in reward probability between the two options. For the learning task, reward probabilities were estimated using the Bayesian optimal learner. Choice behaviour in the biperiden session (pink) and the placebo session (grey) is shown. Solid lines represent mean, shaded areas SEM across participants.

2. Logistic mixed-effects models

Full results of the logistic mixed-effects models for the gambling task (table S1) and the learning task (table S2).

Table S1: Results from logistic mixed-effects models of the gambling task.

	β	**SEM**	***z*(42303)**	***p***
Intercept	0.03	0.03	0.86	.392
Drug	0.03	0.02	1.37	.172
Probability	*2.90*	*0.03*	*95.78*	*< .001*
Magnitude	*1.58*	*0.02*	*73.66*	*< .001*
EV	*0.50*	*0.02*	*30.05*	*< .001*
Alternation bias	*-0.07*	*0.03*	*-2.64*	*.008*
Drug x probability	0.00	0.03	0.12	.902
Drug x magnitude	-0.03	0.02	-1.55	.121
Drug x EV	-0.02	0.02	-1.02	.310
Drug x bias	-0.02	0.03	-0.82	.414

Table S2: Results from logistic mixed-effects models of the learning task.

	β	**SEM**	***z*(42303)**	***p***
Intercept	-0.04	0.02	-1.78	.076
Drug	-0.01	0.01	-0.78	.434
Probability	*1.27*	*0.02*	*80.28*	*< .001*
Magnitude	*0.75*	*0.01*	*51.45*	*< .001*
EV	*-0.04*	*0.02*	*-2.31*	*.021*
Repetition bias	*0.10*	*0.01*	*7.86*	*< .001*
Phase	-0.01	0.01	-0.63	.529
Drug x probability	*-0.05*	*0.02*	*-3.09*	*.002*
Drug x magnitude	0.01	0.01	0.60	.546
Drug x EV	0.02	0.02	1.39	.164
Drug x bias	0.01	0.01	0.49	.625
Phase x probability	*-0.08*	*0.02*	*-4.97*	*< .001*
Phase x magnitude	0.01	0.01	0.92	.359
Phase x EV	0.00	0.02	0.31	.757

Phase x bias	-0.01	0.01	-0.67	.502
Drug x phase x probability	*0.04*	*0.02*	*2.29*	*.022*
Drug x phase x magnitude	0.01	0.01	0.38	.706
Drug x phase x EV	0.01	0.02	0.52	.604
Drug x phase x bias	0.02	0.01	1.60	.110

3. Control measurements

During each session, we acquired the participants' mood, using the Bond and Lader Visual Analogue Scales (BL VAS), and blood pressure as control measurements at three time points: before drug intake (T1), before the MEG measurement (T2) and after the MEG measurement (T3). Before the MEG measurement, we additionally measured participants' executive functions using a trail making test. To control for biperiden effects on these measures, we applied linear mixed-effects models.

3.1. Bond and Lader Visual Analogue Scales

Under biperiden, alertness, calmness, and contentedness were significantly decreased at T3 (see table S3-S5, figure S2).

Table S3: Linear mixed-effects results for the alertness score.

	β	**SEM**	***df***	***t***	***p***
Intercept	*7.90*	*0.23*	*63.31*	*34.22*	*< .001*
Drug	0.13	0.16	209.01	0.85	.396
T2	-0.02	0.16	209.01	-0.10	.917
T3	-0.08	0.16	209.01	-0.54	.592
Drug x T2	-0.06	0.22	209.01	-0.29	.776
Drug x T3	*-1.23*	*0.22*	*209.03*	*-5.55*	*< .001*

Table S4: Linear mixed-effects results for the calmness score.

	β	**SEM**	***df***	***t***	***p***
Intercept	*7.75*	*0.28*	*97.87*	*27.77*	*< .001*
Drug	-0.06	0.26	208.82	-0.24	.815
T2	-0.16	0.26	208.82	-0.63	.531
T3	-0.18	0.26	208.82	-0.71	.479
Drug x T2	0.07	0.37	208.75	0.19	.848
Drug x T3	*-0.84*	*0.37*	*208.75*	*-2.29*	*.023*

Table S5: Linear mixed-effects results for the contentedness score.

	β	**SEM**	***df***	***t***	***p***
Intercept	*8.53*	*0.20*	*67.45*	*43.31*	*< .001*
Drug	-0.03	0.14	209.07	-0.18	.858
T2	-0.16	0.14	209.07	-1.17	.244
T3	-0.14	0.14	209.07	-1.00	.316
Drug x T2	-0.02	0.20	209.04	-0.11	.914
Drug x T3	*-0.55*	*0.20*	*209.04*	*-2.79*	*.006*

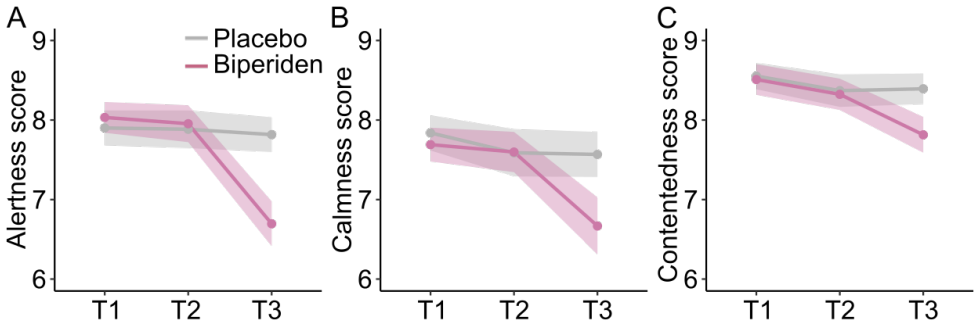

Figure S2: Results of the BL VAS. Scores for **A** alertness, **B** calmness, and **C** contentedness were acquired before drug intake (T1), before MEG measurement (T2), and after MEG measurement (T3) in both the placebo session (grey) and biperiden session (pink). All three measures were reduced at T3 under biperiden. Solid lines represent mean, shaded areas SEM across participants.

3.2. Blood pressure

There were no significant effects of biperiden on systolic and diastolic blood pressure. Instead, the heart rate was decreased at T3 under biperiden (see table S6-S8, figure S3).

Table S6: Linear mixed-effects results for the diastolic blood pressure.

	β	**SEM**	***df***	***t***	***p***
Intercept	*124.48*	*1.69*	*92.78*	*73.76*	*< .001*
Drug	-2.69	1.53	208.04	-1.76	.080
T2	*-3.64*	*1.52*	*208.21*	*-2.39*	*.018*
T3	-0.62	1.52	208.21	-0.41	.684
Drug x T2	3.41	2.15	208.04	1.59	.114
Drug x T3	-3.57	2.15	208.04	-1.66	.099

Table S7: Linear mixed-effects results for the systolic blood pressure.

	β	**SEM**	***df***	***t***	***p***
Intercept	*78.29*	*1.31*	*87.14*	*59.82*	*< .001*
Drug	-1.62	1.14	208.07	-1.42	.158
T2	*-5.31*	*1.14*	*208.22*	*-4.67*	*< .001*
T3	*-2.47*	*1.14*	*208.22*	*-2.17*	*.031*
Drug x T2	1.27	1.61	208.07	0.79	.430
Drug x T3	-2.45	1.61	208.07	-1.53	.129

Table S8: Linear mixed-effects results for the heart rate.

	β	**SEM**	***df***	***t***	***p***
Intercept	*72.49*	*1.60*	*69.72*	*45.25*	*< .001*
Drug	1.83	1.19	208.05	1.55	.123
T2	-2.13	1.18	208.15	-1.81	.072
T3	*-11.83*	*1.18*	*208.15*	*-10.03*	*< .001*
Drug x T2	-3.07	1.67	208.05	-1.84	.067
Drug x T3	*-8.97*	*1.67*	*208.05*	*-5.39*	*< .001*

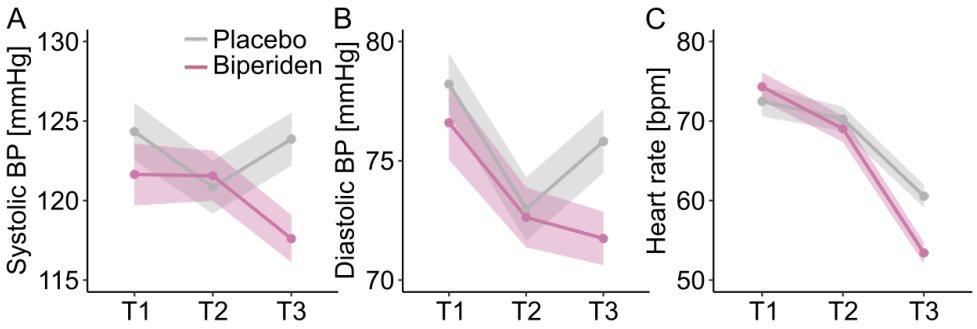

Figure S3: Results of the measurement of the blood pressure. Before drug intake (T1), before MEG measurement (T2), and after MEG measurement (T3) we measured **A** systolic blood pressure (BP), **B** diastolic BP, and **C** heart rate in both placebo session (grey) and biperiden session (pink). Biperiden significantly reduced heart rate at T3. Solid lines represent mean, shaded areas SEM across participants.

3.3. Trail making test

Participants' timing during the trail making test was not affected by biperiden (β = -0.03, *SEM* = 0.84, t_{214} = 0.037, p = .971; figure S4).

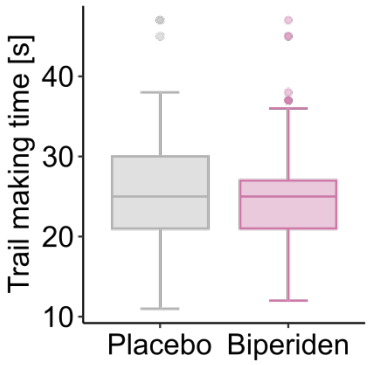

Figure S4: Results of the trail making test before MEG measurement (T2). Trail making time was acquired in the placebo session (grey) and in the biperiden session (pink).

4. Logistic mixed-effects models: independence of biperiden effect

After biperiden administration, we observed a significantly decreased sensitivity to estimated reward probabilities in the learning task, specifically in the volatile phase. To control if this biperiden effect is independent of other measures, we conducted logistic mixed-effects models for the volatile phase with several control measures (heart rate, alertness, calmness, and contentedness quantified by the BL VAS, order of stable and volatile phase, session days and order of medication).

Table S9: Logistic mixed-effects results of the interaction effect of medication and estimated probability difference in the volatile phase considering control measures.

Control measure	*β*	SEM	*N obs*	*z*	*p*
Heart rate	*-0.12*	*0.03*	*16702*	*-4.42*	*< .001*
Alertness	*-0.06*	*0.03*	*16902*	*-2.27*	*.023*
Calmness	*-0.08*	*0.03*	*16902*	*-2.95*	*.003*
Contentedness	*-0.07*	*0.03*	*16902*	*-2.76*	*.006*
Phase order	*-0.08*	*0.03*	*17102*	*-3.33*	*< .001*
Session days	*-0.08*	*0.03*	*17102*	*-3.35*	*< .001*
Order of medication	*-0.08*	*0.03*	*17102*	*-3.32*	*< .001*

5. Bayesian hierarchical models: biperiden-specific results

5.1. Gambling task

For the gambling task, we found no credible effect of biperiden-specific shifts in the Bayesian hierarchical models (figure S5, table S10).

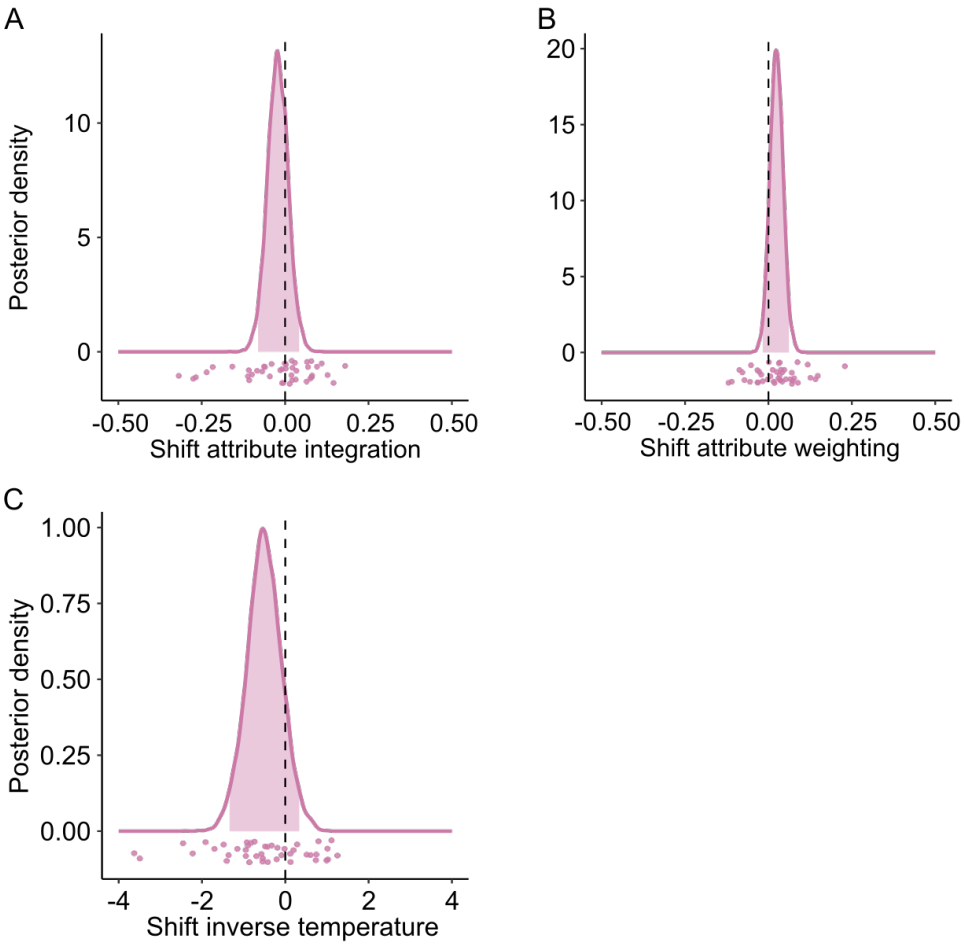

Figure S5. Biperiden-induced shifts in the gambling task from the Bayesian hierarchical model. Density of posterior predictive distributions of the biperiden-specific shift in **A** attribute integration $s_{\omega mult}$, **B** attribute weighting $s_{\omega P}$, and **C** inverse temperature s_{ζ}. Positive shifts represent an increase under biperiden relative to placebo. Shaded areas represent the 95 %-HDI of the posterior predictive distribution and points single-subject means.

Table S10. Group-level parameter estimates of the gambling task. Median (Mdn), standard deviation (SD), and lower and upper bounds of the 95 %-HDI interval are given. The model consisted of estimates for the value construction ω_{mult}, the attribute weighting ω_P, the inverse temperature ζ, and the corresponding biperiden-specific shifts on these parameters $s_{\omega mult}$, $s_{\omega P}$, and s_ζ.

Parameter	**Mdn**	**SD**	**2.5 %**	**97.5 %**
ω_{mult}	*0.38*	*0.07*	*0.25*	*0.52*
ω_P	*0.81*	*0.03*	*0.75*	*0.87*
ζ	*14.70*	*0.92*	*12.92*	*16.53*
$s_{\omega mult}$	-0.02	0.03	-0.08	0.04
$s_{\omega P}$	0.02	0.02	-0.02	0.06
s_ζ	-0.51	0.42	-1.33	0.33

5.2. Learning task: stable phase

For the stable phase of the learning task, we found no credible effect of biperiden-specific shifts in the Bayesian hierarchical models (figure S6, table S11).

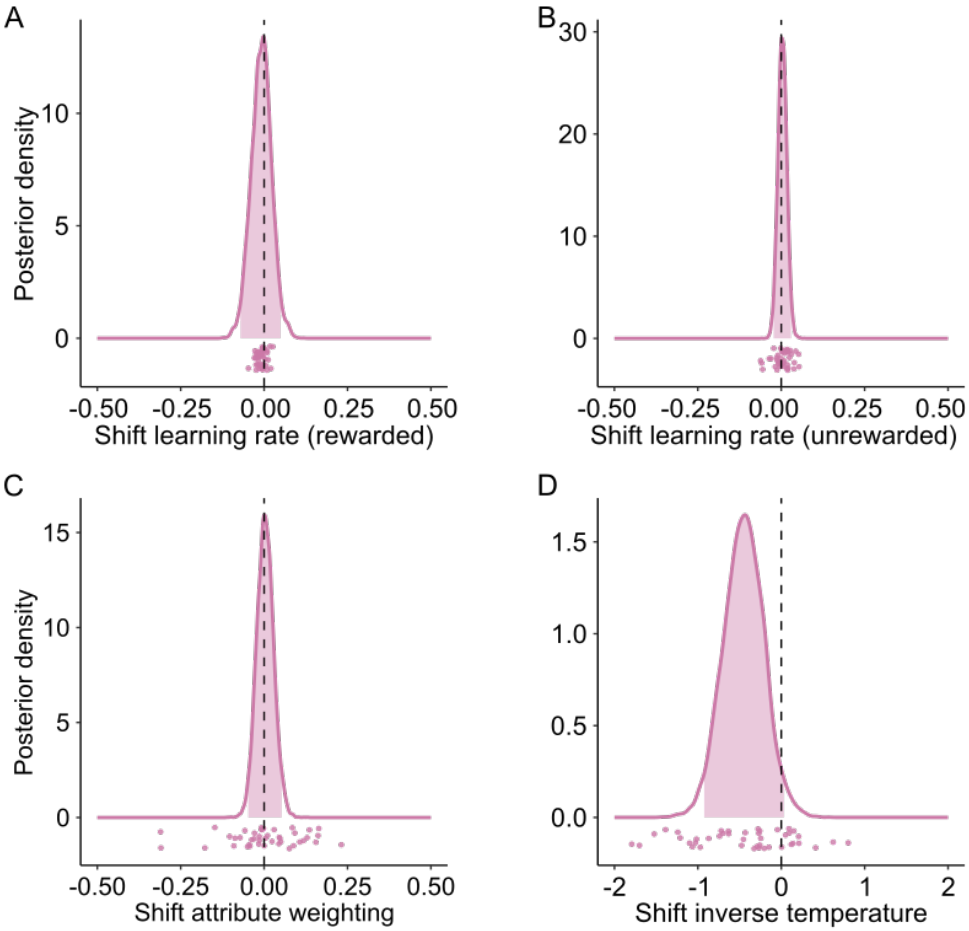

Figure S6. Biperiden-induced shifts in the stable phase of the learning task from the Bayesian hierarchical model. Density of posterior predictive distributions of the biperiden-specific shift in **A**, **B** learning rate in rewarded $s_{\lambda r}$ and unrewarded trials $s_{\lambda u}$, **C** attribute weighting $s_{\omega P}$, and **D** inverse temperature s_{ζ}. Positive shifts represent an increase under biperiden relative to placebo. Shaded areas represent the 95 %-HDI of the posterior predictive distribution and points single-subject means.

Table S11. Group-level parameter estimates of the stable phase of the learning task. Median (Mdn), standard deviation (SD), and lower and upper bounds of the 95%-HDI interval are given. The model consisted of estimates for the learning rate of rewarded choices λ_r, the learning rate of unrewarded choices λ_u, the attribute weighting ω_P, the inverse temperature ζ, and the corresponding biperiden-specific shifts on these parameters $s_{\lambda r}$, $s_{\lambda u}$, $s_{\omega P}$, and s_{ζ}.

Parameter	**Mdn**	**SD**	**2.5 %**	**97.5 %**
λ_r	*0.60*	*0.04*	*0.52*	*0.68*
λ_u	*0.25*	*0.02*	*0.21*	*0.30*
ω_P	*0.59*	*0.04*	*0.50*	*0.66*
ζ	*5.58*	*0.43*	*4.80*	*6.47*
$s_{\lambda r}$	-0.01	0.03	-0.07	0.04
$s_{\lambda u}$	0.00	0.01	-0.02	0.03
$s_{\omega P}$	0.00	0.03	-0.05	0.05
s_{ζ}	-0.45	0.24	-0.92	0.03

6. Exclusion criteria

In our study, healthy male participants conducted behavioural tasks after an oral dose of a placebo and the cholinergic antagonist biperiden. Concurrently, MEG was recorded and after the MEG sessions, an anatomical MRI was acquired. To ensure both MEG/MRI compatibility and a good health status of the participants, we had a preceding screening session. Participants with indications found in List S1 were excluded from the study.

List S1. Exclusion criteria.

- Weight < 60 kg or > 90 kg
- BMI < 18 or > 28
- Systolic blood pressure > 140 mmHg
- Diastolic blood pressure > 90 mmHg
- BDI score > 12
- Impaired vision (and no contact lenses)
- Lactose intolerance
- Regular/recent use of drugs (incl. alcohol, cigarettes)
- Psychiatric diseases
- Neurological diseases
- Hepatic dysfunction
- Renal dysfunction
- Diseases of the cardiovascular system
- Epilepsy
- Glaucoma
- Thyrotoxicosis
- Gastrointestinal diseases
- Diabetes
- Metal implants
- Claustrophobia

7. Bayesian hierarchical models: posterior predictive checks

To assess whether the Bayesian hierarchical models could capture participants' behaviour, we conducted posterior predictive checks for the choice task and both phases of the learning task. Therefore, we simulated 500 datasets based on subject-level estimates of the parameters per participant. Then, we correlated the probability of high EV choice, high magnitude choice, and high probability choice for simulated and raw data.

7.1. Gambling task

For the gambling task, the model could capture participants' behaviour (figure S5A-C). Additionally, simulated behaviour and participants' behaviour were highly significant (figure S7D-F, table S12).

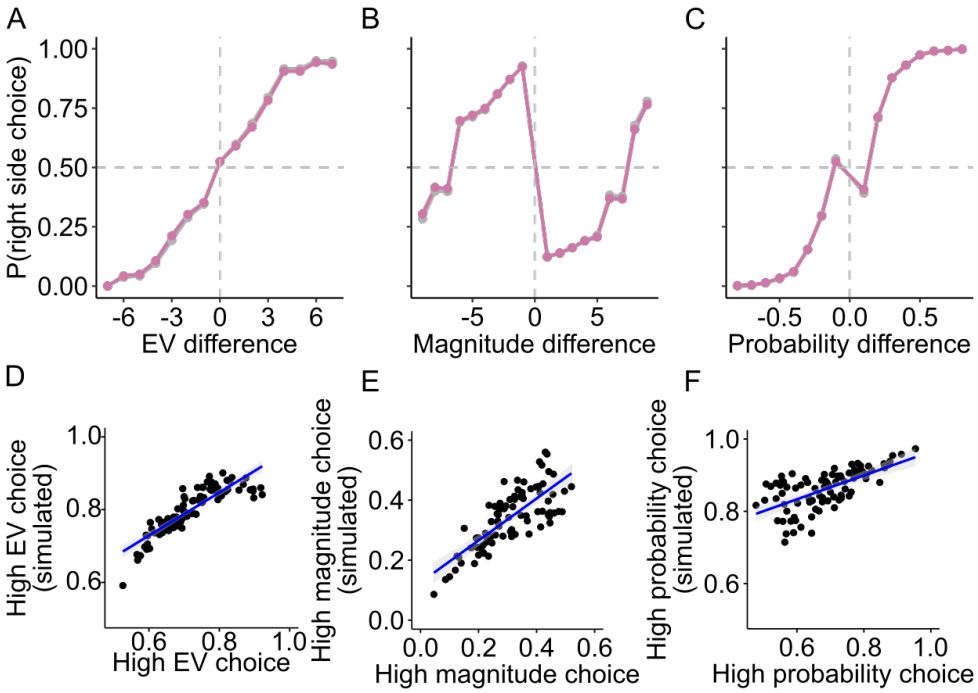

Figure S7: Posterior predictive checks for the gambling task. Probability of simulated right-side choice as a function of difference in **A** EV, **B** reward magnitude, and **C** reward probability between the two options. Choice behaviour was simulated for the biperiden session (pink) and placebo session (grey). Solid lines represent mean, shaded areas SEM across simulations. Correlation between participants' and simulated choices for **D** high EV option, **E** high magnitude option, and **F** high probability option.

Table S12: Posterior predictive checks for the gambling task. Correlation coefficients of participants' choices and simulated choices for EV, reward magnitude and estimated reward probability.

Task parameter	*R*	*p*
EV	*0.9*	*< .001*
Magnitude	*0.76*	*< .001*
Probability	*0.66*	*< .001*

7.2. Learning task: stable phase

The model for the stable phase of the learning task could capture participants' behaviour (figure S8A-C). Additionally, simulated behaviour and participants' behaviour were highly significant (figure S8D-F, table S13).

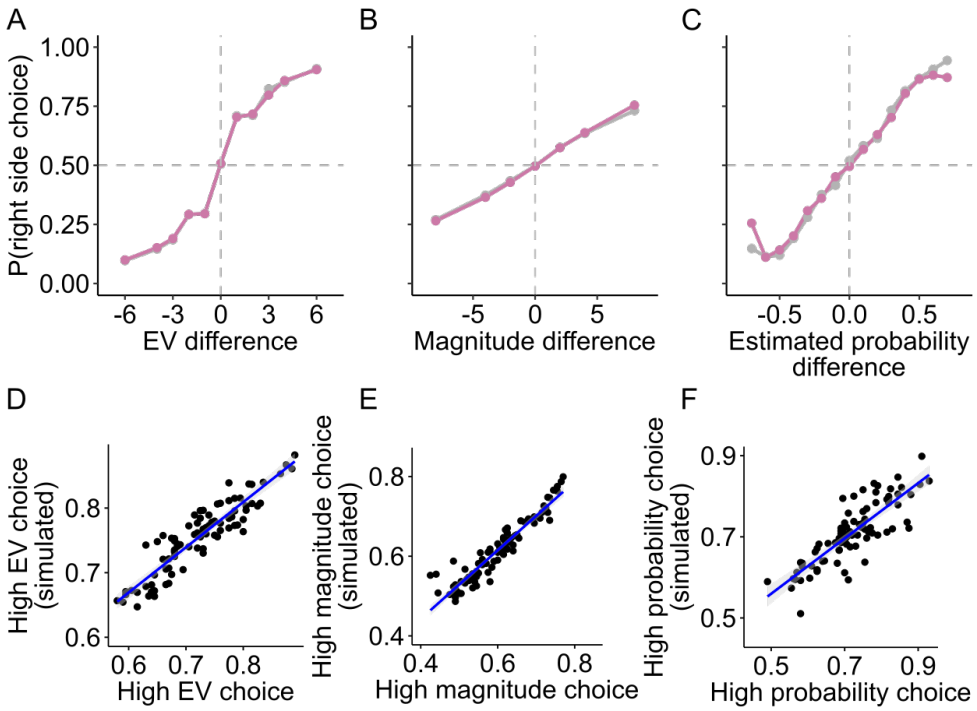

Figure S8: Posterior predictive checks for the stable phase of the learning task. Probability of simulated right-side choice as a function of difference in **A** EV, **B** reward magnitude, and **C** estimated reward probability between the two options. Choice behaviour was simulated for the biperiden session (pink) and placebo session (grey). Solid lines represent mean, shaded areas SEM across simulations. Correlation between participants' and simulated choices for **D** high EV option, **E** high magnitude option, and **F** high probability option.

Table S13: Posterior predictive checks for the stable phase of the learning task. Correlation coefficients of participants' choices and simulated choices for EV, reward magnitude and estimated reward probability.

Task parameter	*R*	*p*
EV	*0.91*	*< .001*
Magnitude	*0.94*	*< .001*
Probability	*0.83*	*< .001*

7.3. Learning task: volatile phase

The model for the stable phase of the learning task could capture participants' behaviour (figure S9A-C). Additionally, simulated behaviour and participants' behaviour were highly significant (figure S9D-F, table S14).

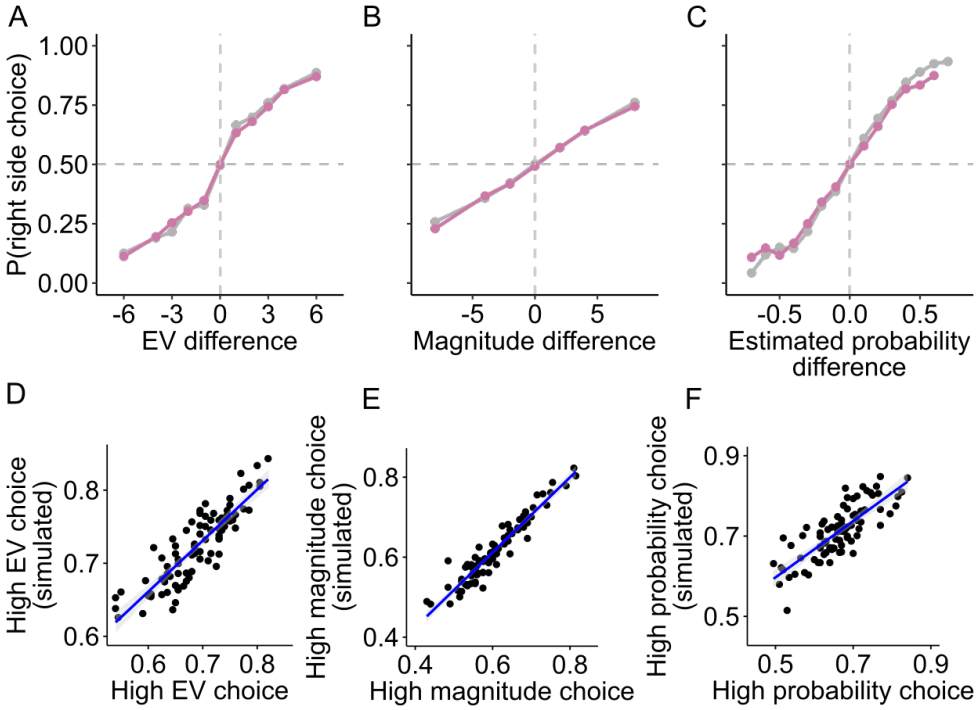

Figure S9: Posterior predictive checks for the volatile phase of the learning task. Probability of simulated right-side choice as a function of difference in **A** EV, **B** reward magnitude, and **C** estimated reward probability between the two options. Choices based on estimated reward probabilities were decreased under biperiden (pink) compared to placebo (grey). Solid lines represent mean, shaded areas SEM across simulations. Correlation between participants' and simulated choices for **D** high EV option, **E** high magnitude option, and **F** high probability option.

Table S14: Posterior predictive checks for the volatile phase of the learning task. Correlation coefficients of participants' choices and simulated choices for EV, reward magnitude and estimated reward probability.

Task parameter	***R***	***p***
EV	*0.87*	*< .001*
Magnitude	*0.93*	*< .001*
Probability	*0.80*	*< .001*

Study II: Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations

Reproduced from

Kurtenbach H, Ort E, Froböse MI, & Jocham G
Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations
PLOS Computational Biology **18(12)**, e1010201 (2022)

Digital Object Identifier (DOI): `https://doi.org/10.1371/journal.pcbi.1010201`

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations

Hannah Kurtenbach*, Eduard Ort, Monja Isabel Froböse, Gerhard Jocham

Biological Psychology of Decision Making, Institute of Experimental Psychology, Heinrich Heine University Düsseldorf, Germany

* hannah.kurtenbach@hhu.de

OPEN ACCESS

Citation: Kurtenbach H, Ort E, Froböse MI, Jocham G (2022) Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations. PLoS Comput Biol 18(12): e1010201. https://doi.org/10.1371/journal.pcbi.1010201

Editor: Stefano Palminteri, Ecole Normale Superieure, FRANCE

Received: May 13, 2022

Accepted: November 24, 2022

Published: December 8, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010201

Data Availability Statement: All data and codes are made available on OSF (https://osf.io/nxhd5/).

Abstract

Performance during instrumental learning is commonly believed to reflect the knowledge that has been acquired up to that point. However, recent work in rodents found that instrumental performance was enhanced during periods when reinforcement was withheld, relative to periods when reinforcement was provided. This suggests that reinforcement may mask acquired knowledge and lead to impaired performance. In the present study, we investigated whether such a beneficial effect of removing reinforcement translates to humans. Specifically, we tested whether performance during learning was improved during non-reinforced relative to reinforced task periods using signal detection theory and a computational modelling approach. To this end, 60 healthy volunteers performed a novel visual go/no-go learning task with deterministic reinforcement. To probe acquired knowledge in the absence of reinforcement, we interspersed blocks without feedback. In these non-reinforced task blocks, we found an increased *d'*, indicative of enhanced instrumental performance. However, computational modelling showed that this improvement in performance was not due to an increased sensitivity of decision making to learnt values, but to a more cautious mode of responding, as evidenced by a reduction of a general response bias. Together with an initial tendency to act, this is sufficient to drive differential changes in hit and false alarm rates that jointly lead to an increased *d'*. To conclude, the improved instrumental performance in the absence of reinforcement observed in studies using asymmetrically reinforced go/no-go tasks may reflect a change in response bias rather than unmasking latent knowledge.

Author summary

It appears plausible that we can only learn and improve if we are told what is right and wrong. But what if feedback overshadows our actual expertise? In many situations, people learn from immediate feedback on their choices, while the same choices are also used as a measure of their knowledge. This inevitably confounds learning and the read-out of learnt

Funding: The work was supported by a European Research Council grant (ERC-CoG 771432) to GJ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

associations. Recently, it was suggested that rodents express their true knowledge of a task during periods when they are *not* rewarded or punished during learning. During these periods, animals displayed improved performance. We found a similar improvement of performance in the absence of feedback in human volunteers. Using a combination of computational modelling and a learning task in which humans' performance was tested with and without feedback, we found that participants adjusted their response strategy. When feedback was not available, participants displayed a reduced propensity to act. Together with an asymmetric availability of information in the learning environment, this shift to a more cautious response mode was sufficient to yield improved performance. In contrast to the rodent study, our results do not suggest that feedback masks acquired knowledge. Instead, it supports a different mode of responding.

Introduction

In everyday life it is crucial to learn whether an action leads to reward or punishment. This adaptive behaviour has been extensively investigated in animal and human experiments and formally captured using reinforcement learning models [1–4]. In these models, the expected value of an action is updated using prediction errors, which reflect the discrepancy between obtained and expected rewards, in order to optimize future choices. Most learning tasks measure task performance while feedback is provided, which inevitably confounds learning with instrumental performance. To decouple learning and instrumental performance, some studies feature a learning phase and a later probe phase in which knowledge is tested in the absence of feedback. These studies show that different neural mechanisms underlie learning and expression of knowledge [5–9]. However, in these studies, acquired knowledge is usually tested *after* the learning performances has reached a plateau. In contrast, little is known about what happens when knowledge is tested without reinforcement during the learning process *prior to* participants reaching asymptotic performance.

During perceptual learning tasks, absence of feedback resulted in impaired [10], or unchanged [11] performance in humans. These results contrast with a recent rodent study in the domain of associative learning: Omitting feedback during early learning improved performance. Notably, performance deteriorated again when reinforcement was reintroduced, suggesting that reinforcement masked the underlying knowledge acquired by the animals [12].

The present study investigates whether this finding, that has been observed in rodents, extends to human learning. Specifically, we asked whether healthy volunteers' performance benefits similarly from omitting reinforcement during instrumental learning. To this end, closely following Kuchibhotla and colleagues [12], we adopted a go/no-go task that required participants to learn, by trial and error, to respond to go stimuli to obtain reward and to withhold responding to no-go stimuli to avoid punishment (monetary wins and losses, respectively). Crucially, reinforced trials were interleaved with multiple blocks in which participants were instructed to continue responding as previously, but no reinforcement was delivered (probe blocks). Similar to the pattern observed in rodents [12], we found that performance, as quantified by the sensitivity index *d'*, was improved in probe blocks, relative to reinforced blocks. However, computational modelling revealed that this pattern did not result from an increased sensitivity to acquired values. Instead, the behavioural pattern in the present paradigm could be completely explained by a mere reduction of an overall propensity to respond. Together with an initial tendency to act (as reflected in a positive initialization of value estimates), this change in overall response bias is sufficient to cause asymmetric changes to hit

and false alarm rates that jointly lead to an increased sensitivity index *d'*. Altogether, these results support the notion that omission of reinforcement may improve instrumental performance, however, rather than unmasking latent associative knowledge, this is due to a change in the overall propensity to act.

Results

Task

Based on recent findings in rodents [12], we hypothesized that the performance of humans in an instrumental learning task increases during non-reinforced compared to reinforced periods. Therefore, we designed a visual go/no-go reinforcement learning task (Fig 1A) consisting of reinforced trials which were interleaved with five probe blocks of non-reinforced trials. We used twelve greebles as stimuli [13]. Half of them were randomly assigned as go options, while the other half was assigned as no-go options. On each trial, one of the twelve stimuli was presented and participants had to learn, from trial and error, to perform a button press for go stimuli and to withhold responding for no-go stimuli. We used a rather high number of stimuli to be learnt by participants in order to evoke slow, incremental learning (Fig 1B). Participants obtained reward (monetary gain) for responding to go stimuli and punishment (monetary loss) for responding to no-go stimuli. Withholding a response resulted in no feedback (and neither monetary gain nor loss). This asymmetric reinforcement schedule follows the design used by Kuchibhotla and colleagues [12] and other work in rodents [14,15]. During probe trials, participants were instructed to continue choosing as they would do during reinforced trials, while reinforcement was temporarily omitted. We performed two experiments: an original study ($N = 30$) and a replication in an independent sample ($N = 30$). The main results were similar across the two studies; therefore, here we report the results of the pooled sample (see Section 1 in S1 Appendix for a separate presentation).

Analysis approaches: Signal detection theory and computational modelling

To assess effects of the removal of reinforcement on task performance, we present two approaches. First, we aimed to replicate the results of Kuchibhotla and colleagues' work [12] based on signal detection theory (SDT). To this end, we computed the sensitivity index *d'*, representing the difference in the relative frequencies of hits and false alarms (button presses to go and no-go stimuli, respectively) [16]. To compute measures from SDT, it is necessary to consider windows of several trials. Importantly, this approach can introduce artefacts in learning paradigms, due to its insensitivity to the general rising trend in performance. Specifically, during learning mean performance on earlier trials is intrinsically lower than mean performance on later trials, irrespective of any manipulation, such as feedback removal. Consequently, *d'* implicitly disadvantages earlier trials compared to later ones, unless performance has reached stable levels. For this reason, we next present a computational modelling approach, which averts this issue by providing trial-by-trial estimates. Conclusions are primarily based on this second approach, presented in the paragraph "Computational modelling reveals a shift in action bias".

Sensitivity index *d'* is increased when reinforcement is omitted

The SDT measure *d'* indicated a gradual increase in participants' performance (Fig 1B), confirming successful acquisition of the correct associations over time. This increase is mostly driven by no-go trials: While the initial go-response probability for no-go trials is high, participants learn to withhold responding over the course of the experiment (Fig 1C). To statistically

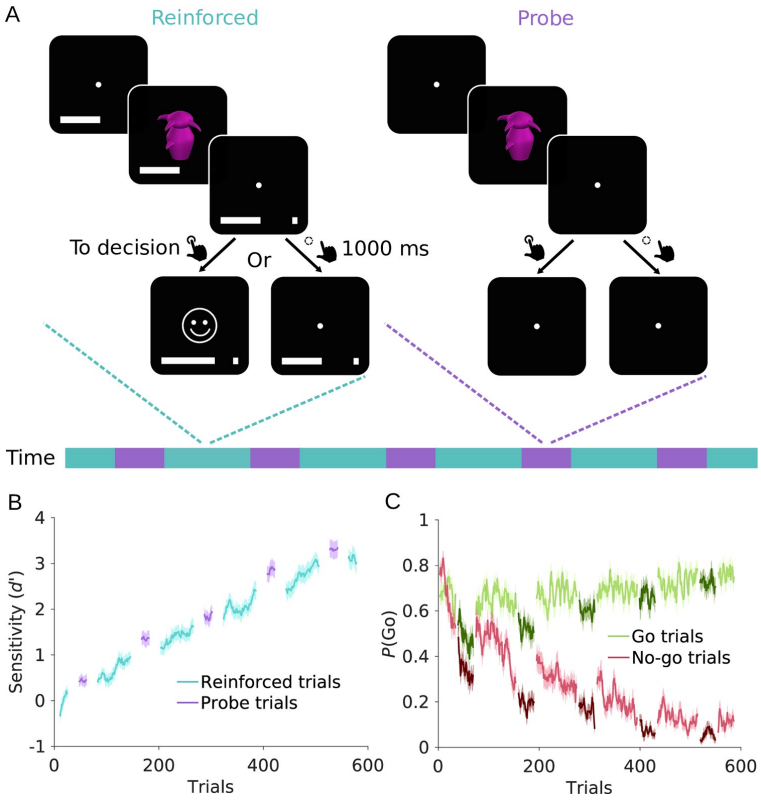

Fig 1. Task structure and participants' behaviour. (A) Schematic of the go/no-go learning task. On each trial, a fixation cross was presented for 1000–1600 ms. Then, participants were presented with one stimulus for 500 ms and had 1000 ms to decide whether to perform a go (button press) or no-go (no button press) response. Blocks of reinforced trials alternated with probe blocks (illustrated in the timeline). On reinforced trials (cyan), a go response resulted in reward or punishment (monetary win or loss, indicated by a smiley or frowny, respectively), depending on whether the stimulus was a go or no-go stimulus. No-go responses resulted in no feedback, and in neither reward nor punishment. A progress bar at the bottom of the screen displayed cumulative reward (rewards increased the bar, punishments shrank it). On probe block trials (purple), participants were required to respond as during reinforced blocks, but no feedback following responses was provided. (B) Sensitivity index *d'*, separately for reinforced (cyan) and probe trials (purple). (C) Time course of go-response probabilities, *P*(Go), for go trials (green) and no-go trials (red). Darker shades of green and red indicate probe trials. Solid lines in B and C represent mean, shaded areas SEM across participants.

https://doi.org/10.1371/journal.pcbi.1010201.g001

assess the change from reinforced to probe blocks, we compared performance of the 36 trials in a probe block with performance in the 36 trials before each probe block (pre-probe trials). Across the entire experiment, *d'* was indeed higher for probe compared to pre-probe trials ($\Delta d'$ = 0.47 ± 0.53, mean ± SEM, t_{59} = 6.94, $p < .001$, Cohen's d = 0.90, Fig 2A). This increase in *d'* was driven by a more pronounced reduction of false alarm rate (*FAR*) than hit rate (*HR*; ΔFAR–ΔHR = -0.07 ± 0.08, t_{59} = -6.42, $p < .001$, Cohen's d = -0.83). However, both measures decreased significantly in probe compared to pre-probe trials (ΔHR = -0.08 ± 0.06, t_{59} = -9.67, $p < .001$, Cohen's d = -1.25; ΔFAR = -0.15 ± 0.08, t_{59} = -15.12, $p < .001$, Cohen's d = -1.95, Fig

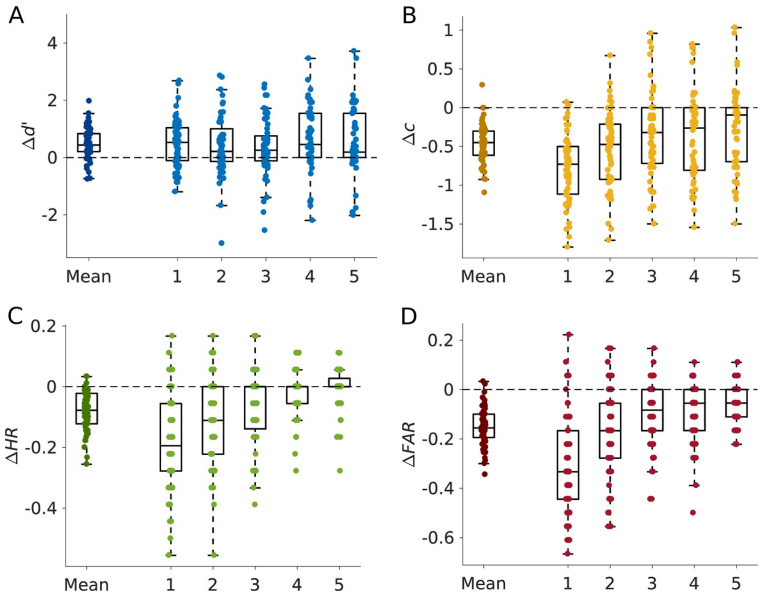

Fig 2. Behavioural results, expressed as difference between probe trials and preceding reinforced trials. Results are shown both for the mean across all five probe blocks (left) and separately for each probe block. Points reflect individual participants' behaviour. (A) The sensitivity index *d'* increased in probe compared to reinforced trials. (B) The negative bias criterion *c* decreased on probe blocks, indicating a reduced propensity to act on probe trials. (C), (D) Both hit rate (*HR*, C) and false alarm rate (*FAR*, D) decreased on probe blocks, but the decrease in *FAR* was more pronounced than the decrease in *HR*, which lead to the increase in *d'* represented in (A).

https://doi.org/10.1371/journal.pcbi.1010201.g002

2C and 2D), while in [12] only a decrease in false alarm rate during probe trials was reported. Once reinforcement was reinstated, *d'* significantly decreased again ($\Delta d' = 0.15 \pm 0.47$, $t_{59} = 2.40$, $p = .020$, Cohen's $d = 0.31$, see Fig G in S1 Appendix). Note that this effect was only evident in Experiment 2 when analysing the two experiments separately (see Section 1 in S1 Appendix). Again, the decrease in *d'* was driven by a significant increase in both hit and false alarm rate ($\Delta HR = -0.09 \pm 0.07$, $t_{59} = -7.92$, $p < .001$, Cohen's $d = -1.02$; $\Delta FAR = -0.10 \pm 0.10$, $t_{59} = -9.80$, $p < .001$, Cohen's $d = -1.27$, see Fig G in S1 Appendix).

In rodents, the removal of feedback improved performance only during early learning [12]. Therefore, we hypothesized that the increase in *d'* is strongest for early probe blocks. Contrary to this, there was no effect of time on the change in *d'* over probe blocks ($F(4, 59) = 0.74$, $p = .568$, $\eta^2 = 0.35$, Fig 2A), despite time effects on hit and false alarm rates (hit rate: $F(4, 59) = 24.18$, $p < .001$, $\eta^2 = 0.01$; false alarm rate: $F(4, 59) = 32.32$, $p < .001$, $\eta^2 = 0.29$, Fig 2C and 2D). Post hoc tests confirmed a significant increase in *d'* from pre-probe to probe trials for all five probe blocks (all $t_{59} \geq 2.74$, $p \leq .008$, see Table I in S1 Appendix). Thus, the increase in *d'* was not specific to early learning. The comparison of probe trials with post-probe trials yielded similar results (see Table J in S1 Appendix).

In addition to the sensitivity index *d'*, we also quantified the change in response bias of participants between probe and reinforced trials using the bias criterion *c* from SDT [16,17]. The bias criterion decreased from pre-probe to probe ($\Delta c = -0.47 \pm 0.25$, $t_{59} = -14.56$, $p < .001$, Cohen's $d = -1.88$, Fig 2B) and increased again in post-probe trials ($\Delta c = -0.42 \pm 0.32$, $t_{59} =$

-10.22, $p < .001$, Cohen's d = -1.32, see Fig G in S1 Appendix), thus, go-responding was reduced during probe trials, but increased when reinforcement was re-introduced again. Furthermore, there was an effect of time on the change of bias criterion from pre-probe to probe trials ($F(4, 59) = 9.21$, $p < .001$, $\eta^2 = 0.14$, Fig 2B), indicating that the reduced go-responding in probe trials diminished over the experiment.

In summary, both hit and false alarm rates decreased during probe compared to reinforced blocks, leading to a reduced response bias c. However, the decrease in false alarm rates was more pronounced compared to the decrease in hit rates, which further resulted in an increased sensitivity index d'.

Computational modelling reveals a shift in action bias

Due to the confound with SDT-based parameters in learning experiments described above, those results cannot be used here to distinguish between a real effect of reinforcement removal and an artefactually introduced effect. To overcome this issue, we used computational modelling. Unlike measures like d' which require consideration of several trials, reinforcement learning models provide value estimates for all stimuli for each trial [18–20]. We used variants of Q-learning with a delta update rule and softmax action selection. Two key parameters are at the heart of these reinforcement learning models: a learning rate and a softmax choice temperature. The learning rate determines the extent to which prediction errors are used to update value estimates, hence, governing the speed of learning. The softmax choice temperature determines how sensitive choices are to acquired value: At higher temperatures, participants' choices are increasingly stochastic, and large values are required to select the correct choice with high probability, while at low softmax temperatures, values slightly greater/lower than zero are sufficient to reliably select the go/no-go action, respectively. While in multi-alternative decisions, the temperature governs the balance between exploration and exploitation, in our paradigm with a single option per trial and deterministic reinforcement, the temperature can be used as an index of choice sensitivity. Critically, since the temperature is fitted based on trial-wise value estimates, it is not subject to the issues d' entails.

To test whether choice sensitivity in probe trials was indeed improved compared to reinforced trials or whether improved performance resulted from a non-specific change in action bias, we set up and compared four different models: a *baseline model*, a *temperature model*, a *bias model* and a *full model*. In all four models, learning rates α were set to a fixed value of 0.06 because learning rate and softmax temperature are strongly correlated for deterministic task structures (see methods for detailed reasoning) [21]. To rule out that the results are specific to this particular choice of learning rate, we re-ran all analyses across a wide range of learning rates and obtained an identical pattern of results (see Section 3.1. in S1 Appendix).

The baseline model included four free parameters: one single softmax temperature τ and a bias term b for both block types, one initial value estimate Q_0 for each of the twelve stimuli, and a decay parameter θ. On reinforced trials in which the go action was selected (and feedback received), values were updated using a delta update rule. On probe trials in which the go action was selected (and no feedback received), no changes were applied to value estimates. During both reinforced and probe trials without a go action, we assumed that values were subject to passive forgetting [22,23] via diffusion towards zero governed by the decay parameter θ. In addition, the softmax choice rule contained a bias term b that indicates participants' overall propensity to respond, independent of the option's current value. The temperature model was based on the baseline model, but it featured separate temperatures τ_R and τ_P, for reinforced and probe trials, respectively. Similarly, the bias model was based on the baseline model, but now, instead of the temperature, we allowed the bias parameter b to be different for reinforced

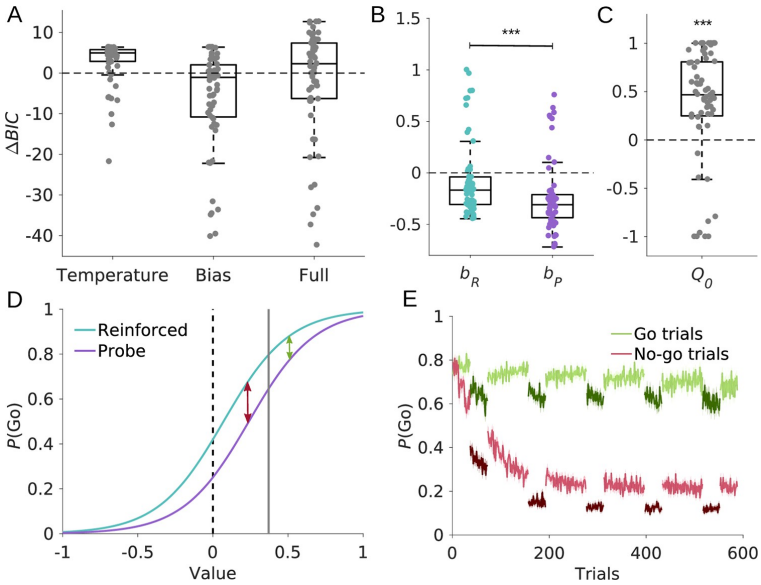

Fig 3. Computational modelling results. (A) Comparison of the Bayesian information criterion (*BIC*) relative to the baseline model. Negative *BIC* differences indicate a decrease in BIC relative to the baseline model and hence better fit. Conversely, a positive *BIC* difference indicates worse fit. The bias model provided the best fit. (B) The bias model contained two separate bias parameters, b_R and b_P, for reinforced and probe blocks, respectively. The bias is reduced on probe compared to reinforced trials. (C) Initial estimates Q_0 of option values. On average, estimates were initialized with positive values. (D) Softmax choice probabilities to select an option as a function of its value. The sigmoids for reinforced and probe trials were generated using the mean fitted parameters. This figure illustrates how a reduction in response bias together with a positive value initialization resulted in the increase in *d'* observed in behaviour. Solid vertical grey line indicates average Q_0. As values of go stimuli were acquired (shifting rightwards from the vertical line), the difference in action probabilities between probe and reinforced trials became smaller (green arrow). Conversely, as values of no-go stimuli were acquired (shifting leftwards from the vertical line), the difference became more pronounced (red arrow), thus leading to a stronger reduction in false alarm rates. (E) Time course of simulated go-response probabilities. The probability *P*(Go) for go trials (green) and no-go trials (red) was simulated based on the bias model. Darker shades of green and red indicate probe trials. Solid lines represent mean, shaded areas SEM across simulations.

https://doi.org/10.1371/journal.pcbi.1010201.g003

and probe trials (b_R and b_P). Finally, the full model incorporated both separate temperature parameters τ_R and τ_P and separate bias parameters b_R and b_P.

Contrary to our expectation, the temperature model performed the worst (*BIC* = 566.90 ± 137.25, median ± SEM, Fig 3A), followed by the full model (*BIC* = 563.48 ± 132.10), outperformed even by the baseline model (*BIC* = 561.56 ± 137.39). The best fitting model was the bias model (*BIC* = 557.86 ± 131.99). Closer analysis of the bias model showed that the response bias in reinforced trials, b_R, was higher compared to the bias in probe blocks, b_P (Δb = -0.17 ± 0.12, t_{59} = -10.92, $p < .001$, Cohen's d = -1.41, Fig 3B). Thus, participants had a reduced propensity to act during probe blocks.

One might argue that the differences in behaviour between reinforced and probe blocks are subtle, such that the improvement in model fit conferred by two separate temperatures in the full model did not survive punishment by the Bayesian information criterion. We therefore explored the full model and tested whether there were differences in either temperatures, τ_R and τ_P, or bias parameters, b_P and b_R, or both. Again, we found that the bias parameter b_R was

significantly higher compared to b_P (Δb = -0.16 ± 0.13, t_{59} = -9.61, $p < .001$, Cohen's d = -1.24), whereas temperatures τ_P and τ_R did not differ significantly ($\Delta\tau$ = -0.01 ± 0.06, t_{59} = -1.65, p = .104, Cohen's d = -0.21). Thus, even in the full model, there is no evidence for a change in decision temperature.

To check the sensibility of the fitting procedure, we performed parameter recovery on simulated data sets generated using the fitted parameters from the best-fitting bias model and tested whether we could recover the ground-truth parameters based on these simulated data. Results showed successful recovery of all model parameters, as evidenced by the high correlations between fitted and recovered parameters (see section *Parameter recovery* and Section 3.2. in S1 Appendix). Next, we validated the winning model. Due to the confound in the SDT-based analyses of the behavioural data, any model which includes learning generates a difference between pre-probe and probe trials, hence, examining qualitative difference between reinforced and probe data for model validation is not warranted. Thus, we used the simulated go-response probabilities for go and no-go trials to validate which parameters are necessary to recapitulate the patterns observed in participants' behaviour (see Figs J and K in S1 Appendix). We found that only the winning model that included separate bias parameters for blocks with and without reinforcement could replicate the observed difference in go-response probabilities between reinforced and probe trials (see Fig 3E and Fig L in S1 Appendix). Taken together, parameter recovery and model validation indicate that our model with two different bias parameters and one fixed softmax temperature provided a plausible account of participants' behaviour in our experiment.

The increase in the sensitivity index *d'* from reinforced to probe blocks, resulted from a differential reduction in hit versus false alarm rates. It may appear surprising that a mere change in response bias is sufficient to drive such differential changes. However, this arises naturally as a consequence of the sigmoid shape of the softmax choice function, together with a positive initialization of value estimates. This effect is shown in Fig 3D depicting the softmax choice functions for the reinforced and probe blocks (based on the fitted values for τ, b_P and b_R). For the vast majority (50 of 60) of our participants, we found positive estimates for initial values (Q_0 = 0.37 ± 0.58, t_{59} = 4.94, $p < .001$, Cohen's d = 0.64, Fig 3C), reflecting participants' tendency to act (i.e., providing a go-response) on the first trials of the experiment. Thus, the initial value estimate is already shifted from zero (dashed line) to higher values (grey line). The difference between the two curves describes the reduced go-response probability for probe trials compared to reinforced trials. This difference is smaller for go stimuli than for no-go stimuli; During the acquisition of values for go stimuli, both curves quickly converge towards 1 (green arrow), while the exact opposite happens for no-go stimuli. These likewise start at a relatively high positive value, but because they are updated in the opposite direction during learning, the difference between the two curves first increases before decreasing again when converging towards -1. Thus, a positive value initialization together with a decrease in action bias results in enhanced instrumental performance during probe blocks, without any change in choice sensitivity to acquired value.

Discussion

When evaluating learning success, instrumental performance is measured during the learning process, conflating measures of learning with proficiency in expressing the acquired knowledge. However, it is well known that learning of action-outcome association relies, in part, on different neural substrates than expression of instrumental performance contingent upon these associations. Specifically, some neural mechanisms required for learning are not involved in the expression of learnt behaviour and vice versa [6,7,8,24,25,26,27]. To disentangle these

two concepts, the phases designed to assess learning versus expression of task performance are usually separated by a considerable delay in these studies. This implies that behaviour during the test phase relies on long-term consolidation of memories. Alternatively, to obtain a pure measure of an agents' current learning success, one option is to omit reinforcement/feedback during the learning process, which yielded inconsistent results in previous studies [10,11,12,28].

We aimed to reconcile the apparent contradictory results by investigating whether removing reinforcement unmasks latent associative knowledge during instrumental learning. Healthy humans performed an instrumental go/no-go learning task with reinforcement in which blocks without reinforcement were interspersed. Replicating previous rodent work [12], we first found that the sensitivity index *d'* was enhanced during blocks in which reinforcement was omitted. However, these findings based on signal detection theory are confounded with trial number, as measures like *d'* are computed over windows of several trials. This is problematic for dynamic processes like learning, where earlier sets of trials are inherently disadvantaged compared to later sets, unless learning has reached a plateau. To avoid this confound, we therefore used reinforcement learning models to investigate the mechanism driving the apparent change in choice sensitivity during non-reinforced trials. Such models have the advantage that they provide a point estimate for the stimulus value on each trial. Contrary to our expectations, better performance in non-reinforced trials did not result from an increased choice sensitivity, as would be reflected in a decreased softmax choice temperature. Instead, our modelling results suggest that the change in *d'* can fully be accounted for by a decrease in a bias parameter (reflecting participants' overall propensity to act), together with a positive value initialization (reflecting participants' tendency to act on the first trials). First, a model with only a single softmax temperature but with two separate bias parameters (for reinforced and non-reinforced blocks, respectively) provided the best fit to participants' choices. Second, this model successfully recapitulated the behavioural patterns we observed. Third, even when we explored a full model (with two separate bias and temperature parameters), we found that, while bias parameters still differed significantly between non-reinforced and reinforced blocks, temperatures did not differ. Altogether our results suggest that the omission of feedback led to an adjustment of response strategy rather than enhancing the expression of latent task knowledge.

The effects of removing feedback may be dependent on the domain of learning being studied. In the domain of perceptual learning, one study reported impaired performance without feedback [10], whereas others reported no performance differences between reinforced and non-reinforced trials in a similar paradigm [11,28]. In the domain of associative learning, one recent rodent study suggests a beneficial effect of omitting reinforcement on performance in an instrumental learning task [12]. We based our learning paradigm on this latter work and also found enhanced performance during non-reinforced trials (measured using the sensitivity index *d'*) in our sample of healthy volunteers. Therefore, despite interpreting the findings differently, we conceptually replicated the rodent work in humans.

Due to challenges inherent to translational work, one might argue that differences in task environments between the rodent tasks used by Kuchibhotla and colleagues' [12] and our human task could lead to results relying on different mechanisms. For example, extinction is well known to occur when animals' actions are not reinforced (as in the probe trials) [29,30]. To prevent extinction effects to confound results, only short probe phases were used in the rodent study [12], making it unlikely that their effects are influenced by extinction. Because rodents did not know that their knowledge is tested in probe trials, we aimed to adapt instructions accordingly. To avoid that participants try to perform better in probe trials, we instructed them to continue responding to the task in non-reinforced trials as they did in reinforced trials.

Finally, these findings hinge on the specific task structure. Our task encouraged an asymmetric response pattern: Performing an action (i.e. go response) is the only possibility to learn and explore possible outcomes for specific stimuli in our task, as refraining from responding yielded no feedback. Therefore, go responses provide a gain in information (information bonus), making them advantageous relative to no-go responses. This information bonus may account for the two main effects our computational model revealed: On the one hand, the information bonus is greatest early in the experiment, because values are not reliable learnt yet. This is reflected in the positive value initialization in our computational model. On the other hand, information can only be obtained in reinforced trials, resulting in a more pronounced general bias to withholding responses in non-reinforced trials. The resulting behaviour is in line with research about curiosity, suggesting that a lack of information makes individuals curious and facilitates information seeking behaviour [31]. Different tasks that do not encourage go responses for optimal performance are unlikely to result in a beneficial effect of reinforcement removal. Specifically, environments that do not favour any specific response (e.g., situations where the outcomes of both choice options are always presented) will result in a neutral value initialization, because participants do not need to perform a go response in order to gain information. Any change in response bias (if present) would affect hit and false alarm rates to the same extent, thus resulting in no difference in performance. Following the same logic in reverse, environments favouring no-go responses should result in a negative value initialization, thus resulting in the opposite pattern: a reduced performance during non-reinforced trials.

Likewise, we assume that an adaptive response bias would also manifest in foraging-related tasks. In previous work, we have shown that the average rate of responding is dependent on the local average reward rate in the environment, even when rewards were not contingent on participants' choices [32]. Furthermore, similar effects may be expected for learning environments characterized by high levels of volatility, where contingencies between choices and outcomes change frequently. In such situations, an agent would benefit from a high response probability on reinforced trials. Conversely, the cost of a false alarm would be greater than the cost of missing an opportunity in non-reinforced trials, resulting in decreased go-response probabilities. Thus, at high levels of environmental volatility, the effect of feedback removal on the response bias might be even more pronounced, as frequent contingency changes further increase the informative value of go-responses during reinforced periods. Therefore, we suggest that the effects of feedback removal on instrumental performance are highly dependent on the particular characteristics of the task at hand.

In conclusion, we found that omitting feedback during learning may indeed improve instrumental performance. However, our results show that this improvement results from a shift in participants' overall bias to act, rather than from unmasking of task knowledge.

Materials and methods

Ethics statement

All studies were approved by the Ethics Committee for Noninvasive Research on Humans of the Heinrich Heine University Düsseldorf (reference OR01-2020-01).

Participants

Participants were recruited from the local student community of the Heinrich Heine University, Germany. Each experiment was run with healthy participants with normal or corrected-to-normal vision and no history of neurological diseases. Participants signed a written informed consent prior to participation and received course credit or monetary compensation

for their participation. Participants were only included if they succeeded at learning during the task. For that reason, we set a performance criterion of $d' \geq 1$, which participants needed to exceed across all reinforced trials of the second half of the experiment. We a priori defined two criteria to constrain the sample size for each experiment: (1) collecting data of at least 30 participants fulfilling the performance criterion, and (2) using a Bayesian stopping rule, meaning that we set out to acquire as many participants as needed to find strong evidence either against or in favour of an overall change in d' between reinforced and probe trials ($BF_{01} > 10$ or $BF_{10} > 10$, respectively). For Experiment 1, 46 participants volunteered to take part in the study and 16 were excluded as their behaviour did not fulfil the predefined performance criterion. In Experiment 2, 39 participants completed the task, 9 of which did not pass the criterion and were therefore excluded. For both experiments, we found strong evidence in favour of an overall change in d' between reinforced and probe trials after inclusion of 30 participants, such that data acquisition in both experiments was stopped after 30 included participants. This resulted in $N = 30$ participants (mean age = 21.2 ± 4.0, 22 female and 8 male) for Experiment 1 and $N = 30$ participants (mean age = 24.5 ± 5.0, 9 female and 21 male) for Experiment 2. As both experiments yielded qualitatively similar results, we pooled the data from both studies for further analyses. Results are reported separately for the two studies in the supplementary materials (see Section 1 in S1 Appendix).

Behavioural task

As paradigm we employed a visual go/no-go learning task written and presented with PsychoPy (version 3.1.5) [33]. Stimuli were presented on an Asus PG248Q LCD display (24", 1920x1080, 60 Hz refresh rate) at a viewing distance of 80 cm. Each trial started with the presentation of a fixation cross, spanning 0.72˚ visual angle, for a pseudo-randomly selected duration of 1000 ms to 1600 ms to keep the participants' attention to the centre of the screen. Stimuli spanned the central 3.58˚ visual angle of the screen and were presented for 500 ms. We used greebles (Greebles 2.0) [13] as stimuli. Greebles are three-dimensional objects which are usually used for object and face recognition. Based on the body shape, greebles are classified into so-called families, while three other features vary for each exemplar, such that similarity between exemplars of the same families is larger than between exemplars across families. To drive slow learning, we used twelve greebles that were evenly sampled from three families. For each participant, half of the greebles of each family were pseudo-randomly assigned to be go stimuli while the other half were no-go stimuli to make sure that learning is similarly difficult across participants. They had to learn to perform a button press for go stimuli and to withhold a response for no-go stimuli. The response could be administered during stimulus presentation and within 1000 ms after stimulus offset. The duration of this response window was well sufficient to perform a go-response (see Fig M in S1 Appendix). Feedback was immediately presented after button press consisting of a smiley or frowny (spanning 1.93˚ visual angle each) for correct and incorrect actions, respectively. Every correct button press was rewarded with 2 cents, which was indicated by an increase of the progress bar, while every incorrect button press led to a 2 cents deduction and a decrease of the progress, respectively. To design the task comparable to previous animal experiments, no feedback was delivered (and neither money won nor lost), when no button was pressed. The task consisted of 588 trials, with each of the twelve stimuli presented 49 times in pseudo-randomized order with the constraints that each one is presented once in twelve trials and that two consecutive stimuli were always different. Reinforced trials were interspersed by five probe blocks (36 trials each). Participants were instructed to respond as they had previously done during the reinforced blocks, but they no longer received feedback for their choices and the progress bar disappeared. The paradigm

was performed in five phases of 120 trials. Participants were encouraged to rest between the phases as long as they needed.

Behavioural analyses

Data analyses were conducted using MATLAB (MATLAB Version R2016b, Massachusetts: The Mathworks Inc.). Statistical significance testing, including the computation of the Bayes Factor of the main effect, was done in JASP (JASP Team (2019). JASP (Version 0.11.1) [macOS]). We calculated the sensitivity index *d'* as:

$$d\prime = z(HR) - z(FAR) \quad [1]$$

With the *z*-scored hit rate *z(HR)* and the *z*-scored false alarm rate *z(FAR)* [16]. As ceiling performances cannot be *z*-scored, we used $HR = 1 - \frac{1}{2N}$ as correction for *HR = 1*, and $HR = \frac{1}{2N}$ as correction for *HR = 0*, with the number of trials *N*. Because we calculated *d'* for a small number of trials, the use of corresponding trials for correction may result in underestimated *d'* values [34], so we used a correction allowing *HR* and *FAR* to be approximately 0 or 1. The negative bias criterion *c* was calculated as:

$$c = \frac{1}{2}(z(HR) + z(FAR)) \quad [2]$$

For comparison of *d'* between probe and reinforced trials, we defined the 36 trials (the length of a probe block) before reinforcement removal as pre-probe trials, and the 36 trials after reinforcement was reinstated as post-probe trials, and computed the difference between probe and pre-probe trials, and between probe and post-probe trials. When visualising the learning curve, *d'* was computed within a sliding window of 21 trials. To further specify the effects of *d'* and the bias criterion, we also analysed hit and false alarm rates separately for reinforced and probe trials.

Changes in behaviour between reinforced and probe blocks were analysed using one-sample Student's *t*-test comparing their difference against zero. The *t*-tests examined the null hypothesis that there is no difference in behaviour between reinforced and probe blocks at a significance level of $\alpha = 0.05$. To test whether the change in behaviour from reinforced to probe blocks differed across successive probe blocks, we subjected the differences in *d'* (probe—pre-probe, and analogously for hit and false alarm rates) to repeated measures ANOVAs (Greenhouse-Geisser correction to adjust for lack of sphericity when $\varepsilon < 1$).

Go-response probabilities *P*(Go) were computed with a sliding window of five trials and averaged over participants for visualization. We used a smaller window size compared to the *d'* learning curves, because it was sufficient to obtain a good resolution for go-probability.

Reinforcement learning models

Computational modelling was performed in MATLAB (MATLAB Version R2016b, Massachusetts: The Mathworks Inc.). Altogether, we tested four models, in each of which values for chosen stimuli were updated using a delta update rule:

$$Q_{i,t+1} = Q_{i,t} + \alpha(r_t - Q_{i,t}) \quad [3]$$

Where $Q_{i,t}$ is the value for stimulus *i* presented on trial *t*, α is the learning rate and r_t is the observed outcome on trial *t*. On trials during reinforced blocks in which subjects performed a go-response, r_t was either -1 or 1, depending on whether the chosen stimulus was a go- or no-go stimulus, respectively. When participants gave a response during probe trials, the value of

the corresponding stimulus was not updated, i.e. $Q_{i,t+1} = Q_{i,t}$. In line with the non-monotonic plasticity hypothesis [22], we have recently shown that associations of unchosen stimuli are weakened [23]. Therefore, on trials during which participants performed a no-go response, we assumed passive forgetting of the displayed option governed by a decay parameter θ (for reinforced and probe trials):

$$Q_{i,t+1} = \theta Q_{i,t} \quad [4]$$

Initial Q-values Q_0 for the first trial were also treated as a free parameter. The learning rate α was not treated as a free parameter and instead fixed at $\alpha = 0.06$ for all participants (see section *Fixed learning rate* for explanation). Choices were modelled using a softmax choice rule. The four models differed with regard to the bias and temperature terms contained in their respective softmax choice rules:

Baseline model. Choices in the baseline model are generated using a softmax choice rule that contains a single temperature τ and bias term b:

$$p_t = \frac{1}{1 + e^{\frac{-(Q_{i,t}+b)}{\tau}}} \quad [5]$$

Where b is a general bias to act and τ the temperature parameter determining the stochasticity of action selection. Importantly, the softmax choice rule used the same parameters for reinforced and probe trials, thus, the baseline model did not discriminate between block types. Altogether, for this model, four parameters were thus fit: the initial Q-value Q_0, the decay θ, the general bias b and the softmax temperature τ.

Temperature model. Like the baseline model, the temperature model also contained a single bias parameter b, but it allowed for the temperature τ to vary between reinforced and probe blocks:

$$p_t = \frac{1}{1 + e^{\frac{-(Q_{i,t}+b)}{\tau_k}}} \quad [6]$$

Where $\tau_k = \tau_R$ for reinforced trials and $\tau_k = \tau_P$ for probe trials. Thus, for this model, five parameters were fit: the initial Q-value Q_0, the decay θ, the general bias b, the softmax temperature τ_R for reinforced trials and the softmax temperature τ_P for probe trials.

Bias model. Here, instead of allowing the temperature to vary between reinforced and probe trials, we now allowed for separate bias parameters:

$$p_t = \frac{1}{1 + e^{\frac{-(Q_{i,t}+b_k)}{\tau}}} \quad [7]$$

Where $b_k = b_R$ for reinforced trials and $b_k = b_P$ for probe trials. Thus, for this model, five parameters were fit: the initial Q-value Q_0, the decay θ, the general bias b_R for reinforced trials, the general bias b_P for probe trials and the softmax temperature τ.

Full model. This model is a combination of the temperature and bias model: It used both a separate temperature and a separate bias for reinforced and probe trials:

$$p_t = \frac{1}{1 + e^{\frac{-(Q_{i,t}+b_k)}{\tau_k}}} \quad [8]$$

Thus, for this model six parameters were fit: the initial Q-value Q_0, the decay θ, the general bias b_R for reinforced trials, the general bias b_P for probe trials, the softmax temperature τ_R for reinforced trials and the softmax temperature τ_P for probe trials.

Model comparison. Models were fit by minimizing the negative log likelihood *NLL*:

$$NLL = -\sum_t log\ (p_t) \quad [9]$$

Where p_t is a vector containing, for each trial t, the model's probability to select the choice performed by the participant. We used unconstrained non-linear optimization implemented in Matlab's function *fmincon*. To minimize the risk of finding local optima, we started optimization from 1000 random starting points for each participant. To account for the different number of parameters, we used the Bayesian information criterion for model comparison:

$$BIC = 2NLL + n_{param} log(n_t) \quad [10]$$

Where n_{param} and n_t are the number of parameters and trials, respectively. A lower *BIC* score indicates a better model fit.

Model validation

To test whether the best-fitting model provides a good account of participants' behaviour, we tested whether we could replicate the behavioural results using simulated datasets. Because the comparison of pre-probe and probe trials confounds performance and trial number, the replication of the SDT-based behavioural analysis using simulated data is not suited for model validation. Instead, we visually inspected simulated data for the baseline model, the temperature model and the bias model. To this end, we simulated 500 datasets per participant based on these models, using the parameter combination fitted for the respective participant. Then, we computed go-response probabilities for go and no-go trials to investigate how single parameters of the different models changed these probabilities and ultimately, if a bias parameter for both reinforced and probe blocks is necessary to describe the patterns found in behaviour (see Section 3.3. in S1 Appendix). For visualization, go-response probabilities were averaged over the number of simulations and simulated participants.

Parameter recovery

In order to test the reliability of fitted parameters, we performed a parameter recovery. We used the 500 simulated datasets per participant from the model validation and fitted the bias model in the same way as described above for the experimental data. For each participant and each parameter, we compared the original model fit with the synthetic model fit (see Fig I in S1 Appendix). The high correlation coefficients (all $\rho > 0.99$, see Table L in S1 Appendix) indicated successful recovery for all parameters.

Fixed learning rate

The deterministic task design gives rise to a strong anti-correlation between learning rate and softmax temperature, thus, both parameters could not be estimated by the models independently [21]. Note however, that it is not plausible to assume different learning rates for probe and reinforced blocks, as no learning rate is applied to the probe trials. Instead, it was our goal to test whether the sensitivity of choices to acquired values changes between reinforced and probe trials. Therefore, we fit the models using a fixed learning rate of $\alpha = 0.06$ for all participants. In order to ascertain that the results are independent of this particular choice of learning rate, we used six different learning rates evenly log-spaced between 0.01 and 0.20. With this set of learning rates, we fitted all four models and compared them. We found that the bias model performed best, while the temperature model was the worst fitting model, independent of the learning rate (see Section 3.1. in S1 Appendix).

Supporting information

S1 Appendix. Further Analyses of behavioural and computational data. In Section 1, we present results separately for the two experiments. In Section 2, we show further analyses of the pooled dataset. In Section 3, analyses of the computational modelling, in particular the control analysis for the fixed learning rate, the parameter recovery and the model validation, are presented. In Section 4, we analysed reaction times.
(PDF)

Acknowledgments

We thank Judith Geusen, Christina Kalinicenko, Joshua Saal and Georg Schäfer for their support during data acquisition. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

Author Contributions

Conceptualization: Hannah Kurtenbach, Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

Data curation: Hannah Kurtenbach, Eduard Ort.

Formal analysis: Hannah Kurtenbach, Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

Funding acquisition: Gerhard Jocham.

Investigation: Hannah Kurtenbach.

Methodology: Hannah Kurtenbach, Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

Project administration: Hannah Kurtenbach, Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

Software: Hannah Kurtenbach, Eduard Ort, Gerhard Jocham.

Supervision: Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

Validation: Hannah Kurtenbach, Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

Visualization: Hannah Kurtenbach.

Writing – original draft: Hannah Kurtenbach.

Writing – review & editing: Eduard Ort, Monja Isabel Froböse, Gerhard Jocham.

References

1. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci. 2007 Sep; 10(9):1214–21. https://doi.org/10.1038/nn1954 PMID: 17676057
2. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron. 2011 Mar; 69(6):1204–15. https://doi.org/10.1016/j.neuron.2011.02.027 PMID: 21435563
3. Lee D, Seo H, Jung MW. Neural Basis of Reinforcement Learning and Decision Making. Annu Rev Neurosci. 2012 Jul 21; 35(1):287–308. https://doi.org/10.1146/annurev-neuro-062111-150512 PMID: 22462543
4. Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. Orbitofrontal Cortex as a Cognitive Map of Task Space. Neuron. 2014 Jan; 81(2):267–79. https://doi.org/10.1016/j.neuron.2013.11.005 PMID: 24462094

5. Sutherland RJ, McDonald RJ, Hill CR, Rudy JW. Damage to the hippocampal formation in rats selectively impairs the ability to learn cue relationships. Behav Neural Biol. 1989 Nov; 52(3):331–56. https://doi.org/10.1016/s0163-1047(89)90457-3 PMID: 2590146
6. Smith-Roe SL, Kelley AE. Coincident Activation of NMDA and Dopamine D $_1$ Receptors within the Nucleus Accumbens Core Is Required for Appetitive Instrumental Learning. J Neurosci. 2000 Oct 15; 20(20):7737–42.
7. Baldwin AE, Holahan MR, Sadeghian K, Kelley AE. W-Methyl-D-Aspartate Receptor-Dependent Plasticity Within a Distributed Corticostriatal Network Mediates Appetitive Instrumental Learning. Behav Neurosci. 2000; 114(1):84–98.
8. Shiflett MW, Brown RA, Balleine BW. Acquisition and Performance of Goal-Directed Instrumental Actions Depends on ERK Signaling in Distinct Regions of Dorsal Striatum in Rats. J Neurosci. 2010 Feb 24; 30(8):2951–9. https://doi.org/10.1523/JNEUROSCI.1778-09.2010 PMID: 20181592
9. Smittenaar P, Chase HW, Aarts E, Nusselein B, Bloem BR, Cools R. Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection—learning or performance?: Dopamine and value-based choice. Eur J Neurosci. 2012 Apr; 35(7):1144–51.
10. Herzog MH, Fahle M. The role of feedback in learning a vernier discrimination task. Vision Res. 1997 Aug; 37(15):2133–41. https://doi.org/10.1016/s0042-6989(97)00043-6 PMID: 9327060
11. Petrov AA, Dosher BA, Lu ZL. Perceptual learning without feedback in non-stationary contexts: Data and model. Vision Res. 2006 Oct; 46(19):3177–97.
12. Kuchibhotla KV, Hindmarsh Sten T, Papadoyannis ES, Elnozahy S, Fogelson KA, Kumar R, et al. Dissociating task acquisition from expression during learning reveals latent knowledge. Nat Commun. 2019 Dec; 10(1):2151. https://doi.org/10.1038/s41467-019-10089-0 PMID: 31089133
13. Gauthier I, Tarr MJ. Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. J Exp Psychol Hum Percept Perform. 2002; 28(2):431–46. https://doi.org/10.1037//0096-1523.28.2.431 PMID: 11999864
14. Berditchevskaia A, Cazé RD, Schultz SR. Performance in a GO/NOGO perceptual task reflects a balance between impulsive and instrumental components of behaviour. Sci Rep. 2016 Jun; 6(1):27389.
15. Reinert S, Hübener M, Bonhoeffer T, Goltstein PM. Mouse prefrontal cortex represents learned rules for categorization. Nature. 2021 May 20; 593(7859):411–7. https://doi.org/10.1038/s41586-021-03452-z PMID: 33883745
16. Green DM, Swets JS. Signal Detection Theory and Psychophysics. Wiley; 1966.
17. Young ME, Sutherland SC, McCoy AW. Optimal go/no-go ratios to maximize false alarms. Behav Res Methods. 2018 Jun; 50(3):1020–9. https://doi.org/10.3758/s13428-017-0923-5 PMID: 28664243
18. Sutton RS, Barto AG. Reinforcement learning: an introduction. Cambridge, Mass: MIT Press; 1998. 322 p. (Adaptive computation and machine learning).
19. Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M. Adaptive Coding of Action Values in the Human Rostral Cingulate Zone. J Neurosci. 2009 Jun 10; 29(23):7489–96. https://doi.org/10.1523/JNEUROSCI.0349-09.2009 PMID: 19515916
20. Klein TA, Ullsperger M, Jocham G. Learning relative values in the striatum induces violations of normative decision making. Nat Commun. 2017 Dec 22; 8(1):16033. https://doi.org/10.1038/ncomms16033 PMID: 28631734
21. Bennett D, Niv Y, Langdon AJ. Value-free reinforcement learning: policy optimization as a minimal model of operant behavior. Curr Opin Behav Sci. 2021 Oct; 41:114–21. https://doi.org/10.1016/j.cobeha.2021.04.020 PMID: 36341023
22. Ritvo VJH, Turk-Browne NB, Norman KA. Nonmonotonic Plasticity: How Memory Retrieval Drives Learning. Trends Cogn Sci. 2019 Sep; 23(9):726–42. https://doi.org/10.1016/j.tics.2019.06.007 PMID: 31358438
23. Luettgau L, Tempelmann C, Kaiser LF, Jocham G. Decisions bias future choices by modifying hippocampal associative memories. Nat Commun. 2020 Dec; 11(1):3318. https://doi.org/10.1038/s41467-020-17192-7 PMID: 32620879
24. Kelley AE, Smith-Roe SL, Holahan MR. Response-reinforcement learning is dependent on *N*-methyl- D -aspartate receptor activation in the nucleus accumbens core. Proc Natl Acad Sci. 1997 Oct 28; 94 (22):12174–9.
25. Andrzejewski ME, Spencer RC, Kelley AE. Instrumental learning, but not performance, requires dopamine D1-receptor activation in the amygdala. Neuroscience. 2005 Jan; 135(2):335–45. https://doi.org/10.1016/j.neuroscience.2005.06.038 PMID: 16111818
26. Corbit LH, Balleine BW. The role of prelimbic cortex in instrumental conditioning. Behav Brain Res. 2003 Nov; 146(1–2):145–57. https://doi.org/10.1016/j.bbr.2003.09.023 PMID: 14643467

27. Ostlund SB, Balleine BW. Differential Involvement of the Basolateral Amygdala and Mediodorsal Thalamus in Instrumental Action Selection. J Neurosci. 2008 Apr 23; 28(17):4398–405. https://doi.org/10.1523/JNEUROSCI.5472-07.2008 PMID: 18434518

28. Haddara N, Rahnev D. The Impact of Feedback on Perceptual Decision-Making and Metacognition: Reduction in Bias but No Change in Sensitivity. Assoc Psychol Sci. 2022 Jan; 33(2):259–75. https://doi.org/10.1177/09567976211032887 PMID: 35100069

29. Pavlov IP. Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex. Oxf Univ Press Lond. 1927; 142.

30. Bouton ME. Context and Behavioral Processes in Extinction. Learn Mem. 2004 Sep; 11(5):485–94. https://doi.org/10.1101/lm.78804 PMID: 15466298

31. van Lieshout LLF, Traast IJ, de Lange FP, Cools R. Curiosity or savouring? Information seeking is modulated by both uncertainty and valence. Verguts T, editor. PLOS ONE. 2021 Sep 24; 16(9):e0257011. https://doi.org/10.1371/journal.pone.0257011 PMID: 34559816

32. Jocham G, Brodersen KH, Constantinescu AO, Kahn MC, Ianni AM, Walton ME, et al. Reward-Guided Learning with and without Causal Attribution. Neuron. 2016 Apr; 90(1):177–90. https://doi.org/10.1016/j.neuron.2016.02.018 PMID: 26971947

33. Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, et al. PsychoPy2: Experiments in behavior made easy. Behav Res Methods. 2019 Feb; 51(1):195–203. https://doi.org/10.3758/s13428-018-01193-y PMID: 30734206

34. Hautus MJ. Corrections for extreme proportions and their biasing effects on estimated values of d′. Behav Res Methods Instrum Comput. 1995 Mar; 27(1):46–51.

S1 Appendix: Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations

Contents

1. Results separately for the two experiments

We acquired two independent data sets (each *N* = 30), which we pooled for the main analyses. Here, we present the results separately for Experiment 1 (Fig A-C and Table A-D) and Experiment 2 (Fig D-F and Table E-H).

1.1. Experiment 1

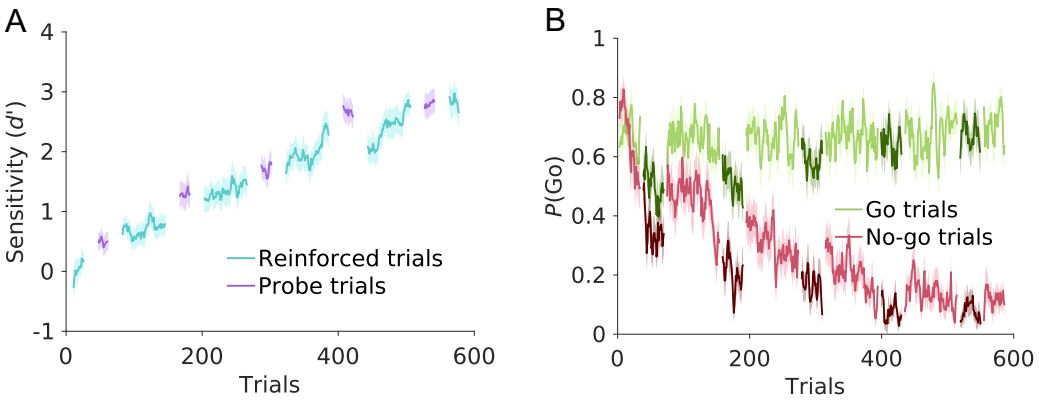

Fig A. Average learning performance for Experiment 1 (*N* = 30).
(A) Sensitivity index *d'*, separately for reinforced (cyan) and probe trials (purple). Solid lines represent mean performance, shaded areas SEM across participants. (B) Time course of go-response probabilities, *P*(Go), for go trials (green) and no-go trials (red). Darker shades of green and red illustrate probe trials. Solid lines represent mean, shaded areas SEM across participants.

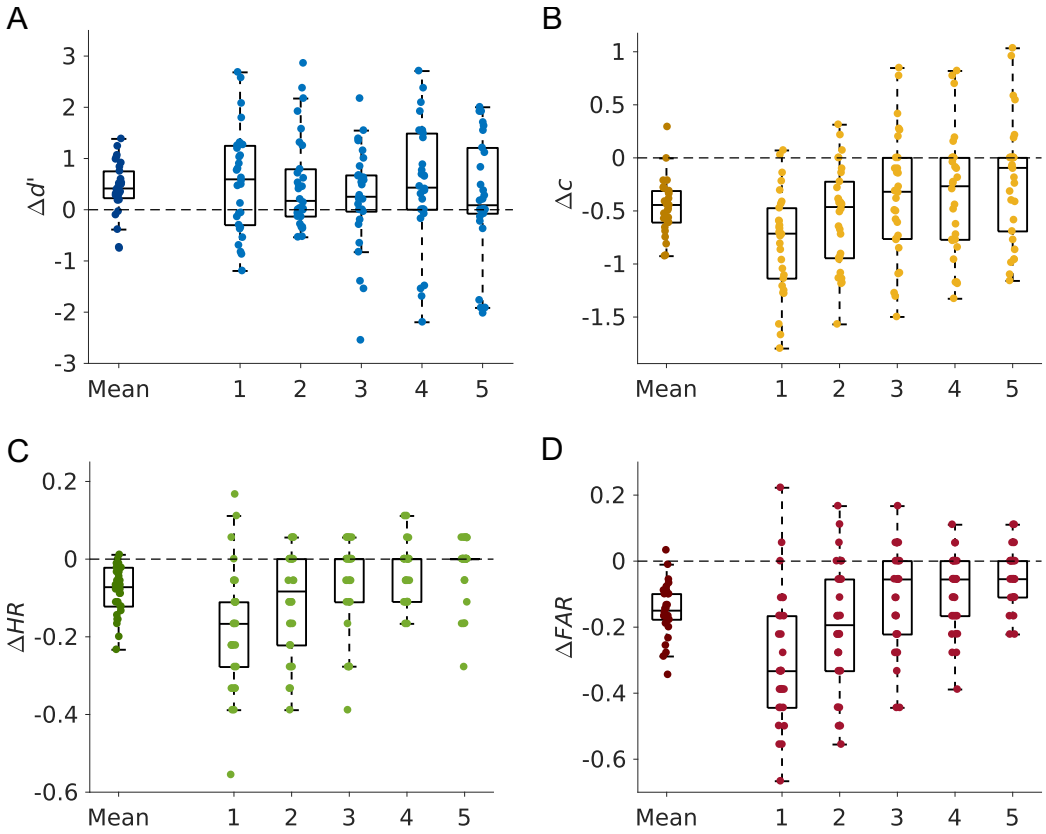

Fig B. Behavioural results, expressed as difference between probe trials and preceding reinforced trials for Experiment 1 (*N* = 30).
Results are shown both for the mean across all five probe blocks (left) and separately for each probe block. (A**)** The sensitivity index *d'* increased in probe compared to reinforced trials. (B) The negative bias criterion *c*, decreased in probe blocks, indicating a reduced propensity to act on probe trials. (C), (D) Both hit rate (*HR*, C) and false alarm rate (*FAR*, D) decreased in probe blocks.

Table A. Difference between probe and pre-probe trials for Experiment 1.
Statistical comparisons were performed using Student's *t*-test.

	Mean	*SEM*	t_{29}	*p*	Cohen's *d*
d'	0.42	0.50	4.57	<.001	0.83
c	-0.46	0.25	-9.84	<.001	-1.807
HR	-0.08	0.06	-7.11	<.001	-1.309
FAR	-0.15	0.08	-10.16	<.001	-1.86

Table B. Effect of time on the difference between probe and pre-probe trials for Experiment 1.
Statistical comparisons were performed using repeated measures ANOVA with Greenhouse-Geisser correction, where appropriate.

	F(4, 29)	*p*	η^2
d'	0.58	.675	0.02
c	5.47	<.001	0.16
HR	10.15	<.001	0.26
FAR	14.09	<.001	0.33

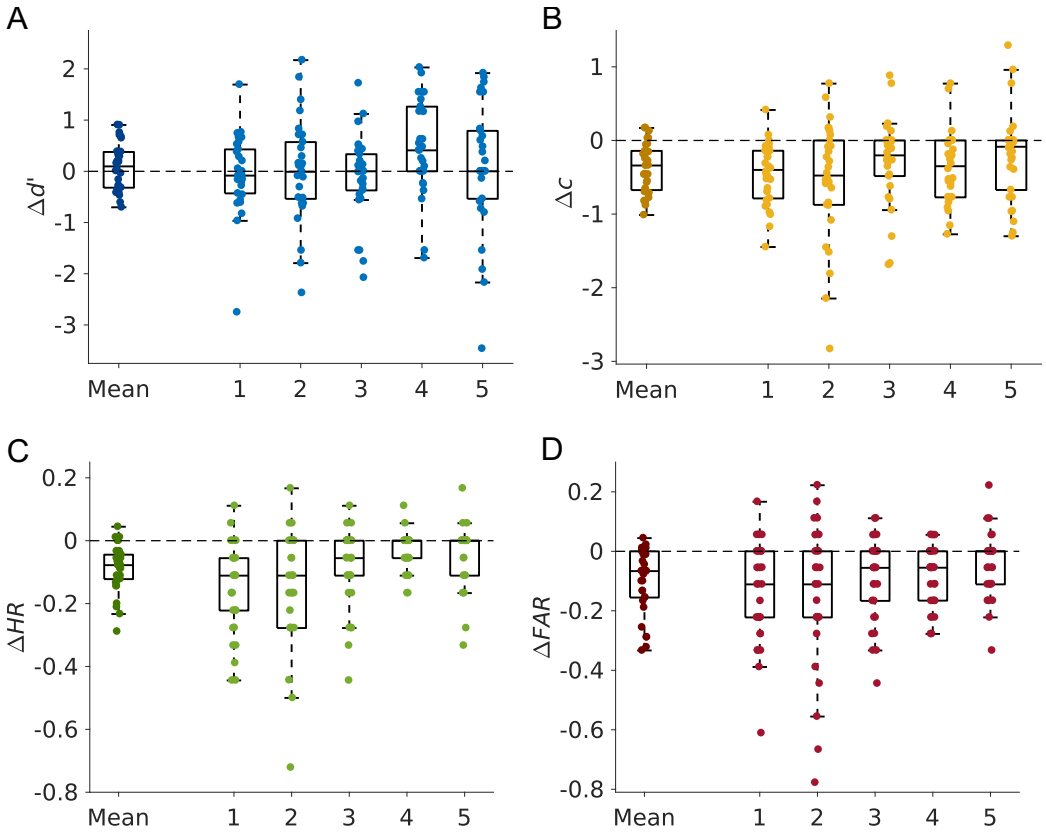

Fig C. Behavioural results, expressed as difference between probe trials and subsequent reinforced trials for Experiment 1 (*N* = 30).
Results are shown both for the mean across all five probe blocks (left) and separately for each probe block. (A) The sensitivity index *d'* was not significantly different in probe compared to reinforced trials. (B) The negative bias criterion *c*, decreased in probe blocks, indicating a reduced propensity to act on probe trials. (C), (D) Both hit rate (*HR*, C) and false alarm rate (*FAR*, D) decreased in probe blocks.

Table C. Difference between probe and post-probe trials for Experiment 1.
Statistical comparisons were performed using Student's *t*-test.

	Mean	*SEM*	t_{29}	*p*	Cohen's *d*
d'	0.08	0.46	0.90	.375	0.17
c	-0.38	0.34	-6.05	<.001	-1.11
HR	-0.09	0.08	-6.41	<.001	-1.17
FAR	-0.10	0.11	-4.89	<.001	-0.89

Table D. Effect of time on the difference between probe and post-probe trials for Experiment 1.
Statistical comparisons were performed using repeated measures ANOVA with Greenhouse-Geisser correction, where appropriate.

	F(4, 29)	*p*	η^2
d'	1.82	.130	0.06
c	1.76	.141	0.06
HR	5.83	.001	0.17
FAR	3.32	.028	0.10

1.2. Experiment 2

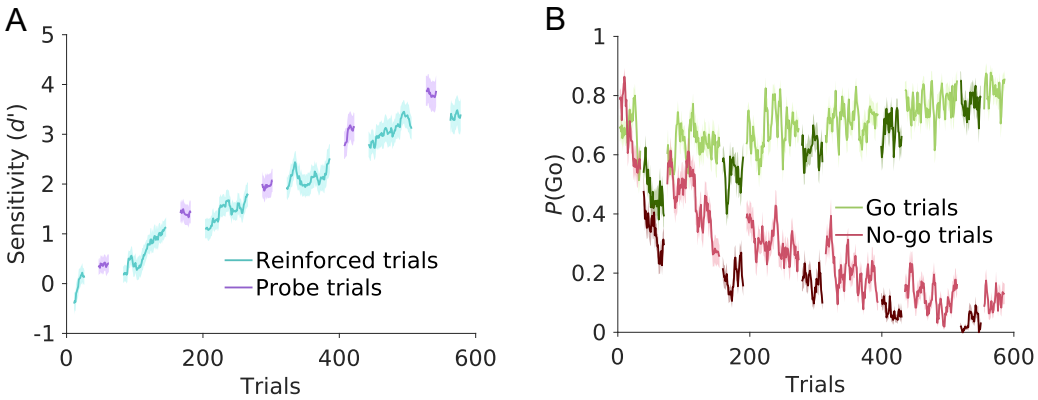

Fig D. Average learning performance for Experiment 2 (*N* = 30).
(A) Sensitivity index *d'*, separately for reinforced (cyan) and probe trials (purple). Solid lines represent mean performance, shaded areas SEM across participants. (B) Time course of go-response probabilities, *P*(Go), for go trials (green) and no-go trials (red). Darker shades of green and red illustrate probe trials. Solid lines represent mean, shaded areas SEM across participants.

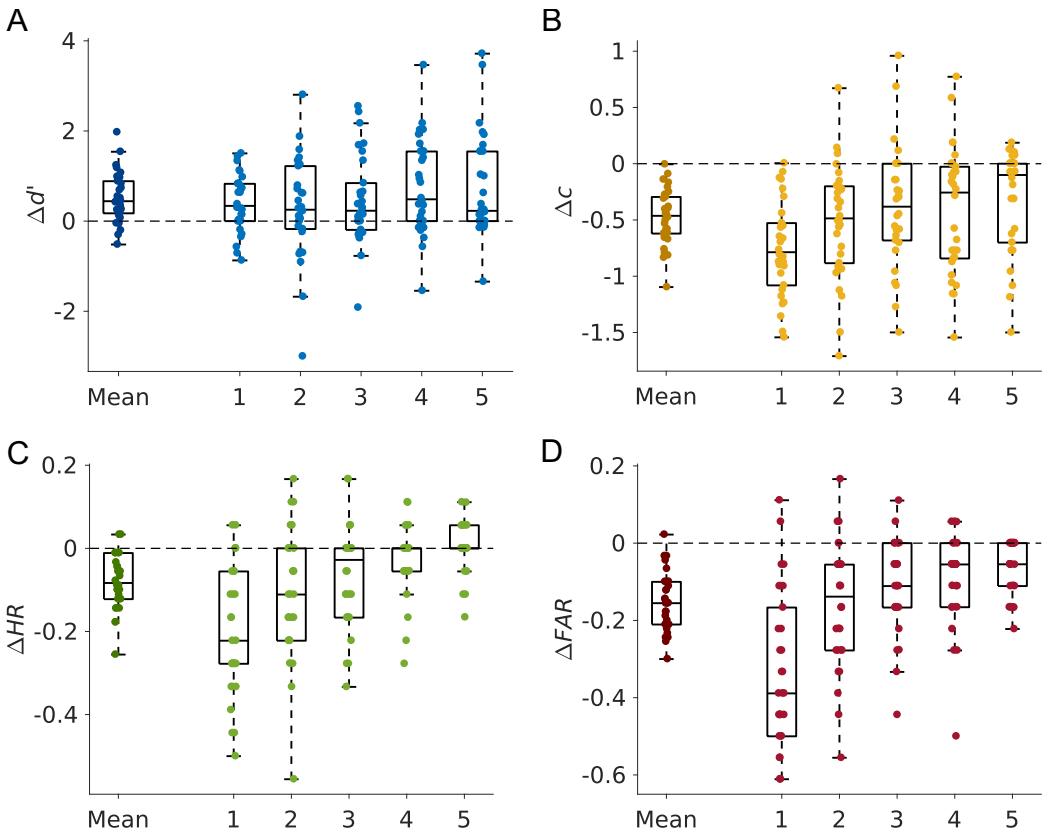

Fig E. Behavioural results, expressed as difference between probe trials and preceding reinforced trials for Experiment 2 (*N* = 30).
Results are shown both for the mean across all five probe blocks (left) and separately for each probe block. (A) The sensitivity index *d'* increased in probe compared to reinforced trials. (B) The negative bias criterion *c*, decreased in probe blocks, indicating a reduced propensity to act on probe trials. (C), (D) Both hit rate (*HR*, C) and false alarm rate (*FAR*, D) decreased in probe blocks.

Table E. Difference between probe and pre-probe trials for Experiment 2.
Statistical comparisons were performed using Student's *t*-test.

	Mean	*SEM*	t_{29}	*p*	Cohen's *d*
d'	0.52	0.55	5.19	<.001	0.95
c	-0.48	0.25	-10.61	<.001	-1.94
HR	-0.08	0.07	-6.48	<.001	-1.18
FAR	-0.15	0.08	-11.08	<.001	-2.02

Table F. Effect of time on the difference between probe and pre-probe trials for Experiment 2.
Statistical comparisons were performed using repeated measures ANOVA with Greenhouse-Geisser correction, where appropriate.

	F(4, 29)	*p*	η^2
d'	1.44	.227	0.05
c	3.72	.007	0.11
HR	13.91	<.001	0.32
FAR	18.81	<.001	0.39

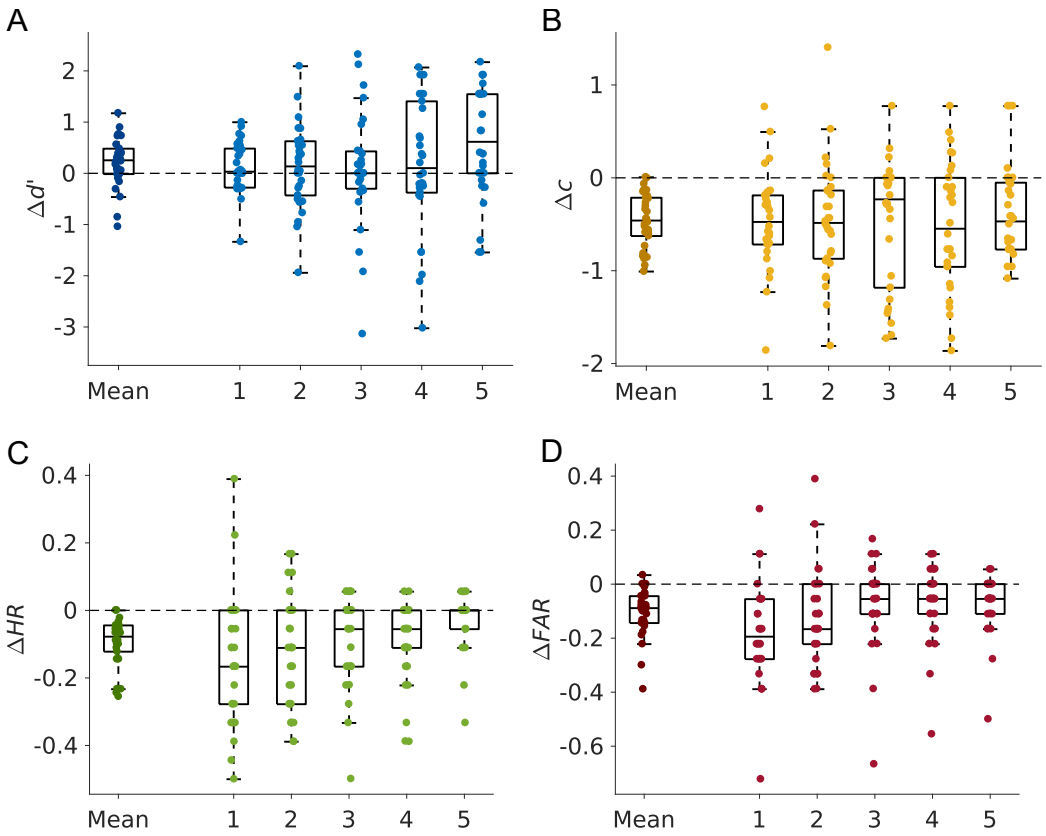

Fig F. Behavioural results, expressed as difference between probe trials and subsequent reinforced trials for Experiment 2 (*N* = 30).
Results are shown both for the mean across all five probe blocks (left) and separately for each probe block. (A) The sensitivity index *d'* increased in probe compared to reinforced trials. (B) The negative bias criterion *c*, decreased in probe blocks, indicating a reduced propensity to act on probe trials. (C), (D) Both hit rate (*HR*, C) and false alarm rate (*FAR*, D) decreased in probe blocks.

Table G. Difference between probe and post-probe trials for Experiment 2.
Statistical comparisons were performed using Student's *t*-test.

	Mean	*SEM*	t_{29}	*p*	Cohen's *d*
d'	0.22	0.48	2.47	.020	0.45
c	-0.46	0.29	-8.64	<.001	-1.58
HR	-0.09	0.07	-7.40	<.001	-1.35
FAR	-0.11	0.09	-6.47	<.001	-1.18

Table H. Effect of time on the difference between probe and post-probe trials for Experiment 2.
Statistical comparisons were performed using repeated measures ANOVA with Greenhouse-Geisser correction, where appropriate.

	F(4, 29)	*p*	η^2
d'	1.70	.155	0.06
c	0.24	.915	0.01
HR	3.13	.040	0.10
FAR	3.50	.018	0.11

The results for both experiments are very similar. The comparison of pre-probe and probe trials yield identical results for both experiments: *d'*, bias criterion, hit and false alarm rate change significantly and there is a significant effect of time for these parameters except for *d'*. The results of the comparison of post-probe and probe trials are similar for both experiments, with the exception that *d'* decreased significantly in post-probe trials in Experiment 2, but not in Experiment 1.

2. Behavioural analysis of the pooled data set

Here, we report the post hoc tests for the difference in the sensitivity *d'* between probe and pre-probe trials (Table I). Additionally, we analysed the difference between probe and post-probe trials of the pooled data (Fig G and Table J).

Table I. Post hoc tests for the difference in *d'* of probe and pre-probe trials.

Block	t_{59}	*p*	Cohen's *d*
1	4.44	<.001	0.57
2	3.03	.004	0.39
3	2.74	.008	0.35
4	4.35	<.001	0.56
5	3.43	.001	0.44

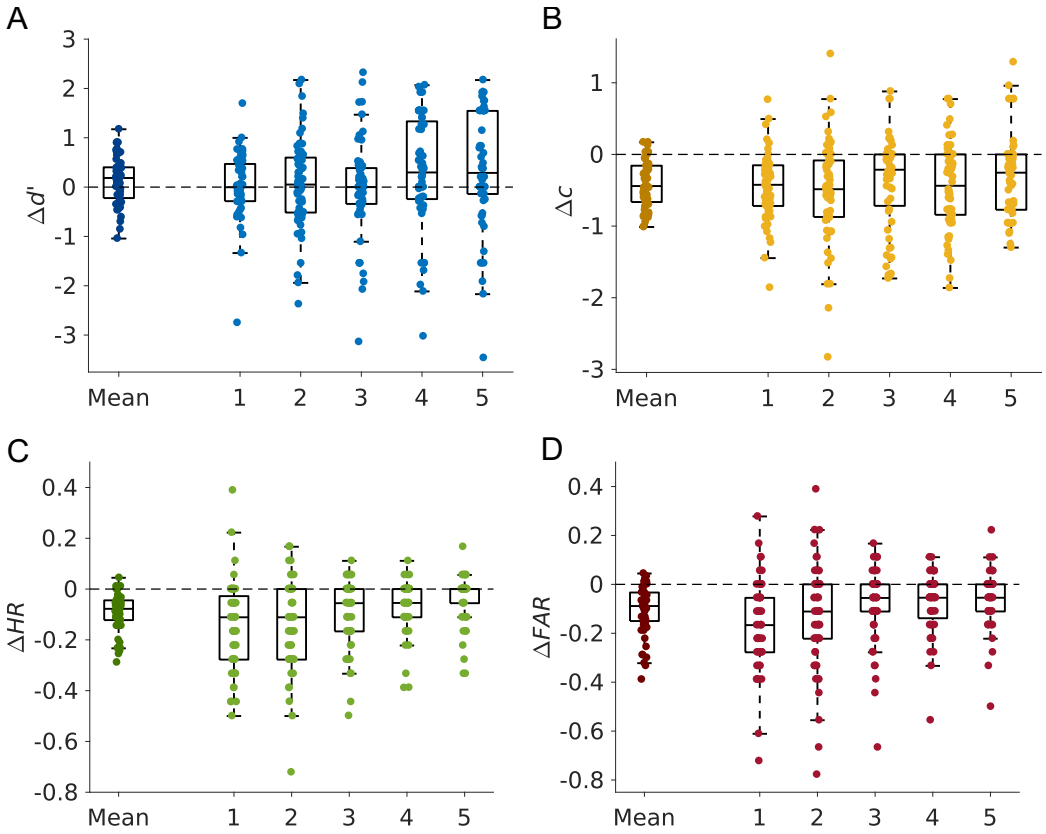

Fig G. Behavioural results, expressed as difference between probe trials and subsequent reinforced trials for all participants (*N* = 60).
Results are shown both for the mean across all five probe blocks (left) and separately for each probe block. (A) The sensitivity index *d'* increased in probe compared to reinforced trials. (B) The negative bias criterion *c*, decreased in probe blocks, indicating a reduced propensity to act on probe trials. (C), (D) Both hit rate (*HR*, C) and false alarm rate (*FAR*, D) decreased in probe blocks.

Table J. Effect of time on the difference between probe and post-probe trials for all participants.
Statistical comparisons were performed using repeated measures ANOVA with Greenhouse-Geisser correction, where appropriate.

	$F(4, 59)$	p	η^2
d'	1.99	.097	0.03
c	1.22	.304	0.02
HR	7.66	.001	0.12
FAR	6.21	.001	0.10

3. Analysis of modelling results

We compared all models with five additional learning rates evenly log-spaced between 0.01 and 0.2 to verify that that the results are not dependent on the learning rate $\alpha = 0.06$, which we chose for the main analysis (Table K and Fig H). The results of the recovery for all free parameters of the bias model are shown in this part (Fig I and Table L).

3.1. Model comparison with a set of fixed learning rates

Table K. Comparisons of baseline, temperature, bias and full model with varying learning rates *α*. Parameters were fit for all participants (N = 60), and *BIC*s (mean ± SEM) were calculated for model comparison. Learning rate α = 0.06 which was used for main analyses is bold.

α	Baseline model	Temperature model	Bias model	Full model
0.01	512.47 ± 134.65	515.60 ± 134.34	504.87 ± 129.09	509.71 ± 129.17
0.018	518.10 ± 135.23	521.07 ± 134.82	510.76 ± 129.67	515.45 ± 129.71
0.033	526.96 ± 135.96	529.89 ± 135.65	520.08 ± 130.47	524.67 ± 130.53
0.06	**539.42 ± 137.39**	**542.22 ± 137.25**	**533.25 ± 131.99**	**537.56 ± 132.10**
0.11	555.32 ± 138.50	557.98 ± 138.56	550.02 ± 133.19	554.08 ± 133.32
0.2	573.18 ± 138.01	575.43 ± 137.39	568.69 ± 133.17	572.56 ± 132.85

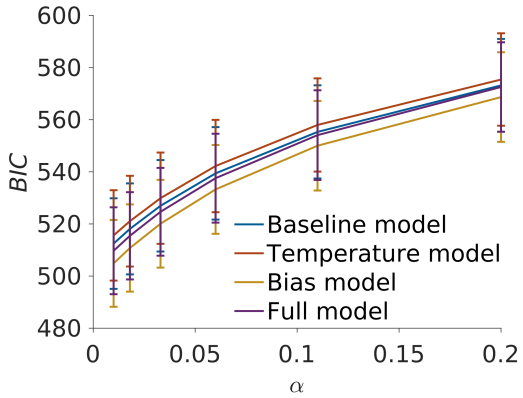

Fig H. Model comparison with different learning rates.
Comparison of mean *BIC*s for six different learning rates, error bars represent SEM. For all learning rates, the bias model provided the best fit. The full model fitted second best, followed by the baseline model. The temperature model performed the worst.

3.2. Parameter recovery for the bias model

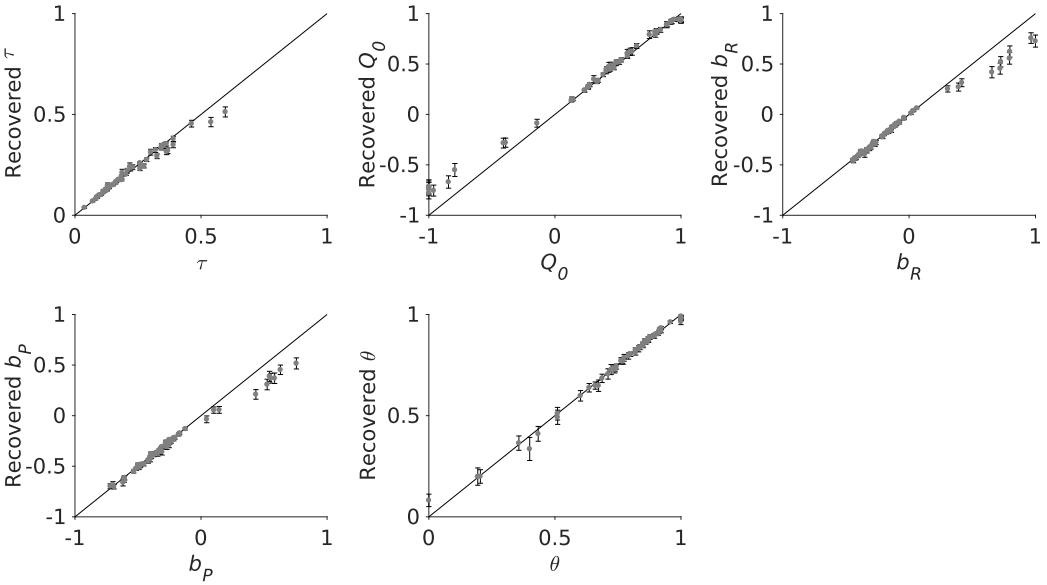

Fig I. Parameter recovery for all free parameters of the bias model.
Parameters fitted to participants' behaviour are plotted against the recovered parameters. Error bars represent 95% Cousineau-Morey confidence intervals.

Table L. Correlation coefficients for the recovery of each parameter.

Free Parameter	ρ	p
τ	0.991	<.001
Q_0	0.998	<.001
b_R	0.994	<.001
b_P	0.995	<.001
θ	0.997	<.001

3.3. Model validation

Because of the artefacts of the SDT analysis, we did not have quantitative measures for the model validation. Therefore, we plotted the go-response probabilities based on model simulations and inspected visually whether the goodness of model fit supports the *BIC* outcomes. First, we validated whether the individual parameters of the baseline model are necessary to describe the general behaviour (Fig J). Second, we compared different types of forgetting (Fig K). Third, we validated whether the temperature or bias model could reproduce the behavioural change in probe trials (Fig L).

3.3.1. Individual parameters

First, we ignored the probe trials and checked whether all parameters in the baseline model are needed to describe the empirical behaviour: Participants' go-response probabilities for both go and no-go trials started high with the probability for go trials staying high and the probability for no-go trials decreasing over time.

We started with a model containing two free parameters: a softmax temperature and a general bias. This simple model was not suitable to describe the empirical data as the go-

response probabilities start relatively low and the probabilities for go and no-go trials increased and decreased over time, respectively (*BIC* = 608.47 ± 118.63, mean ± SEM, Fig J.A). Adding an initial *Q*-value improved the fit compared to the simplest model, but was still not able to reproduce the participants' behaviour (*BIC* = 595.62 ± 119.28, Fig J.B). The same applies to a model containing of a softmax temperature, a general bias and a decay parameter (*BIC* = 591.77 ± 123.25, Fig J.C). In this model, we assumed a decay of learnt values towards zero when no go-response is performed. For the baseline model, we combined a softmax temperature, a general bias, an initial *Q*-value and a decay parameter and this model is able to reproduce the behaviour described above (*BIC* = 539.42 ± 137.39, Fig J.D).

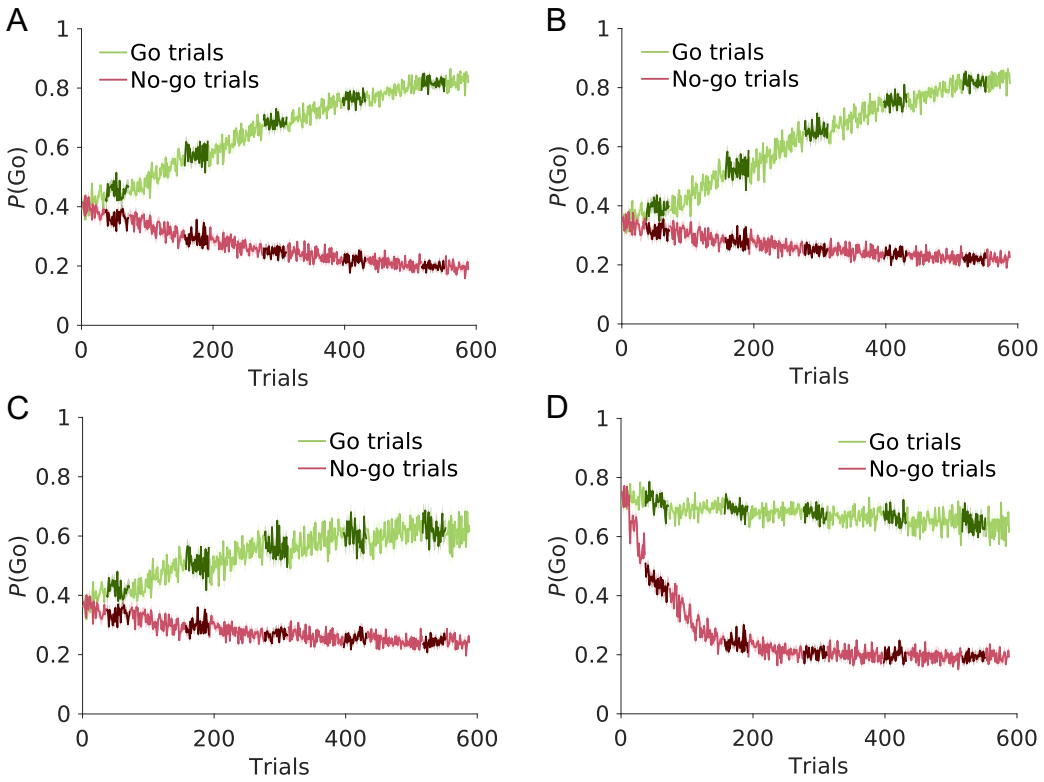

Fig J. Set up of the baseline model.
Time course of simulated go-response probabilities, *P*(Go), for go trials (green) and no-go trials (red). Darker shades of green and red illustrate probe trials. Solid lines represent mean, shaded areas SEM across simulations. Simulations are based on (A) a softmax temperature and a general bias, (B) a softmax temperature, a general bias and an initial *Q*-value, (C) a softmax temperature, a general bias and a decay parameter and (D) the complete baseline model (softmax temperature, general bias, initial *Q*-value, decay parameter).

3.3.2. Types of forgetting

There are several ways to implement a decay of option values due to forgetting. We implemented two different ways of forgetting and compared it to our baseline model. First, we set up a model in which values decay towards the initial *Q*-value. The model worsened again and due to a low initial *Q*-value, the go-response probabilities start low with the probabilities for no-go trials staying low and the probabilities for go trials increasing over time, which is not in line with the participants' behaviour (*BIC* = 547.74 ± 116.46, Fig K.A).

Another approach for the decay parameter is to implement forgetting when no feedback for the go-response is received (instead of forgetting after no-go-responses). Again, this model performed worse compared to the baseline model and could not capture the participants' behaviour (*BIC* = 600.70 ± 117.24, Fig K.B).

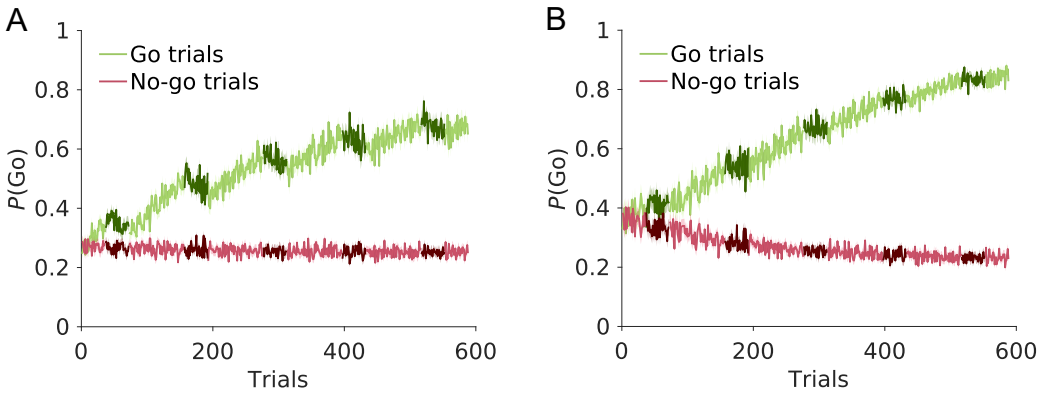

Fig K. Variations of the baseline model.
Time course of simulated go-response probabilities, *P*(Go), for go trials (green) and no-go trials (red). Darker shades of green and red illustrate probe trials. Solid lines represent mean, shaded areas SEM across simulations. Simulations are based on (A) a model with values decaying towards the initial *Q*-value instead of zero and (B) a model with values decaying when no feedback is given instead of no response decay.

3.3.3. Modelling the behaviour in probe trials

In probe trials, participants' go-response probabilities for both go and no-go trials decreased. Based on the baseline model, we now implemented two models differentiating between reinforced and probe trials; In the temperature model fitted a softmax temperature separately for each trial type. It performed worse than the baseline model and the decrease in go-response probabilities for both go and no-go trials is not comparable to the observed behaviour (*BIC* = 542.22 ± 137.25, Fig L.A). The bias model, which fitted a general bias separately for each trial type, performed better than the baseline model and looks comparable to the participants' behaviour (*BIC* = 533.25 ± 131.99, Fig L.B).

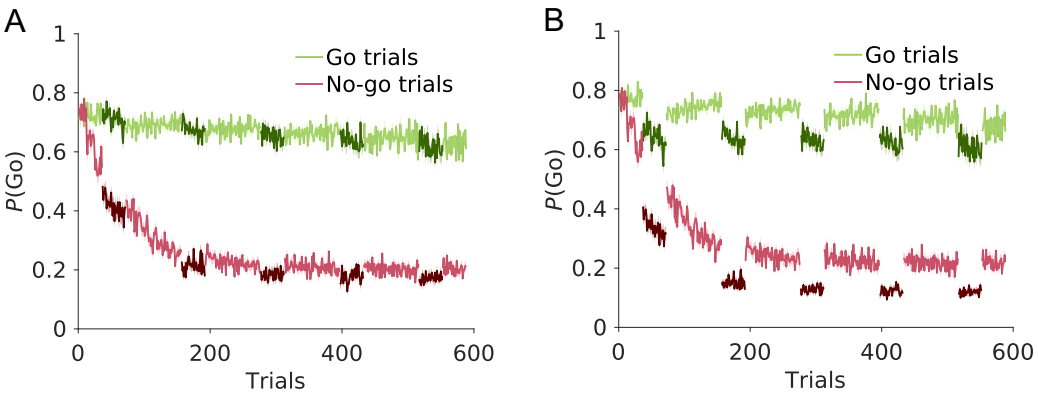

Fig L. Simulations of changed behaviour in probe trials.
Time course of simulated go-response probabilities, *P*(Go), for go trials (green) and no-go trials (red). Darker shades of green and red illustrate probe trials. Solid lines represent mean, shaded areas SEM across simulations. Simulations are based on (A) the temperature model and (B) the bias model.

4. Control analysis for reaction times

To ensure that the response window was long enough for participants to administer a go-response, we plotted the distribution of participants' reaction times (RTs) in all trials.

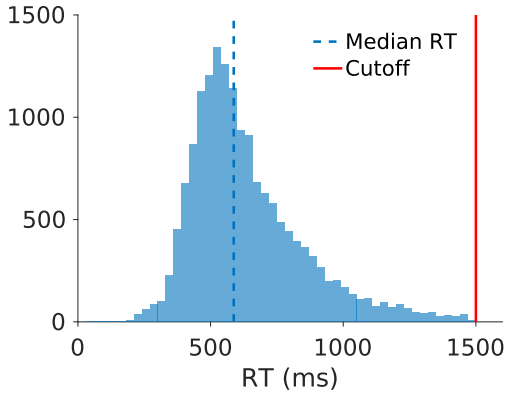

Fig M. Distribution of participants' RTs.
The cutoff of 1500 ms relative to stimulus onset marks the end of the response window.

Danksagung

An dieser Stelle möchte ich mich gerne bei allen Menschen bedanken, die diese Arbeit ermöglicht haben und mich während dieser intensiven Zeit unterstützt haben.

Zuerst möchte ich mich bei Prof. Dr. Gerhard Jocham bedanken. Vielen Dank, dass ich von Beginn an Teil Deiner neuen Arbeitsgruppe in Düsseldorf sein durfte und Du mir ein so vielseitiges und spannendes Projekt zugetraut hast. Deine Begeisterung für die Wissenschaft hat mich von Beginn an motiviert und durch Deine Betreuung und Unterstützung konnte ich sowohl auf wissenschaftlicher als auch auf persönlicher Ebene über mich hinauswachsen.

Weiterhin bedanke ich mich bei PD Dr. Jan Hirschmann. Während meiner Masterarbeit hast Du mir die Freude an der wissenschaftlichen Arbeit vermittelt. Danke, dass Du mich darüber hinaus während meiner Promotionszeit mit Deinem umfangreichen Wissen unterstützt hast.

Vielen Dank an Prof. Dr. Tobias Kalenscher, dass Du die Rolle meines Mentors übernommen hast.

Ich danke meinen Kollegen für die wunderbare und motivierende Arbeitsatmosphäre. Allen voraus bedanke ich mich bei Monja I. Froböse, Ph.D., und Eduard Ort, Ph.D., mit denen ich die gesamte Doktorandenzeit über eng zusammengearbeitet habe. Vielen Dank für eure Unterstützung sowohl bei fachlichen als auch persönlichen Fragen und für eure Ermutigung während der gesamten Zeit. Vielen Dank an Ana Antonia Dias Maile und Dr. Mani Erfanian Abdoust, dass wir unsere Doktorandenzeit gemeinsam durchlaufen haben und uns dabei stets gegenseitig unterstützen konnten. Thank you, Dr. Anna Marzecová, for your help with any issues and for offering such a precise perspective. Dankeschön an Dr. Luca F. Kaiser und Dr. Lennart Lüttgau, dass ihr mich beim Einstieg unterstützt habt und mich

mit eurer Begeisterung für die Psychologie motiviert habt. Thanks to Dr. Lina Skora and Dr. Felix Ball for your helpful comments in group meetings and during lunch. Des Weiteren bedanke ich mich bei Christiane Heil für die Hilfe bei allen organisatorischen Fragen und bei Judith Geusen, Christina Kalinicenko, Joshua Saal, Georg Schäfer, Ana Antonia Dias Maile, Paul Höchter, Marlene Hüsken, Kouta Sasaki, Helena Schmidt, Hanin Alejel und Helena El Kholy für die Hilfe bei der Datenerhebung.

Ebenso gilt mein Dank allen Kollaborationspartnern während meiner Promotionszeit, insbesondere dem gesamten Institut um Prof. Dr. Alfons Schnitzler für die Ermöglichung der Datenerhebung und die Unterstützung bei der Analyse. Vielen Dank an PD Dr. Markus Butz für Deinen unermüdlichen Einsatz am MEG und Deinen ansteckenden Optimismus.

Ich bedanke mich bei allen, die sich die Zeit genommen haben, an meinen Studien teilzunehmen, und diese Arbeit dadurch erst ermöglicht haben.

In gleicher Weise gilt mein Dank allen Kollegen und Freunden, die meine Dissertation Korrektur gelesen haben und mir mit wertvollen Tipps geholfen haben.

Zu guter Letzt möchte ich mich bei meiner Familie und meinen Freunden bedanken. Danke, dass ich mich immer auf euch verlassen kann und ihr mir abseits von fachlichen Themen Unterstützung und Ablenkung geboten habt. Besonders bedanken möchte ich mich bei Jens, Danke für Deinen Rat und Deine Geduld.

Eidesstattliche Versicherung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Düsseldorf, _____ _____