Computational Mechanisms of Social Decision Making and Learning

Jamal Esmaily

Graduate School of
Systemic Neurosciences
LMU Munich

Dissertation der Graduate School of Systemic Neurosciences der Ludwig-Maximilians-Universität München

January 2025

Supervisor:
Dr. Bahador Bahrami
Dept. of General and Experimental Psychology, LMU Munich

First Reviewer: Dr. Bahador Bahrami
Second Reviewer: Prof. Dr. Zhuanghua Shi
External Reviewer: Dr. Bernhard Spitzer

Date of Submission: Jan 15, 2025
Date of Defense: May 13, 2025

I dedicate this thesis to the “eager young minds of tomorrow”

این رساله دکتری را تقدیم می کنم به "ذهن های مشتاق جوان فردا"

Abstract

Understanding the mechanisms of decision-making and learning is fundamental to advancing cognitive neuroscience and improving collective decision-making strategies. While extensive research has elucidated aspects of individual decision-making, significant gaps remain in our understanding of the computational and neurobiological processes underlying social decision-making and how decision-making strategies are learned over time. To address these gaps, we conducted two complementary studies that leverage computational modelling, neurobiological data, and behavioural analysis.

In our first study, we focused on the neural mechanisms underlying social decision-making, specifically how collaborators align their confidence judgments. Despite the importance of confidence communication in social contexts, the computational basis for this process has remained elusive. We developed a neurobiological model supported by EEG, eye-tracking, and behavioural data to investigate confidence matching during perceptual decision-making. By combining psychophysical tasks, neural data, and computational modelling, we demonstrated how humans utilize information about a collaborator's confidence to adjust their own decisions and confidence levels, providing a robust framework for predicting and validating confidence alignment in collaborative tasks.

Studying social decision-making using methods designed for individual decision-making poses unique challenges. Social decision-making often requires a large number of subjects (N~30) and individual decision making computational models requires subjects extensive training to achieve reliable results, making interdisciplinary studies like ours particularly demanding. During this process, we noticed that the computational mechanisms of training itself are poorly understood, with many studies discarding training data as noisy or irrelevant. This gap motivated our second study, which aimed to explore how decision-making strategies are learned. To address this, we developed a reinforcement learning (RL) framework that models perceptual decision-making as a dynamic process where the decision boundary is optimized over time. Our model learns to balance the cost of waiting with external rewards, offering a computational tool to study the evolution of decision thresholds and learning dynamics.

Together, these studies provide a comprehensive exploration of both social decision-making mechanisms and the learning processes that shape decision strategies. The results offer insights into how humans align confidence in collaborative settings and how they refine decision boundaries through experience. These findings may have broad implications for real-world applications, such as improving teamwork in high-stakes environments (e.g., medical diagnostics, or financial trading) and developing training programs that enhance decision-making efficiency. By advancing our understanding of these complex processes, this research lays the groundwork for more effective individual and collaborative decision-making strategies.

Acknowledgments

I would like to express my gratitude to my supervisors for their invaluable guidance and support. I am also profoundly thankful to my lovely partner, Farinoosh, and my family for their unwavering love, patience, and encouragement throughout this journey.

Table of Contents

1 General Introduction

Decision-making is a critical cognitive process that involves choosing between different courses of action based on their potential outcomes. It is a complex process that encompasses everything from simple daily choices to high-stakes decisions in business and politics. Social decision-making adds another layer of complexity, as it involves navigating interactions with others, understanding their intentions, and predicting their behaviour. This type of decision-making is essential in environments where cooperation, competition, and communication play a crucial role, such as in team projects, negotiations, or any group-based activities. Similarly, learning is a mechanism by which we acquire new knowledge or refine existing knowledge and skills through experience. It enables us to adapt to new situations, solve problems, and make better decisions based on past experiences.

What happens in the brain during these seemingly trivial and everyday phenomena has been a key focus of cognitive neuroscience (Gold & Shadlen, 2007; Schultz et al., 1997; Shadlen & Kiani, 2013). We have achieved great successes in understanding brain mechanisms in these cognitive processes (Ratcliff, 1978). Despite these advances, there remains a substantial gap in our comprehension of the brain mechanisms behind real-life decision-making and learning (Huettel, 2010). However, understanding such complex systems often benefits from an initial focus on simplified scenarios, allowing for a clearer grasp of the fundamental processes before scaling to more intricate situations (Summerfield & Miller, 2023). This incremental approach is the cornerstone of neuroscience research. Consistent with this methodology, this thesis examines simplified scenarios of social decision-making and learning in the hope that these findings will enhance our understanding of the broader mechanisms involved.

1.1 Perceptual Decision Making

Perceptual decision making is a fundamental cognitive process by which individuals interpret and act upon sensory information from the environment (Gold & Shadlen, 2007; T. Hanks & Summerfield, 2017). This process underlies our ability to make judgments about the world around us, from the simple (e.g., determining the colour of a traffic light) to the complex (e.g., interpreting a facial expression). The perceptual decision-making process typically begins with sensory input—data collected by our sensory organs, like eyes and ears, from the external environment. This sensory information is then processed by the brain's relevant sensory areas, which decode and analyse the data to extract meaningful patterns and details (Shadlen & Kiani, 2013).

One of the long-term objectives of cognitive neuroscience is to elucidate the neural mechanisms that underlie decision formation (Beck et al., 2008; Roitman & Shadlen, 2002). The most agreed upon mechanism of perceptual decision making suggests that perceptual decision making involves the integration of sensory data with prior knowledge (Ratcliff, 1978; Ratcliff & McKoon, 2008). This process may stop once a predefined threshold is reached (Ratcliff, 1978; Ratcliff et al., 2006, 2009). When the quality of signal (stimulus) is high, we expect that the integration process finished sooner. In contrast, when the sensory signal is noisier, the time that is required to make as accurate decision increases. For example, imagine a scenario in which you are driving a car in foggy road conditions. The fog significantly reduces visibility, making the sensory input—what you can see of the road and its surroundings—much noisier and less reliable. In such a scenario, your brain requires more time to integrate this unclear sensory data mainly because information per unit of time is low. This additional processing time is necessary to reach the threshold of certainty needed to make safe driving decisions, such as determining the safe speed or identifying the right moment to make a turn. This processing time could be significantly reduced in a clear, sunny day in which the sensory signal is stronger. The relationship between accuracy and processing time may suggest that subjects accumulate information to improve performance.

One of the simplest forms of decision making is decision between two (rather than more) options (Britten et al., 1992; Shadlen & Newsome, 2001). Studying these simple decisions may help us to understand human (or animal) decision making better and generalize our understanding to more complex and realistic scenarios (T. Hanks & Summerfield, 2017). One popular class of these simple decision are Two-alternative forced choice (2AFC) tasks (Hautus et al., 2011; Roitman & Shadlen, 2002). They are commonly used in experimental neuroscience to measure an individual's ability to discriminate between two different stimuli (Hautus et al., 2011, 2011). There are numerous instances of these tasks in different modalities (Chancel & Ehrsson, 2020; Ganea, 2021; García-Pérez & Alcalá-Quintana, 2020; Morris, 2022; Simen et al., 2009).

Britten et al (1992) introduced a Random Dot Motion (RDM) discrimination task, a simple and effective 2AFC task. Generally, subject is required to decide whether the majority of dots are moving toward option 1 (here: up) or option 2 (here: down). The portion of the dots that moves in a "same" direction is called coherence (correlation) level. The higher this value, the easier the task. We normally expect to see higher accuracy and lower RT with increasing coherence level. This task has been instrumental in understanding basic mechanism of perceptual decision making such hierarchical, discrete, social, and sequential decision making (Bang et al., 2020; Purcell & Kiani, 2016b, 2016a; Resulaj et al., 2009).

1.1.1 Brain and Perceptual Decision Making

The simplicity and effectiveness of the RDM task encouraged many researchers to delve deeper into how this task parameters affects neural activities, EEG signals and eye data (Kelly & O'Connell, 2013; O'Connell et al., 2012; Roitman & Shadlen, 2002; Urai et al., 2017). In one the earliest attempts, Britten et al (1992), recorded from neurons in extra striate cortex (areas MT and MST) while the monkey did the RDM task. Given that RDM is a "visual" discrimination task and also the rich history of vision neuroscience (Haan & Cowey, 2011; Mishkin & Ungerleider, 1983), MT (medial temporal) opted as a proper candidate for encoding RDM task information. They indeed showed that MT is actively participating in the discrimination of motion based on coherence levels. MT activities were found to be predictive of subject decision-making when the subject was indeed instructed to make a decision. However, during passive decision-making, MT continued to encode sensory information, suggesting that decision-making processes themselves are not encoded within MT. Furthermore, there was no indication that MT is involved in evidence accumulation. The MT neurons simply showed an initial stimulus dependant raise in the beginning of the stimulus presentations and kept their activities persistence afterwards. This indicates a distinct role for MT in sensory processing rather than active decision-making or the integration of decision-relevant information. These results implied that there may be other areas in the brain that indeed "accumulate" the evidence that has been precisely encoded in MT.

To this end, Shadlen and Newsome (Roitman & Shadlen, 2002; Shadlen & Newsome, 2001) recoded neural activities in the lateral intraparietal area (LIP). Many neurons in the lateral intraparietal area (LIP) respond to visual stimuli that are the target of a planned saccadic eye movement (Colby et al., 1996; Gnadt & Andersen, 1988; Platt & Glimcher, 1997). When the direction of random-dot motion instructs the choice of a target for a saccade, LIP activity modulates in a way that predicts the monkey's eye movement response. The gradual evolution of activity during motion viewing and its dependence on the difficulty of the discrimination suggests that neurons in LIP may represent the accumulation of visual information about motion leading to the formation of the monkey's decision. This result was hailed as a neural signature of evidence accumulation, thereby strengthening the concept of evidence accumulation in perceptual decision making literature. This notion, however, was challenged by some other studies (Katz et al., 2016; Latimer et al., 2015; Stine et al., 2020). They argue that ramping activity observed at the neural population level does not necessarily imply evidence accumulation within individual neurons. They showed that a population of neurons, each exhibiting *pulse-like* activity, can collectively produce a ramping pattern of evidence accumulation. Therefore, a single neuron does not necessarily accumulate evidence on its own.

Besides the electrophysiology recording, other studies used EEG and eye tracking to understand the decision making processes. Kelly et al (Kelly & O'Connell, 2013; O'Connell et al., 2012) located an important area of the brain responsible for perceptual decision making and evidence accumulation. The component identified by this group, Centro Parietal Positivity (CPP), shows a significant relationship to coherence level. Interestingly, they showed that EEG signals indeed show a ramping activity proportional to coherence level, signifying a signature for evidence accumulation. They also showed that these pattern of activity is indeed decision dependant and has not modulation to coherence level once subjects is passively viewing the stimulus (O'Connell et al., 2012). Other studies also provided evidence that there is correlation between CPP and pupil and microsaccades (van Kempen et al., 2019). Together, these studies showed that, evidence accumulation is not limited to LIP neuron and could be detected via other measurement too.

1.1.2 Confidence in Perceptual Decision Making

An integral aspect of decision making process is the confidence with which these decisions are made (Kiani et al., 2014; O'Connell et al., 2018). Perceptual decision confidence refers to the subjective probability that one's choice or judgment is correct, reflecting a meta-cognitive assessment of the decision-making process itself. Confidence plays a critical role in behaviours and has significant implications for learning, reasoning, and error correction.

Unlike conventional objective measures of decision-making such as accuracy and reaction time, decision confidence is a subjective measure. Researchers have made numerous attempts to model and explain decision confidence using objective measures such as task difficulty, accuracy, and reaction time (Kepecs et al., 2008; Kiani et al., 2014; Zylberberg et al., 2016). While these measures offer some explanatory power, the subjective nature of confidence adds layers of complexity. This complexity indicates that confidence formation likely incorporates additional cognitive and perceptual factors, making it a distinct and multifaceted component of decision-making (Hagura et al., 2023; Lee & Daunizeau, 2021; Turner et al., 2021). As a result, ongoing research continues to explore the nuanced mechanisms behind confidence, aiming to uncover how subjective certainty emerges and influences behaviour (Esmaily et al., 2024).

Several data modalities were used to enhance our understanding of decision confidence (Balsdon et al., 2021; Gherman & Philiastides, 2018; Vafaei Shooshtari et al., 2019). Kiani et al (Kiani & Shadlen, 2009) investigated how confidence in perceptual decision-making is represented in the brain, particularly within the parietal cortex of rhesus monkey. They trained monkeys to make perceptual decisions about motion direction and allowed them to opt out for a guaranteed smaller reward if they were uncertain. The study found that neural activity in the parietal cortex not only reflected the monkeys' decisions but also their uncertainty levels, suggesting that these neurons encode both the choice and the certainty associated with that choice. This insight contributes to understanding the neural basis of confidence and decision-making processes. Confidence was also investigated via other data modalities such as EEG and eye tracking as well (Pisauro et al., 2017; Vafaei Shooshtari et al., 2019). Microsaccades, the small unvoluntary movement of gaze, were also linked to perceptual confidence as well (Loughnane et al., 2018; van Kempen et al., 2019). In another important study (Urai et al., 2017) showed that pupil data can encode decision confidence after feedback and in inter trial intervals. Given their non-invasive nature, these EEG and pupil components could help us to understand formation of perceptual decision confidence in humans.

1.1.3 Computational Models of Decision Making

Decision-making processes have long fascinated computational neuroscience (O'Connell et al., 2018). Some approaches aim to provide descriptive explanations of these processes, with the Drift Diffusion Model (DDM) being a well-known example (Ratcliff, 1978; Ratcliff et al., 2009). Other approaches focus on mechanistic explanations, incorporating brain and neural mechanisms to model decision-making in a manner that mirrors how the brain likely operates (Wang, 2002, 2008; Wong & Wang, 2006). These models strive to replicate the underlying neural activity and interactions, offering insights into the biological basis of decision-making. In the following, we will describe DDM and one simple mechanistic model of decision making that is based on attractor neural network. Both models are used in this thesis. Note that these models are not exclusive; there are also other competing models that attempt to explain the decision-making process (Dubreuil et al., 2022; Latimer et al., 2015; Mastrogiuseppe & Ostojic, 2018; Stine et al., 2020). The models mentioned in this thesis are not, by any means, the definitive or only models of how the brain makes decisions. Yet, of course, enhance our understanding of decision making and learning processes in the brain.

1.1.3.1 Drift Diffusion Model

The Drift Diffusion Model (DDM) is a mathematical model in the family of sequential sampling models that used to explain decision-making processes, particularly in two-choice tasks (Ratcliff, 1978). It describes how information is accumulated over time until a decision threshold is reached. The key elements of the DDM include drift rate, decision threshold, and starting point. It defines as follows:

$$dv(t) = \mu\, dt + \eta\, dW(t) \qquad (1)$$

where $v(t)$ is the decision variable at time t, μ is the drift rate, representing the average rate of evidence accumulation. η is the noise parameter usually set to 1 and $dW(t)$ is a Wiener process (standard Brownian motion). The decision is made when $v(t)$ reaches one of the boundaries: $v(t) = B$, where B - sometimes also denoted by a - is the decision threshold. The initial value $v(0) = z$ (usually set to $z = B/2$) representing the starting point of the decision process.

The characteristics of drift term $\mu\, dt$ and diffusion term $\eta\, dW(t)$ determines the stochastic dynamics of v over time. In the general form of Fokker Planck equation both drift and diffusion term are partial differential equations that could change over both time and space. DDM utilized the 1D version of Fokker Planck equation while both drift and diffusion term are time-independent (scaler rather than PDEs). Although this treatment may be deemed as a massive simplification, yet DDM has been extremely successful in explaining various decision making behaviors (T. Hanks & Summerfield, 2017; Kiani et al., 2014; Kiani & Shadlen, 2009; O'Connell et al., 2018; Purcell & Kiani, 2016b, 2016a).

In the decision making scenario, the stochastic process $v(t)$ is terminated when the boundaries is hit. Having gone through "first passage" calculation of $v(t)$ to B and also setting $\mu = Kc$ where K is drift coefficient and c is task difficulty, we can define the response time (RT) and accuracy as follows: RT is characterized as the first time $v(t) = B$. Meanwhile, accuracy is defined as the probability that $v(t) = B$ correctly reflects the true decision outcome. This formulation integrates both the dynamics of decision variables and the influence of task difficulty on the decision-making process. Accuracy and RT could be solved analytically as follows (Ratcliff, 1978):

$$Accuracy(c) \ = \ 1 \, / \, (1 \ + \ e^{-2KcB}) \qquad (2)$$

$$RT(c) = \frac{B}{KC} tanh(KcB) \qquad (3)$$

Parameters of DDM (K, B) have a measurable effect on the output accuracy and reaction time. Higher values of K lead to faster and more accurate decision-making. Higher thresholds also increase the accuracy and RT. The profile of effect of each parameter one accuracy and RT has been nicely shown in (Palmer et al., 2005). By adjusting the drift rate, decision threshold, and starting point, the DDM can model various types of decision-making behaviours, capturing both the accuracy and reaction time distributions observed in empirical data.

1.1.3.2 Neural Attractor Model

DDM model is a descriptive model that is usually agnostic about neural dynamics and mechanisms (Ratcliff & McKoon, 2008). While it excels in illustrating decision-making processes at a cognitive level, it does not delve into the detailed neuronal dynamics that govern these processes. In contrast, neurobiological models fill this gap by providing a deeper understanding of the neuronal activity. This comprehensive detailing of neural mechanisms establishes a foundational bridge from the microscale to the macroscale of cognitive phenomena, setting the stage for more integrated models in computational neuroscience (O'Connell et al., 2018; Wong & Wang, 2006).

It all started with The Hodgkin-Huxley (HH) model (Hodgkin & Huxley, 1952), developed in 1952 and honoured with a Nobel Prize, that significantly sparked interest in neural dynamics. This model provided a mathematical framework for understanding how neurons generate and transmit electrical signals through action potentials. By meticulously describing the ionic mechanisms that underlie the electrical activity of neurons, specifically through the dynamic properties of sodium and potassium ion channels, it laid the foundation for the field of computational neuroscience. This model not only deepened our understanding of the physiological basis of neural activity but also influenced the development of various applications ranging from drug discovery to the design of neural prosthetics and the simulation of neural networks, underscoring its profound impact on both theoretical and applied neuroscience.

The Hodgkin-Huxley model employs four-dimensional nonlinear differential equations to capture the dynamics of ion channels and changes in membrane voltage over time (Hodgkin & Huxley, 1952). The equations illustrate how the conductance of potassium and sodium ions varies with both membrane voltage and time, influencing ion flow across the membrane and thereby affecting the neuron's electrical state. The four-dimensional nature of the Hodgkin-Huxley model, however, complicates its scalability. As a result, researchers have sought to develop approximated and reduced versions of the model, while preserving its essential properties, such as the ability to produce specific spike patterns like tonic and phasic firing. Prominent examples of these simplified models are Izhikevich (Izhikevich, 2003) and LIF models (Burkitt, 2006). The Leaky Integrate-and-Fire (LIF) model is one of the most popular reduced versions of the Hodgkin-Huxley model and continues to be widely used (Gerstner & Kistler, 2002). This one-dimensional model forms the foundation of many large-scale spiking neural networks (Gerstner et al., 2014; Gerstner & Kistler, 2002).

Large scale spiking neural network has found its way into perceptual decision making as well (Wang, 2008; Wimmer et al., 2015, 2016; Wong & Wang, 2006). The most notable one is 2002 model of Wang (Wang, 2002). The paper introduces a biophysically realistic model that uses excitatory and inhibitory interactions within networks of spiking neurons to support and simulate decision-making dynamics. Wang's model specifically illustrates how neural circuits can maintain decision-related information over time through synaptic reverberation, even in the absence of continuous input. This framework helps to understand the neural basis of decision-making by showing how cortical circuits may encode and manipulate decision variables, providing a link between neural activity patterns and cognitive functions related to perceptual decisions.

This insightful model by Wang, although providing deep understanding, proved challenging to interpret and directly correlate with behavioural data due to its complexity and numerous free parameters (Wong & Wang, 2006). In response to these challenges, subsequent research by Wong & Wang (2006) utilized mean field theory alongside a set of plausible approximations to simplify Wang's model into a more manageable form. This resulted in a reduced model described by only two nonlinear ordinary differential equations. Specifically, the model simulates the average firing rates of two neural populations that are crucial in the accumulation of information during perceptual decision-making tasks. When inputs proportional to the stimulus coherence levels are introduced to the network, a competitive interaction emerges between two units, each representing alternative choices. This competitive race continues until the firing rates of one of the units reach a high-firing-rate attractor state, at which point the decision favoured by that unit is selected. This streamlined model not only simplifies the interpretation and fitting of behavioural data but also retains the core dynamics essential for understanding decision-making processes in neural circuits. The detail of this model is described in project one (Esmaily et al., 2023).

1.1.3.3 Computational Models of Decision Confidence

The basic models used for decision-making typically focus on modelling reaction time and accuracy but do not include a measure of confidence (Ratcliff & McKoon, 2008). To incorporate decision confidence, these models require modification. The common approach in the majority of studies that attempt to model confidence is to calculate the distance between their measure of "evidence" and the decision boundary (Kepecs et al., 2008; Lee et al., 2023; Wei & Wang, 2015). The concept of evidence varies across different frameworks (firing rate, one (Kepecs et al., 2008) vs accumulated samples (Lee et al., 2023) from a normal distribution), each with its own strengths and limitations. The notion of a decision threshold, however, remains largely consistent across all models, usually represented as a scalar predefined value. The literature on confidence computation can generally be divided into two categories: statistical-based models and neural (brain-plausible) models. Given that the nature of confidence is probabilistic, it is not surprising that confidence has been more extensively modelled within statistical and probabilistic frameworks (Zylberberg et al., 2012). These models typically leverage principles of (Bayesian) statistics to quantify and predict confidence levels in decision-making scenarios (Kepecs & Mainen, 2012; Pouget et al., 2016).

In one of the early efforts, Kepects et al. introduced a mechanism for computing confidence (Kepecs et al., 2008). They assume a normal distribution for the stimulus evidence (s, the mean is determined by stimulus strength) and the boundary (b, mean is 0 and could change if there is a bias). In each trial, one sample would be randomly drawn from each distribution (s_i and b_i). A choice is then calculated by comparing the two samples ($s_i < b_i$), and a confidence value is estimated by calculating the distance between them $|s_i - b_i|$. If the stimulus has a large mean (indicating a very clear stimulus), then, on average, the distance to b_i would also be large. Conversely, with a very noisy stimulus, the scenario changes; the random sample from s_i is likely to be very close to b_i, resulting in a smaller distance and thus a lower confidence scenario. Given that s_i and bi are stochastic, this framework offers a stochastic measurement of confidence.

In order to incorporate decision confidence, sequential sampling models typically use a framework called “race” model. In this model, each choice has an accumulator and each race to reach a predefined decision threshold. Confidence computation, however, has a similar approach where they compare accumulated evidence to the decision threshold. If the comparison is made at the moment of decision, then it effectively involves comparing the evidence for the less favoured option (the "loser accumulator") with the decision threshold (Ratcliff & Starns, 2009, 2013). This occurs because the winning accumulator has already reached the decision threshold, indicating that a decision has been made in its favour. Naturally, when the "loser accumulator" accumulates significantly less evidence, the gap to the decision threshold is greater. According to these models, a larger gap suggests higher confidence in the decision. These models also implicitly account for the effect of response time (RT). A higher RT allows the loser accumulator to gather more evidence, thus reducing the distance to the fixed threshold and implying lower confidence in the decision. However, the explicit relationship between RT and confidence is not clearly defined in these models. Most importantly, these models do not treat RT as a causal element in the formation of confidence, suggesting that while RT influences the dynamics of decision-making, it is not seen as directly shaping confidence.

Kiani et al (Kiani et al., 2014) tried to identify a formal and explicit relationship of RT and confidence. They proposed a model where confidence was determined by both decision accuracy and, crucially, RT. In their model, RT appeared on the right-hand side of the equation, suggesting a causal relationship with confidence. To strengthen this view, they conducted a psychophysical experiment where the stimulus duration was deliberately extended while the amount of evidence remained constant. They observed a decrease in confidence in the prolonged version of the task, suggesting that RT may play a causal role in the formation of confidence. Note that other models of confidence may not predict lower confidence in this task.

Neural models also follow a similar approach in modelling decision confidence (Wei & Wang, 2015). In these frameworks, the firing rate of neurons is typically considered to encode "evidence" for decision-making. Confidence in these models can be computed by taking the difference (or some nonlinear function of difference) between the firing rate of the less favoured (loser) population and a designated decision threshold. The greater this difference, the higher the confidence. This method effectively uses the magnitude of neural activity as a proxy for the strength of evidence supporting a decision, linking neural responses directly to confidence formation.

1.2 Social decision making

Social perceptual decision-making investigates how groups of individuals integrate sensory information to make collective decisions (Bahrami et al., 2010, 2012; Mahmoodi et al., 2022). The brain's ability to integrate multiple streams of sensory input, assess the reliability of information, consider the perspectives of others, and reach a consensus is crucial for many aspects of social behaviour and cooperation. This field of study bridges aspects of cognitive neuroscience, social psychology, and neurobiology to understand how the brain supports interactive decision-making among individuals. The big promise in social neuroscience is that by mapping the pathways and mechanisms through which our brains undertake joint decision-making, neuroscience not only elucidates fundamental aspects of human cognition but also provides insights into the nature of social interactions and their impact on individual and group behaviour (Mojzisch & Krug, 2008).

Again, for simplicity, many studies have focused on understanding two-person (dyadic) or paired decision-making (Bahrami et al., 2010; Najar et al., 2020). This represents the most basic form of social decision-making, allowing researchers to better control extraneous variables compared to larger groups. In the meantime, other studies attempt to scale up the social setting by increasing the group size gradually, making the environment progressively more reflective of real-life situations (Barrera-Lemarchand et al., 2024; Navajas et al., 2018). The dyadic studies offer precise control and detailed mechanistic insights and larger group studies, though less controllable, provide a more realistic reflection of complex social interactions in natural settings. This thesis focuses on dyadic setting.

A large portion of the literature dedicated to exploring the neural mechanisms that underpin these collective processes (Arabadzhiyska et al., 2022; Moore et al., 2021; Y. Shen & Zhou, 2021). Neural investigations into joint perceptual social decision-making are motivated by the need to understand how humans, as inherently social beings, process information not in isolation but in concert with others (Deaner et al., 2005; Klein et al., 2009; Shepherd et al., 2006). Key to this research is the study of specific neural circuits and regions involved in decision-making, such as the prefrontal cortex, which is known for its role in complex cognitive behaviours including planning, reasoning, and social interaction (Labutina et al., 2024; Rilling & Sanfey, 2011; Tremblay et al., 2017).

Many investigations have sought to uncover the underlying neural mechanisms. Techniques such as EEG, eye-tracking, and neural imaging have been pivotal in these explorations (Konovalov & Ruff, 2022; Moore et al., 2021; Rojas et al., 2020; J. Shen et al., 2022; Valsangiacomo, 2023). Brain synchronization, in particular, has emerged as a core area of interest within studies of social interactions and decision-making (Hasson et al., 2012; Luft et al., 2022; Mukamel et al., 2005). This phenomenon, where neural activities in different regions of the brain or between individuals temporally align, creating coordinated brain wave patterns, is believed to be critical for various cognitive processes including perception, attention, memory, and social interactions. It is deemed to facilitate efficient communication between brain regions, allowing complex cognitive functions to be executed smoothly and effectively.

In social contexts, brain synchronization between individuals is often examined using hyperscanning techniques, which reveal how humans align their neural activities during activities such as communication, empathy, and cooperative tasks. This area has gained considerable attention; however, it is not without controversy. Some research has raised serious doubts about the results and implications of brain synchronization studies, questioning the reliability and interpretability of these findings (Nam et al., 2020; Shadlen & Movshon, 1999).

The primary social phenomenon this thesis focuses on is confidence matching (Bang et al., 2017). In 2017, Bang et al. (Bang et al., 2017) explored how pairs of individuals in group decision-making settings adapt their expressed confidence levels to align with each other, a process termed "confidence matching." This heuristic strategy involves both parties matching their levels of certainty and uncertainty. The robustness of their findings was demonstrated by applying various tasks across different demographics (UK, Iran), with confidence matching observed universally

1.2.1 Computational Approaches in Social Decision Making

Computational approaches have primarily focused on adapting individual computational frameworks to social settings. Many studies have utilized the Drift Diffusion Model (DDM) to compare the fitted parameters between isolated and social contexts. Although these studies provide valuable insights, they often lack a clear understanding of the mechanisms by which parameters in isolated settings transform under social influences. Some research has advanced further by linking individual models to simulate the impact of one person's decisions on another's parameters (Tump et al., 2020, 2022, 2024). For instance, one study (Tump et al., 2020) connected a person's decision to the drift rate of another, effectively making one person's decision a function of another's behaviour. However, due to the vast number of possible parameter combinations, it is typically impractical to test every potential pair to determine the most data-supported combination. Consequently, researchers often rely on plausible intuitions about the connections and combinations to test their assumptions against the data. Although this approach is not complete, it significantly enhances our understanding of the computational mechanisms underlying social decision-making especially in the larger group sizes.

Although a diverse array of computational models for social decision-making has been proposed with varying degrees of success, these models are often descriptive and fall short in explaining the underlying neural mechanisms driving these processes. To address this gap, in this thesis, we introduce a brain-plausible computational model based on attractor neural networks. This model is designed to capture the communication of uncertainty in social decision-making processes. Our proposed model may provides mechanistic insights into how uncertainty is represented and shared between agents during social interactions.

1.3 Learning

Beside decision making, understanding how learning occurs is a fundamental aspect of cognitive science. Learning encompasses the biological processes by which new information is acquired, processed, and solidified within the brain. It involves the alteration of neural circuits in response to experiences, a phenomenon that can lead to lasting changes in behaviour.

In neuroscience, the study of learning is intimately connected to the principles of reinforcement learning (RL) and the role of dopamine as a neuromodulator (O'Doherty et al., 2003; Schultz et al., 1997). Reinforcement learning, a key concept in both machine learning and cognitive neuroscience, involves learning to make decisions based on the rewards or punishments received from previous actions (Collins & Cockburn, 2020; Eckstein et al., 2021; Subramanian et al., 2022). In the brain, dopamine is central to the RL process, acting as a signal for reward anticipation and influencing the synaptic plasticity that underlies learning (Schultz et al., 1997).

Dopamine neurons, primarily located in the midbrain areas such as the ventral tegmental area (VTA) and substantia nigra, are activated in response to rewarding stimuli or events that are better than expected (Garritsen et al., 2023; Hegarty et al., 2013). This activation releases dopamine in target areas like the striatum and prefrontal cortex, which are crucial for decision-making and learning. The dopamine signal helps to reinforce behaviours that lead to rewards through a process known as synaptic potentiation. This mechanism enhances the connections between neurons that are active during successful actions, making it more likely that these actions will be repeated in the future.

The interplay between dopamine signalling and synaptic plasticity forms the neural basis for learning behaviours that maximize rewards, a direct parallel to the computational models used in RL algorithms. Understanding how dopamine influences learning and decision-making not only sheds light on fundamental brain functions but also has implications for addressing neuropsychiatric disorders such as addiction and schizophrenia, where dopamine signalling is often disrupted. Thus, the neuroscience of learning, through the lens of RL and dopamine, offers a convergence of theory, experimental science, and clinical relevance (Garritsen et al., 2023).

1.3.1 Reinforcement Learning

Reinforcement learning (RL) is a powerful computational framework that teaches agents to make a sequence of decisions by interacting with an environment in order to maximize a notion of cumulative reward (Sutton & Barto, 2018). This framework has been utilized in neuroscience studies to explain the learning process in humans and animals. Neuroscientific research has identified correlations between neural activity and the computational principles of RL suggesting that the brain may also utilize RL as a model for learning (O'Doherty et al., 2003; Schultz et al., 1997). The three fundamental components of RL—states, actions, and rewards—define the structure of the learning problem and guide the agent's learning process.

In reinforcement learning, a state represents the current situation of the environment (Sutton & Barto, 2018). It encompasses all the information necessary for the agent to make a decision. States can be discrete, like the positions on a chessboard, or continuous, like the speed and position of a car. The set of all possible states is known as the state space. The complexity of the state space can significantly affect the difficulty of the learning task, especially if the state space is very large or infinite.

Actions are the set of all possible moves or decisions an agent can make in a given state. Just like states, actions can be either discrete (e.g., turning left or right) or continuous (e.g., varying the pressure applied to a pedal). The choice of action at each step is based on the agent's policy, which is essentially a strategy for selecting actions based on the current state and the knowledge the agent has acquired so far. The goal of the agent is to learn a policy that maximizes the cumulative reward over time.

Rewards are immediate feedback provided to the agent after it takes an action in a specific state. The reward function is crucial as it shapes the learning and behaviour of the agent by telling it what is good or bad. It is a scalar feedback signal that indicates the benefit of the action taken in the current state. The agent's objective is to maximize the total amount of reward it receives in the long run, which involves not just seeking immediate rewards but also considering the long-term consequences of actions.

Together, states, actions, and rewards create a feedback loop that the agent uses to learn from its experiences. The agent observes the state of the environment, takes an action based on its current policy, receives a reward and observes the new state resulting from its action, and then updates its policy based on the experience gained. This ongoing process, driven by the interaction of these three elements, allows the agent to improve its behaviour over time and adapt to complex, dynamic environments.

Within RL, there are two primary approaches: model-based and model-free. Model-based algorithms rely on constructing and utilizing a model of the environment to make decisions. Model-free learning is a subset of reinforcement learning techniques where the agent learns to make decisions based solely on the experience it gains through interacting with the environment, without any explicit knowledge or modelling of the environment's dynamics. This approach contrasts with model-based learning, where the agent develops a model of how the environment behaves and uses this model to make decisions. In model-free learning, the agent relies entirely on the experiences it accumulates from its actions and the resulting outcomes. It updates its policy based on the rewards it receives and the states it encounters. This direct approach allows the agent to adapt its strategy incrementally, focusing on practical outcomes rather than theoretical predictions. We will focus on model-free algorithms, as there is substantial evidence suggesting that human and non-human learning often aligns more closely with model-free methods (Miranda et al., 2020).

The core of model-free learning involves the estimation of value functions or the direct learning of the policy:

(1) Value Function-Based Methods: These methods estimate the value of being in a given state (or taking a particular action in a state). The most common algorithms include Q-learning and SARSA, which learn action-value functions that tell the agent how good it is to perform a particular action in a specific state.

(2) Policy-Based Methods: These methods optimize the policy directly without explicitly maintaining a value function. Techniques like Policy Gradient methods (Sutton et al., 1999) fall into this category, where the policy is adjusted directly based on the gradient of expected reward. Here we focused on Q-learning algorithm (Watkins & Dayan, 1992).

Q-learning is a popular model-free off-policy reinforcement learning algorithm that uses Temporal Difference (TD, especially TD(0)) learning to estimate the value of state-action pairs. It enables an agent to learn optimal policies without requiring a model of the environment. At the core of Q-learning is the Q-value or action-value function, $Q(s, a)$, which estimates the expected utility of taking action a in state s and then following the optimal policy thereafter. The goal of Q-learning is to learn the Q-value function that accurately reflects these utilities, allowing the agent to make optimal decisions by simply favouring the action with the highest Q-value in any given state. Q-learning is a type of Temporal Difference (TD) learning, which is characterized by its use of incomplete episodes for learning—i.e., the agent doesn't need to wait until the end of an episode to update its value estimates. Instead, it updates its estimates based on the reward received and the estimated value of the subsequent state, effectively learning from every step taken in the environment. This progression into the future can be expanded arbitrarily; for instance, Q-learning considers only one step into the future, which is why it is categorized as TD(0), or temporal difference learning with one look-ahead.

The algorithm starts with an arbitrary Q-value function, typically initialized to zero. At each time step, the agent chooses an action a_t from the current state s_t based on a policy derived from the current Q-value (commonly a SoftMax policy). Then, the model executes the action, observes the reward r_t, and transitions to the new state s_{t+1}. The algorithm then updates the Q-value for the state-action pair $Q(s_t, a_t)$ using the observed reward and the maximum Q-value of the next state. The update rule is given by:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \varepsilon\big(r_t + \gamma\, Max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\big) \qquad (4)$$

Where ε is the learning rate and γ is the discount factor. Under certain conditions (such as visiting all state-action pairs infinitely often and proper tuning of the learning rate), Q-learning is guaranteed to converge to the optimal action-value function (Watkins & Dayan, 1992).

1.4 Intersection of Decision Making and Learning

Decision-making models primarily focus on how we integrate evidence from a noisy environment to make optimal or suboptimal decisions. These models often operate under the assumption that the decision-making strategy remains constant during learning. To validate this assumption, many studies extensively train their subjects to ensure there is no learning effect influencing decision-making strategies (Beck et al., 2008; T. Hanks et al., 2014; T. D. Hanks et al., 2011). In essence, traditional decision-making models are not designed to account for learning processes.

On the other hand, learning models are concerned exclusively with how decisions evolve over time as a result of learning. These models focus on the mechanisms of learning rather than how evidence and information from the environment are initially gathered and utilized. Typically, learning models employ very clear and non-noisy stimuli to ensure that the process of information processing and evidence accumulation does not complicate the understanding of learning dynamics. This distinct focus highlights the separation in the objectives and methodologies between decision-making and learning models within cognitive and computational neuroscience.

One of the fundamental characteristics of models like the Drift Diffusion Model (DDM) is the presence of a decision threshold, a concept critical for explaining a wide range of behaviours. One notable assumption in nearly all the decision-making models discussed thus far is that the decision threshold is explicitly defined. This term must be clearly established before applying the model, and more importantly constant in all trials. Although there have been some attempts to elucidate the neural mechanisms of decision boundaries through large spiking neural networks (see Lo & Wang (2006) for examples), the process by which we arrive at a decision threshold remains largely unknown. In other words, the dynamics of how decision boundaries evolve during learning are not well understood.

Consequently, there is a need for a model that can access an implicit, time-dependent decision boundary, where time refers to trials or training progress. Addressing this gap is the primary focus of Project II, which aims to develop a computational framework that integrates the principles of both decision-making and learning. This model seeks to harness the advantages of each realm, facilitating a deeper understanding of how decision thresholds adapt and evolve through ongoing learning and decision-making processes. By bridging these two areas, Project II hopes to provide more comprehensive insights into the dynamic interplay between learning and decision-making, ultimately enhancing our ability to model and predict complex cognitive behaviours.

2 Aim of the Thesis

The aim of this thesis is to enhance our understanding of human decision-making, particularly in social settings, and the processes of learning in perceptual decision making. In the first project, we develop a brain-plausible computational model to explain how confidence matching arises during social decision-making. Our attractor neural network framework was backed by evidence in pupil and EEG data. One of the main challenges in this project is obtaining a sufficient number of subjects who have undergone extensive training, as this is crucial for model validation. Training of the subjects took a significant amount of time and effort while the mechanisms underlying this training phase in perceptual decision-making are not well understood, presenting a gap in the research.

This gap motivated the second project of the thesis, which aims to propose a model that attempts to elucidate the training phase of perceptual decision-making. Understanding this phase is critical because it involves learning how to process and integrate sensory information to make accurate decisions. By modelling the training phase, we hope to uncover the neural and cognitive mechanisms that facilitate the transition from novice to expert decision-makers. This model will not only provide insights into the training dynamics but may also help optimize training protocols for decision making studies.

3 Project (1)

"Confidence is contagious. So is lack of confidence."

— *Vince Lombardi*

This chapter includes the research article "Interpersonal alignment of neural evidence accumulation to social exchange of confidence," published in *eLife*. The article investigates the neurocomputational mechanisms underlying confidence matching in group decision-making. In this work, we comprehensively studied how confidence is communicated between individuals and how it shapes individual decisions. Using an interdisciplinary approach, we examined human decision-making in social contexts through psychophysics, eye tracking, EEG, and computational modelling.

We found that individuals tended to adjust their confidence in line with that of their partners, even though their objective accuracy remained unchanged. Interestingly, reaction times also shifted in accordance with reported confidence. To explain these non-trivial patterns, we introduced a biologically plausible attractor neural network model.

Importantly, we systematically tested and validated the model's core assumptions using eye-tracking and EEG data. First, we assessed whether participants' internal beliefs changed based on their partner's confidence, using pupil data as a proxy for internal belief. Next, we used EEG data to estimate the rate of evidence accumulation, a neural measure not directly observable in behavioural data. Both analyses supported our computational model's predictions.

Together, this work provides a computational framework for understanding how confidence is communicated and aligned during group decision-making.

DOI: https://doi.org/10.7554/eLife.83722

Contributions:

Jamal Esmaily (JE), Sajjad Zabbah (SZ), Reza Ebrahimpour (RE), Bahador Bahrami (BB)

The author of this thesis is the first author of this manuscript.

JE contributions: Conceptualization, Data curation, Software, Formal analysis, Validation, Visualization, Methodology, Writing – original draft, Writing – review and editing.

SZ contribution: Conceptualization, Data curation, Software, Methodology, Writing – original draft, Writing – review and editing.

RE contribution: Conceptualization, Software, Supervision, Funding acquisition, Validation, Investigation, Methodology, Project administration, Writing – review and editing.

BB contribution: Conceptualization, Resources, Supervision, Funding acquisition, Validation, Investigation, Visualization, Writing – original draft, Project administration, Writing – review and editing.

Interpersonal alignment of neural evidence accumulation to social exchange of confidence

Jamal Esmaily[1,2,3]*, Sajjad Zabbah[4,5,6], Reza Ebrahimpour[7]*†, Bahador Bahrami[1,8]*†

[1]Department of General Psychology and Education, Ludwig Maximillian University, Munich, Germany; [2]Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Islamic Republic of Iran; [3]Graduate School of Systemic Neurosciences, Ludwig Maximilian University Munich, Munich, Germany; [4]School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Islamic Republic of Iran; [5]Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom; [6]Max Planck UCL Centre for Computational Psychiatry and Aging Research, University College London, London, United Kingdom; [7]Institute for Convergent Science and Technology, Sharif University of Technology, Tehran, Islamic Republic of Iran; [8]Centre for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

Abstract Private, subjective beliefs about uncertainty have been found to have idiosyncratic computational and neural substrates yet, humans share such beliefs seamlessly and cooperate successfully. Bringing together decision making under uncertainty and interpersonal alignment in communication, in a discovery plus pre-registered replication design, we examined the neurocomputational basis of the relationship between privately held and socially shared uncertainty. Examining confidence-speed-accuracy trade-off in uncertainty-ridden perceptual decisions under social vs isolated context, we found that shared (i.e. reported confidence) and subjective (inferred from pupillometry) uncertainty dynamically followed social information. An attractor neural network model incorporating social information as top-down additive input captured the observed behavior and demonstrated the emergence of social alignment in virtual dyadic simulations. Electroencephalography showed that social exchange of confidence modulated the neural signature of perceptual evidence accumulation in the central parietal cortex. Our findings offer a neural population model for interpersonal alignment of shared beliefs.

Editor's evaluation

This important study examines how humans use information about the confidence of collaborators to guide their own perceptual decision making and confidence judgements. The study addresses this question with a combination of psychophysics, electrophysiological modeling, and computational modelling that provides a compelling validation of a computational framework that can be used to derive and test theory-based predictions about how collaborators use communication to align their confidence and thereby optimize their collective performance.

***For correspondence:** jimi.esmaily@gmail.com (JE); ebrahimpour@sharif.edu (RE); bbahrami@gmail.com (BB)

†These authors contributed equally to this work

Competing interest: The authors declare that no competing interests exist.

Funding: See page 23

Received: 26 September 2022
Preprinted: 09 November 2022
Accepted: 09 November 2023
Published: 21 December 2023

Reviewing Editor: Redmond G O'Connell, Trinity College Dublin, Ireland

Introduction

We communicate our confidence to others to share our beliefs about uncertainty with them. However, numerous studies have shown that even the same verbal or numerical expression of confidence can have very different meanings for different people in terms of the underlying uncertainty (**Ais et al., 2016**; **Navajas et al., 2017**; **Fleming et al., 2010**). Similar inter-individual diversity has been found at the neural level (**Fleming et al., 2010**; **Sinanaj et al., 2015**; **Baird et al., 2013**). Still, people manage to cooperate successfully in decision making under uncertainty (**Bahrami et al., 2010**; **Austen-Smith and Banks, 1996**). What computational and neuronal mechanisms enable people to converge to a *shared meaning* of their confidence expressions in interactive decision making despite the extensively documented neural and cognitive diversity? This question drives at the heart of recent efforts to understand the neurobiology of how people adapt their communication to their beliefs about their interaction partner (**Stolk et al., 2016**). A number of studies have provided compelling empirical evidence of brain-to-brain coupling that could underlie adaptive communication of shared beliefs (**Silbert et al., 2014**; **Honey et al., 2012**; **Hasson et al., 2004**; **Dikker et al., 2014**; **Konvalinka et al., 2010**). These works remain, to date, mostly observational in nature. Plausible neuro-computational mechanism(s) accounting for how interpersonal alignment of beliefs may arise from the firing patterns of decision-related neural populations in the human brain are still lacking (**Hasson and Frith, 2016**; **Wheatley et al., 2019**). Using a multidisciplinary approach, we addressed this question at behavioral, computational, and neurobiological levels.

By sharing their confidence with others, joint decision makers can surpass their respective individual performance by reducing uncertainty through interaction (**Bahrami et al., 2010**; **Sorkin et al., 2001**). Recent works showed that during dyadic decision making, interacting partners adjust to one another by matching their own average confidence to that of their partner (**Bang et al., 2017**). Such confidence matching turns out to be a good strategy for maximizing joint accuracy under a range of naturalistic conditions, e.g., uncertainty about the partner's reliability. However, at present there is no link connecting these socially observed emergent characteristics of confidence sharing with the elaborate frameworks that shape our understanding of confidence in decision making under uncertainty (**Navajas et al., 2017**; **Fleming et al., 2010**; **Pouget et al., 2016**; **Adler and Ma, 2018**; **Aitchison et al., 2015**).

Theoretical work has shown that sequential sampling can, in principle, provide an optimal strategy for making the best of whatever uncertain, noisy evidence is available to the agent (**Heath, 1984**). These models have had great success in explaining the relationship between decision reaction time (RT) and accuracy under a variety of conditions ranging from perceptual (**Hanks and Summerfield, 2017**; **Gold and Shadlen, 2007**) to value-based decisions (**Ruff and Fehr, 2014**) guiding the search for the neuronal mechanisms of evidence accumulation to boundary in rodent and primate brains (**Schall, 2019**). The relation between RT and accuracy, known as speed-accuracy trade-off, has been recently extended to a three-way relationship in which choice confidence is guided by *both* RT and probability (or frequency) of correct decision (**Pouget et al., 2016**; **Kiani et al., 2014**; **Vickers, 1970**). Critically, these studies have all focused on decision making in *isolated individuals* deciding privately (**Wheatley et al., 2019**). Little is known about how these computational principles and neuronal mechanisms can give rise to socially shared beliefs about uncertainty.

To bridge this gap, we examined confidence-speed-accuracy trade-off in social vs isolated context in humans. We combined a canonical paradigm (i.e. dynamic random dot motion [RDM]) extensively employed in psychophysical and neuroscientific studies of speed-accuracy-confidence trade-off (**Hanks and Summerfield, 2017**; **Gold and Shadlen, 2007**; **Kelly and O'Connell, 2013**) with interactive dyadic social decision making (**Bahrami et al., 2010**; **Bang et al., 2017**). We replicated the emergence of confidence matching and obtained pupillometry evidence for shared subjective beliefs in our social implementation of the random dot paradigm and we observed a novel pattern of confidence-speed-accuracy trade-off specifically under the social condition. We constructed a neural attractor model that captured this trade-off, reproduced confidence matching in virtual social simulations and made neural predictions about the coupling between neuronal evidence accumulation and social information exchange that were born out by the empirical data.

Results

We used a discovery-and-replication design to investigate the computational and neurobiological substrates of confidence matching in two separate steps: 12 participants (4 female) were recruited in study 1 (discovery) and 15 (5 female, age: 28 (mean) ± Std (7)) in study 2 (replication, second study was pre-registered: https://osf.io/5zces). In each study, participants reported the direction of a random-dot

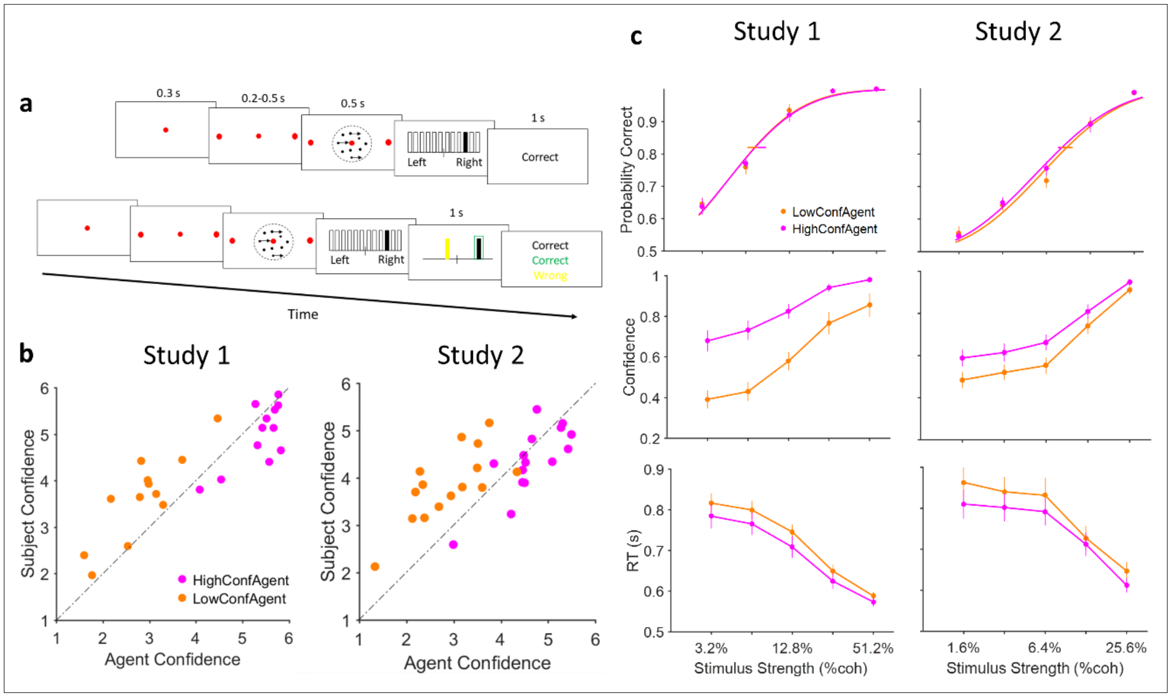

Figure 1. Experiment paradigm and behavioral results. (**a**) Timeline of trials in isolated (top) and social (bottom) conditions. After stimulus presentation, subjects reported their decision and confidence simultaneously by clicking on 1 of the 12 vertical bars. In the social condition, decision and confidence of participant (white in the experiment, here black for illustration purpose) and partner (yellow) were color coded. (**b**) Confidence matching. Participants confidence against agent confidence show a significant relation in both studies (linear regression p<0.001 for both studies). (**c**) Under social condition, when participants were paired with high (magenta) vs low (dark orange) confidence partner, accuracy (top panel) did not change (horizontal lines, 68% confidence interval of bootstrap test with 10,000 repetitions) but confidence (middle panel) and reaction time (RT) (bottom panel) were altered. Curves fitted to the accuracy data are Weibull cumulative distribution function. Error bars are standard error of the mean (SEM) across subjects.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Accuracy and confidence of the computer generated partners (CGPs).

Figure supplement 2. Statistical analysis of the confidence matching effect.

Figure supplement 3. Examination of the hypothesis that the partner's confidence at trial *t* modulates the participant behavior at trial *t*+1.

Figure supplement 4. Summary of debriefing results of the second study.

motion stimulus and indicated their confidence (***Figure 1a***) while EEG and eye tracking data were recorded, simultaneously. After an extensive training procedure (see Materials and methods for the recruitment), participants reached a stable behavioral (accuracy and RT) performance level. Then, two experimental sessions were conducted: first a private session (200 trials) in which participants performed the task alone; then a social session (800 trials for study 1 and 400 for study 2) in which they performed the task interactively together with a partner (implied to be another participant in a neighboring lab room).

In every trial (***Figure 1a***), after fixation for 300 ms was confirmed by closed-loop real-time eye tracking, two choice-target points appeared at 10° eccentricity corresponding to the two possible motion directions (left and right). After a short random delay (200–500 ms, truncated exponential distribution), a dynamic RDM (see ***Shadlen and Newsome, 2001***) was centrally displayed for 500 ms in a virtual aperture (5° diameter). At the end of the motion sequence, the participant indicated the direction of motion and their confidence on a 6-point scale by a single mouse click. A horizontal line intersected at midpoint and marked by 12 rectangles (6 on each side) was displayed. Participants moved the mouse pointer – initially set at the midpoint – to indicate their decision (left vs right of midpoint) and confidence by clicking inside one of the rectangles. Further distance from the midpoint indicated more confidence. RT was calculated as the time between the onset of the motion stimulus sequence and the onset of deviation of the mouse pointer (see Materials and methods for more details) (***Resulaj et al., 2009***) at the end of stimulus presentation.

In the isolated trials, the participant was then given visual feedback for accuracy (correct or wrong). In the social trials (***Figure 1a***, bottom panel), after the response, participants proceeded to the social stage. Here, the participants' own choice and confidence as well as that of their partner were displayed coded by different colors (white for participants; yellow for partners). Joint decision was automatically arbitrated in favor of the decision with higher confidence. Finally, three distinct color-coded feedback messages (participant, partner, and joint decision) were displayed.

Participants were instructed to try to maximize the joint accuracy of their social decisions. In order to achieve joint benefit, confidence should be expressed such that the decision with higher probability of correct outcome dominates (***Bahrami et al., 2010***). For this to happen, the participant needs to factor in the partner's behavior and adjust her confidence accordingly. For example, if the participant believes that her decision is highly likely to be correct, her confidence should be expressed such that joint decision is dominated by the partner only if the probability that the partner's decision is correct is even higher (and not, for example, if the partner expressed a high confidence habitually). This social modulation of one's confidence in a perceptual decision comprises the core of our model of social communication of uncertainty.

Following from an earlier study (***Bang et al., 2017***), for each block the participants were led to believe that they were paired with a new, anonymous human partner. In reality, in separate blocks, they were paired with four computer generated partners (henceforward, CGPs; see Materials and methods) constructed and tuned to parameters obtained from the participant's own behavior in the isolated session: (1) high accuracy and high confidence (HAHC; i.e. this CGP's decisions were more likely to be more confident as well as more accurate); (2) high accuracy and low confidence (HALC); (3) low accuracy and high confidence (LAHC); and (4) low accuracy and low confidence (LALC) (see Materials and methods for details). For study 2, we used two CGPs (HCA and LCA) while the agent accuracy was similar to those of participants (***Bang and Fleming, 2018***) (Wilcoxon rank sum, p=0.37, *df* = 29, *zval* = 0.89). See ***Figure 1—figure supplement 1*** for confidence and accuracy data of CGPs. Each participant completed 4 blocks of 200 trials cooperating with a different CGP in each block. Our questionnaire results also confirmed that our manipulation indeed worked (***Figure 1—figure supplement 4***) and more importantly none of the subject suspected their partners was an artificial one.

Having observed the confidence matching effect in both studies (***Figure 1b***), a permutation analysis confirmed that this effect did not arise trivially from mere pairing with any random partner (***Bang et al., 2017***; ***Figure 1—figure supplement 2***). The difference between the participant's confidence and that of their partner was smaller in the social (vs isolated) condition (***Figure 1—figure supplement 2***) consistent with the prediction that participants would match their average confidence to that of their partner in the social session (***Bang et al., 2017***).

Having established the socially emergent phenomenon of confidence matching in the dynamic RDM paradigm, we then proceeded to examine choice speed, accuracy, and confidence under social

Table 1. Details of statistical results in behavioral data (***Figure 1***).

	Response	Regressors	Estimate	SE	CI	t-Stat	p-Value	Total number
Study 1	Accuracy (HC vs LC)	Coherency	0.007	0.0006	[0.006 0.008]	11.57	<0.001	9600
		Condition	–0.002	0.021	[–0.045 0.04]	–0.1	0.92	9600
	Confidence (HC vs LC)	Coherency	0.0475	0.0008	[0.046 0.049]	56.5	<0.001	9600
		Condition	1.361	0.03	[1.31 1.42]	46.4	<0.001	9600
	RT (HC vs LC)	Coherency	–0.005	0.0001	[–0.005 –0.004]	–44.4	<0.001	9600
		Condition	0.029	0.004	[–0.035 –0.021]	7.85	<0.001	9600
Study 2	Accuracy (HC vs LC)	Coherency	0.0209	0.0016	[0.017 0.024]	13.23	<0.001	6000
		Condition	–0.0092	0.0296	[–0.067 0.049]	–0.31	0.76	6000
	Confidence (HC vs LC)	Coherency	0.1011	0.1011	[0.097 0.106]	47.47	<0.001	6000
		Condition	0.496	0.037	[0.42 0.56]	13.32	<0.001	6000
	RT (HC vs LC)	Coherency	–0.009	0.0003	[–0.01 –0.008]	–26.22	<0.001	6000
		Condition	0.0363	0.006	[0.024 0.048]	6.12	<0.001	6000

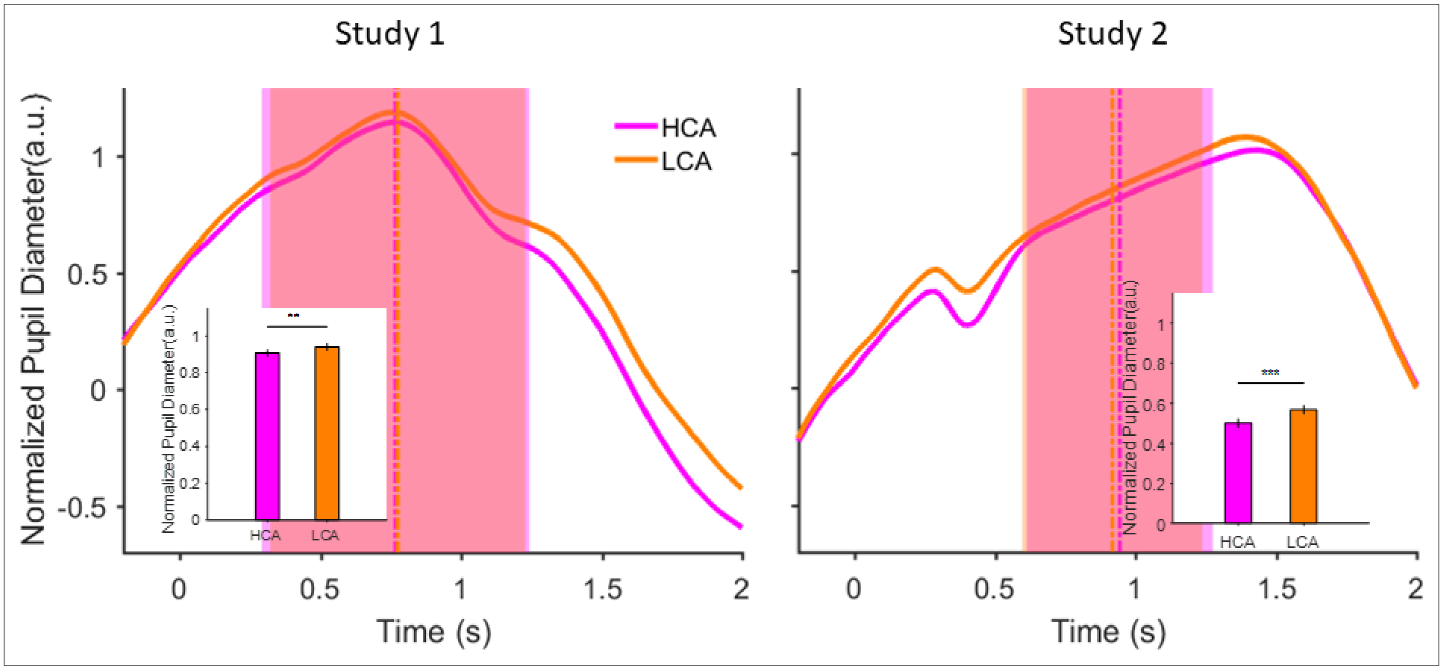

Figure 2. Pupil size during inter-trial interval (ITI) under pairing conditions in the social context when participant was paired with a high (HCA) or low confidence (LCA) agent. Normalized pupil diameter aligned to start of ITI period (*t*=0). Vertical dashed lines show average ITI duration. The shaded areas are one standard deviation of ITI period in each condition. Inset shows grand average (mean) pupil size during ITI under the two social conditions. Error bars are 95% confidence interval across trials. (**) indicates p<0.01 and (***) shows p<0.001. In the interest of clarity, signals were smoothed using an averaging filter.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Pupil size correlates with participant's own confidence in the isolated condition.

Figure supplement 2. Time series analysis of pupil size during inter-trial interval.

conditions (***Figure 1c***). We observed that when participants were paired with a high (vs low) confidence partner, there was no significant difference in accuracy between the social conditions (p=0.92, p=0.75 for study 1 and 2 respectively, generalized linear mixed model [GLMM], see Supplementary materials for details of the analysis [***Table 1***], ***Figure 1c*** top-left panel); confidence, however, was significantly higher (p<0.001 for both studies, ***Table 1***, ***Figure 1c*** middle panel) and RTs were significantly faster (p<0.001 for both, ***Table 1***, ***Figure 1c*** bottom panel) in the HCA vs LCA.

This pattern of dissociations of speed and confidence from accuracy is non-trivial because the expectations of the standard sequential sampling models would be that a change in confidence should be reflected in change in accuracy (***Pouget et al., 2016***; ***Sanders et al., 2016***). Many alternative mechanistic explanations are, in principle, possible. The rich literature on sequential sampling models in the random-dot paradigm permit articulating the components of such intuitive explanations as distinct computational models and comparing them by formal model comparison (see further below).

In order to assess the impact of social context on the participants' level of subjective uncertainty and rule out two important alternative explanations of confidence matching, we next examined the pupil data. Several studies have recently established a link between state of uncertainty and baseline (i.e. non-luminance mediated) variations in pupil size (**Bang et al., 2017**; **Wei and Wang, 2015**; **Nassar et al., 2012**; **Eldar et al., 2013**; **Murphy et al., 2014**; **Urai et al., 2017**). If the impact of social context on confidence were truly reflective of a similar change in the participant's belief about uncertainty, then we would expect the smaller pupil size when paired with high (HCA) vs low confidence agent (LCA) indicating lower subjective uncertainty. Alternatively, if confidence matching were principally due to pure imitation (***Rendell et al., 2011***; ***Iacoboni, 2009***) or due to some form of social obligation in agreeing with others (e.g. normative conformity [***Stallen and Sanfey, 2015***]) without any change in belief, we would expect the pupil size to remain unaffected by pairing condition under social context. We found that during the inter-trial interval (ITI), pupil size was larger in the blocks where participants

Table 2. Details of statistical results in pupil data (***Figure 2***).

	Response	Regressors	Estimate	SE	CI	t-Stat	p-Value	Total number
Study 1	Pupil	Condition	–0.038	0.011	[–0.06 –0.01]	–3.30	<0.001	8390
Study 2	Pupil	Condition	–0.066	0.015	[–0.09 –0.04]	–4.37	<0.001	5842

were paired with LCA (vs HCA) (***Figure 2***, GLMM analysis, p<0.01 and p<0.001 for study 1 and 2 respectively, see Supplementary materials for details of the analysis; ***Table 2***). We have added a time series analysis that demonstrates the temporal encoding of experimental conditions in the pupil signal during ITI (see ***Figure 2—figure supplement 2***). It is important to bear in mind that pupil dilation has been linked to other factors such as mental effort (***Lee and Daunizeau, 2021***), level of surprise (***Kloosterman et al., 2015***), and arousal level (***Murphy et al., 2014***) as well. These caveats notwithstanding, the patterns of pupil dilation within the time period of ITI that are demonstrated and replicated here, are consistent with the hypothesis that participants' subjective belief was shaped by interactions with differently confident partners. To support this conclusion further, we provide supplementary evidence linking the participant's own confidence to pupil size (***Figure 2—figure supplement 1***).

To arbitrate between alternative explanations and develop a neural hypothesis for the impact of social context on decision speed and confidence, we constructed a neural attractor model (***Wong and Wang, 2006***), a variant from the family of sequential sampling models of choice under uncertainty (***Bogacz et al., 2006***). Briefly, in this model, noisy sensory evidence was sequentially accumulated by two competing mechanisms (red and blue in ***Figure 3a*** left) that raced toward a common pre-defined decision boundary (***Figure 3a*** right) while mutually inhibiting each other. Choice was made as soon as one mechanism hits the boundary. This model has accounted for numerous observations of perceptual and value-based decision-making behavior and their underlying neuronal substrates in human (***Hunt et al., 2012***) and non-human primate (***Wei and Wang, 2015***) brain. Following previous works (***Wei and Wang, 2015***; ***Balsdon et al., 2020***; ***Rolls et al., 2010***; ***Atiya et al., 2019***) we defined model confidence as the time-averaged difference between the activity of the winning and losing accumulators (corresponding to the shaded gray area between the two accumulator traces in ***Figure 3a*** right, for the model simulation see ***Figure 3—figure supplement 2***) during the period of stimulus presentation (from 0 to 500 ms). Importantly, this definition of confidence is consistent with recent findings that computations of confidence continue *after* a decision has been made as long as sensory evidence is available (***Ruff and Fehr, 2014***; ***Balsdon et al., 2020***; ***van Kempen et al., 2019***; ***Moran et al., 2015***). We also demonstrate that our results do not depend on this specific formulation and also replicate with another alternative method (***Vickers, 1979***)(see ***Figure 3—figure supplement 3***).

Earlier works that demonstrated the relationship between decision uncertainty and pupil-related, global arousal state in the brain (***Murphy et al., 2014***; ***Urai et al., 2017***) guided our modeling hypothesis. We modeled the social context as a global, top-down additive input (***Figure 3a***; W_x) in the attractor model. This input drove both accumulator mechanisms equally and positively. The impact of this global top-down input is illustrated in ***Figure 3a*** right: with a positive top-down drive (W_x>0), the winner (thick blue) and the loser (thick red) traces both rise faster compared to zero top-down drive (dotted lines). The model's counterintuitive feature is that the surface area between the winning and losing accumulator is larger in the case of positive (dark gray shading) versus zero (light gray shading) top-down input. Model simulations show that when 0<W_x, this difference in surface area leads to faster RTs and higher confidence but does not change accuracy because it does not affect the decision boundary. These simulation results are consistent with our behavioral findings comparing HCA vs LCA conditions (***Figure 1c***).

We formally compared our model to three alternative, plausible models of how social context may affect the decision process. Without loss of generality, we used data from study 2 to fit the model. The first model hypothesized that partner's confidence dynamically modulated the decision bound (***Balsdon et al., 2020***) (parameter *B* in ***Equation 21***). In this model, the partner's higher confidence reduced the threshold for what counted as adequate evidence, producing the faster RTs under HCA (***Figure 1***.c). The second model proposed that partner's confidence changed non-decision time (NDT) (***Stine et al., 2020***; ***Equation 22***). Here, pairing with high confidence partner would not have any impact on perceptual processing but instead, non-specifically decrease RTs across all coherence levels without affecting accuracy. Finally, in the third model, the stimulus-independent perceptual gain (***Eldar***

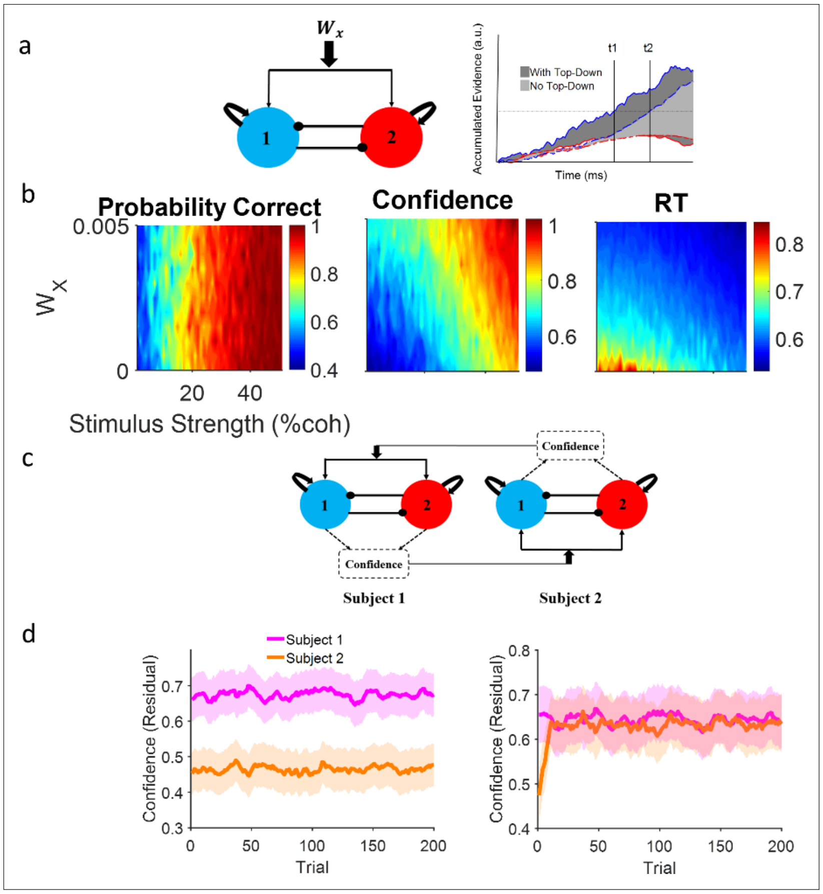

Figure 3. Neural attractor model. (**a**) Left: A common top-down (W_x) current drives both populations, each selective for a different choice alternative. Right: A schematic illustration of the impact of a positive top-down drive on accumulator dynamics. Confidence corresponds to the shaded area between winning (blue) and losing (red) accumulators. Solid lines and dark gray shade: positive top-down drive; dashed lines and light gray shade: zero top-down drive. With positive top-down current, the winner hits the bound earlier (*t*1 vs *t*2) and the surface area between the competing accumulator traces is larger (dark vs light gray). (**b**) Systematic examination of the impact of W_x on model behavior. Left panel: Accuracy does not depend on the top-down current but confidence (middle) and reaction time (RT) (right) change accordingly. Colors indicate different levels of top-down current. Each curve is the average of 10,000 simulations of the model given the top-down current. (**c**) Dynamic coupling in simulated dyadic interaction. Virtual dyads were constructed by feeding one model's confidence in previous trial to the other model as top-down drive and vice versa. (**d**) Left: Unconnected virtual dyad members (W_x = 0) simulate the isolated condition. Right: When the virtual dyad members are connected with top-down drive proportional to one another's confidence in previous trial, dyad members' confidence converge over time. In the isolated condition, confidence matching is not observed even though the pair receive the exact same sequence of stimuli. Shadowed areas of the confidence interval 95% resulted from 50 parallel simulations and curves were smoothed by an averaging filter for clearer illustration. The correlation with coherence has been removed from the confidence values via residual analysis (see ***Figure 3—figure supplement 1*** confidence values).

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Confidence matching without removing the correlation with the shared stimulus coherence.

Figure supplement 2. The effect of top-down current on the attractor network.

Figure supplement 3. Model performance regarding different confidence representations.

Figure supplement 4. Model comparison.

Figure supplement 5. Model vs data.

Figure supplement 6. The speed of confidence matching.

Figure supplement 7. Model falsification.

Figure supplement 8. Model predictions for confidence matching are not sensitive to linearity assumptions.

et al., 2013; **Li et al., 2018**) parameter of input current (parameter μ_0 in **Equation 23**) was modulated by partner confidence. Here, higher partner confidence increased the perceptual gain (as if increasing the volume of the radio) leading to increased confidence and decreased RT (**Figure 1c**) and would be consistent with the pupillometry results. In each model, in the social condition, the parameter of interest was linearly modulated by the confidence of the partner in the previous trial. Importantly, in **Figure 1—figure supplement 3**, we show that empirically, such trial-by-trial dependence is observed in confidence and RTs data in both study 1 and 2. Formal model comparison showed that our top-down additive current model was superior to all three alternatives (see **Figure 3—figure supplement 4**).

Having shown that a common top-down drive can qualitatively reproduce the impact of social context on speed-accuracy-confidence and quantitatively excel other alternatives in fitting the observed behavior, we then used the winning model to simulate our interactive social experiment virtually (**Figure 3c**). We simulated one decision maker with high confidence (subject 1 in **Figure 3d**) and another one with low confidence (subject 2). To simulate subject 1, we slightly increased the excitatory and the inhibitory weights. The opposite was done to simulate subject 2 (see Materials and methods for details). We then paired the two simulated agents by feeding the confidence of each virtual agent (from trial *t*–1) (**Bang et al., 2017**) as top-down input to the other virtual agent (in trial *t*).

Using this virtual social experiment, we simulated the dyadic exchanges of confidence in the course of our experiment and drew a prediction that could be directly tested against the empirical behavioral data. Without any fine-tuning of parameters or any other intervention, confidence matching emerged spontaneously when two virtual agents with very different confidence levels in isolated condition (**Figure 3d** left) were paired with each other as a dyad (**Figure 3d** right). Importantly, the model could be adapted to show different speed of matching as well (see **Figure 3—figure supplement 6**). However, for simplicity we presented the simplest case in the main text.

To identify the neural correlates of interpersonal alignment of belief about uncertainty, we note that previous works using non-invasive electrophysiological recordings in humans engaged in motion discrimination (**Twomey et al., 2016**; **Stolk et al., 2013**) have identified the signature, accumulate-to-bound neural activity characteristic of evidence accumulation in the sequential sampling process. Specifically, these findings show a centropareital positivity (CPP) component in the event-related potential that rises with sensory evidence accumulation across time. The exact correspondence between the neural CPP and elements of the sequential sampling process are not yet clear (**O'Connell et al., 2018**). For example, CPP could have resulted from the spatial superposition of the electrical activity of both accumulators or be the neural activity corresponding to the difference in accumulated evidence. These caveats notwithstanding, consistent with the previous literature, we found that in the isolated condition, our data replicated those earlier findings: **Figure 4a** shows a clear CPP event-related potential whose slope of rise was strongly modulated by motion coherence (GLMM, p<0.001 and p=0.01 for study 1 and 2 receptively, see **Supplementary file 1d** and **Figure 4—figure supplement 2** for more details). Importantly, we have added the response-locked analysis of the CPP signals (see **Figure 4—figure supplement 4**). We do see that the response-locked CPP waveforms converge to one another for high vs low coherence trials at the moment of the response.

Our model hypothesized that under social condition, a top-down drive – determined by the partner's communicated confidence in the previous trial – would modulate the rate of evidence accumulation (**Figure 3a**). We tested if the CPP slope were larger *within* every given coherence bin when the participant was paired with an HCA (vs LCA). Indeed, the data demonstrated a larger slope of CPP rise under HCA vs LCA (**Figure 4c**, study 1 for the social condition p=0.15 but for the second study p<0.01, see **Tables 3 and 4** for more details). These findings demonstrate that interpersonal alignment of confidence is associated with a modulation of neural evidence accumulation – as quantified by CPP – by the social exchange of information (also see **Figure 4—figure supplement 3**). It is important to note a caveat here before moving forward. These data show that both CPP and confidence are different between the HCA and LCA conditions. However, due to the nature of our experimental design, it would be premature to conclude from them that CPP *contributes causally to* the alignment of subjectively held beliefs or behaviorally expressed confidence. Put together with the behavioral confidence matching (**Figure 1b**) and the pupil data (**Figure 2**) our findings suggest that some such neural-social coupling could be the underlying basis for the construction of a shared belief about uncertainty.

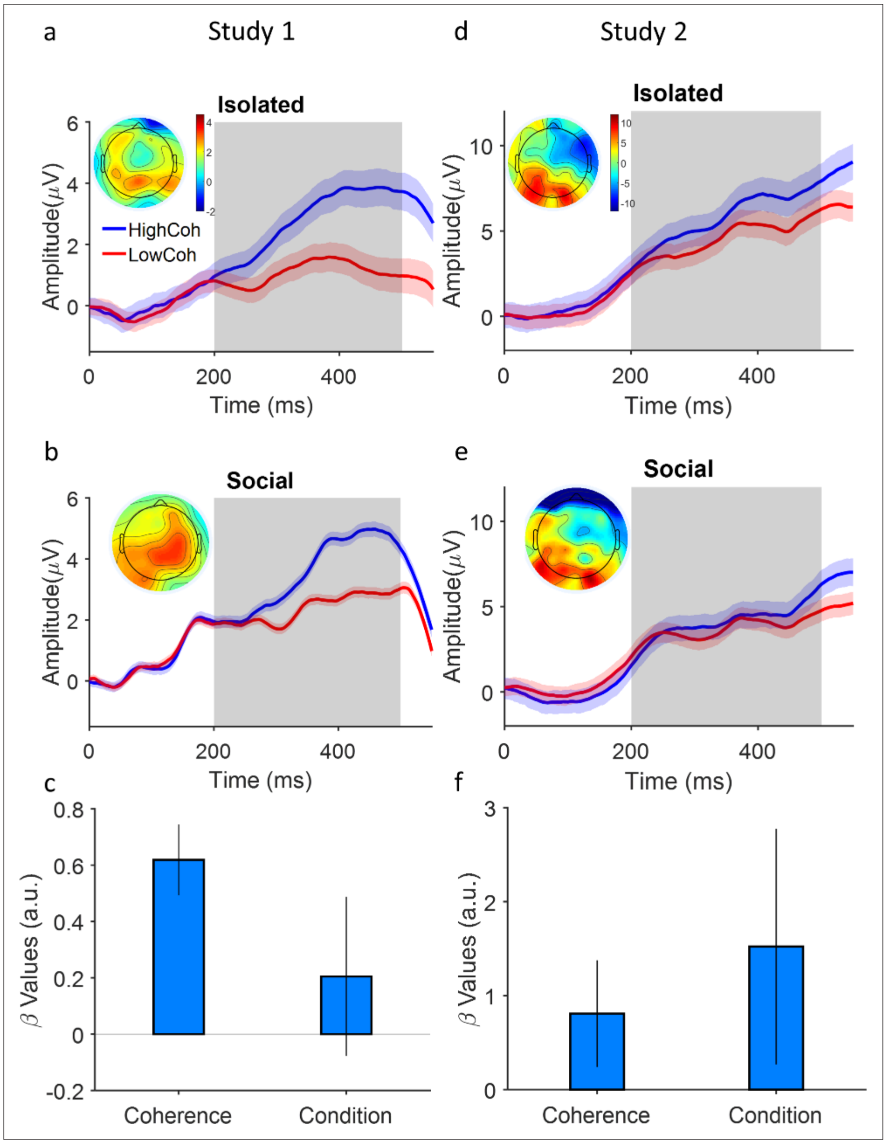

Figure 4. Coupling of neural evidence accumulation to social exchange of information. (**a**) Centroparietal positivity (CPP) component in the isolated condition: event-related potentials are time-locked to stimulus onset, binned for high and low levels of coherency (for study 1, low: 3.2%, 6.4%, 12.8%; high: 25.6% and 51.2%; for study 2 (**d**), low: 1.6%, 3.2%, 6.4%; high: 12.8%, 25.6%) and grand averaged across centropatrial electrodes (see Materials and methods). Inset shows the topographic distribution of the EEG signal averaged across the time window indicated by the gray area. (**b**) CPP under social condition. Conventions the same as panel (a). (**c**) A generalized linear mixed model (GLMM) model showed the significant relation of centroparietal signals to levels of coherency and social condition (high confidence agent [HCA] vs low confidence agent [LCA]). Error bars are 95% confidence interval over the model's coefficient estimates. Signals were smoothed by an averaging filter; shaded areas are SEM across trials.

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Electrode placement in each study.

Figure supplement 2. Relation of EEG signals from centropartial area of the brain to coherence levels and social conditions.

Figure supplement 3. Simulated slope of the accumulator activity in our computational model in low confidence agent (LCA) and high confidence agent (HCA) conditions.

Figure supplement 4. Response-locked EEG signal separated for high vs low coherence levels.

Figure supplement 5. Power calculation (Monte Carlo simulation) for EEG slope effect (***Figure 4*** in the main manuscript).

Table 3. Details of statistical results in EEG data (***Figure 4***).

	Response	Regressors	Estimate	SE	CI	t-Stat	p-Value	Total number
Study 1	EEG slope	Coherency	0.62	0.065	[0.49. 074]	9.64	<0.001	6492
		Condition	0.2	0.14	[-0.07 0.49]	1.42	0.15	6492
Study 2	EEG slope	Coherency	0.8	0.29	[0.24 1.37]	2.8	<0.01	5367
		Condition	1.52	0.63	[0.27 2.77]	2.39	0.017	5367

Discussion

We brought together two so-far-unrelated research directions: confidence in decision making under uncertainty and interpersonal alignment in communication. Our approach offers solutions to important current problems in each.

For decision science, we provide a model-based, theoretically grounded neural mechanism for going from individual, idiosyncratic representations of uncertainty (***Navajas et al., 2017***; ***Fleming et al., 2010***) to socially transmitted confidence expressions (***Bahrami et al., 2010***; ***Bang et al., 2017***) that are seamlessly shared and allow for successful cooperation. The social-to-neuronal coupling mechanism that we borrowed from the communication literature (***Hasson and Frith, 2016***; ***Wheatley et al., 2019***) is crucial in this new understanding of the neuronal basis of relationship between subjectively private and socially shared uncertainty.

For communication science, by examining perceptual decision making under uncertainty in social context, we created a laboratory model in which the goal of communication was to arrive at a shared belief about uncertainty (rather than creating a look-up table for the meaning of actions [***Stolk et al., 2016***; ***Silbert et al., 2014***; ***Honey et al., 2012***]). In this way, we could employ the extensive theoretical, behavioral, and neurobiological body of knowledge in decision science (***Pouget et al., 2016***; ***Adler and Ma, 2018***; ***Aitchison et al., 2015***; ***Hanks and Summerfield, 2017***; ***Gold and Shadlen, 2007***; ***Ruff and Fehr, 2014***; ***Schall, 2019***; ***Kiani et al., 2014***; ***Kelly and O'Connell, 2013***; ***Shadlen and Newsome, 2001***; ***Resulaj et al., 2009***; ***Sanders et al., 2016***; ***Wei and Wang, 2015***; ***Eldar et al., 2013***; ***Urai et al., 2017***; ***Yeung and Summerfield, 2012***; ***Fleming and Daw, 2017***; ***Kiani and Shadlen, 2009***) to construct a mechanistic neural hypothesis for interpersonal alignment.

Over the past few years, the efforts to understand the 'brain in interaction' have picked up momentum (***Wheatley et al., 2019***; ***Frith and Frith, 1999***). A consensus emerging from these works is that, at a conceptual level, successful interpersonal alignment entails the mutual construction of a shared cognitive space between brains (***Stolk et al., 2015***; ***Wheatley et al., 2019***; ***Friston and Frith, 2015***). This would allow interacting brains to adjust their internal dynamics to converge on shared beliefs and meanings (***Hasson and Frith, 2016***; ***Gallotti and Frith, 2013***). To identify the neurobiological substrates of such shared cognitive space, brain-to-brain interactions need to be described in terms of information flow, i.e., the impact that interacting partners have on one another's brain dynamics (***Wheatley et al., 2019***).

The evidence for such information flow has predominantly consisted of demonstrations of alignment of brain-to-brain activity (i.e. synchrony at macroscopic level, e.g. fMRI BOLD signal) when people process the same (simple or complex) sensory input (***Honey et al., 2012***; ***Hasson et al., 2004***; ***Breveglieri et al., 2014***; ***Mukamel et al., 2005***; ***Hasson and Honey, 2012***) or engage in complimentary communicative (***Silbert et al., 2014***) roles to achieve a common goal. More recently, dynamic coupling (rather than synchrony) has been suggested as a more general description of the nature of brain-to-brain interaction (***Hasson and Frith, 2016***). Going beyond the intuitive notions of synchrony and coupling, to our knowledge, no computational framework – grounded in the principles of neural

Table 4. Details of statistical results in EEG data (***Figure 4—figure supplement 2*** top row).

	Response	Regressors	Estimate	SE	CI	t-Stat	p-Value	Total number
Study 1	EEG slope	Coherency	0.02	0.005	[0.01 0.03]	4.48	<0.001	1523
Study 2	EEG slope	Coherency	0.06	0.02	[0.01 0.11]	2.54	<0.01	2822

computing – has been offered that could propose a plausible quantitative mechanism for these empirical observations of brain-to-brain coupling.

Combining four different methodologies, the work presented here undertook this task. Behaviorally, our participants engaged in social perceptual decision making under various levels of sensory and social uncertainty (**Bahrami et al., 2010**; **Bang et al., 2017**). Emergence of confidence matching (**Figure 1b**) showed that participants coordinated their decision confidence with their social partner. Pupil data (**Figure 2**) suggested that participant's belief about uncertainty was indeed shaped by the social coordination. A dissociation (**Figure 1c**) of decision speed and confidence from accuracy was reported that depended on the social context. This trade-off, as well as the emergence of confidence matching, was successfully captured by a neural attractor model (**Figure 3**) in which two competing neural populations of evidence accumulators – each tuned to one choice alternative – were driven by a common top-down drive determined by social information. This model drew predictions for behavior (**Figure 3d**) and neuronal activity (**Figure 4**, **Figure 4—figure supplements 1–5**) that were born out by the data. Social exchange of information modulated the neural signature of evidence accumulation in the parietal cortex.

Although numerous previous works have employed sequential sampling models to explain choice confidence, the overwhelming majority (**Pouget et al., 2016**; **Aitchison et al., 2015**; **Hanks and Summerfield, 2017**; **Gold and Shadlen, 2007**; **Ruff and Fehr, 2014**; **Schall, 2019**; **Kiani et al., 2014**; **Sanders et al., 2016**; **Kiani and Shadlen, 2009**; **Krajbich and Rangel, 2011**) have opted for the drift diffusion family of models. Neural attractor models have so far been rarely used to understand confidence (**Rolls et al., 2010**; **Atiya et al., 2019**; **Wang, 2002**). Our attractor model is a reduced version (**Wong and Wang, 2006**) of the original biophysical neural circuit model for motion discrimination (**Wang, 2002**). The specific affordances of attractor models allowed us to implement social context as a sustained, tonic top-down feedback to both accumulator mechanisms. More importantly, we were able to simulate social interactive decision making by virtually pairing any given two instances of the model (one for each member of a dyad) with each other: the confidence produced by each in a given trial served as top-down drive for the other in the next trial. Remarkably, a shared cognitive space about uncertainty (i.e. confidence matching) emerged spontaneously from this simulated pairing without us having to tweak any model parameters.

At a conceptual level, deconstructing the social communication of confidence into a comprehension and a production process (**Silbert et al., 2014**) is helpful. Comprehension process refers to how socially communicated confidence is incorporated in the recipient brain and affects their decision making. Production process refers to how the recipient's own decision confidence is constructed to be, in turn, socially expressed. It is tempting to attribute the CPP neural activity in the parietal cortex to the production process. Comprehension process, in turn, could be the top-down feedback from prefrontal brain areas previously implicated in confidence and metacognition (**Fleming et al., 2010**; **Fleming and Daw, 2017**; **De Martino et al., 2017**) to the parietal cortex. However, we believe that our neural attractor model in particular and the empirical findings do not lend themselves easily to this conceptual simplification. For example, the evidence accumulation process can be a part of the production (because confidence emerges from the integrated difference between accumulators) as well as the comprehension process (because the rate of accumulation is modulated by the received social information). As useful as it is, the comprehension/production dichotomy's limited scope should be recognized. Instead, armed with the quantitative framework of neural attractor models (for each individual) and interactive virtual pairing (to simulate dyads), future studies can now go beyond the comprehension/production dichotomy and examine the neuronal basis of interpersonal alignment with a model that have a strong footing in biophysical realities of neural computation.

Several limitations apply to our study. We chose different sets of coherence levels for the discovery (experiment 1) and replication (experiment 2). This choice was made deliberately. In experiment 1 we included a very high coherence (51%) level to optimize the experimental design for demonstrating the CPP component in the EEG signal. In experiment 2, we employed peri-threshold coherence levels in order to focus on behavior around the perceptual threshold to strengthen the model fitting and model comparison. This trade-off created some marginal differences in the observed effect sizes in the neural data across the two studies. The general findings were in good agreement.

The main strength of our work was to put together many ingredients (behavioral data, pupil and EEG signals, computational analysis) to build a picture of how the confidence of a partner, in the context of joint decision making, would influence our own decision process and confidence evaluations. Many of the effects that we describe here are well described already in the literature but putting them all together in a coherent framework remains a challenge. For example, our study did not directly examine neural alignment between interaction partners. We measured the EEG signal one participant at a time. The participant interacted with an alleged (experimenter-controlled) partner in any given trial. Our experimental design, however, permitted strict experimental control and allowed us to examine the participants' social behavior (i.e. choices and confidence), pupil response, and brain dynamics as they achieved interpersonal alignment with the partner. Moreover, while the hypotheses raised by our neural attractor model did examine the nature of brain dynamics involved in evidence accumulation under social context, testing these hypotheses did not require hyper-scanning of two participants at the same time. We look forward to future studies that use the behavioral and computational paradigm described here to examine brain-to-brain neural alignment using hyper-scanning.

We have interpreted our findings to indicate that social information, i.e., partner's confidence, impacts the participants' beliefs about uncertainty. It is important to underscore here that, similar to real life, there are other sources of uncertainty in our experimental setup that could affect the participants' belief. For example, under joint conditions, the group choice is determined through the comparison of the choices and confidences of the partners. As a result, the participant has a more complex task of matching their response not only with their perceptual experience but also coordinating it with the partner to achieve the best possible outcome. For the same reason, there is greater outcome uncertainty under joint vs individual conditions. Of course, these other sources of uncertainty are conceptually related to communicated confidence, but our experimental design aimed to remove them, as much as possible, by comparing the impact of social information under high vs low confidence of the partner.

Our study brings together questions from two distinct fields of neuroscience: perceptual decision making and social neuroscience. Each of these two fields have their own traditions and practical common sense. Typically, studies in perceptual decision making employ a small number of extensively trained participants (approximately 6–10 individuals). Social neuroscience studies, on the other hand, recruit larger samples (often more than 20 participants) without extensive training protocols. We therefore needed to strike a balance in this trade-off between number of participants and number of data points (e.g. trials) obtained from each participant. Note, for example, that each of our participants underwent around 4000 training trials. Importantly, our initial study (N=12) yielded robust results that showed the hypothesized effects nearly completely, supporting the adequacy of our power estimate. However, we decided to replicate the findings in a new sample with N=15 participants to enhance the reliability of our findings and examine our hypothesis in a stringent discovery-replication design. In ***Figure 4—figure supplement 5***, we provide the results of a power analysis that we applied on the data from study 1 (i.e. the discovery phase). These results demonstrate that the sample size of study 2 (i.e. replication) was adequate when conditioned on the results from study 1.

Finally, one natural limitation of our experimental setup is that the situation being studied is very specific to the design choices made by the experimenters. These choices were made in order to operationalize the problem of social interaction within the psychophysics laboratory. For example, the joint decisions were not an agreement between partners (***Bahrami et al., 2010***; ***Bahrami et al., 2012***). Instead, following a number of previous works (***Bang et al., 2017***; ***Bang et al., 2020***), joint decisions were automatically assigned to the most confident choice. In addition, partner's confidence and choice were random variables drawn from a distribution prespecified by the experimenter and therefore, by design, unresponsive to the participant's behavior. In this sense, one may argue that the interaction partner's behavior was not 'natural' since they did not react to the participant's confidence communications (note however that the partner's response times and accuracy were not entirely random but matched carefully to the participant's behavior prerecorded in the individual session). How much of the findings are specific to these experimental setting and whether the behavior observed here would transfer to other real-life settings is an open question. For example, it is plausible that participants may show some behavioral reaction to the response time variations since there is some evidence indicating that for binary choices like here, response times also systematically communicate uncertainty to others (***Patel et al., 2012***). Future studies could examine the degree to which the results might be paradigm-specific.

Materials and methods

Participants

A total of 27 participants (12 in experiment 1 and 15 in experiment 2; 10 females; average age: 24 years; all naïve to the purpose of the experiment) were recruited for a two-session experiment – isolated and social session. All subjects reported normal or corrected-to-normal vision. The participants did several training sessions in order to become familiar with the procedure and reach a consistent pre-defined level of sensitivity (see Materials and methods for more details).

Recruitment

Participants volunteered to take part in the experiment in return for course credit for study 1. For study 2, a payment of 80,000 Toman equivalent to 2.5€ per session was made to each participant. On the experiment day, participants were first given the task instructions. Written informed consent was then obtained. The experiments were approved by the local Ethics Committee at Shaheed Rajaei University's Department of computer engineering.

Task design

In the isolated session, each trial started with a red fixation point in the center of the screen (diameter 0.3°). Having fixated for 300 ms (in study 1, for a few subjects with eye monitoring difficulty this period shortened), two choice-target points appeared at 10° eccentricity corresponding to the two possible motion directions (left and right) (***Figure 1***). After a short random delay (200–500 ms, truncated exponential distribution), a dynamic RDM stimulus was displayed for 500 ms in a virtual aperture (5° diameter) centered on the initial fixation point. These motion stimuli have been described in detail elsewhere (***Shadlen and Newsome, 2001***). At the end of the motion stimulus a response panel (see ***Figure 1a***) was displayed on the screen. This response panel consisted of a horizontal line extending from left to the right end of the display, centered on the fixation cross. On each side of the horizontal line, six vertical rectangles were displayed side by side (***Figure 1a***) corresponding to six confidence levels for each decision alternative. The participants reported the direction of the RDM stimulus and simultaneously expressed their decision and confidence using the mouse.

The rectangles on the right and left of the midpoint corresponded to the right and left choices, respectively. By clicking on the rectangles further the midpoint participants indicated higher confidence. In this way, participant indicated their confidence and choice simultaneously (***Kiani et al., 2014***; ***Mahmoodi et al., 2015***) For experiment 1, response time was defined as the moment that the marker deviated (more than one pixel) from the center of the screen. However, in order to rule out the effect of unintentional movements, for the second study we increased this threshold to one degree of visual angle. The participants were informed about their accuracy by a visual feedback presented in the center of the screen for 1 s (correct or wrong).

In the social session, the participants were told they were paired with an anonymous partner. In fact, they were paired with a CGP tailored to the participant's own behavior in their isolated session. The participants did not know about this arrangement. Stimulus presentation and private response phase were identical to the isolated session. After the private response, the participants were presented with a social panel right (***Figure 1***). In this panel, the participant's own response (choice and confidence) were presented together with that of their partner for 1 s. The participant and the partner responses were color-coded (white for participants; yellow for partners). Joint decision was determined by the choice of the more confident person and displayed in green. Then, three distinct color-coded feedbacks were provided.

In both isolated and social sessions, the participants were seated in an adjustable chair in a semi-dark room with chin and forehead supported in front of a CRT display monitor (first study: 17 inches; PF790; refresh rate, 85 Hz; screen 164 resolution, 1024×768; viewing distance, 57 cm, second study: 21 inches; Asus VG248; refresh rate, 75 Hz; screen resolution, 1024×768; viewing distance, 60 cm). All the code was written in PsychToolbox (***Brainard, 1997***; ***Kleiner et al., 2007***; ***Pelli, 1997***).

Training procedure

Each participant went through several training sessions (on average 4) to be trained on RDM task. They first trained in a response-free (i.e. RT) version of the RDM task in which motion stimulus was discontinued as soon as the participant responded. They were told to decide about the motion direction of dots as fast and accurately as possible (*Kiani et al., 2014*). Once they reached a *stable* kevel of accuracy and RT, they proceeded to the main experiment. Before participating in the main experiment, they performed another 20–50 trials of warm-up. Here, the stimulus duration was fixed and responses included confidence report. For the social sessions, participants were told that in every block of 200 trials, they would be paired with a different person, seated in another room, with whom they would collaborate. They were also instructed about the joint decision scheme and were reminded that the objective in the social task was to maximize collective accuracy. Data from training and warm-up trials were included in the main analysis.

Procedure

Each participant performed both the isolated and the social task. In the isolated session, they did one block containing 200 trials. Acquired data were employed to construct four computer partners for the first study and two partners for the second study. We used the procedure introduced in a previous works to generate CGPs (*Bang et al., 2017*; *Bang et al., 2022*). In the first study, the four partners were distinguished by their level of average accuracy and overall confidence: HAHC, HALC, LAHC, and finally LALC. For the second study partners only differed in confidence: HCA and LCA. Each participant performed one block of 200 trials for each of the paired partners – 800 overall for study 1 and 400 overall for study 2.

In the social session, participants were told to try to maximize the joint decision success (*Bang et al., 2017*). They were told that their payment bonus depended on by their joint accuracy (*Bang et al., 2020*). While performing the behavioral task, EEG signals and pupil data were also recorded.

Computer generated partner

In study 1, following *Bang et al., 2017*, four partners were generated for each participant tuned to the participant's own behavioral data in the isolated session. Briefly, we created four simulated partners by varying their mean accuracy (high or low) and mean confidence (high or low). First, in the isolated session, the participant's sensory noise (σ) and a set of thresholds that determined the distribution of their confidence responses were calculated (see Materials and methods also). Simulated partner's accuracy was either high ($0.3\times\sigma$) or low ($1.2\times\sigma$). Mean confidence of simulated partners were also set according to the participant's own data. For low confidence simulated partner, average confidence was set to the average of participant's confidence in the low coherence (3.2% and 6.4%) trials. For the high confidence simulated partners, mean confidence was set to the average confidence of the participant in the high coherence (25.6% and 51.2%) trials. RTs were chosen randomly by sampling from a uniform random distribution (from 0.5 to 2 s). Thus, in some trials the participant needed to wait for the partner's response.

Having thus determined the parameters of the simulated partners, we then generated the sequence of trial-by-trial responses of a given partner using the procedure introduced by *Bang et al., 2017*. To produce the trial-by-trial responses of a given partner, we first generated a sequence of coherence levels with given directions (+ for rightward and – for leftward directions). Then we created a sequence of random values (sensory evidence), drawn from a Gaussian distribution with mean of coherence levels and variance of σ (sensory noise). Then, via applying the set of thresholds taken from the participant's data in isolated condition, we mapped the sequence of random values into trial-by-trial responses to generate a partner with a given confidence mean. Finally, to simulate lapses of attention and response errors, we randomly selected a response (from a uniform distribution over 1–6) on 5% of the trials (see *Figure 1—figure supplement 1* for the accuracy and confidence of the generated partners).

For study 2, we used the same procedure as study 1 and simulated two partners. These partners' accuracy was similar to the participant but each had a different confidence means (high confidence and low confidence partners). Therefore, we kept the σ constant and only change the confidence. For low confidence simulated partner, average confidence was set to the average of participant's confidence in the low coherence (1,6%, 3.2%, and 6.4%) trials. For the high confidence simulated partners, mean confidence was set to the average confidence of the participant in the high coherence (12.8% and 25.6%) trials.

Signal detection theory model for isolated sessions

In study 1 and 2, we simulated 4 and 2 artificial partners, respectively. We followed the procedure described by *Bang et al., 2017*. Briefly, working with the data from the isolated session, the sensory noise (σ) and response thresholds (θ) for each participant were calculated using a signal detection theory model. In this model, the level of sensory noise (σ) determines the participant's sensitivity and a set of 11 thresholds determines the participant's response distribution, which indicate both decision (via its sign) and confidence within the same distribution (see below).

On each trial, the sensory evidence, *x*, is sampled from a Gaussian distribution, $x \in N(s, \sigma^2)$. The mean, *s*, is the motion coherence level and is drawn uniformly from the set $s \in S = \{-0.512, -0.256, -0.128, -0.064, -0.032, 0.032, 0.064, 0.128, 0.256, 0.512\}$ (for the second study $S = \{-0.256, -0.128, -0.064, -0.032, -0.016, 0.016, 0.032, 0.064, 0.128, 0.256\}$). The sign of *s* indicates the correct direction of motion (right = positive) and its absolute value indicates the motion coherency. The standard deviation, σ, describes the level of sensory noise and is the same for all stimuli. We assumed that the internal estimate of sensory evidence (*z*) is equal to the raw sensory evidence (*x*). If *z* is largely positive, it denotes high probability of choosing right direction and vice versa for largely negative values.

To determine the participant's sensitivity and the response thresholds, first, we calculated the distribution of responses (*r*, ranging from –6 to 6, where the participant's confidence was ($c = |r|$), and her decision was determined by the sign of *r*). ***Equation 1*** shows the response distribution.

$$p_i = \begin{cases} \mathrm{p}\left(z \le \theta_{-6}\right) & i = -6 \\ \mathrm{p}\left(\theta_{\mathrm{i}-1} < z \le \theta_{\mathrm{i}}\right) & -6 < i \le -1 \ or \ \ 2 \le i < 6 \\ \mathrm{p}\left(\theta_{-1} < z \le \theta_1\right) & i = 1 \\ \mathrm{p}\left(z > \theta_5\right) & i = 6 \end{cases} \tag{1}$$

Using θ and σ, we mapped *z* to participants response (*r*). We found thresholds θ_i over *S* where $i = -6,–5, -4,–3, -2, –1, 1, 2, 3, 4, 5$ such that:

$$\sum_{j \le i} pj = \frac{1}{10} \sum_{s \epsilon S} \Phi\left(\frac{\theta i - s}{\sigma}\right) \tag{2}$$

where Φ is the Gaussian cumulative density function. For each stimulus, $s \in S$, the predicted response distribution, $p\left(r = i|s\right)$, calculated by *S*3:

$$\mathrm{p}\left(r = i|s\right) = \begin{cases} \left(\frac{\theta_{-6} - s}{\sigma}\right) & i = -6 \\ \left(\frac{\theta_i - s}{\sigma}\right) - \left(\frac{\theta_{i-1} - s}{\sigma}\right) & -6 < i < 6 \\ 1 - \left(\frac{\theta_5 - s}{\sigma}\right) & i = 6 \end{cases} \tag{3}$$

From here, the model's accuracy could be calculated by *S*4:

$$a_{agent} = \frac{\sum_{s \in S.s>0} \sum_{i=1}^{6} p_{i.s} + \sum_{s \in S.s<0} \sum_{i=-6}^{-1} p_{i.s}}{10} \tag{4}$$

Given participant's accuracy, we could find a set of θ and σ.

Confidence estimation

Once we had determined θ and σ, we could produce a confidence landscape with a specific mean. In order to generate one high confidence and another low confidence partner, we needed to alter mean confidence by modifying the θ. There could be an infinite number of confidence distribution with the desired mean. We were interested in the maximum entropy distribution that satisfied two constraints: mean confidence should be specified, and the distribution must sum to 1. Using Lagrange multiplier (λ) the response distribution was calculated as:

$$p_{\mathrm{i}} = \frac{e^{i\lambda}}{\sum_{j=1}^{6} e^{i\lambda}} \tag{5}$$

with λ chosen by solving the constraint

$$c=\frac{\sum_{j=1}^{6} je^{j\lambda}}{\sum_{j=1}^{6} e^{j\lambda}} \tag{6}$$

We transformed confidence distributions (1–6) to response distributions (−6 to −1 and 1–6) by assuming symmetry around 0. ***Figure 1—figure supplement 1*** shows the accuracy and confidence of generated agents.

Computational model

We employed a previously described attractor network model (***Wong and Wang, 2006***) which is itself the reduced version of an earlier one (***Wang, 2002***) inspired by the mean field theory. The model consists of two units simulating the average firing rates of two neural populations involved in information accumulation during perceptual decisions (***Figure 3a***). When the network is given inputs proportional to stimulus coherence levels, a competition breaks out between two alternative units. This race would continue until firing rates of one of the two units reaches the high-firing-rate attractor state at which point the alternative favored by the unit is chosen. The details of this model have been comprehensively described elsewhere (***Wong and Wang, 2006***).

Each unit was selective to one choice (***Equations 7; 8***) and received an input as follows:

$$x_1 = J_{N11}S_1 - J_{N12}S_2 + I_0 + I_1 + I_{noise1} \tag{7}$$

$$x_2 = J_{N22}S_2 - J_{N21}S_1 + I_0 + I_2 + I_{noise2} \tag{8}$$

where J_{N11} and J_{N22} indicated the excitatory recurrent connection of each population and J_{N12} and J_{N21} showed the mutual inhibitory connection values. For the simulation in ***Figure 3b*** we set the recurrent connections to 0.3157 nA and inhibitory ones to 0.0646 nA. I_0 indicated the effective external input which was set to 32.55 nA. I_{noise1}/I_{noise2} stood for the internal noise in each population unit. This zero mean Gaussian white noise was generated based on the time constant of 2 ms and standard deviation of 0.02 nA. I_1/I_2 indicated the input currents proportional to the motion coherence level such that:

$$I_1 = J_{A.ext}\mu_0\left(1+\frac{c}{100}\right) \tag{9}$$

$$I_2 = J_{A.ext}\mu_0\left(1-\frac{c}{100}\right) \tag{10}$$

where $J_{A.ext}$ was the average synaptic coupling from the external source and set to 0.0002243 (nA Hz^{-1}), c was coherence level and μ_0 , a.k.a. perceptual gain, was the input value when the coherence was zero (set to 45.8 Hz).

S_1 and S_2 were variables representing the synaptic current of either population and were proportional to the number of active NMDA receptors. Whenever the main text refers to accumulated evidence, we refer to S_1 and S_2 variables. Dynamics of these variables were as follows:

$$\frac{dS_1}{dt} = -\frac{S_1}{\tau_s} + \left(1-S_1\right)\gamma H\left(x_1\right) \tag{11}$$

$$\frac{dS_2}{dt} = -\frac{S_2}{\tau_s} + \left(1-S_2\right)\gamma H\left(x_2\right) \tag{12}$$

where τ_s , the NMDA receptor delay time constant, was set to 100 ms, γ set to 0.641 and the time step, dt, was set to 0.5 ms. Dynamical ***Equations 11; 12*** were solved using forward Euler method (***Wong and Wang, 2006***). (H), the generated firing rates of either populations, was calculated by:

$$H\left(x\right) = \frac{ax-b}{1-e^{-d\left(ax-b\right)}} \tag{13}$$

where a, b, and d were set to 270 Hz nA^{-1}, 108 Hz, and 0.154 s, respectively. These constants indicated the input-output relationship of a neural population.

The model's choice in each trial was defined as the accumulated evidence of either population that first touched a threshold, and the decision time was defined as the time when the threshold was touched. Notably, the decision threshold was set to $S_{threshold}$ = 0.32. Moreover, the confidence was defined as the area between two accumulators (S_1 and S_2 in ***Equations 11; 12***), in the time span of 0–500 ms, which was defined as:

$$Confidence = \left| \int_0^{500} \left(S_1 - S_2\right)\ dt \right| \tag{14}$$

which was normalized by following logistic function (***Wei and Wang, 2015***):

$$Normalized\ Confidence = b_1 + \frac{a}{e^{(kConfidence - b_0)}} \tag{15}$$

where the values of b_1 , *a*, *k*, and b_0 were set to 1.32, –0.99, 5.9, and 0.16 respectively for model on entire trials of subjects in isolated sessions; *confidence* is calculated in ***Equation 14*** in time period of [0–500]ms.

In line with previous studies, we calculated the absolute difference between accumulators (***Equation 14***; ***Wei and Wang, 2015***; ***Rolls et al., 2010***). In this formulation, confidence is calculated from model activity during the stimulus duration (***Atiya et al., 2019***). Notably, in our confidence definition, we integrated the accumulators' difference even when the winning accumulator hit the threshold (post-decision period) (***Balsdon et al., 2020***; ***Navajas et al., 2016***; ***Yu et al., 2015***). This formulation of confidence provided a successful fit to subjects' behaviors (***Figure 3—figure supplement 5***). To demonstrate that our key findings do not depend on this specific formulation, we implemented another alternative method (***Vickers, 1979***) and showed qualitatively similar results (***Figure 3—figure supplement 3***) are obtained.

We calibrated the model to the data from the isolated condition to identify the best fitting parameters that would describe the participants' behavior in isolation. In this procedure decision threshold, inhibitory and excitatory connections, NDT (set 0.27 s) and μ_0 were considered as the model variables (see ***Supplementary file 1h*** for parameter values).

In order to explain the role of social context on participant's behavior, we added a new input current to the model. Importantly we kept all other parameters of the model identical to the best fit to the participants' behavior in the isolated situation:

$$x_1 = J_{N11}S_1 - \ J_{N12}S_2 + \ I_0 + I_1 + I_{noise1} + W_x \tag{16}$$

$$x_2 = J_{N22}S_2 - \ J_{N21}S_1 + \ I_0 + I_2 + I_{noise2} + W_x \tag{17}$$

In order to evaluate the effect of W_x on the RT, accuracy, and confidence, we simulated the model while systematically varying the values of W_x (***Figure 3b***).

Having established the qualitative relevance of W_x in providing a computational hypothesis for the impact of social context, then we defined W_x proportional to the confidence of partner as follows:

Table 5. Details of statistical results for the impact of previous trial (***Figure 1—figure supplement 3***).

	Response	Regressors	Estimate	SE	CI	t-Stat	p-Value	Total number
Study 1	Accuracy (HC vs LC)	Coherency	0.007	0.0006	[0.006 0.008]	11.58	<0.001	9600
		Conf (*t*–1)	–0.0017	0.005	[–0.01 0.01]	–0.28	0.77	9600
	Confidence (HC vs LC)	Coherency	0.047	0.001	[0.045, 0.049]	54.7	<0.001	9600
		Conf (*t*–1)	0.32	0.008	[0.3 0.33]	38.31	<0.001	9600
	RT (HC vs LC)	Coherency	–0.005	0.0001	[–0.0048 0.0044]	–44.36	<0.001	9600
		Conf (*t*–1)	–0.0055	0.001	[–0.007 –0.003]	–5.44	<0.001	9600
Study 2	Accuracy (HC vs LC)	Coherency	0.02	0.002	[0.02 0.024]	13.23	<0.001	6000
		Conf (*t*–1)	0.003	0.008	[–0.012 0.018]	0.37	0.7	6000
	Confidence (HC vs LC)	Coherency	0.1	0.002	[0.097 0.0106]	47.2	<0.001	6000
		Conf (*t*–1)	0.09	0.01	[0.07 0.11]	8.6	<0.001	6000
	RT (HC vs LC)	Coherency	–0.009	0.0003	[–0.001 –0.008]	–26.2	<0.001	6000
		Condition	0.005	0.001	[0.001 0.008]	2.98	<0.01	6000

$$W_x = \alpha . C_{partner(t-1)} \tag{18}$$

where t was the trial number. The model inputs were identical to isolated situation expect for the top-down current of W_x which indicates the social input where α was a normalization factor (or coupling coefficient) and $C_{partner(t-1)}$ indicates the partner's confidence in the previous trial. Thus, we added a social input based on the linear combination of the partner's confidence in the previous trial. Importantly the model performance is not sensitive to linearity assumptions (see ***Figure 3—figure supplement 8***). Notably, the behavioral effect reported in the main script is also evident respect to the confidence of the agent in the previous trial (***Figure 1—figure supplement 3*** and ***Table 5***).

For simulations reported in ***Figure 3d***, we created high and low confident models by altering the inhibitory and excitatory connections of the original model. For the high confident model, excitatory and inhibitory connections were set to 0.3392 and 0.0699. For the low confident model excitatory and inhibitory connections were set to 0.3163 and 0.0652 respectively. For the simulation of social interaction (***Figure 4f***), we coupled two instances of the model using ***Equation 20*** with α set to –0.0008 and 0.005 for high confident and low confident models, respectively. We ran the parallel simulations 50 times and reported the average results.

In order to remove the effect of coherence levels from models' confidence, we measured the residuals of models' confidence after regressing out the impact of coherence. Using this simple regression model:

$$\text{Model Confidence} = \beta_0 + \beta_1 \text{Coh} + \epsilon \tag{19}$$

where *Coh* is the motion coherence level and ϵ is the error term, we removed the information explainable by motion coherence levels from confidence data as following. Confidence residuals were therefore:

$$\text{Confidence Residuals} = \beta_0 + \epsilon \tag{20}$$

All the simulations of model in the text – and parameters reported in the method – are related to the model calibrated on the collapsed data of all subjects (*n*=3000 for isolated sessions of study 2).

Alternative formulations for confidence in the computational model

In our main model, confidence is formalized by ***Equation 14***. We calculated the integral of difference between the losing and the winning accumulator during the stimulus presentation. This value would then be fed into a logistic function (***Equation 15***) to produce the final confidence reported by the model (***Figure 3b*** middle panel). To demonstrate the generality of our findings, we used another alternative (but similar) formulation in the previous literature for confidence representation. In ***Figure 3—figure supplement 3***, we compare the resulting 'raw' confidence values (i.e. confidence values before they are fed to ***Equation 15***).

Alternative formulations for confidence are:

1. For comparison we plot our main formulation (***Equation 14***) in ***Figure 3—figure supplement 3a***.
2. By calculating the difference between winning and losing accumulator at the END of stimulus duration (***Navajas et al., 2016***; ***Figure 3—figure supplement 3b***, we call this End method).

Our simulations showed that our formulation (***Figure 3—figure supplement 3a***) shows an expected modulation to top-down currents. ***Figure 3—figure supplement 3b*** also shows a similar pattern which indicates our results are not different from End method. Therefore, our computational results could be generalized to different confidence representation methods.

Model comparison

For model comparison, we used the fitted parameters from the isolated session (study 2 only without loss of generality). The model parameters for the isolated condition were extracted for each participants in their own respective isolated session (*n*=3000 across all participants). Then we compared all 'alternative' models with a 'single free parameter' to determine the model with the best account to behavioral data in social sessions (*n*=6000 across all participants). We considered three alternative models for the comparison. Note that in all models *a* is the normalization factor and the free parameter.

Bound model

We hypothesized that partner's confidence modulates the participant's decision boundary according to:

$$B = B_{Isolated} + aConf_{t-1} \tag{21}$$

B determines the threshold applied on the solution of the ***Equations 11; 12*** (see Materials and methods). $B_{Isolated}$ denotes the threshold in the isolated model. In this model, in social condition the bound depends on the value of the agent's confidence in the previous trial. Note that the optimum value of *a*, normalization or coupling factor, is most likely to be negative since it generates lower RTs in social vs isolated situation.

NDT model

We hypothesized that NDT would be modulated by confidence of agent in the previous trial. Here,

$$NDT = NDT_{Isolated} + aConf_{t-1} \tag{22}$$

$NDT_{Isolated}$ was the NDT fitted on the isolated data. Similarly, the optimum *a* was expected to be negative.

Gain model

We hypothesized that social information modulated the perceptual gain defined as:

$$\mu_0 = \mu_{0_{Isolated}} + aConf_{t-1} \tag{23}$$

where μ_0 denotes the input value of the model when motion coherence is zero (***Equations 9; 10***, Materials and methods) and $\mu_{0_{Isolated}}$ was calculated based on isolated data. If *a* is positive, then μ_0 would be greater under social condition vs isolated condition, which in turn generates lower RTs and higher confidence.

In order to incorporate the accuracy, RT, and confidence in model comparison, we calculated the RT distribution of trials in each of the 12 confidence levels, 6 for left decision (−6 to –1) and 6 for right decision (1–6). The RT in each level was further divided into two categories (***Ratcliff and McKoon, 2008***) (less than 700 ms and larger than 700 ms). We tried to maximize the likelihood of behavioral RT distribution in each response level (confidence and choice) given the model structure and parameters. The probability matrix was defined as follows:

$$Pmat = \left[p_i\left(RT < 700\right), p_i\left(RT > 700\right)\right] \qquad -6 \le i \le 6 \tag{24}$$

where *i* is confidence levels ranging from –6 to 6. Note, the probability was calculated based on all trials in our behavioral data set (6000 trials). The model's probability matrix was also calculated in a similar manner. Hence, we derived a probability matrix of 12 response levels and 2 RT bins. The likelihood function was defined as follows:

$$JointPmat = |Pmat_{Behave} - Pmat_{Model}| \tag{25}$$

$$Cost = \sum_{i=1}^{12}\sum_{j=1}^{2} JointPmat_{(i,j)} \tag{26}$$

Since we used similar parameters for the models (all models had one free parameter, *a*) we could directly compare cost values corresponding to each model. The model with the lowest cost is the preferred model; the parameters were found via MATLAB *fmincon* function. As is often the case, there was some variability across participants (see ***Figure 3—figure supplement 4***). To strengthen the conclusions about model comparison, we also provide evidence from a model falsification exercise that we performed. We simulated the models between two different social conditions (HCA and LCA) to see which model could, in theory, follow the behavioral pattern (***Figure 1c***). Indeed, we attempted to numerically *falsify* the alternative models. ***Figure 3—figure supplement 7*** shows the alternative model fails to reproduce the effect observed in ***Figure 1c***.

Eye monitoring and pupilometery

In both studies, the eye movements were recorded by an EyeLink 1000 (SR- Research) device with a sampling rate of 1000 Hz which was controlled by a dedicated host PC. The device was set in a desktop and pupil-corneal reflection mode while data from the left eye was recorded. At the beginning of each block, for most subjects, the system was recalibrated and then validated by 9-point schema presented on the screen. One subject was showed a 3-point schema due to the repetitive calibration difficulty. Having reached a detection error of less than 0.5°, the participants were led to the main task. Acquired eye data for pupil size were used for further analysis. Data of one subject in the first study was removed from further analysis due to storage failure.

Pupil data were divided into separate epochs and data from ITI were selected for analysis. ITI interval was defined as the time between offset of trial (*t*) feedback screen and stimulus presentation of trial (*t*+1). Then, blinks and jitters were detected and removed using linear interpolation. Values of pupil size before and after the blink were used for this interpolation. Data was also mid-pass filtered using Butterworth filter (second order, [0.01, 6] Hz) (***van Kempen et al., 2019***). The pupil data was

Table 6. The rate of trial rejection of eye tracking (only data of social) and EEG data (visual inspection) per participant.

	Participants	Eye tracking rejection % (social)	EEG trial rejection % (visual)
	1	12.25	4.6
	2	12.87	31.1
	3	0.5	22.1
	4	4	14.8
	5	1.37	34.4
	6	0	4.6
	7	7.75	8.8
	8	0.37	24.4
	9	6.37	7.6
	10	0	46
	11	0.12	NA
Study 1 (Discovery)	12	NA	NA
	1	0	4
	2	1.25	1
	3	5.75	8.5
	4	0.5	3
	5	1	16
	6	1.5	2.5
	7	0	0.5
	8	1.5	9
	9	0	2
	10	1	4
	11	1	7.5
	12	0.5	0
	13	0.75	10.5
	14	2.5	12
Study 2 (Replication)	15	14.75	4.5

z-scored and then was baseline corrected by removing the average of signal in the period of [–1000 0] ms interval (before ITI onset). Importantly, trials with ITI >3 s were excluded from analysis (365 out of 8800 for study 1 and 128 out 6000 for study 2; also see ***Table 6*** and Selection criteria for data analysis in Supplementary materials).

EEG signal recording and preprocessing

For the first study, a 32-channel eWave32 amplifier was used for recording which followed the 10–10 convention of electrode placement on the scalp (for the locations of the electrodes, see ***Figure 4—figure supplement 1***; right mastoid as the reference). The amplifier, produced by ScienceBeam (http://www.sciencebeam.com/), provided a 1 K sampling rate (***Vafaei Shooshtari et al., 2019***). For the second study we used a 64-channel amplifier produced by LIV team (http://lliivv.com/en/) with 250 Hz sampling rate (see the electrode placement in ***Figure 4—figure supplement 1***).

Raw data were analyzed using EEGLAB software (***Delorme and Makeig, 2004***). First, data were notch filtered in the range of 45–55 Hz in order to remove the line noise. Using an FIR filter in the range of 0.1–100 Hz, high-frequency noise was also removed from data. Artifacts were removed by visual inspection using information from independent component analysis. Noisy trials were also removed by avisual inspection. Noisy channels were interpolated using EEGLAB software. The signals were divided into distinct epochs aligned to stimulus presentation ranging from 100 ms pre-stimulus onset until 500 ms post-stimulus offset. After preprocessing, EEG data in the designated epochs that had higher (lower) values than 200 (–200) μV were excluded from analysis (see ***Table 6*** and Materials and methods for detailed data analysis) (***Kelly and O'Connell, 2013***). We used CP1, CP2, Cz, and Pz electrodes for further analysis. In the first study, EEG recording was not possible in two participants due to unresolvable impedance calibration problems in multiple channels.

Relation of CPP to coherence and social condition

Activities of centroparietal area of the brain is shown to be modulated with coherence level. Here, we showed that CPP activities are statistically related to the coherence levels (***Figure 4—figure supplement 2***, top-row) in both studies. Furthermore, we tested how much this relationship is dependent to social condition (HCA, LCA, ***Figure 4—figure supplement 2***, bottom-row). Our analysis showed that the slope (respect to coherence levels) is different in HCA vs LCA (also see ***Table 6***). Notably, this effect is in line with our neural model prediction (see ***Figure 4—figure supplement 3***, next section).

Selection criteria for data analysis

The data included in both studies could be classified into three main categories: behavioral, eye tracking, and EEG. For the behavioral analysis, data from all participants were included. In study 1, eye tracking data from one participant was lost due to storage failure. For pupil analysis, we excluded the trials with ITI longer than 3 s (~4% of trials in study 1 and ~2% for study 2).

Table 7. Generalized linear mixed model (GLMM) including interaction terms (p-values are reported).

	Response	Coherence	Condition (LC vs HC)	Condition* coherence
Study 1	Accuracy	p<0.001	p=0.92	p=0.96
	Confidence	p<0.001	p<0.001	p<0.001
	RT	p<0.001	p<0.001	p<0.05
	Pupil	p=0.43	p=0.20	p=0.31
	EEG slope	p<0.01	p=0.15	p=0.91
Study 2	Accuracy	p<0.001	p=0.75	p=0.87
	Confidence	p<0.001	p<0.001	p<0.001
	RT	p<0.001	p<0.001	p=0.34
	Pupil	p=0.35	p=0.06	p=0.17
	EEG slope	p=0.62	p<0.05	p=0.68

Table 8. Attractor model's parameters.

Parameter	Parameter value	Reference, remarks
JN,ii	0.3157 nA	Calibrated based on pool of isolated data, also fitted on individual subjects' data
JN,ij	0.0646 nA	Calibrated based on pool of isolated data, also fitted on individual subjects' data
μ_0	45.8 Hz	Calibrated based on pool of isolated data, also fitted on individual subjects' data
NDT	0.27 s	Calibrated based on pool of isolated data, also fitted on individual subjects' data
Bound	0.32 nA	Calibrated based on pool of isolated data, also fitted on individual subjects' data
a (***Equation 15***)	–0.99	Calibrated based on pool of isolated data, also fitted on individual subjects' data
b_0 (***Equation 15***)	1.32	Calibrated based on pool of isolated data, also fitted on individual subjects' data
b_1 (***Equation 15***)	–0.165	Calibrated based on pool of isolated data, also fitted on individual subjects' data
k (***Equation 15***)	5.9	Calibrated based on pool of isolated data, also fitted on individual subjects' data
I_0	0.3255 nA	From ***Wang, 2002***; ***Wong and Wang, 2006***
$J_{A,ext}$	0.00022 nA Hz^{-1}	From ***Wang, 2002***; ***Wong and Wang, 2006***
τ_s	0.1 s	From ***Wang, 2002***; ***Wong and Wang, 2006***
dt	0.0005 s	From ***Wang, 2002***; ***Wong and Wang, 2006***
a (***Equation 13***)	270 (V nC)$^{-1}$	From ***Wang, 2002***; ***Wong and Wang, 2006***
b (***Equation 13***)	108 Hz	From ***Wang, 2002***; ***Wong and Wang, 2006***
d (***Equation 13***)	0.154 s	From ***Wang, 2002***; ***Wong and Wang, 2006***
γ	0.641	From ***Wang, 2002***; ***Wong and Wang, 2006***
Noise_std	0.025	From ***Wang, 2002***; ***Wong and Wang, 2006***
I_noise	0.02	From ***Wang, 2002***; ***Wong and Wang, 2006***

We also analyzed brain data of participants in both studies. For the ERP analysis, we excluded trials with an absolute amplitude greater than 200 microvolts (overall less than 1% for both trials) as this data was deemed as outlier. Moreover, noisy trials and ICA components (around 5% of components in study 2) were rejected by visual inspection. Noisy electrodes were also interpolated (~8% of electrodes in study 2); see ***Table 6*** for more details. In study 1, EEG data from two participants were lost due to a technical failure. All data (behavioral, eye tracking, and EEG) for study 2 were properly stored, saved, and made available at https://github.com/JimmyEsmaily/ConfMatch (copy archived at ***Esmaily, 2023***; ***MathWorks Inc, 2023***).

Statistical analysis

For hypothesis testing, we employed a number of GLMM. Unless otherwise stated, in our mixed models, participant was considered as random intercept. Details of each model is described in ***Tables 1–6*** in the Supplementary materials. This approach enabled us to separate the effects of coherency and partner confidence. For RT and confidence, we assumed that the data is normality distributed. For the accuracy data we assumed the distribution is Poisson. We used a maximum likelihood method for fitting. All p-values reported in the text were drawn from the GLMM method, unless stated otherwise. For completeness, for each analysis we have added interaction terms as well (see ***Tables 7 and 8***).

Permutation test to confirm confidence matching

A key null hypothesis ($\mathrm{p}(\vartheta)$ where ϑ is the measure of interest: confidence matching) that we ruled out was that confidence matching was forced by the experimental design limitations and, therefore, would be observed in any random pairing of participants within our joint decision making setup. To reject this hypothesis, we performed a permutation test following ***Bang et al., 2017*** (see their Supplementary Figure 3 for further details). For each participant and corresponding CGP pair, we defined $|c_1 – c_2|$ where c_i is the average confidence of participant *i* in a given pair. We then estimated the null distribution for this variable by randomly re-pairing the participant with other participants and computing the mean confidence matching for each such re-paired set (total number of sets 1000). In ***Figure 1—figure supplement 2*** (bottom row), the red line shows the empirically observed mean of confidence matching in our data. The null distribution is shown in black. Proportion of values from the null distribution that were less than the empirical mean was P~0.

In addition, we defined an index for measuring the confidence matching (***Figure 1—figure supplement 2***, first row): $\Delta m = \left| C_{isolated(Subject)} - C_{agent} \right| - \left| C_{social(Subject)} - C_{agent} \right|$. The larger the Δm the higher is the confidence matching. Although we did not observe a significant effect of Δm, we showed that this index is significantly different from zero in the HCA condition.

Acknowledgements

JE and BB were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (819040 – acronym: rid-O). BB was supported by the NOMIS foundation and Templeton Religion Trust.

Additional information

Funding

Funder	Grant reference number	Author
European Research Council	819040 - acronym: rid-O	Jamal Esmaily

Open access funding provided by Max Planck Society. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Jamal Esmaily, Conceptualization, Data curation, Software, Formal analysis, Validation, Visualization, Methodology, Writing – original draft, Writing – review and editing; Sajjad Zabbah, Conceptualization, Data curation, Software, Methodology, Writing – original draft, Writing – review and editing; Reza Ebrahimpour, Conceptualization, Software, Supervision, Funding acquisition, Validation, Investigation, Methodology, Project administration, Writing – review and editing; Bahador Bahrami, Conceptualization, Resources, Supervision, Funding acquisition, Validation, Investigation, Visualization, Writing – original draft, Project administration, Writing – review and editing

Author ORCIDs

Jamal Esmaily https://orcid.org/0000-0001-5529-6732
Reza Ebrahimpour https://orcid.org/0000-0002-7013-8078
Bahador Bahrami https://orcid.org/0000-0003-0802-5328

Ethics

Human subjects: Both experiments were approved by the local ethics committee at Faculty of computer engineering at Shahid Rajeie and also Iran University in Tehran, Iran (ethics application approval date and/or number: 5769). Written informed consent was obtained from all participants. The consent form was in the local Farsi language and did not include a "consent to publish" because data were anonymised, individual identity information was completely removed from the them and none of the experimental hypotheses involved the exact identification of the individual data from any participant. Participants received a fixed monetary compensation for their contribution.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.83722.sa1
Author response https://doi.org/10.7554/eLife.83722.sa2

Additional files

Supplementary files

- MDAR checklist
- Supplementary file 1. This file contains supplementary tables that contains details of statistical analysis.

Data availability

Data that supports the findings of the study can be found here: https://osf.io/v7fqz/.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Bahrami B, Esmaily J	2020	Neurobiology of Confidence Matching	https://osf.io/v7fqz/	Open Science Framework, v7fqz

References

Adler WT, Ma WJ. 2018. Limitations of proposed signatures of Bayesian confidence. *Neural Computation* **30**:3327–3354. DOI: https://doi.org/10.1162/neco_a_01141, PMID: 30314423

Ais J, Zylberberg A, Barttfeld P, Sigman M. 2016. Individual consistency in the accuracy and distribution of confidence judgments. *Cognition* **146**:377–386. DOI: https://doi.org/10.1016/j.cognition.2015.10.006, PMID: 26513356

Aitchison L, Bang D, Bahrami B, Latham PE. 2015. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLOS Computational Biology* **11**:e1004519. DOI: https://doi.org/10.1371/journal.pcbi.1004519, PMID: 26517475

Atiya NAA, Rañó I, Prasad G, Wong-Lin KF. 2019. A neural circuit model of decision uncertainty and change-of-mind. *Nature Communications* **10**:2287. DOI: https://doi.org/10.1038/s41467-019-10316-8, PMID: 31123260

Austen-Smith D, Banks JS. 1996. Information aggregation, rationality, and the condorcet Jury Theorem. *American Political Science Review* **90**:34–45. DOI: https://doi.org/10.2307/2082796

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. 2010. Optimally Interacting Minds. *Science* **329**:1081–1085. DOI: https://doi.org/10.1126/science.1185718

Bahrami B, Olsen K, Bang D, Roepstorff A, Rees G, Frith C. 2012. What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B* **367**:1350–1365. DOI: https://doi.org/10.1098/rstb.2011.0420

Baird B, Smallwood J, Gorgolewski KJ, Margulies DS. 2013. Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of Neuroscience* **33**:16657–16665. DOI: https://doi.org/10.1523/JNEUROSCI.0786-13.2013, PMID: 24133268

Balsdon T, Wyart V, Mamassian P. 2020. Confidence controls perceptual evidence accumulation. *Nature Communications* **11**:1753. DOI: https://doi.org/10.1038/s41467-020-15561-w, PMID: 32273500

Bang D, Aitchison L, Moran R, Herce Castanon S, Rafiee B, Mahmoodi A, Lau JYF, Latham PE, Bahrami B, Summerfield C. 2017. Confidence matching in group decision-making. *Nature Human Behaviour* **1**:0117. DOI: https://doi.org/10.1038/s41562-017-0117

Bang D, Fleming SM. 2018. Distinct encoding of decision confidence in human medial prefrontal cortex. *PNAS* **115**:6082–6087. DOI: https://doi.org/10.1073/pnas.1800795115

Bang D, Ershadmanesh S, Nili H, Fleming SM. 2020. Private-public mappings in human prefrontal cortex. *eLife* **9**:e56477. DOI: https://doi.org/10.7554/eLife.56477, PMID: 32701449

Bang D, Moran R, Daw ND, Fleming SM. 2022. Neurocomputational mechanisms of confidence in self and others. *Nature Communications* **13**:4238. DOI: https://doi.org/10.1038/s41467-022-31674-w, PMID: 35869044

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* **113**:700–765. DOI: https://doi.org/10.1037/0033-295X.113.4.700, PMID: 17014301

Brainard DH. 1997. The Psychophysics Toolbox. *Spatial Vision* **10**:433–436 PMID: 9176952.

Breveglieri R, Galletti C, Dal Bò G, Hadjidimitrakis K, Fattori P. 2014. Multiple aspects of neural activity during reaching preparation in the medial posterior parietal area V6A. *Journal of Cognitive Neuroscience* **26**:878–895. DOI: https://doi.org/10.1162/jocn_a_00510, PMID: 24168224

Delorme A, Makeig S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods* **134**:9–21. DOI: https://doi.org/10.1016/j.jneumeth.2003.10.009, PMID: 15102499

De Martino B, Bobadilla-Suarez S, Nouguchi T, Sharot T, Love BC. 2017. Social information is integrated into value and confidence judgments according to its reliability. *The Journal of Neuroscience* **37**:6066–6074. DOI: https://doi.org/10.1523/JNEUROSCI.3880-16.2017, PMID: 28566360

Dikker S, Silbert LJ, Hasson U, Zevin JD. 2014. On the same wavelength: predictable language enhances speaker-listener brain-to-brain synchrony in posterior superior temporal gyrus. *The Journal of Neuroscience* **34**:6267–6272. DOI: https://doi.org/10.1523/JNEUROSCI.3796-13.2014, PMID: 24790197

Eldar E, Cohen JD, Niv Y. 2013. The effects of neural gain on attention and learning. *Nature Neuroscience* **16**:1146–1153. DOI: https://doi.org/10.1038/nn.3428, PMID: 23770566

Esmaily J. 2023. Confmatch. https://archive.softwareheritage.org/swh:1:dir:a3de79d590798c99a8dadd948ab6f2eb268304c3;origin=https://github.com/JimmyEsmaily/ConfMatch;visit=swh:1:snp:1dfaeff8fcb5056fbcd830194ce0dc527324a064;anchor=swh:1:rev:2adfb563eb24c137213d7163754d9e146fa42c50. Software Heritage.

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. 2010. Relating introspective accuracy to individual differences in Brain Structure. *Science* **329**:1541–1543. DOI: https://doi.org/10.1126/science.1191883

Fleming SM, Daw ND. 2017. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review* **124**:91–114. DOI: https://doi.org/10.1037/rev0000045, PMID: 28004960

Friston K, Frith C. 2015. A Duet for one. *Consciousness and Cognition* **36**:390–405. DOI: https://doi.org/10.1016/j.concog.2014.12.003, PMID: 25563935

Frith CD, Frith U. 1999. Interacting Minds--A Biological Basis. *Science* **286**:1692–1695. DOI: https://doi.org/10.1126/science.286.5445.1692

Gallotti M, Frith CD. 2013. Social cognition in the we-mode. *Trends in Cognitive Sciences* **17**:160–165. DOI: https://doi.org/10.1016/j.tics.2013.02.002, PMID: 23499335

Gold JI, Shadlen MN. 2007. The neural basis of decision making. *Annual Review of Neuroscience* **30**:535–574. DOI: https://doi.org/10.1146/annurev.neuro.29.051605.113038, PMID: 17600525

Hanks TD, Summerfield C. 2017. Perceptual decision making in Rodents, Monkeys, and Humans. *Neuron* **93**:15–31. DOI: https://doi.org/10.1016/j.neuron.2016.12.003

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. 2004. Intersubject synchronization of cortical activity during natural vision. *Science* **303**:1634–1640. DOI: https://doi.org/10.1126/science.1089506

Hasson U, Honey CJ. 2012. Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage* **62**:1272–1278. DOI: https://doi.org/10.1016/j.neuroimage.2012.02.004, PMID: 22348879

Hasson U, Frith CD. 2016. Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society B* **371**:20150366. DOI: https://doi.org/10.1098/rstb.2015.0366

Heath RA. 1984. Random-walk and accumulator models of psychophysical discrimination: A critical evaluation. *Perception* **13**:57–65. DOI: https://doi.org/10.1068/p130057, PMID: 6473053

Honey CJ, Thesen T, Donner TH, Silbert LJ, Carlson CE, Devinsky O, Doyle WK, Rubin N, Heeger DJ, Hasson U. 2012. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**:423–434. DOI: https://doi.org/10.1016/j.neuron.2012.08.011, PMID: 23083743

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. DOI: https://doi.org/10.1038/nn.3017

Iacoboni M. 2009. Imitation, empathy, and mirror neurons. *Annual Review of Psychology* **60**:653–670. DOI: https://doi.org/10.1146/annurev.psych.60.110707.163604, PMID: 18793090

Kelly SP, O'Connell RG. 2013. Internal and external influences on the rate of sensory evidence accumulation in the human brain. *The Journal of Neuroscience* **33**:19434–19441. DOI: https://doi.org/10.1523/JNEUROSCI.3355-13.2013, PMID: 24336710

Kiani R, Shadlen MN. 2009. Representation of confidence associated with a decision by Neurons in the Parietal Cortex. *Science* **324**:759–764. DOI: https://doi.org/10.1126/science.1169405

Kiani R, Corthell L, Shadlen MN. 2014. Choice certainty is informed by both evidence and decision time. *Neuron* **84**:1329–1342. DOI: https://doi.org/10.1016/j.neuron.2014.12.015, PMID: 25521381

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C. 2007. What's new in psychtoolbox-3. *Perception* **36**:1–16.

Kloosterman NA, Meindertsma T, van Loon AM, Lamme VAF, Bonneh YS, Donner TH. 2015. Pupil size tracks perceptual content and surprise. *The European Journal of Neuroscience* **41**:1068–1078. DOI: https://doi.org/10.1111/ejn.12859, PMID: 25754528

Konvalinka I, Vuust P, Roepstorff A, Frith CD. 2010. Follow you, Follow me: continuous mutual prediction and adaptation in Joint Tapping. *Quarterly Journal of Experimental Psychology* **63**:2220–2230. DOI: https://doi.org/10.1080/17470218.2010.497843

Krajbich I, Rangel A. 2011. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *PNAS* **108**:13852–13857. DOI: https://doi.org/10.1073/pnas.1101328108, PMID: 21808009

Lee DG, Daunizeau J. 2021. Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife* **10**:e63282. DOI: https://doi.org/10.7554/eLife.63282, PMID: 33900198

Li V, Michael E, Balaguer J, Herce Castañón S, Summerfield C. 2018. Gain control explains the effect of distraction in human perceptual, cognitive, and economic decision making. *PNAS* **115**:E8825–E8834. DOI: https://doi.org/10.1073/pnas.1805224115, PMID: 30166448

Loughnane GM, Newman DP, Tamang S, Kelly SP, O'Connell RG. 2018. Antagonistic interactions between Microsaccades and evidence accumulation processes during decision formation. *The Journal of Neuroscience* **38**:2163–2176. DOI: https://doi.org/10.1523/JNEUROSCI.2340-17.2018, PMID: 29371320

Mahmoodi A, Bang D, Olsen K, Zhao YA, Shi Z, Broberg K, Safavi S, Han S, Nili Ahmadabadi M, Frith CD, Roepstorff A, Rees G, Bahrami B. 2015. Equality bias impairs collective decision-making across cultures. *PNAS* **112**:3835–3840. DOI: https://doi.org/10.1073/pnas.1421692112, PMID: 25775532

MathWorks Inc. 2023. Data analysis was performed using MATLAB. R2016a. GitHub. https://www.mathworks.com/products/matlab/data

Moran R, Teodorescu AR, Usher M. 2015. Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology* **78**:99–147. DOI: https://doi.org/10.1016/j.cogpsych.2015.01.002, PMID: 25868113

Mukamel R, Gelbard H, Arieli A, Hasson U, Fried I, Malach R. 2005. Coupling between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. *Science* **309**:951–954. DOI: https://doi.org/10.1126/science.1110913

Murphy PR, Vandekerckhove J, Nieuwenhuis S. 2014. Pupil-linked arousal determines variability in perceptual decision making. *PLOS Computational Biology* **10**:e1003854. DOI: https://doi.org/10.1371/journal.pcbi.1003854, PMID: 25232732

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI. 2012. Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience* **15**:1040–1046. DOI: https://doi.org/10.1038/nn.3130, PMID: 22660479

Navajas J, Bahrami B, Latham PE. 2016. Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences* **11**:55–60. DOI: https://doi.org/10.1016/j.cobeha.2016.05.005

Navajas J, Hindocha C, Foda H, Keramati M, Latham PE, Bahrami B. 2017. The idiosyncratic nature of confidence. *Nature Human Behaviour* **1**:810–818. DOI: https://doi.org/10.1038/s41562-017-0215-1, PMID: 29152591

O'Connell RG, Shadlen MN, Wong-Lin KF, Kelly SP. 2018. Bridging Neural and computational viewpoints on perceptual Decision-Making. *Trends in Neurosciences* **41**:838–852. DOI: https://doi.org/10.1016/j.tins.2018.06.005

Patel D, Fleming SM, Kilner JM. 2012. Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B* **279**:4853–4860. DOI: https://doi.org/10.1098/rspb.2012.1847

Pelli DG. 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* **10**:437–442 PMID: 9176953.

Pouget A, Drugowitsch J, Kepecs A. 2016. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience* **19**:366–374. DOI: https://doi.org/10.1038/nn.4240, PMID: 26906503

Ratcliff R, McKoon G. 2008. Drift Diffusion Decision Model: Theory and data. *Neural Computation* **20**:873–922. DOI: https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Rendell L, Fogarty L, Hoppitt WJE, Morgan TJH, Webster MM, Laland KN. 2011. Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences* **15**:68–76. DOI: https://doi.org/10.1016/j.tics.2010.12.002, PMID: 21215677

Resulaj A, Kiani R, Wolpert DM, Shadlen MN. 2009. Changes of mind in decision-making. *Nature* **461**:263–266. DOI: https://doi.org/10.1038/nature08275, PMID: 19693010

Rolls ET, Grabenhorst F, Deco G. 2010. Decision-Making, Errors, and Confidence in the Brain. *Journal of Neurophysiology* **104**:2359–2374. DOI: https://doi.org/10.1152/jn.00571.2010

Ruff CC, Fehr E. 2014. The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience* **15**:549–562. DOI: https://doi.org/10.1038/nrn3776

Sanders JI, Hangya B, Kepecs A. 2016. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* **90**:499–506. DOI: https://doi.org/10.1016/j.neuron.2016.03.025, PMID: 27151640

Schall JD. 2019. Accumulators, Neurons, and Response Time. *Trends in Neurosciences* **42**:848–860. DOI: https://doi.org/10.1016/j.tins.2019.10.001, PMID: 31704180

Shadlen MN, Newsome WT. 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* **86**:1916–1936. DOI: https://doi.org/10.1152/jn.2001.86.4.1916, PMID: 11600651

Silbert LJ, Honey CJ, Simony E, Poeppel D, Hasson U. 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *PNAS* **111**:E4687–E4696. DOI: https://doi.org/10.1073/pnas.1323812111, PMID: 25267658

Sinanaj I, Cojan Y, Vuilleumier P. 2015. Inter-individual variability in metacognitive ability for visuomotor performance and underlying brain structures. *Consciousness and Cognition* **36**:327–337. DOI: https://doi.org/10.1016/j.concog.2015.07.012, PMID: 26241023

Sorkin RD, Hays CJ, West R. 2001. Signal-detection analysis of group decision making. *Psychological Review* **108**:183–203. DOI: https://doi.org/10.1037/0033-295x.108.1.183, PMID: 11212627

Stallen M, Sanfey AG. 2015. The neuroscience of social conformity: implications for fundamental and applied research. *Frontiers in Neuroscience* **9**:337. DOI: https://doi.org/10.3389/fnins.2015.00337, PMID: 26441509

Stine GM, Zylberberg A, Ditterich J, Shadlen MN. 2020. Differentiating between integration and non-integration strategies in perceptual decision making. *eLife* **9**:e55365. DOI: https://doi.org/10.7554/eLife.55365, PMID: 32338595

Stolk A, Verhagen L, Schoffelen J-M, Oostenveld R, Blokpoel M, Hagoort P, van Rooij I, Toni I. 2013. Neural mechanisms of communicative innovation. *PNAS* **110**:14574–14579. DOI: https://doi.org/10.1073/pnas.1303170110, PMID: 23959895

Stolk A, D'Imperio D, di Pellegrino G, Toni I. 2015. Altered communicative decisions following ventromedial prefrontal lesions. *Current Biology* **25**:1469–1474. DOI: https://doi.org/10.1016/j.cub.2015.03.057, PMID: 25913408

Stolk A, Verhagen L, Toni I. 2016. Conceptual Alignment: How Brains Achieve Mutual Understanding. *Trends in Cognitive Sciences* **20**:180–191. DOI: https://doi.org/10.1016/j.tics.2015.11.007, PMID: 26792458

Twomey DM, Kelly SP, O'Connell RG. 2016. Abstract and Effector-Selective Decision Signals Exhibit Qualitatively Distinct Dynamics before Delayed Perceptual Reports. *The Journal of Neuroscience* **36**:7346–7352. DOI: https://doi.org/10.1523/JNEUROSCI.4162-15.2016, PMID: 27413146

Urai AE, Braun A, Donner TH. 2017. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications* **8**:14637. DOI: https://doi.org/10.1038/ncomms14637, PMID: 28256514

Vafaei Shooshtari S, Esmaily Sadrabadi J, Azizi Z, Ebrahimpour R. 2019. Confidence Representation of Perceptual Decision by EEG and Eye Data in a Random Dot Motion Task. *Neuroscience* **406**:510–527. DOI: https://doi.org/10.1016/j.neuroscience.2019.03.031, PMID: 30904664

van Kempen J, Loughnane GM, Newman DP, Kelly SP, Thiele A, O'Connell RG, Bellgrove MA. 2019. Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *eLife* **8**:e42541. DOI: https://doi.org/10.7554/eLife.42541, PMID: 30882347

Vickers D. 1970. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics* **13**:37–58. DOI: https://doi.org/10.1080/00140137008931117, PMID: 5416868

Vickers D. 1979. *Decision Processes in Visual Perception* ACADEMIC PRESS INC.

Wang XJ. 2002. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**:955–968. DOI: https://doi.org/10.1016/s0896-6273(02)01092-9, PMID: 12467598

Wei Z, Wang XJ. 2015. Confidence estimation as a stochastic process in a neurodynamical system of decision making. *Journal of Neurophysiology* **114**:99–113. DOI: https://doi.org/10.1152/jn.00793.2014, PMID: 25948870

Wheatley T, Boncz A, Toni I, Stolk A. 2019. Beyond the Isolated Brain: The Promise and Challenge of Interacting Minds. *Neuron* **103**:186–188. DOI: https://doi.org/10.1016/j.neuron.2019.05.009, PMID: 31319048

Wong K-F, Wang X-J. 2006. A recurrent network mechanism of time integration in perceptual decisions. *The Journal of Neuroscience* **26**:1314–1328. DOI: https://doi.org/10.1523/JNEUROSCI.3733-05.2006, PMID: 16436619

Yeung N, Summerfield C. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B* **367**:1310–1321. DOI: https://doi.org/10.1098/rstb.2011.0416

Yu S, Pleskac TJ, Zeigenfuse MD. 2015. Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology. General* **144**:489–510. DOI: https://doi.org/10.1037/xge0000062, PMID: 25844627

4 Project (2)

"Anyone who has never made a mistake has never learned anything new."

— *Albert Einstein*

This chapter includes the research article "Sequential sampling without comparison to boundary through model-free reinforcement learning." This article investigates the computational mechanisms underlying perceptual decision-making during the training phase. We introduced a model-free reinforcement learning (RL) model capable of making perceptual decisions. In our framework, the state is defined as the accumulated evidence up to the current time step. The actions include: choosing "left," "right," or "wait." The reward structure provides positive rewards for correct decisions, negative rewards for incorrect ones, and a small penalty for waiting.

The objective of the RL agent is to learn the optimal policy that determines when to terminate the decision process (by choosing left or right), effectively balancing external reward (accuracy) with the internal cost of waiting. Unlike traditional decision-making models that rely on explicit decision boundaries, our model can arrive at decisions with an implicit boundary learned through experience.

We conducted extensive analyses and tested our model across various scenarios. Our model successfully replicated hallmark behavioural signatures of standard decision-making paradigms, as well as several additional scenarios explored in the literature. Together, this work offers new insights into the often-overlooked initial learning phase of perceptual decision-making, a phase typically discarded or underexplored in existing studies.

DOI: https://arxiv.org/abs/2408.06080

Contributions:

Jamal Esmaily (JE), Rani Moran (RM), Yasser Roudi (YR), Bahador Bahrami (BB)

The author of this thesis is the first author of this manuscript.

JE contributions: Conceptualization, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing – original draft, Writing – review and editing.

RM contribution: Conceptualization, Supervision, Methodology, Validation, Writing – original draft, Writing – review and editing.

YR contribution: Conceptualization, Supervision, Methodology, Validation, Investigation, Writing – original draft, review and editing.

BB contribution: Conceptualization, Resources, Supervision, Funding acquisition, Validation, Investigation, Visualization, Writing – original draft, Project administration, Writing – review and editing.

Sequential sampling without comparison to boundary through model-free reinforcement learning

Jamal Esmaily*[1,2], Rani Moran[3,4,5], Yasser Roudi[6,7], and Bahador Bahrami*[1]

[1]Department of General Psychology and Education, Ludwig Maximilians University Munich, Munich, Germany
[2]Graduate School of Systemic Neurosciences, Ludwig Maximilians University Munich, Munich, Germany
[3]Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, Queen Square Institute of Neurology, London, UK
[4]Wellcome Centre for Human Neuroimaging, University College London, Queen Square Institute of Neurology, London, UK
[5]School of Biological and Behavioural Sciences, Queen Mary University of London, UK
[6]Kavli Institute for Systems Neuroscience, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[7]Department of Mathematics, King's College London, London, United Kingdom

Abstract

Although evidence integration to the boundary model has successfully explained a wide range of behavioral and neural data in decision making under uncertainty, how animals learn and optimize the boundary remains unresolved. Here, we propose a model-free reinforcement learning algorithm for perceptual decisions under uncertainty that dispenses entirely with the concepts of decision boundary and evidence accumulation. Our model learns whether to commit to a decision given the available evidence or continue sampling information at a cost. We reproduced the canonical features of perceptual decision-making such as dependence of accuracy and reaction time on evidence strength, modulation of speed-accuracy trade-off by payoff regime, and many others. By unifying learning and decision making within the same framework, this model can account for unstable behavior during training as well as stabilized post-training behavior, opening the door to revisiting the extensive volumes of discarded training data in the decision science literature.

1 Introduction

Sequential sampling models have had great success in explaining the decision dynamics that govern the relationship between choice reaction time and accuracy under a variety of conditions spanning perceptual [1, 2, 3] , value-based [4, 5] and even moral decisions [6]. The general principle of these models is that to make the best of the noisy, uncertainty-ridden information that an agent (e.g., rodent, monkey, human, etc) gets from its environment, one could accumulate the sequentially arriving noisy samples across time and compare the sum to a certain designated decision criterion. These models have been instrumental in interpreting the neurophysiological investigations of the mechanisms of decision making in humans [7, 8], and non-human animals [1, 9].

These models have often been applied to empirical data collected *after* extensive training when performance has already stabilized at a predefined benchmark level and is unlikely to change with more practice. However, a number of previous works examined the evolution of the drift-diffusion model (DDM) parameters in the course of learning [10, 11, 12, 13, 14, 15, 16]. Typically, these studies fit an instance of DDM to the empirically observed reaction time and choice data across the stages of

*Corresponding authors: jimi.esmaily@gmail.com, bbahrami@gmail.com learning. These results show that decision bounds decrease and/or drift rates increase during learning without explaining what mechanism could be implementing these changes. However, sequential sampling models are often kept agnostic about the process of how the agent comes to learn the decision criteria in the first place leaving open a question of how drift rates and/or thresholds are updated during learning.

With adequate opportunity for practice, agents' behavior stabilizes in a given experimental context, for example under a specific payoff structure that relates choices to accuracy and rewards. Many studies have shown that agents are capable of changing their decision strategy, trading off speed and accuracy, when payoff structure is changed (see [17] for a review; [18, 19];). Similar findings have documented decisional flexibility under other types of changes of context such as frequency of choice categories [20] or the asymmetric physical effort needed for executing different choice options [21]. These behavioral adjustments to change of context have been attributed to the dynamic shifting of the decision boundary raising theoretical and empirical problems.

Theoretically, once the agent has learned a given decision boundary, it is not clear through what mechanism the boundary is subsequently adjusted in response to new contexts. A number of previous studies built on principles of model-based Reinforcement Learning (RL) [22, 23] to derive normative, "ideal-observer" solutions for what the boundary should be in order to maximize rewards or reward-rates, considering cognitive and opportunity costs associated with postponing a decision. These solutions assume that the agent has a "model" representing the statistical structure of the environment (e.g., the distribution of task difficulty). This model provides a transition structure that predicts the prospective amounts of evidence that would be accumulated if the decision were to be postponed. Based on such transition structures the agent can derive the optimal, reward-maximizing choice-threshold(s). This reward optimization problem was solved using the Markov decision process in one case [22] and Dynamic programming in another [23]. The key advantage of such model-based solutions is that they are highly flexible in that when the environment changes, they can quickly update their transition structure and readily recalculate the optimal choice threshold, without needing any elaborate experience with the new environment. However, a limitation of this model-based approach is that it relies on complex calculations that require a deep knowledge of the environment and the task at hand. It is unclear whether animals have access to such knowledge and can perform such demanding calculations. As a result, these approaches leave open the question of how to learn the boundary when such knowledge is not (yet) available.

Empirically, the search for the neurobiological correlates of context-dependent shifting of decision boundaries has faced considerable difficulty. Numerous studies have identified specific neurons in the prefrontal and parietal cortex of various animals that show the hallmarks of the accumulation process during perceptual decision making under uncertainty. These neurons's activity rises during the period of stimulus observation. The rate of this rise is proportional to the strength of the sensory signal and reflects an accumulation of noisy evidence. This is at least the case in macaque area LIP where in trials with similar reaction times, these neurons reach a stereotyped firing rate shortly before action initiation; see [1, 3, 9, 24] for comprehensive review. The boundary shift hypothesis would predict that these stereotyped firing rates should covary with behavioral changes observed in speed-accuracy trade-off. This does not seem to be the case. Instead, Heitz et al. [17] observed several heterogeneous phenomena (e.g., changes of baseline firing rate, sensory gain, and the duration of perceptual processing) in the activity of boundary neurons in the macaque monkey's frontal eye field during speed-accuracy trade-off. They noted that these observations were quite distinct from and in some cases even contradictory to the elegant and parsimonious predictions of a shift in the activity of the boundary neurons. Another study by Hank et al [24] examined the neural activity in the macaque LIP with the hypothesis that this bound changes dynamically in response to different speed-accuracy trade-off conditions. They observed different results: the terminating threshold levels of neural activity were similar across all regimes even though the animal behavior adjusted dynamically to the different regimes.

To address these problems, here we introduce a theory for learning how to make perceptual decisions under uncertainty based on model-free (temporal difference) RL principles. Our approach is simple and minimizes the required foreknowledge of the statistical structure of the environment compared to the model-based approaches discussed before. Our model also employs simpler calculations, but likely at the expense of flexibility [25]. Most importantly, our model dispenses with the concept of a decision boundary altogether.

While standard RL models employ two actions corresponding to the two choice alternatives, our model adds to this an overtly simple innovation: besides the standard two actions, a Wait action permits the agent to *stay undecided* and continue sampling the environment, albeit at a cost. We show that this minimal innovation creates a fundamentally new type of sequential sampling model that learns to make decisions under uncertainty and could dynamically change its strategy in response to changes of environment context. We demonstrate that this simple model reproduces the hallmark features of much more sophisticated evidence accumulation models.

2 Results

2.1 The setup

In Fig. 1, we show a schema of the model tailored to the two-alternative – left vs right decisions– random dot motion discrimination paradigm; see [2, 3] for more details about the the task. In this model, time is defined from the agent's perspective. This description is agnostic to distinct notions of time (e.g., trial number, block number, trial onset, ...) that are meaningful only to the experimenter who is studying the agent. To simplify communication and avoid misunderstandings, in Fig. 1 and the model description that follows, we make this distinction explicit. We use separate notations to refer to the time – that we consider to be discrete – at the level of trial number and specific moments of time *within* each trial, denoted by u and t, respectively.

We consider a Q-learning agent trained over U trials. In each trial u, a random dot motion stimulus is presented whose coherence level c^u is sampled uniformly and independently from the set $\mathcal{C} \subseteq [-1, 1]$. Positive values of coherence indicate rightward motion and negative values indicate leftward motion. At any given time, the agent can choose an action from the set $\mathcal{A} = \{\text{Right}, \text{Left}, \text{Wait}\}$. If either actions Right or Left are chosen, then the agent receives a reward R_{correct} or R_{wrong} depending on whether the decision was correct or wrong, the current trial is terminated and a new trial begins. If the agent chooses the action Wait, it receives R_{wait} and the trial continues. Although for simplicity we refer to all these feedbacks as rewards, their values can indeed be negative and thus embody a cost. In what follows we assume $R_{\text{wrong}} < 0$, $R_{\text{wait}} < 0$ and $R_{\text{correct}} > 0$ unless stated otherwise; we often denote the set of reward values as $\mathbf{R} \equiv [R_{\text{correct}}, R_{\text{wrong}}, R_{\text{wait}}]$. Unless stated otherwise, $\mathbf{R} \equiv [R_{\text{correct}} = 20, R_{\text{wrong}} = -50, R_{\text{wait}} = -1]$

The agent is endowed with a set of states $\mathcal{S} = \{-M, -M+\Delta \cdots, 0, \cdots, M-\Delta, M\}$. Here Δ indicates the resolution of the state of the model. At time t in trial u, the state of the system is denoted by s^u_t system. At the beginning of each trial the system starts at state 0, that is $s^u_0 = 0$. As time progresses from t to $t + \delta t$ within a trial, where δt is the time step, that is while the agent chooses the Wait action, the state of the system is updated as

$$s^u_{t+\delta t} = \lfloor (s^u_t + E_t)\delta t \rceil_{\mathcal{S}}, \tag{1}$$

where $\lfloor x \rceil_{\mathcal{S}}$ indicated the closest element of $\mathcal{S}$ to x, $E_t \sim \mathcal{N}(Kc^u, \sigma^2)$ models noisy sensory evidence received at time t and is taken to be a sample from a normal distribution with mean Kc^u and variance σ^2. Unless otherwise stated, we use $\Delta = 1, \delta t = 1ms$ (Therefore, the units of Reaction Time in our simulation are in milliseconds).

Unless otherwise stated, in all simulations reported here, we have $K = 0.4$, and $\sigma = 1$. In principle, one can absorb K into the range of stimuli coherence $\mathcal{C}$, but to be consistent with previous studies where K had to be fit to data, we employ the above notation. We also reiterate that the Left and Right actions lead to the termination of the existing trial, the start of a new trial, and are thus terminating actions. Before we proceed forward, we would like to note that although in Eq. (1), we used the addition of the current state s^u_t and sensory input E_t for updating the state and for accumulation of evidence, this specific choice is not mandatory and the model can perform reasonably without accumulation too; see section 6.

At any given time t, during the trial u, associated with every state $s \in \mathcal{S}$ and every action $a \in \mathcal{A}$, there is a Q-value denoted by $Q^u_t(s, a)$. At the beginning of each trial, the Q-values for trial u are copied from the end of trial $u - 1$ with $Q^0_0(s, a) = 0$. The resulting Q-table is used to determine, in a given state s, which action, a, is selected at time t in trial u. This is done via a softmax function yielding

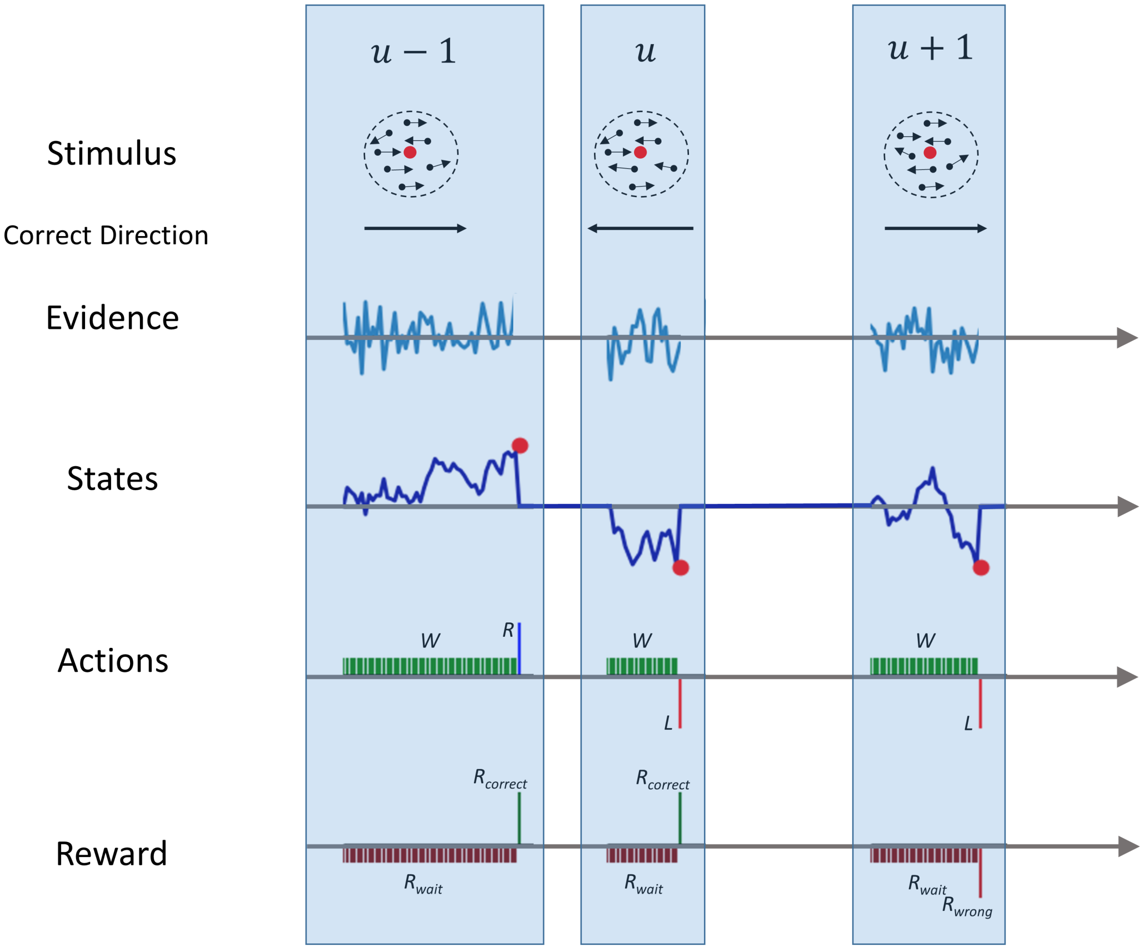

Figure 1. Schematic illustration of the perceptual stimulus, trial structure, and model components. The structure of the task and three consecutive trials ($u-1$, u, $u+1$) are illustrated with time progression taking place from left to right (horizontal arrow) both during the trials and from one trial to the next. The variable widths of the white gaps between trials depict the random duration of inter-trial intervals. We assume that no update happens during these periods. In each trial, a random dot motion stimulus moves towards the left or right. The evidence (E_t in Eq. (1)) is sampled every time the agent chooses to wait (i.e., Wait action) and the state variable is updated by accumulating the evidence. This within-trial updating continues until the agent chooses one of the terminating actions (L, R) at which point the state and evidence variables are then set to zero and remain zero until the beginning of the presentation of the new motion stimulus in the next trial. The states at which these terminating actions are taken are the *terminal state*, indicated by the red circle in the plots denoted by States. At each time point, the agent receives a reward based on the action that it has taken. Unlike sequential sampling models, no comparison to any threshold is explicitly formalized in the model and taken by the model.

probabilistic action selection as follows

$$p_t^u(a|s) = \frac{e^{\beta Q_t^u(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\beta Q_t^u(s,a')}}, \tag{2}$$

where β controls the degree of stochasticity in action selection and is set to 50, unless otherwise stated.

Once an action $a_t^u \in \{\text{Right}, \text{Left}, \text{Wait}\}$ has been taken, the corresponding reward R_t^u has been collected and the transition to the new state has occurred, the Q-table is updated as

$$Q_{t+1}^u(s_t^u, a_t^u) = Q_t^u(s_t^u, a_t^u) + \epsilon \left[R_t^u + \gamma \max_{a \in \mathcal{A}} (Q_t^u(s_{t+1}^u, a)) - Q_t^u(s_t^u, a_t^u) \right], \tag{3}$$

where ϵ is the learning rate (set to 0.1 unless stated otherwise) and γ is the discount factor of the temporal difference (TD) term. Unless otherwise stated, we use $\gamma = 0.9$ except when the trial is terminated and $\gamma = 0$. For the systematic study of the effect of parameters on the terminal state, see S.5.

2.2 Evolution of the Q-table during learning

In this section, we examine the evolution of the Q-table in the course of learning. A more detailed intuition about the dynamic of the model is provided in the Supplemental material (see S.1).

In Fig. 2(a), we see three snapshots of the model Q-table at different stages of training. After 400 trials, there is an island of states around 0, where the Q values of the Wait action (green) are the largest, and those of Right and Left are considerably more negative. As the learning proceeds (middle and right panels), this island expands. At the right boundary of this island, one can see a blue bump (c.f., blue arrows) indicating a number of states for which the corresponding Q-value for the Right action exceeds that for the other two.

Similarly, a red bump is visible on the left side, indicating those states for which the Left action is more likely to be chosen. How quickly in training these bumps appear depends on the learning rate, ϵ; see Fig. S.2. By the end of the training, the Q-value for each of the terminating actions has exceeded that of the Wait action on its corresponding side, that is, red on the left and blue on the right. Note that, for the the peaks to arise stably by the end of the learning phase, the number of available states (i.e., M) should be large enough. With insufficient M, the Wait island (Green lines) expands to the whole range of available states, freezing the agent in a state of perpetual anticipation and paralysis.

We define the state in which a terminating action is chosen as the *terminal state*. If $\beta = \infty$, the terminal states correspond to the peak of the bumps in the Q-table. For lower values of β, those peaks are merely more probable to be a terminal state. In Fig. 2(b)-(c), we see the evolution of the terminal states in the course of training. They start near zero and progressively move away. This trend is not monotonic, implying that the Q-learning algorithm searches for and ends up fluctuating around some Q-table that strikes a balance between the cost of waiting, the costs and benefits of wrong and correct decisions.

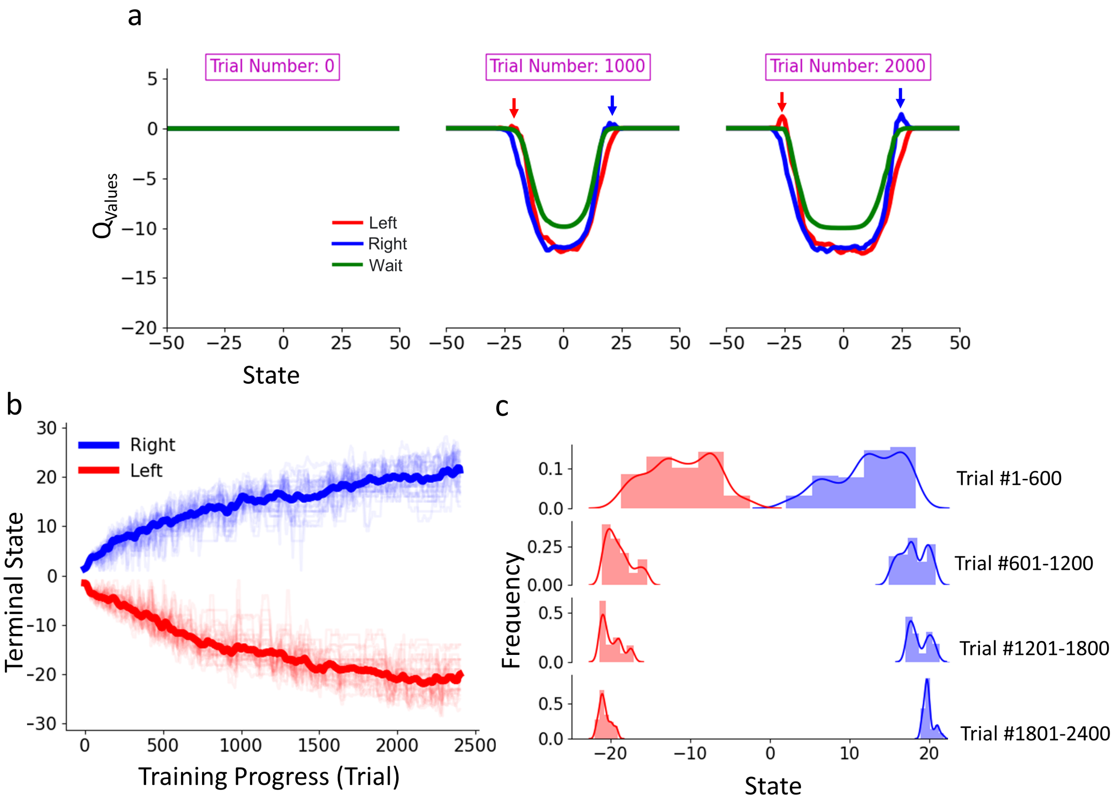

Figure 2. Evolution of the Q-table. (a) Snapshots of Q-values of each action at each state shown at the beginning of the learning (Trial number= 0) where all Q-values are set to zero. The Q-values shown are averaged over 30 simulations with the same parameters. In each trial, u, the coherence level c^u is chosen randomly and with equal probability from the set $\mathcal{C} = [-51.2\%, -25.6\%, ...0, ...25.6\%, 51.2\%]$. As training proceeds, the Q-values associated with the Wait action (green) in the states around zero stay higher but those for the terminating actions (Left, and Right) drop to lower values. The Q-value for each terminating action exceeds that of the Wait on the side corresponding to the correct choice (i.e., red on the left and blue on the right - see arrows). (b) Terminal states initially emerge near zero and then, with training, move away from it toward rightward (blue) and leftward (red). Each thin lines show the results for one of the 30 simulations, and the tick line represents the average over those simulations. (c) Histograms showing the fraction of times that a state had the largest Q-value (when averaged over 30 simulations) during 4 different periods of learning, each comprising 600 trials. As training progresses, the histograms shift away from zero and become narrower in spread; the solid curves are fitted to the histograms. In the simulations reported in this figure, $U = 2400$ trials were used.

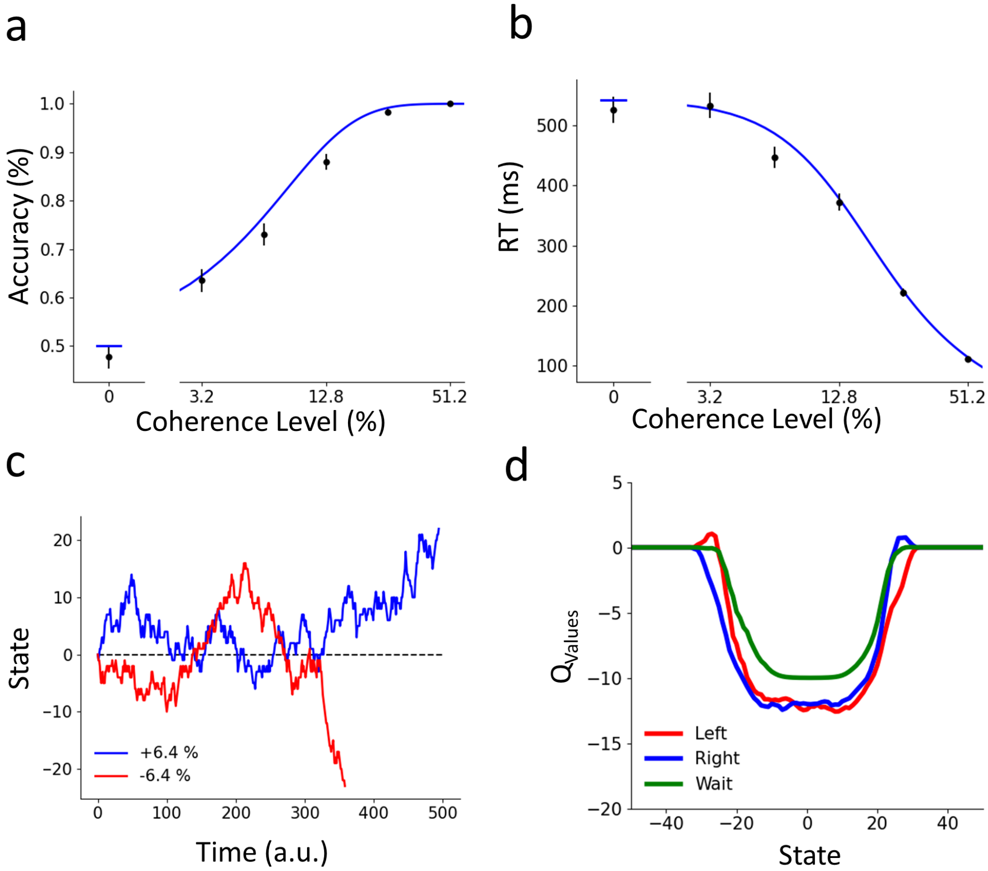

Figure 3. Model performance after training. The psychometric curve showing choice Accuracy (a), and the chronometric curve showing Reaction Time (RT) (b), both plotted as a function of the coherence level, c. Data from the simulations are denoted by black points and the lines in (a) and (b) show Eqs. (4a) and (4b). To plot these lines we fixed B in Eq. (4b) and (4a) to the average of the model's terminal state over 2400 test trials during which the Q-table was left unchanged. The error bars show SEM over these trials. (c) Two examples of the states taken by the model as time progresses through the test trials where Q-table is fixed for two stimuli with opposing directions. (d) Q-values of the trained model for different actions in different states. The bumps appearing at states ~ 20 and ~ -20 indicate the location of the terminal states. All other parameters are the same as Fig. 2.

2.3 Model's behavior in perceptual decision making

Having described some of the main features of the model above, here we showed that the trained RL model's behavior matches the hallmarks of perceptual decision making under uncertainty observed in empirical studies in humans, non-human primates, and rodents; c.f. [1, 26] for reviews of these empirical findings.

Fig. 3(a)-(b) shows the psychometric and chronometric functions that describe the relationship between decision accuracy and reaction time with motion coherence. As can be seen in this figure, increasing coherence increases accuracy and decreases reaction time, replicating extensive previous empirical findings. Perhaps more surprisingly, the simulation results (black symbols) are also in decent agreement with predictions obtained directly from the closed-form solutions to the bounded accumulation process [27]:

$$\mathrm{Accuracy}(c|B,K) = \frac{1}{1+\exp[-2KcB]} \tag{4a}$$

$$\mathrm{RT}(c|B,K) = \frac{B}{Kc}\ \tanh(KcB), \tag{4b}$$

where K and c were already introduced in Eq. (1), and B is the terminal state of the RL model at the end of learning, defined in the section 2.2. Previous empirical studies that employ the sequential sampling models often interpret the behavior using the above two equations for which K and B are fitted to the data. In the language of sequential sampling, K and B are known as the drift rate coefficient and the decision bound/threshold, respectively.

In addition to demonstrating the model's summary behavior through psychometric and chronometric functions, we also studied the inner workings of the model within the course of each trial, as shown in Fig. 3(c). We see the state transitions (Eq. (1)) in two example trials. The blue trace shows a trial in which a weak rightward ($c = +6.4\%$) stimulus was presented to the model. The model took its time, collecting evidence and switching states for a fairly long number of time (≈ 500) steps. In comparison, the red trace shows another trial where a similarly weak but leftward stimulus was presented to the model. Here the model took a shorter time, arriving at the correct terminal state before the 400 time steps. These two examples suggest that our RL model performs similarly to the sequential sampling models. This is particularly remarkable because at no point in the model description and training did we introduce any explicit boundary. This is where our model diverges from the ideal observer approaches [23, 22, 28] that calculate the boundary based on their *a priori* knowledge of the structure of the environment.

In S.6, we demonstrate the replication of a number of other, related empirical observations such as the difference between error and correct reaction times (Fig. S.11), post-error slowing (Fig. S.12) and the impact of volatility on decisions accuracy and reaction times (Fig. S.13). We encourage interested readers to utilize the model code and try out replicating other empirical findings in perceptual decision-making.

3 Decision dynamics during learning

A key problem with previous RL models of perceptual decision making [29, 30] is that they did not produce any predictions about the development of reaction times during training. Our model, however, is naturally apt to address this problem. We, therefore, proceeded to examine how model reaction times and accuracy change in the course of learning.

Fig. 4(a) illustrates the changes in model decision accuracy with progress in learning. Consistent with numerous empirical observations, model accuracy starts around the chance level and progressively improves. We draw a direct comparison between the model behavior in Fig. 4(b) and empirical data reproduced from a previous study [31] in the inset. Adopting that study's terminology, the threshold was defined as the coherence level at which the model performed at 82% accuracy, and the lapse rate was defined as the residual error rate at the highest level of coherence (100%). Our model's choice behavior shows qualitative consistency with those empirical observations.

In Fig. 4(c) the RL model's reaction times in the course of perceptual learning are plotted. Reaction times start fast and slow down with more training. This pattern of reaction times is indeed a direct consequence of the Q-learning algorithm: since the Q-table is initialized to zero, actions have similar values at the onset of learning. Therefore, in the early phases of learning, the terminating Right or Left actions are chosen with a high probability of $\sim 66.6\%$ before any evidence is accumulated. The model learns to wait by committing frequent quick errors and decreasing the value of all three actions for the states around zero, albeit to different degrees. In other words, the consequence of starting the Q-table with such a blank slate is that the model would be barely exposed to the stimulus early in training. Consequently, early trials provide little opportunity for the model to learn the association between coherence and terminal actions. This profile of behavior, however, is different from several previous empirical observations where training usually starts with slower reaction times that progressively get faster [10, 11, 12, 13, 14, 15, 16]. Since this divergence is largely a consequence of the initial symmetry of the actions embodied in the initial blank slate Q-table, in S.3, we offer two alternative solutions to this issue based on the initialization of the Q-table differently.

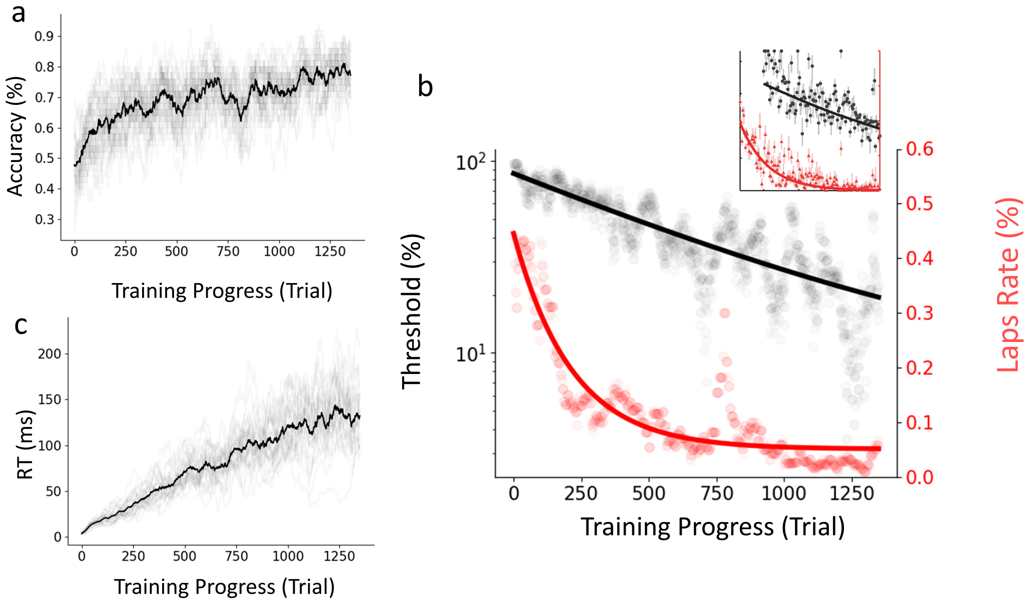

Figure 4. Changes in decision Accuracy and RT during training. (a) Accuracy increases as training progresses. Light grey curves show this for 30 individual simulations, with the same model parameters, smoothed through convolution with a unity array of size of 50. The solid black line shows the average over these simulations. (b) Changes in psychometric threshold (black) and lapse rate (red) during training. The psychometric threshold is defined as the coherence level at which the model performs at 82% accuracy (e.g. α in Weibull CDF [3, 31]) and the lapse rate is the error rate in trials with 100% coherence. Think curves are fits to the data points via similar functions used in [31]. The inset shows empirical data from macaque monkeys [31]. (c) Same as (a) but for the reaction times.

4 The impact of payoff structure on Speed-Accuracy Trade-off

Having examined the dynamics of Speed-Accuracy Trade-off (SAT) during learning, we then proceeded to examine whether the model could flexibly trade-off speed and accuracy under various payoff conditions. This demonstration is critical for two reasons. First, previous reinforcement learning models have never been employed to explain the variations in choice reaction time in response to characteristics of the environment. Second, the extensive previous literature on speed-accuracy tradeoff [32, 33, 34] in experiments involving human [35, 36] and non-human [17] primates provides a strong set of constraints to test our model with.

Here, we focus on previous empirical works in humans demonstrating that increasing the cost of errors relative to the reward for correct choice prolongs reaction times and prioritizes accuracy; conversely, speed was prioritized when the reward for correct choice was increased [35, 37, 36]. These empirical observations were explained by changes of bound in sequential sampling models with a fixed drift rate. Several other studies that examined SAT in humans in a variety of decision tasks have also argued that SAT is best explained by changes in decision bound [38, 39]; c.f. Discussion).

To examine if our modeling framework could account for these empirical observations, we examined the following hypothesis: as long as the cost of waiting is kept low, increasing the cost of mistakes (vs the benefit of correct responses) should tip the balance towards accuracy. We tested our model under various payoff regimes, systematically altering the Cost-Benefit Ratio (CBR), defined as $|R_{\text{wrong}}/R_{\text{correct}}|$, over a wide range without imposing or assuming any decision threshold beforehand. The results in Fig. 5 show that when CBR is high (red curves in Fig. 5(a)-(b)), the model reaction times are longer and accuracy is higher. In contrast, when CBR is low, decisions are faster and mistakes are more frequent (purple curves in Fig. 5(a)-(b). These findings confirm our hypothesis. To further understand how these results arise from the RL dynamics, we investigated the relationship between CBR and the position of terminal states in the Q-table. Fig. 5(c) shows a direct relation between the position of the terminal state and the CBR, depending on the learning rate ϵ. For any given ϵ, the position of

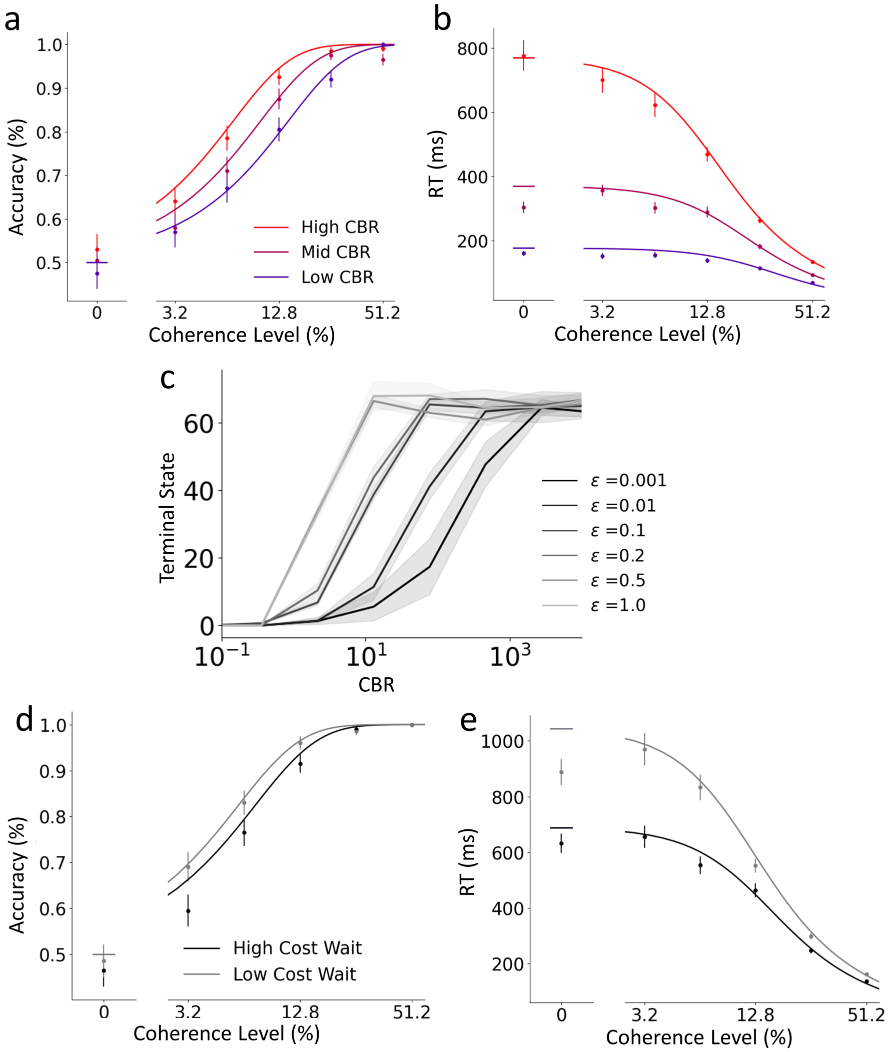

Figure 5. Speed accuracy trade-off (SAT) of the trained model (a) Choice Accuracy and (b) RT for different values of CBR: Large (4), Intermediate (3), and Small (2) indicated by different colors. (c) The terminal state for different values of CBR and different learning rates ϵ. Increasing CBR pushes the terminal state further away from zero, producing the dependence shown in (a)-(b). Curves are smoothed using a moving average filter and $U = 900$. CBR values were changed from 0.01 to 10^5 in equal logarithmic steps, by fixing $R_{\text{correct}} = 20$ and changing R_{wrong}. Choice Accuracy (d) and RT (e) versus coherence level the cost of the Wait action is changed while CBR is kept constant. RT values are smaller and Accuracy is lower compared to the cost of the Wait action, as can be seen by comparing black and gray curves corresponding to $R_{\text{wait}} = -2$ and$R_{\text{wait}} = -1$, respectively. Error bars are SEMs across trials; The simulation involved 1200 trials.

the terminal state remains relatively constant for CBRs beyond a certain critical value. This value obviously depends on M and also other parameters of the model, e.g. the total number of trials (see Fig. S.7 for more details).

One caveat of examining SAT as a function of CBR is that CBR is independent of the cost of the Wait action. However, any plausible mechanistic explanation of SAT should factor in this cost [23]. To examine the impact of changing the cost of the wait (W) action on SAT balance, we tested the hypothesis that with equal cost of error and benefit of correct response, increasing the cost of waiting should prioritize speed. Fig. 5(d)-(e) show that indeed, when waiting is more costly (black curve), reaction times decrease and accuracy is diminished. Reducing the cost of waiting (gray curve in Fig. 5(d)-(e)) reverts the trade-off in favor of accuracy. Together, the results in this section indicate that our model-free RL is able to *learn* how long to wait before committing to a definitive choice in a way that balances the cost of evidence accumulation against the cost and benefit of choice outcomes.

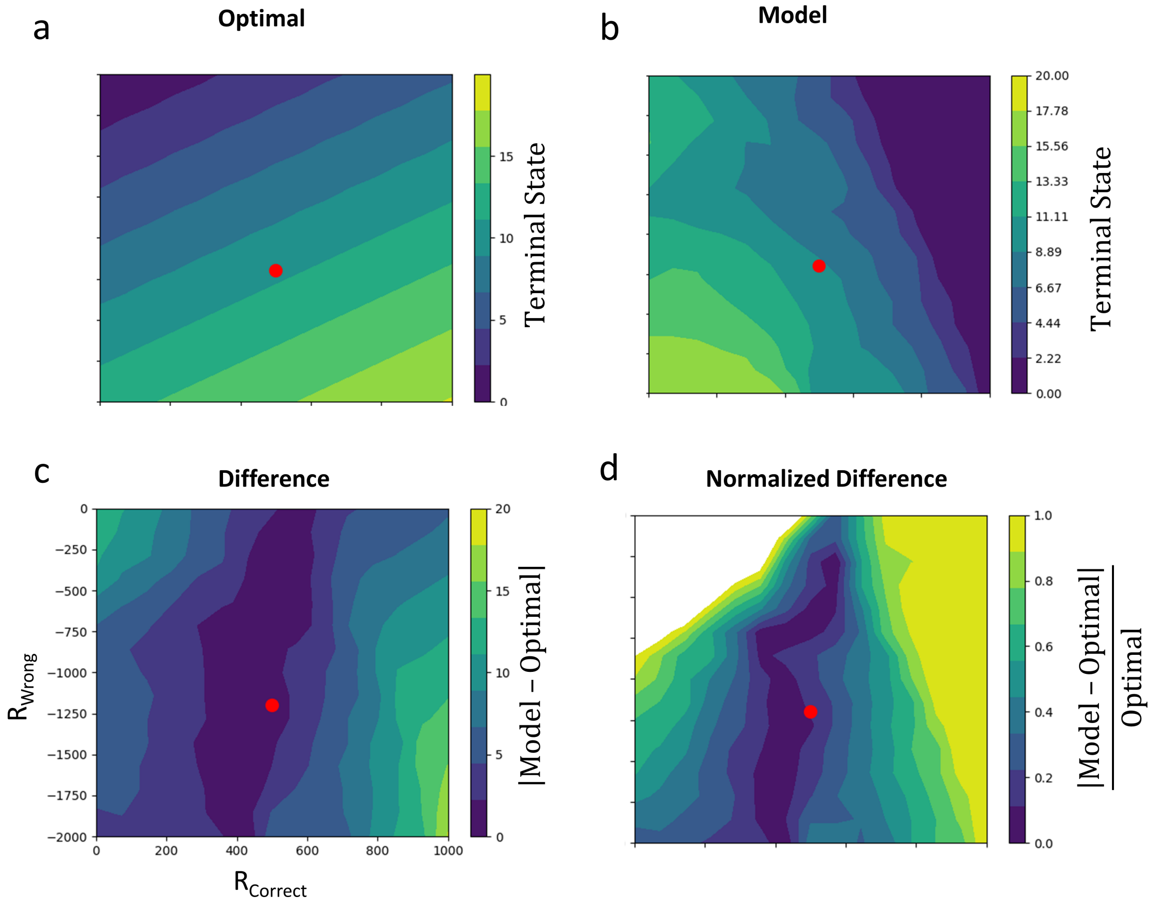

Figure 6. Comparison of our model with the optimal model for$c = 6.4\%$ (a) Optimal terminal state and reward regime for the optimal model Eq. (5). We studied the relation of optimal terminal state (colors) to R_{correct} and R_{wrong}. The red dot denotes a reward regime that has been used in Fig. 7. (b) same as (a) but for models simulations (N=10 iteration, $\epsilon = 0.01, \beta = 50, \Delta = 0.1, \gamma = 1, U = 3000$). (c) Difference between optimal terminal state (a) and model's terminal state (b). (d) same as (c) but with the normalized difference. The payoff regime that we chose (500, -1200) has a small distance to the optimal model.

5 Comparison with Optimizing Expected Reward

Setting a decision threshold on evidence accumulation in perceptual decision making under uncertainty can be thought of as an optimization of a cost function. One reasonable choice for such a function is the expected reward defined as

$$\text{ER}(B, K) = R_{\text{correct}}\text{Accuracy}(c|B, K) - R_{\text{wrong}}[1 - \text{Accuracy}(c|B, K)] - R_{\text{wait}}\text{RT}(c|B, K) \qquad (5)$$

where Accuracy, RT are the same as those in Eq. (4). Since in Eq. (5), B can take continuous values, in the simulations reported in this section, we have also used $\Delta = 0.1$, so as to make the discrimination of the states of the model finer. We consider the case of $\Delta = 1$ which is used in the other results reported so far in S.4. We call the value of B for which $\text{ER}(B, K)$ is maximized as *optimal terminal state*. Note that we used the simple grid search over Eq. (5) to obtain the optimal B.

In Fig. 6, we show the value of the optimal terminal state (Fig. 6(a)) and those reached by the model (Fig. 6(b)), as well as the difference between the two, Fig. 6(c), for different choices of R_{correct} and

R_{wrong} and fixed $R_{\text{wait}} = -1$. It is clear from these figures, that there is a region in the $(R_{\text{correct}}, R_{\text{wrong}})$ plane that the optimal terminal state and that of the model are quiet close to each other (e.g. within 10%). Fig. 7 shows the relationship between the optimal terminal state and that of the model when R_{correct} and R_{wrong} correspond to a point in this region, specifically the point denoted by the red dot in Fig. 6, with $\mathbf{R} = \{500, -1200, -1\}$.

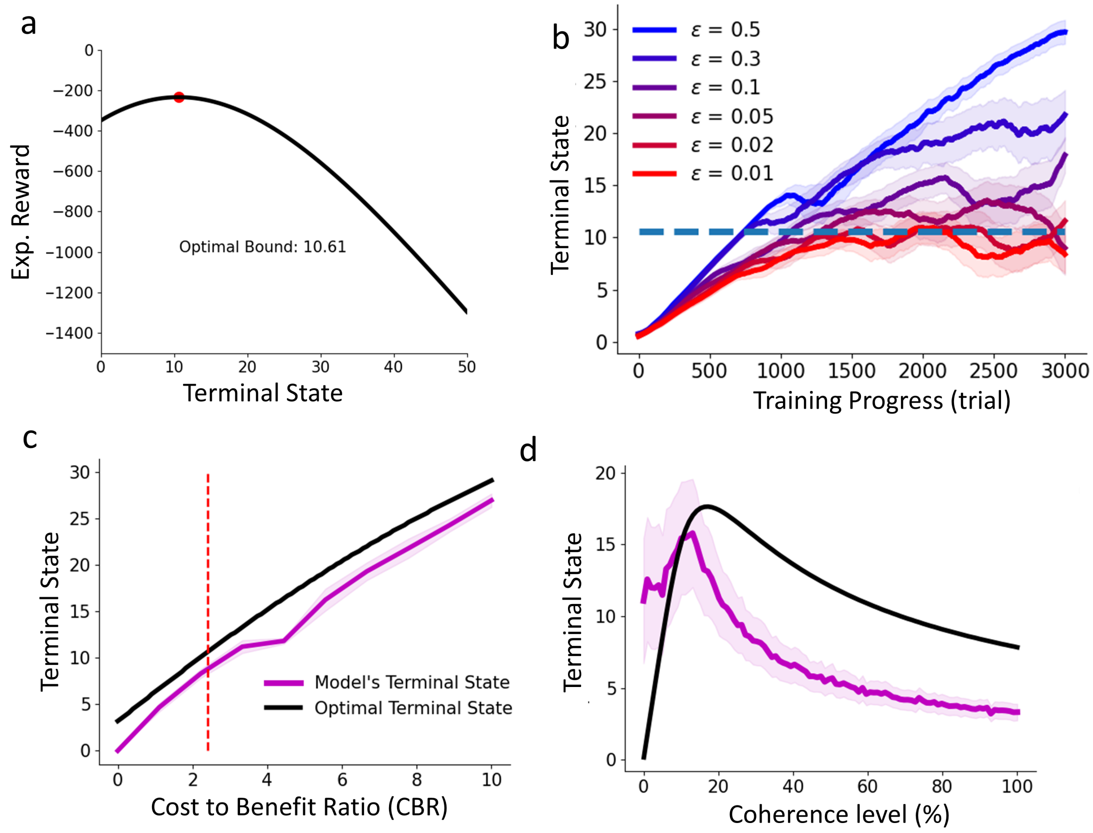

Figure 7. Optimality of the terminal states. (a) Expected Reward (Eq. (5)) as a function of B for $\mathbf{R} = [500, -1200, -1], \gamma = 1$ and $c = 0.064$; the maximum is denoted by the red dot. (b) Position of the terminal state reached by the model for different learning rates ϵ, compared to the optimum (dashed line). (c) Optimal terminal states and those reached by the model with $\epsilon = 0.01$ and after 3000 trials, plotted versus the cost-benefit ratio (CBR). The model's terminal states in taken as the average of the terminal states over the last 300 trials. CBR corresponding to the parameters in (b) are indicated by a dashed line. (d) Similar to (b) but for different coherence levels; $\epsilon = 0.05$ and $U = 6000$ were used. In the case of simulations in (b)-(d) the solid curves are averages and the shaded are the STD over 5 simulations. In all simulations here we used $\Delta = 0.1$.

In Fig. 7(a), we first plot the expected reward as a function of the terminal state B; the optimal terminal state is denoted by a red dot. Fig. 7(b) shows when trained on a fixed coherence c, how the terminal state reached by the model compares with the optimal terminal state. For large values of ϵ (blue curve in Fig. 7(b)), the model overshoots the optimal. This means that, across training, the model keeps accumulating more and more evidence, increasing its reaction time, without the accuracy changing significantly: large ϵ constrains accuracy but not reaction time. The model can, however, arrive at a terminal state close to an optimal one (dashed line) as long as the learning rate ϵ is adequately small. In sum, the model that is not explicitly designed to optimize the expected reward can indeed be close to optimal in the sense of the cost function in Eq. (5) for some pay-off regimes.

In Fig. 7(c) we examine the concordance between the terminal state reached by the model and the optimal one in more detail. Interestingly, when plotted as a function of the cost-benefit ratio, the terminal state found by the the model follows that of the optimal solution closely, with the difference between the two remaining relatively constant as the the cost-benefit ratio changes. Combined with the fact that the actual position of the model and optimal terminal states increase linearly with CBR, for larger CBR, the terminal state of the model can get to only a few percent of the optimal solution; the red dashed line, corresponding to the CBR of the reward values used in Fig. 7(b).

Fig. 7(d) shows how changing the coherence level affects the optimal terminal state and the terminal state reached by the model. We can see that although the concordance is not always great, the terminal state reached by the model follows similar trends as the optimal. Firstly, in both cases, the terminal state initially increases with coherence. They both then reach a maximum, before decreasing with c. Although the terminal state of the model is consistently smaller than the optimal one for larger c increases, they have comparable slopes of decay with c.

In S.4 we show that the results described above are also true when we discretize the states of the model with $\Delta = 1$. We discuss the differences between the two choices of Δ in more detail in that section. We also note that, for the simulations in this section, the coherence level was fixed because the expected reward in Eq. (5) was defined for fixed c. Extending this definition to a more general case in which c changing during training is possible, but finding the corresponding optimum is a non-trivial task. Yet, the results of Fig. 7 discussed here indicate that even in such a scenario, the model does not diverge far from the optimal solution.

6 Deciding without evidence accumulation

Up to this point, our model dispensed with the assumption of a decision boundary. Yet, it still relied on accumulating the moment-to-moment sensory evidence: in Eq. (1) E_t is added to s_t. Reasonable as it may be, it is important to see whether this operation is a necessary condition for the model to function. In this section, we show that indeed it is not. Even when state dynamics that do not involve accumulation of evidence, action selection and Q-learning, namely, Eq. (3)) and Eq. (2), are sufficient for decision making.

We set the dynamics of the state variable to follow *exterma detection* [40, 41], and we have

$$s_t^u = \lfloor E_t \rceil_{\mathcal{S}}. \tag{6}$$

An example of the resulting dynamics is shown in Fig. 8, with $\epsilon = 0.1$, reward set $\mathbf{R} = [20, -50, -1]$ and $M = 100$ and the resolution of state space, as before, is set to $\Delta = 1$. Starting with the same parameters with which the evidence accumulation (Eq. (1)) model yielded reasonable performance previously, we see (Fig. 8(a)-(b)) that, the success of Eq. (6) depends critically on the value of sensory gain parameter K. When K value is in the range used in the evidence accumulating simulations ($K = 0.4$), then E_t in Eq. (6) is rarely sufficiently large to move the model towards higher states. Consequently, for such a small K, the system practically gets stuck in a few states around zero. This change, though, when we consider larger K and the model performs satisfactorily.

Increasing K leads to larger values of E_t for the same coherence level c^μ. Even though the model with Eq. (6) does not accumulate the momentary evidence it receives, it nonetheless achieves very reasonable performance if we set (for example) $K = 5$: the psychometric and chronometric curves shown in Fig. 8(a) and (b), respectively, now reflect the same canonical features that we previously observed in Fig. 3 and in empirical studies. Note that although the accuracy of the extrema detection model matches that of the accumulation model, the reaction time (RT) is consistently longer for extrema detection, especially at lower coherence levels. As can be seen by comparing Fig. 8(c) and e.g. Fig. 3(d), the Q-values associated with the actions in each state after training show the same prominent feature, namely, that terminating states develop on the left and right sides of 0, corresponding to (correct) Left and Right actions. With Eq. (6), the Q-values of the actions, however, depend more strongly on the states compared to Eq. (1).

To compare the the state dynamics of the two approaches quantitatively, we kept all the parameters including the trial sequence identical between the two and simulated 20 simulations for each case. Fig. 8(d), demonstrates that the accuracy of the accumulation model is higher than that of the extrema detection for all values of K. Variability across simulation runs is also larger for extrema detection. For large K, however, these differences diminish. Similarly, Fig. 8(e) shows that the RT of the

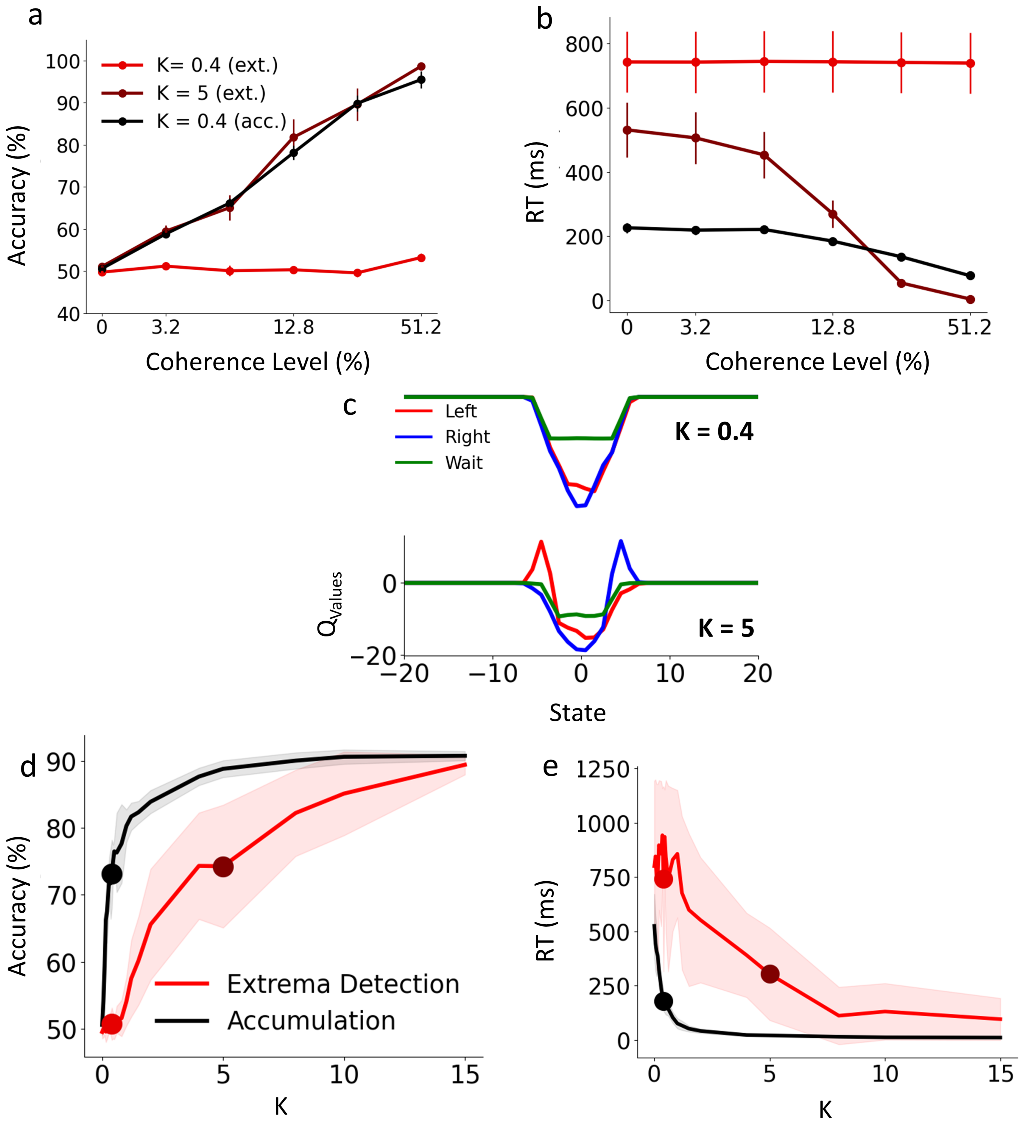

Figure 8. Decision making in the model without accumulating evidence. (a) The Accuracy of the extrema detection model (Eq. (6)) with $K = 0.4$ (light red) and $K = 5$ (dark red) and that of the evidence accumulation model (Eq. (1)) with $K = 0.4$ (black). The extrema detection model with small K (light red) performs near chance level but increasing K leads to accuracy similar, although with more variance, compared to the evidence accumulation model. (b) Similar to (a) but for reaction time. The exterma model with small K stays around the maximum in most simulations (maximum time of stimulus sequence was set to 1000 samples) with few reaching lower values, hence the larger variability. (c) The Q-values after training for the extrema detection model with small K (top) and large K (bottom); compare this to Fig. 3. (d) Systematic study of the effect of K on the accuracy of extrema detection (red) and evidence accumulation models (black), averaged over all tested coherence levels. The dots correspond to the simulations depicted in (a) using the same color conventions. (e) same as (d) but for RT. Error bars in (a) and (b) and the shaded area in (d) and (e) are SEM across 20 simulations, each including 900 training and test trials.

accumulation model, for all values of K, is consistently shorter than those of the extrema detection although this difference shrinks significantly with higher values of K. As a side note for interested readers, increasing the resolution of states by decreasing Δ can produce similar results. This is not depicted here but made available for examination in the shared code.

7 Discussion

Sequential sampling with integration to bound models has been extensively influential in our understanding of decision making under uncertainty [42, 43]. These models, however, have not produced a compelling explanation for two crucial issues. First, it is not clear how, when the animals are introduced to a new task, they learn the decision boundary given that they know very little about the experiment and the requirements of the task. Second, the way these models explain how agents adapt to changes in context is not consistent with empirical evidence. Studies of speed-accuracy trade-off in humans and other animals have shown that once the agent has learned the task and performs adequately, they can adapt to the new context. accumulation to bound models explain this by proposing a change in the decision bound, but empirical investigations in primate brains have not found evidence for such a proposal [17, 44, 32]. By introducing a new, simple Q-learning model we addressed both of these issues.

Two key innovations distinguish our Q-learning algorithm: an action repertoire with an extra, $Wait$ alternative and a state variable that accumulated noisy samples of information from the environment. The combination of these two innovations allowed the model to learn to decide via sequential sampling without, at any point, comparing the accumulated evidence with a decision boundary. The resulting model provides clear answers to the two outlined above. Regarding the first issue, that learning to decide, our model goes beyond earlier works that used model-based approaches for the determination of boundaries in perceptual decision making under uncertainty [23, 36]. Departing from those earlier works, our model-free approach does not require detailed explicit knowledge of the underlying statistical structure of the task. The model qualitatively reproduces the canonical hallmarks of chronometric and psychometric behavior observed in perceptual decision making under uncertainty. Earlier applications of model-free RL to perceptual decision making [29, 30] did not account for reaction times. By naturally accounting for reaction times, our model thus goes one step beyond those earlier models.

Furthermore, our model performed reasonably even *without accumulating* evidence showing that sequential sampling is sufficient, by itself, for perceptual decision making under uncertainty. As long as the internal states are arranged such that larger coherence levels push the model to states further away from zero and do so more rapidly than lower coherence levels, psychophysically plausible decision making needs neither boundary nor accumulation. The key challenge of reinforcement learning here, therefore, is to *exploit* that correspondence between the external noisy sensory information and internal states. How that correspondence is implemented in the agent's brain could be a combination of evolutionary, genetic, and developmental processes. This is a fundamental question that is beyond the scope of the work presented here but calls for revisiting the tenets of the commonly held beliefs about the neural mechanism underlying decision making under uncertainty.

Regarding the second issue, that of context, by studying the effect of CBR, we demonstrated that our model could adapt and trade off speed with an accuracy consistent with the requirements of the payoff table: asymmetric payoff tables with a higher reward for correct or higher punishment for incorrect decisions shifted the model's psychometric and chronometric functions consistent with empirical observations [36]. Moreover, when rightwards and leftward movement stimuli were not present with equal probability, the Q-table dynamically changed to adapt to the new distribution producing faster reaction times for the more likely option. Our model thus successfully adapts to changes in both payoff and probability contexts.

A number of previous studies have utilized evidence accumulation models to investigate the dynamics of value-based choice in reinforcement learning [45, 46, 47]. It is important to highlight the differences between our work and those previous approaches. In short, whereas they relied on sequential sampling models to characterize the within-trial dynamics of choice in an RL task, we relied on RL to characterize how the sequential sampling process evolves within and across trials. In those previous works, the RL agent's aim is to find out which choice option (e.g., "bandit") yields the higher reward by iteratively choosing among the options, comparing the outcome to its expectations, and updating its options' values. At any given trial, the choice is made stochastically through a diffusion process (towards a predefined boundary) whose drift rate is a function of the latest update of the option values. In these models, the rate of evidence accumulation is determined by value learning. Here we opted for a much simpler approach by adopting a constant drift rate (c.f. Eq. (1)) and dispensing altogether with boundary, regulating the evidence accumulation process by model-free RL instead. Integrating the concurrent learning of the drift rate into our model could be a promising avenue for future research.

A key strength of the sequential sampling framework is that it treats decision making under uncertainty as an optimization problem whose aim is to find the decision boundary that maximizes the reward rate given a combination of signal coherence, payoff regime, average accuracy, and average reaction time. Our simulations revealed that the position of the terminal state in the Q-table in some cases matches that of the optimal solution, and in general the way it changes with the parameters of the task, reflects features similar to those of the optimal solution. Optimizing the expected reward could, of course, be a consequence of Q-learning, but to meet the sufficient conditions for this to be the case, each state-action pair must be tried "many times," and the learning rates must be allowed to decay (see [48]). However, our simulations do not meet these conditions: the model does not explore all state-action pairs and uses a constant learning rate. Even if these conditions are met, convergence to optimality can take an extremely long time due to the complex nature and large expanse of the state space. This may explain the consistent underestimation of the decision threshold observed in model simulations. It is thus encouraging to see that while our model is not geared to any of these requirements it still behaves consistently with the optimal solution.

A longstanding tradition in the investigation of perceptual decision making is to have subjects (be they human or non-human animals) undergo extensive training to reach a stable, asymptotic level of performance before they participate in the experiment proper. In this tradition, which includes many of the current authors' previous works too, hundreds and thousands of such training trials are discarded because these data (behavioural and/or physiological) are deemed too unstable for interpretation. Another justification for discarding training data was that, up to now, there was no simple and general model for interpreting behavior during learning in perceptual decisions making under uncertainty. Future studies will be able to use our model to revisit those discarded data and compare them to model predictions.

References

[1] Joshua I. Gold and Michael N. Shadlen. "The Neural Basis of Decision Making". en. In: *Annual Review of Neuroscience* 30.1 (July 2007), pp. 535–574. ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev.neuro.29.051605.113038. URL: https://www.annualreviews.org/doi/10.1146/annurev.neuro.29.051605.113038 (visited on 11/08/2022).

[2] Kh Britten et al. "The analysis of visual motion: a comparison of neuronal and psychophysical performance". en. In: *The Journal of Neuroscience* 12.12 (Dec. 1992), pp. 4745–4765. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.12-12-04745.1992. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.12-12-04745.1992 (visited on 01/24/2022).

[3] Jamie D. Roitman and Michael N. Shadlen. "Response of Neurons in the Lateral Intraparietal Area during a Combined Visual Discrimination Reaction Time Task". en. In: *The Journal of Neuroscience* 22.21 (Nov. 2002), pp. 9475–9489. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.22-21-09475.2002. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.22-21-09475.2002 (visited on 11/08/2022).

[4] Ian Krajbich and Antonio Rangel. "Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions". en. In: *Proceedings of the National Academy of Sciences* 108.33 (Aug. 2011), pp. 13852–13857. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1101328108. URL: https://pnas.org/doi/full/10.1073/pnas.1101328108 (visited on 11/14/2022).

[5] Vincent B. McGinty, Antonio Rangel, and William T. Newsome. "Orbitofrontal Cortex Value Signals Depend on Fixation Location during Free Viewing". en. In: *Neuron* 90.6 (June 2016), pp. 1299–1311. ISSN: 08966273. DOI: 10.1016/j.neuron.2016.04.045. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627316301593 (visited on 11/14/2022).

[6] Hongbo Yu et al. “How peer influence shapes value computation in moral decision-making”. In: *Cognition* 211 (June 2021), p. 104641. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2021.104641. URL: https://www.sciencedirect.com/science/article/pii/S0010027721000603 (visited on 05/27/2024).

[7] S. P. Kelly and R. G. O’Connell. “Internal and External Influences on the Rate of Sensory Evidence Accumulation in the Human Brain”. en. In: *Journal of Neuroscience* 33.50 (Dec. 2013), pp. 19434–19441. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.3355-13.2013. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.3355-13.2013 (visited on 11/14/2022).

[8] Roger Ratcliff, Marios G. Philiastides, and Paul Sajda. “Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG”. en. In: *Proceedings of the National Academy of Sciences* 106.16 (Apr. 2009), pp. 6539–6544. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0812589106. URL: https://pnas.org/doi/full/10.1073/pnas.0812589106 (visited on 11/14/2022).

[9] Bingni W. Brunton, Matthew M. Botvinick, and Carlos D. Brody. “Rats and Humans Can Optimally Accumulate Evidence for Decision-Making”. en. In: *Science* 340.6128 (Apr. 2013), pp. 95–98. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1233912. URL: https://www.science.org/doi/10.1126/science.1233912 (visited on 11/14/2022).

[10] Roger Ratcliff, Anjali Thapar, and Gail McKoon. “Aging, practice, and perceptual tasks: A diffusion model analysis.” en. In: *Psychology and Aging* 21.2 (2006), pp. 353–371. ISSN: 1939-1498, 0882-7974. DOI: 10.1037/0882-7974.21.2.353. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0882-7974.21.2.353 (visited on 11/24/2023).

[11] Javier Masís et al. “Strategically managing learning during perceptual decision making”. en. In: *eLife* 12 (Feb. 2023), e64978. ISSN: 2050-084X. DOI: 10.7554/eLife.64978. URL: https://elifesciences.org/articles/64978 (visited on 11/24/2023).

[12] Naoshige Uchida and Zachary F Mainen. “Speed and accuracy of olfactory discrimination in the rat”. en. In: *Nature Neuroscience* 6.11 (Nov. 2003), pp. 1224–1229. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn1142. URL: https://www.nature.com/articles/nn1142 (visited on 11/24/2023).

[13] Amitai Shenhav, Matthew M. Botvinick, and Jonathan D. Cohen. “The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function”. en. In: *Neuron* 79.2 (July 2013), pp. 217–240. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.07.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627313006077 (visited on 11/24/2023).

[14] Fuat Balci et al. “Acquisition of decision making criteria: reward rate ultimately beats accuracy”. en. In: *Attention, Perception, & Psychophysics* 73.2 (Feb. 2011), pp. 640–657. ISSN: 1943-3921, 1943-393X. DOI: 10.3758/s13414-010-0049-7. URL: http://link.springer.com/10.3758/s13414-010-0049-7 (visited on 11/24/2023).

[15] Charles C. Liu and Takeo Watanabe. “Accounting for speed–accuracy tradeoff in perceptual learning”. en. In: *Vision Research* 61 (May 2012), pp. 107–114. ISSN: 00426989. DOI: 10.1016/j.visres.2011.09.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0042698911003373 (visited on 11/24/2023).

[16] Gilles Dutilh et al. “A diffusion model decomposition of the practice effect”. en. In: *Psychonomic Bulletin & Review* 16.6 (Dec. 2009), pp. 1026–1036. ISSN: 1069-9384, 1531-5320. DOI: 10.3758/16.6.1026. URL: http://link.springer.com/10.3758/16.6.1026 (visited on 11/24/2023).

[17] Richard P. Heitz and Jeffrey D. Schall. “Neural Mechanisms of Speed-Accuracy Tradeoff”. en. In: *Neuron* 76.3 (Nov. 2012), pp. 616–628. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.08.030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627312007672 (visited on 11/11/2022).

[18] Stephen M. Fleming et al. “Effects of Category-Specific Costs on Neural Systems for Perceptual Decision-Making”. In: *Journal of Neurophysiology* 103.6 (June 2010). Publisher: American Physiological Society, pp. 3238–3247. ISSN: 0022-3077. DOI: 10.1152/jn.01084.2009. URL: https://journals.physiology.org/doi/full/10.1152/jn.01084.2009 (visited on 05/27/2024).

[19] Louise Whiteley and Maneesh Sahani. “Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes”. In: *Journal of Vision* 8.3 (Mar. 2008), p. 2. ISSN: 1534-7362. DOI: 10.1167/8.3.2. URL: https://doi.org/10.1167/8.3.2 (visited on 05/27/2024).

[20] T. D. Hanks et al. “Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task”. en. In: *Journal of Neuroscience* 31.17 (Apr. 2011), pp. 6339–6352. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.5613-10.2011. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.5613-10.2011 (visited on 11/08/2022).

[21] Nobuhiro Hagura, Patrick Haggard, and Jörn Diedrichsen. “Perceptual decisions are biased by the cost to act”. In: *eLife* 6 (Feb. 2017). Ed. by Joshua I Gold. Publisher: eLife Sciences Publications, Ltd, e18422. ISSN: 2050-084X. DOI: 10.7554/eLife.18422. URL: https://doi.org/10.7554/eLife.18422 (visited on 05/27/2024).

[22] P. Dayan and N. D. Daw. “Decision theory, reinforcement learning, and the brain”. en. In: *Cognitive, Affective, & Behavioral Neuroscience* 8.4 (Dec. 2008), pp. 429–453. ISSN: 1530-7026, 1531-135X. DOI: 10.3758/CABN.8.4.429. URL: http://link.springer.com/10.3758/CABN.8.4.429 (visited on 02/09/2023).

[23] J. Drugowitsch et al. “The Cost of Accumulating Evidence in Perceptual Decision Making”. en. In: *Journal of Neuroscience* 32.11 (Mar. 2012), pp. 3612–3628. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.4010-11.2012. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.4010-11.2012 (visited on 11/08/2022).

[24] Timothy Hanks, Roozbeh Kiani, and Michael N Shadlen. “A neural mechanism of speed-accuracy tradeoff in macaque area LIP”. en. In: *eLife* 3 (May 2014), e02260. ISSN: 2050-084X. DOI: 10.7554/eLife.02260. URL: https://elifesciences.org/articles/02260 (visited on 11/08/2022).

[25] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018. ISBN: 0-262-35270-2.

[26] Michael N. Shadlen and Roozbeh Kiani. “Decision Making as a Window on Cognition”. en. In: *Neuron* 80.3 (Oct. 2013), pp. 791–806. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.10.047. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627313009999 (visited on 11/08/2022).

[27] Roger Ratcliff. “A Theory of Memory Retrieval”. en. In: (1978), p. 50.

[28] Anya Skatova, Patricia A. Chan, and Nathaniel D. Daw. “Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task”. en. In: *Frontiers in Human Neuroscience* 7 (2013). ISSN: 1662-5161. DOI: 10.3389/fnhum.2013.00525. URL: http://journal.frontiersin.org/article/10.3389/fnhum.2013.00525/abstract (visited on 02/09/2023).

[29] Chi-Tat Law and Joshua I Gold. “Reinforcement learning can account for associative and perceptual learning on a visual-decision task”. en. In: *Nature Neuroscience* 12.5 (May 2009), pp. 655–663. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.2304. URL: http://www.nature.com/articles/nn.2304 (visited on 11/08/2022).

[30] Thorsten Kahnt et al. “Perceptual Learning and Decision-Making in Human Medial Frontal Cortex”. en. In: *Neuron* 70.3 (May 2011), pp. 549–559. ISSN: 08966273. DOI: 10.1016/j.neuron.2011.02.054. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627311002960 (visited on 11/08/2022).

[31] Chi-Tat Law and Joshua I Gold. “Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area”. en. In: *Nature Neuroscience* 11.4 (Apr. 2008), pp. 505–513. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn2070. URL: http://www.nature.com/articles/nn2070 (visited on 11/08/2022).

[32] Richard P. Heitz. “The speed-accuracy tradeoff: history, physiology, methodology, and behavior”. en. In: *Frontiers in Neuroscience* 8 (June 2014). ISSN: 1662-453X. DOI: 10.3389/fnins.2014.00150. URL: http://journal.frontiersin.org/article/10.3389/fnins.2014.00150/abstract (visited on 11/11/2022).

[33] A. Wald. *Sequential Analysis*. New York: Wiley & Sons, 1947.

[34] Mervyn Stone. “Models for choice-reaction time”. In: *Psychometrika* 25.3 (Sept. 1960), pp. 251–260. ISSN: 1860-0980. DOI: 10.1007/BF02289729. URL: https://doi.org/10.1007/BF02289729.

[35] Patrick Simen, Jonathan D. Cohen, and Philip Holmes. “Rapid decision threshold modulation by reward rate in a neural network”. In: *Neural Networks*. Neurobiology of Decision Making 19.8 (Oct. 2006), pp. 1013–1026. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2006.05.038. URL: https://www.sciencedirect.com/science/article/pii/S0893608006001626 (visited on 02/27/2024).

[36] Rafal Bogacz et al. “Do humans produce the speed-accuracy trade-off that maximizes reward rate?” eng. In: *Quarterly Journal of Experimental Psychology (2006)* 63.5 (May 2010), pp. 863–891. ISSN: 1747-0226. DOI: 10.1080/17470210903091643.

[37] Patrick Simen et al. “Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions”. eng. In: *Journal of Experimental Psychology. Human Perception and Performance* 35.6 (Dec. 2009), pp. 1865–1897. ISSN: 1939-1277. DOI: 10.1037/a0016926.

[38] B. A. J. Reddi and R. H. S. Carpenter. “The influence of urgency on decision time.” In: *Nature Neuroscience* 3 (2000). Place: United Kingdom Publisher: Nature Publishing Group, pp. 827–830. ISSN: 1546-1726(Electronic),1097-6256(Print). DOI: 10.1038/77739.

[39] B.A.J. Reddi, K. N. Asrress, and R.H.S. Carpenter. “Accuracy, Information, and Response Time in a Saccadic Decision Task”. en. In: *Journal of Neurophysiology* 90.5 (Nov. 2003), pp. 3538–3546. ISSN: 0022-3077, 1522-1598. DOI: 10.1152/jn.00689.2002. URL: https://www.physiology.org/doi/10.1152/jn.00689.2002 (visited on 11/14/2022).

[40] Gabriel M Stine et al. “Differentiating between integration and non-integration strategies in perceptual decision making”. en. In: *eLife* 9 (Apr. 2020), e55365. ISSN: 2050-084X. DOI: 10.7554/eLife.55365. URL: https://elifesciences.org/articles/55365 (visited on 11/11/2022).

[41] Andrew B. Watson. “Probability summation over time”. In: *Vision Research* 19.5 (Jan. 1979), pp. 515–522. ISSN: 0042-6989. DOI: 10.1016/0042-6989(79)90136-6. URL: https://www.sciencedirect.com/science/article/pii/0042698979901366 (visited on 07/22/2024).

[42] Roger Ratcliff et al. “Diffusion Decision Model: Current Issues and History”. In: *Trends in Cognitive Sciences* 20.4 (Apr. 2016), pp. 260–281. ISSN: 1364-6613. DOI: 10.1016/j.tics.2016.01.007. URL: https://www.sciencedirect.com/science/article/pii/S1364661316000255 (visited on 07/22/2024).

[43] Rani Moran. “Optimal decision making in heterogeneous and biased environments”. eng. In: *Psychonomic Bulletin & Review* 22.1 (Feb. 2015), pp. 38–53. ISSN: 1531-5320. DOI: 10.3758/s13423-014-0669-3.

[44] Roozbeh Kiani, Leah Corthell, and Michael N. Shadlen. “Choice Certainty Is Informed by Both Evidence and Decision Time”. en. In: *Neuron* 84.6 (Dec. 2014), pp. 1329–1342. ISSN: 08966273. DOI: 10.1016/j.neuron.2014.12.015. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627314010964 (visited on 11/08/2022).

[45] Mads Lund Pedersen, Michael J. Frank, and Guido Biele. “The drift diffusion model as the choice rule in reinforcement learning”. en. In: *Psychonomic Bulletin & Review* 24.4 (Aug. 2017), pp. 1234–1251. ISSN: 1069-9384, 1531-5320. DOI: 10.3758/s13423-016-1199-y. URL: http://link.springer.com/10.3758/s13423-016-1199-y (visited on 11/08/2022).

[46] Michael J. Frank et al. “fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning”. en. In: *The Journal of Neuroscience* 35.2 (Jan. 2015), pp. 485–494. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.2036-14.2015. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.2036-14.2015 (visited on 11/08/2022).

[47] Laura Fontanesi, Stefano Palminteri, and Maël Lebreton. “Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling”. en. In: *Cognitive, Affective, & Behavioral Neuroscience* 19.3 (June 2019), pp. 490–502. ISSN: 1531-135X. DOI: 10.3758/s13415-019-00723-1. URL: https://doi.org/10.3758/s13415-019-00723-1 (visited on 06/17/2024).

[48] Christopher J. C. H. Watkins and Peter Dayan. “Q-learning”. en. In: *Machine Learning* 8.3 (May 1992), pp. 279–292. ISSN: 1573-0565. DOI: 10.1007/BF00992698. URL: https://doi.org/10.1007/BF00992698 (visited on 06/06/2024).

[49] Christopher J. Burke et al. “Neural mechanisms of observational learning”. In: *Proceedings of the National Academy of Sciences* 107.32 (Aug. 2010). Publisher: Proceedings of the National Academy of Sciences, pp. 14431–14436. DOI: 10.1073/pnas.1003111107. URL: https://www.pnas.org/doi/full/10.1073/pnas.1003111107 (visited on 02/27/2024).

[50] Anis Najar et al. “The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning”. en. In: *PLOS Biology* 18.12 (Dec. 2020). Publisher: Public Library of Science, e3001028. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001028. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001028 (visited on 02/27/2024).

[51] Caroline J. Charpentier, Kiyohito Iigaya, and John P. O'Doherty. "A Neuro-computational Account of Arbitration between Choice Imitation and Goal Emulation during Human Observational Learning". en. In: *Neuron* 106.4 (May 2020), 687–699.e7. ISSN: 08966273. DOI: 10.1016/j.neuron.2020.02.028. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627320301513 (visited on 02/27/2024).

[52] Hatim A. Zariwala et al. "The Limits of Deliberation in a Perceptual Decision Task". en. In: *Neuron* 78.2 (Apr. 2013), pp. 339–351. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.02.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627313001682 (visited on 11/14/2022).

[53] Braden A. Purcell and Roozbeh Kiani. "Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex". en. In: *Neuron* 89.3 (Feb. 2016), pp. 658–671. ISSN: 08966273. DOI: 10.1016/j.neuron.2015.12.027. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627315011290 (visited on 11/08/2022).

[54] P. Cisek, G. A. Puskas, and S. El-Murr. "Decisions in Changing Conditions: The Urgency-Gating Model". en. In: *Journal of Neuroscience* 29.37 (Sept. 2009), pp. 11560–11571. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.1844-09.2009. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1844-09.2009 (visited on 11/08/2022).

[55] Gouki Okazawa and Roozbeh Kiani. "Neural Mechanisms that Make Perceptual Decisions Flexible". en. In: *Annual Review of Physiology* 85.1 (Feb. 2023), annurev–physiol–031722–024731. ISSN: 0066-4278, 1545-1585. DOI: 10.1146/annurev-physiol-031722-024731. URL: https://www.annualreviews.org/doi/10.1146/annurev-physiol-031722-024731 (visited on 12/05/2022).

[56] David Thura and Paul Cisek. "The Basal Ganglia Do Not Select Reach Targets but Control the Urgency of Commitment". en. In: *Neuron* 95.5 (Aug. 2017), 1160–1170.e5. ISSN: 08966273. DOI: 10.1016/j.neuron.2017.07.039. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627317306876 (visited on 11/14/2022).

[57] Gordon D. Logan and Matthew J. C. Crump. "Cognitive Illusions of Authorship Reveal Hierarchical Error Detection in Skilled Typists". In: *Science* 330.6004 (Oct. 2010). Publisher: American Association for the Advancement of Science, pp. 683–686. DOI: 10.1126/science.1190483. URL: https://doi.org/10.1126/science.1190483 (visited on 11/11/2022).

[58] Curtis L. Baker and Isabelle Mareschal. "Chapter 12 Processing of second-order stimuli in the visual cortex". In: *Progress in Brain Research*. Vol. 134. Vision: From Neurons to Cognition. Elsevier, Jan. 2001, pp. 171–191. DOI: 10.1016/S0079-6123(01)34013-X. URL: https://www.sciencedirect.com/science/article/pii/S007961230134013X (visited on 03/08/2024).

[59] Ariel Zylberberg, Christopher R Fetsch, and Michael N Shadlen. "The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision". en. In: (), p. 31.

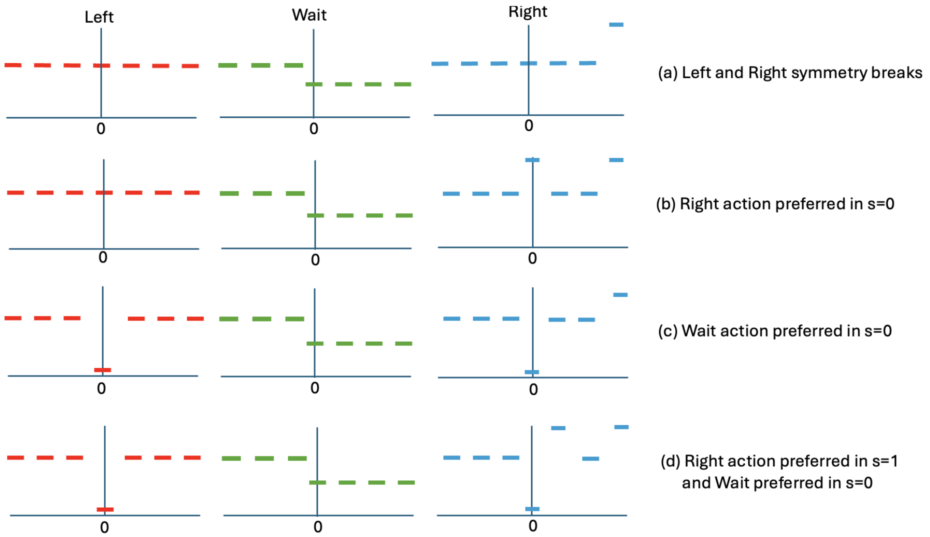
Left
Wait
Right
0
(a) Left and Right symmetry breaks
(b) Right action preferred in s=0
(c) Wait action preferred in s=0
(d) Right action preferred in s=1
and Wait preferred in s=0

Figure S.1. Cartoon depicting different stages of the Q-table for the toy model. The Q-values of the different actions are shown for each state for each of the stages of the dynamics discussed in the text.

Supplementary Material

S.1 A toy example scenario

To gain some insight into why the emergence of the terminal states discussed in section 2.2 makes sense, it is imperative to consider the simplest version of the model: when at each trial u, the coherence level c^u can be either +1 or -1 and $\sigma = 0$, $\epsilon = 1$, $\gamma = 0$ and $\beta = \infty$. In this case, the Q-table evolves deterministically and the actions are also taken deterministically, thus making understanding the evolution of the Q-values considerably easier. As will be shown below, in this case after a number of trials the Q-table converges to a configuration, whereby in state 0 the Wait action will have the largest Q-value while in states -1 and 1, Left and Right actions will have the largest Q-values, respectively. Starting from state, 0, the agent will start with Wait then move to state 1 or state -1 depending on the coherence level, and then respectively choose the Right or the Left action.

The Left/Right Symmetry of the Q-table breaks. Without loss of generality, suppose $c^1 = 1$. The agent makes a Right or Left actions by chance after a number, Γ, of Wait actions. Since the Q-values are initialized at zero, this happens with probability $(1/3)^{\Gamma+1}$ and the trial ends with the agent in state $s_\Gamma = \Gamma$. Assuming $\Gamma \geq 1$, at the end of the first trial, we have $Q(s,:) = (0, 0, R_{\text{wait}})$ for $s = 0, \cdots, \Gamma-1$. For $s_\Gamma = \Gamma$ at which the terminating action is taken, two possibilities exist: $Q(\Gamma,:) = (R_{\text{correct}}, 0, 0)$ with probability 1/2 when the final action is correct (Right), or $Q(\Gamma,:) = (0, R_{\text{wrong}}, 0)$ with probability 1/2, when it is incorrect (Left). In both cases the *Right* action ends up being the preferred action in $s = \Gamma$. The important point now is that the symmetry between between $s > 0$ and $s < 0$ is now broken, as a consequence of performing the first trial. This is shown in Fig. S.1a.

One of Left or Right actions becomes the preferred action at $s = 0$. At the beginning of this trial, $Q(0, \text{Wait}) < Q(0, \text{Left}) = Q(0, \text{Right}) = 0$, so Left or Right actions will be chosen with equal chance, and the trial ends. If $c^2 = 1$, taking the Left action is incorrect and turns $Q(0, \text{Left}) = R_{\text{wrong}}$, while taking the Right actions is correct, turning $Q(0, \text{Right}) = R_{\text{correct}}$. In both cases, the Q-values of the remaining two actions do not change, and, since $R_{\text{wrong}} < R_{\text{wait}} < R_{\text{correct}}$, the Right action will have the largest Q-value at state zero; see Fig. S.1b. Similarly, if $c^2 = -1$, then the Left action will have the largest Q-value at state zero.

At the beginning of the next trial, then, the Right action will always be immediately taken. Note that if we had $\Gamma = 0$ in Trial 1, the agent would end up in this situation at the beginning of Trial $u = 2$ and the following would ensue.

Wait becomes the preferred action at $s = 0$. Suppose that $c^2 = 1$. If $c^3 = 1$, this will be the correct decision, and $Q(0, \text{Right})$ will further increase. This continues to be the case in all subsequent trials until a trial with $c = -1$ is encountered. When the first $c = -1$ trial is observed, the Right decision that the agent takes in state zero is incorrect, and thus $Q(0, \text{Right})$ decreases. After a number of trials, which depends on how many of the preceding trials had $c = 1$, as well as how large $|R_{\text{wrong}}| - |R_{\text{correct}}|$ is, eventually, $Q(0, \text{Right}) < Q(0, \text{Wait})$. If $c^3 = -1$, this happens at the third trial itself.

In other words, after a sufficient number of trials, not only the symmetry between $s > 0$ and $s < 0$ is broken ($Q(s, \text{Wait}) < Q(s, \text{Right}) = Q(s, \text{Left})$ for $s = 1, \cdot, \Gamma - 1$, and $Q(s, :) = 0$ for $s = -\Gamma, \cdots, 1$), the Wait action is now the preferred action at $s = 0$; see Fig. S.1c. This is also the case if $c^2 = -1$.

Right action becomes preferred at $s = 1$. In the trial that follows the formation of this structure in the Q-table, the agent first Waits, and if $c = 1$, moves to state $+1$. Here it will take Right and Left actions with equal chance leading to $Q(1, \text{Right}) = R_{\text{correct}} > Q(1, \text{Left}) > Q(1, \text{Wait})$, or $Q(1, \text{Right}) = 0 > Q(1, \text{Wait}) > Q(1, \text{Left}) = R_{\text{wrong}}$: the Right action will end up having the largest Q-value in state $s = 1$; see Fig. S.1d. If $c = -1$, the agent moves to state -1, where all actions have Q-value zero, and then everything will be similar to the case of Trial $u = 1$ above but with Right and Left actions reversed. Eventually, the Left action becomes the preferred action in state -1.

Given a sufficient number of trials with right- or left-wards stimuli with similar frequency, converges to a Q-table where the largest Q-values for the -1, 0, and 1 correspond to those of the Left, Wait, and Right actions, respectively.

S.2 Effect of the Learning Rate ϵ and number of states M

The trajectory that Q-values take during training and consequently the development of the terminal states shown in Fig. 2 (see section 2.2) is of course dependent on the learning rate, ϵ. Fig. S.2 this dependence, showing that for the model with a higher learning rate, the states that become the terminal states are further to the right or left of those with smaller learning rates, of course, is such states exist. Naturally, then, the accuracy and reaction time achieved by the model after a fixed number of trials depends on both M and ϵ as shown in Fig. S.3 for $U = 900$. For any given learning rate, ϵ, average reaction time (Fig. S.3(a)) and average accuracy (Fig. S.3(c)) quickly asymptotic as M is increased. Fig. S.3(b)-(d) show the same data re-plotted with the learning rate, ϵ, on the x-axis and the number of states M determining the color code. For very small values of M, average reaction times show a nonlinear relationship to the learning rate and accuracy remains very low. As M is increased to $\gtrsim 40$, however, average reaction time shows progressively more linear relationship to the learning rate.

S.3 SAT During learning: Observation Learning and Time Dependent Waiting

As discussed in section 3, when the Q-values are all initialized at zero, the reaction times at the beginning of training are small while they increase as training progresses. This pattern is not consistent with several previous empirical studies and in this section, we propose two alternative ways to resolve this inconsistency.

S.3.1 Observational learning

To explain this inconsistency between empirical observations and theoretical observations, we reiterate that our RL model starts with a blank slate, i.e., without any experience of what the stimulus consists of and how it relates to action and outcome. This is not the case in empirical studies of decision making neither in human nor in animal experiments. Human participants are often given a set of "instructions" explaining to them, often verbally, what they are about to experience and what they are expected to do. Moreover, in nearly all psychophysical experiments, before starting the experiment, participants get a few "warm up" trials in which the experimenter demonstrates the stimuli, the correct choices, the actions, and the corresponding rewards. In animal studies too, experiments are preceded by extensive and elaborate acclimatization and habituation procedures to familiarize the animal with the environment and its affordances such as the stimuli, and the available actions. These procedures are necessary to reduce the animal's anxiety in the novel environment and to draw its attention to

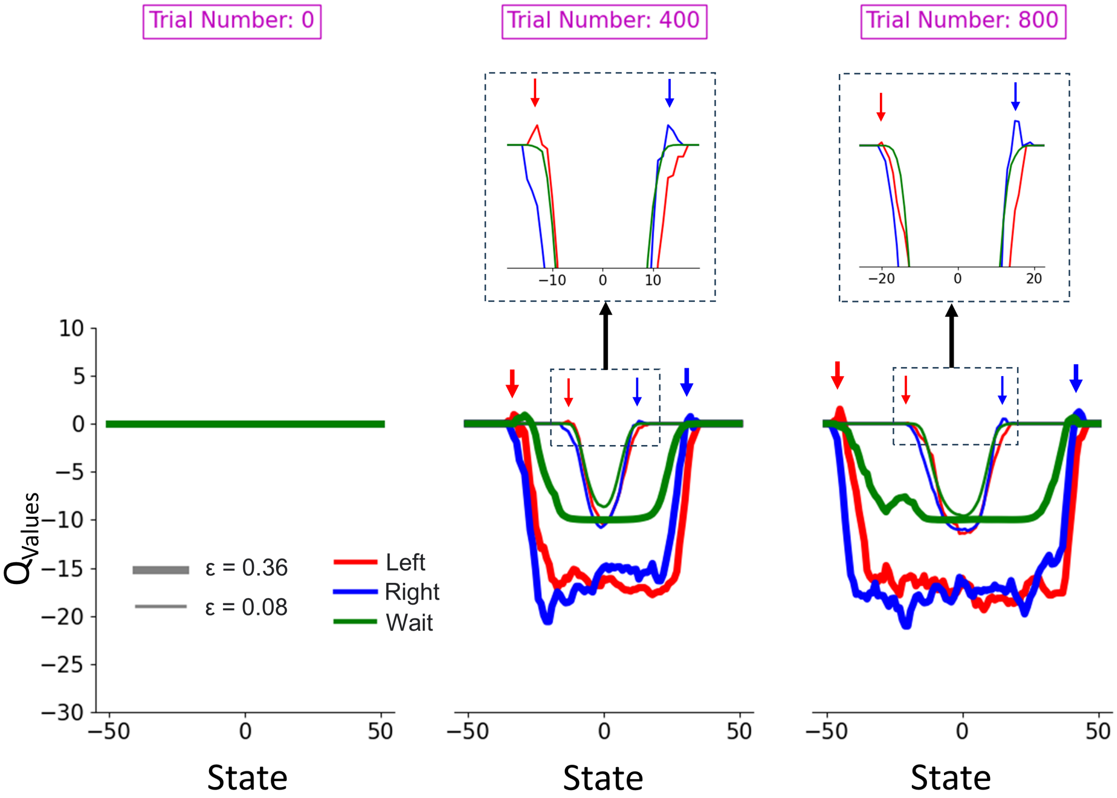

Figure S.2. The effect of learning rate on q-values during learning. The thick lines are for $\epsilon = 0.36$. The thin lines indicate q-values under $\epsilon = 0.08$ (other parameters are kept the same). For clarity of terminal state positions, the zoomed-in versions are depicted on top of the middle and right panels. For a given number of training trials, larger ϵ spans wider over state space. Model parameters are $U = 900$.

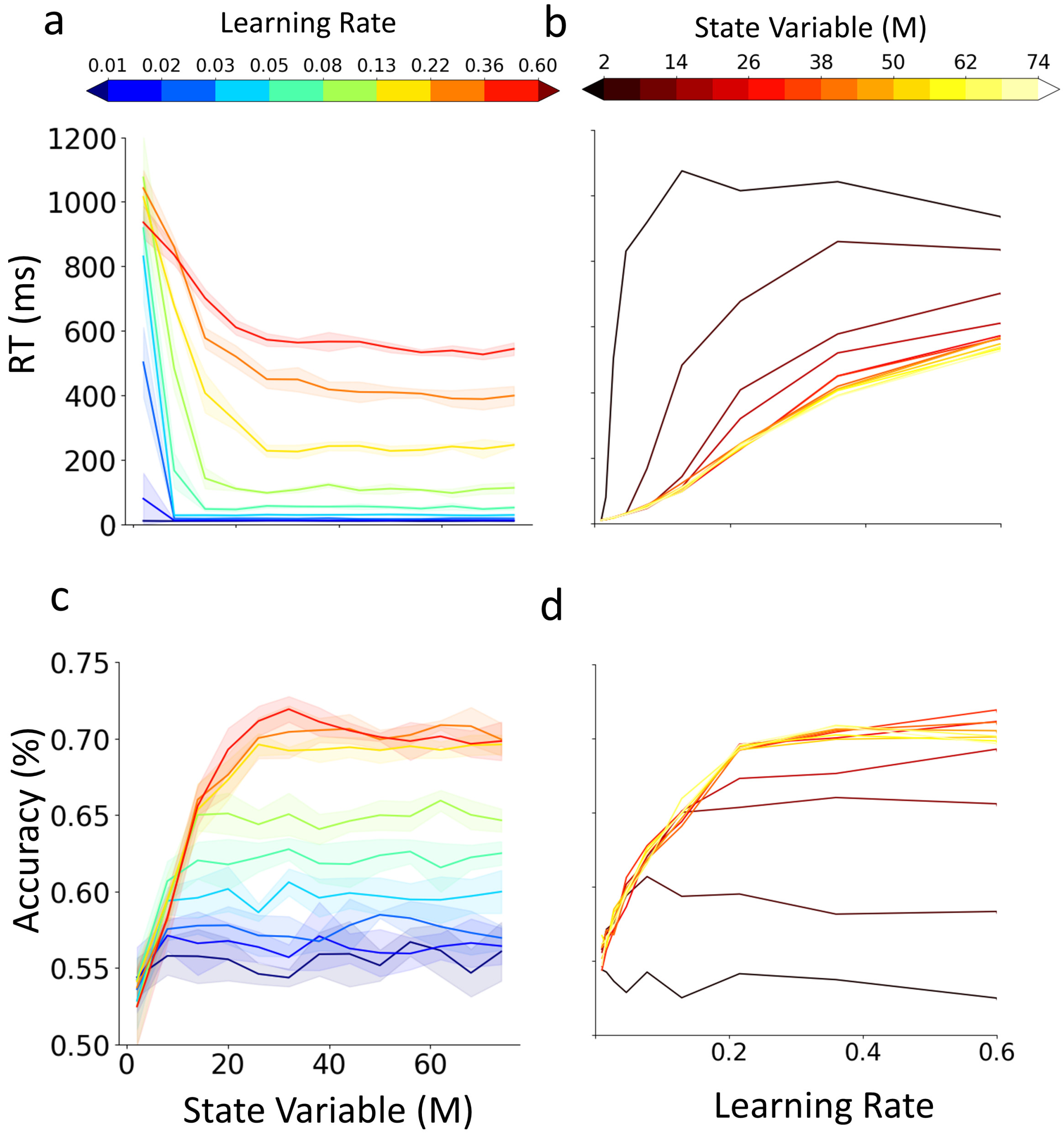

Figure S.3. Effect of learning rate and the available number of states on Model's behaviour (RT and Accuracy) during training (a) Average RT (y axis) decreases with increasing the number of available states (x axis). The exact shape of this relationship depends on the learning rate (color bar). (b) Average RT (y axis) does not have a monotonic relationship to learning rates (x-axis). Here, colors represent the number of available states. (c) Same as (a) but for accuracy. (d) Same as (b) but for accuracy. In both (a) and (c), the shaded areas are standard deviations over interactions (n=5). Here, we used a version of the model with no TD and with hard max (namely, $\gamma = 0, \beta = \infty$)

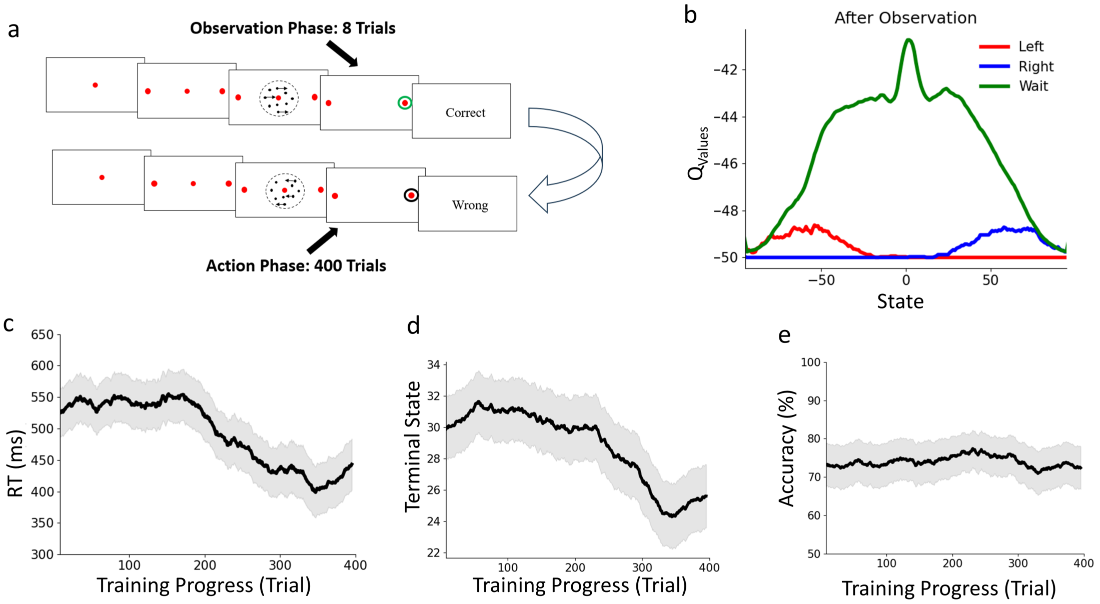

Figure S.4. Model behaviour following a period of observational learning. (a) Observational Learning paradigm. Following [49] the model *observes* the stimulus and the behaviour of a demonstrator in 8 trials. In each trial, a highly coherent stimulus (51.2%) is presented for a fixed duration of 300 time steps followed by the execution of the correct action and reception of the corresponding reward. The model updates its state and Q-table accordingly. (b) Q-value at the end of observational learning, averaged over 100 simulations. The Q-values at the end of the observational learning is used as the initial condition for the model going through 400 trials of learning similar to those the main text (Eqs. (1), (2), (3)). (c) The reaction times, (d) terminal states and (e) choice Accuracy as learning progresses. In (c)-(e) Solid curves depicts averages over 50 simulations, each smoothed as in Fig. 4, and grey areas indicate SEM. The rewards set used for this analysis is $\mathbf{R} = \{100, -50, -1\}$.

the relevant components of the experimental setup. These procedures are often specific to the lab, animal species, experimental question, and the experimenter's personal style and experience. The impact of these informal procedures, both in humans as well as other animals, on behavior is unknown and often ignored under the convenient assumption that the scientific question of the study must be more interesting and important than issues as trivial as the impact of instructions. Indeed, our own formulation of the RL model described so far did not take this issue into account. To close this gap, we simulate the instructions that human participants receive in psychophysical experiments on decision making by giving the model a short "warm-up" with observational learning before embarking on its instrumental learning task.

In addition to learning by doing, one can acquire new behaviors, skills, or information by watching the actions and outcomes of others. In the RL framework, the observational learner can update its value function (for example, the Q-table in our formulation here) by observing a demonstrator's chosen actions and their outcomes [49, 50, 51]. We had the model with the blank slate initial Q-table observe a demonstrator going through several demonstration trials of the random dot motion task Fig. S.4(a). In each trial, the stimulus had high coherence (%51.2). The demonstrator chose the Wait action at consecutive time steps of the trial while the observing learner updated its state variable and Q-table. When the stimulus duration came to its end, the demonstrator chose the correct terminal action and received the reward R_{correct}, again with the observing learner updating its Q-table. Fig. S.4(b) shows the observer's Q-table after 8 demonstration trials, each of them comprised of 400-time steps. This *instructed* model was then tasked with learning to make its own decisions. Here, the model started with long reaction times and became progressively faster without losing accuracy (Fig. S.4(e)), via lowering the terminal state (Fig. S.4(d)).

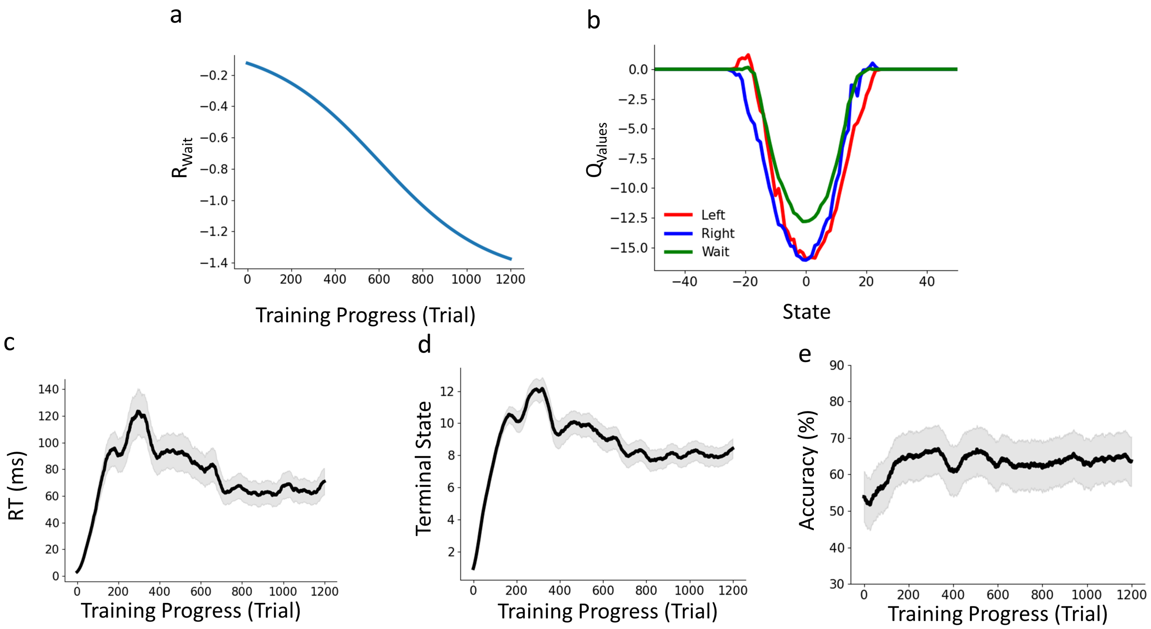

Figure S.5. Effect of time-dependant cost of waiting. (a) Cost of waiting increases (that is $R_{\text{wait}} < 0$ decreases) as trials progress (Eq. (7)). (b)-(e) Same as Fig. S.4(b)-(e) but for the case time-dependent R_{wait} and and 50 simulations.

S.3.2 Time Dependant Waiting

Another plausible solution for the discrepancy between the results of Fig. 4 is making the cost of the Wait action, R_{wait} time-dependent, e.g. by using the following

$$R_{\text{wait}}(u) = \frac{R^{\infty}_{\text{wait}}}{1 + e^{\lambda(\tau - u)}} \tag{7}$$

where R^{∞}_{wait} defines the asymptotic value of $R_{\text{wait}}(u)$ as time approaches infinity, λ determines how quickly $R_{\text{wait}}(u)$ approaches this value and τ determines the point at which $R_{\text{wait}}(u)$ is halfway to R^{∞}_{wait}.

In the simulations in this section, we used $R^{\infty}_{\text{wait}} = -1.5$, $\lambda = 0.004$ and $\tau = 600$, resulting in $R_{\text{wait}}(u)$ as plotted in Fig. S.5(a) with the cost of waiting at the start of the experiment to be $R_{\text{wait}}(0) = -0.12$.

In the beginning, since the cost of waiting is not significant, the model tends to move toward higher states as lower states are perceived as less rewarding. This results in an increase in both response time (RT) and accuracy. This behavior is also typically observed in the base model (see Fig. 4). However, as the cost of waiting progressively increases (Fig. S.5(a)), the model ceases to advance to larger states. Instead, it is enforced to make decisions more quickly by reducing the terminal state. Consequently, after an initial rise, the model begins to decide faster (see Fig. S.5(c)). This reduction in the terminal state, however, is not substantial enough to affect accuracy. Therefore, after an initial rise, accuracy remains unchanged (see Fig. S.5(e)).

S.4 Comparison with optimal terminal state for $\Delta = 1$

In section 5 of the main text, we defined optimal terminal state as the value of B that maximizes $ER(B, K)$ in Eq. (5) with other parameters fixed. Since Eqs. (4), and Eq. (5), are often defined over a continuous range for B, in that section we considered the case of of finer discretization of the state space by assuming $\Delta = 0.1$ in Eq. (1).

Here in Fig. S.6, we show what happens if the discretization is not as fine: we choose $\Delta = 1$, a value that we have used in our other simulations. Everything here is the same as Fig. 7, except for Δ. One can see that, overall, the results do not qualitatively change.

Quantitatively, however, there are some small differences. Firstly, as expected, the terminal state of the model reaches the optimal after fewer learning trials (Fig. S.6(b) vs Fig. 7(b)). Perhaps more noteworthy is the fact that when plotted versus CBR, the variability in the terminal state reached by different simulations of the model is larger for $\Delta = 1$ than $\Delta = 0.1$ (shaded magenta Fig. S.6(c) vs Fig. 7(c) but that the average is still close but mostly smaller than the optimal terminal state. Since, similar to the corresponding figure in the main text (Fig. 7(b)-(c)), these results are for $c = 0.064$, in Fig. S.6(d) we also compare the terminal state reached by the model with the optimal terminal state for different coherence levels. Although the overall pattern is the same as Fig. 7(d), with $\Delta = 0.1$, the model terminal state typically is at higher values and in fact overshoots that of the optimal states for a range of c.

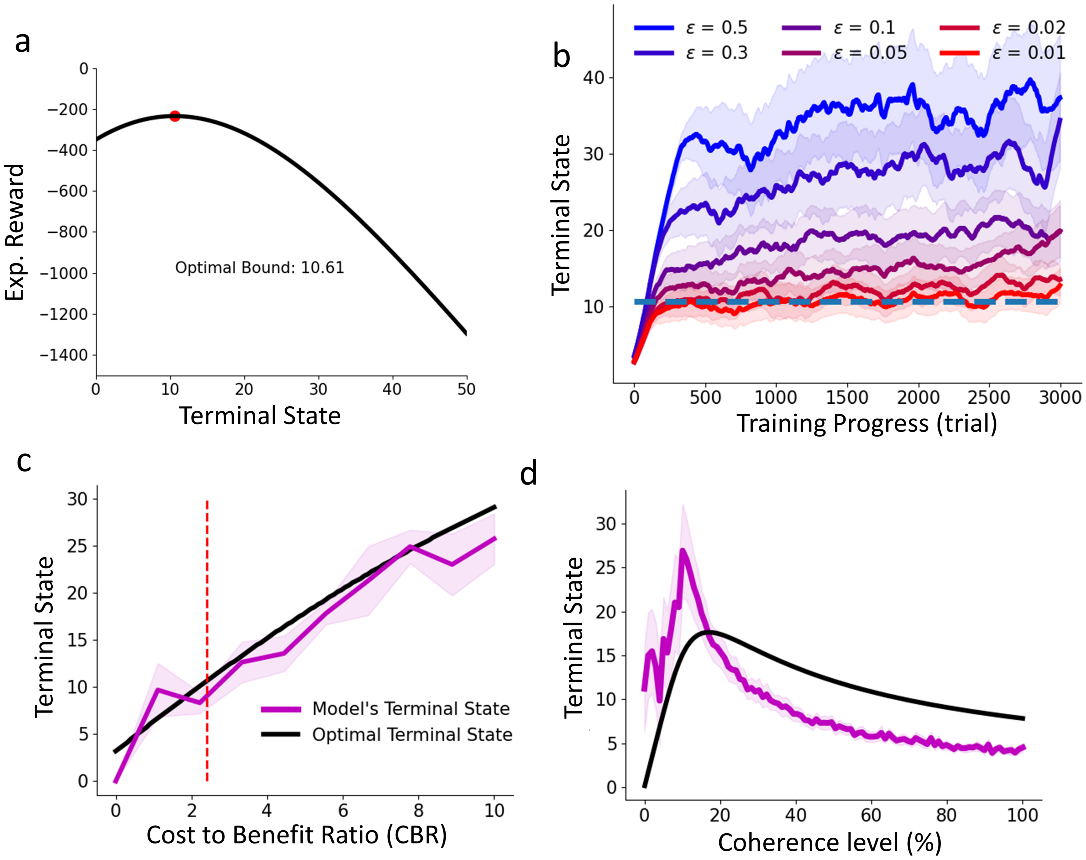

Figure S.6. Optimality of the terminal states for $\Delta = 1$. Everything is the same as Fig. 7, except that we now have $\Delta = 1$.

S.5 The effect of various parameters on the terminal state

For the majority of the analyses shown in this work, we used parameter values reported in section 2.1. In Fig. S.8, we report a systematic study of the impact of each parameter on the terminal state reached by the model. We used a fixed set of parameters for each analysis while systemically changing the parameter of interest. The number of training trials was fixed at $U = 900$ and the terminal states at the last 100 of these were averaged.

S.6 Replication of previous empirical findings

Similar to any other cognitive computational model, our model could serve as a virtual participant in numerous empirical studies found in the literature of perceptual and value-based decision making and test whether one could ”replicate” those findings. Importantly, our model could be tested at the level

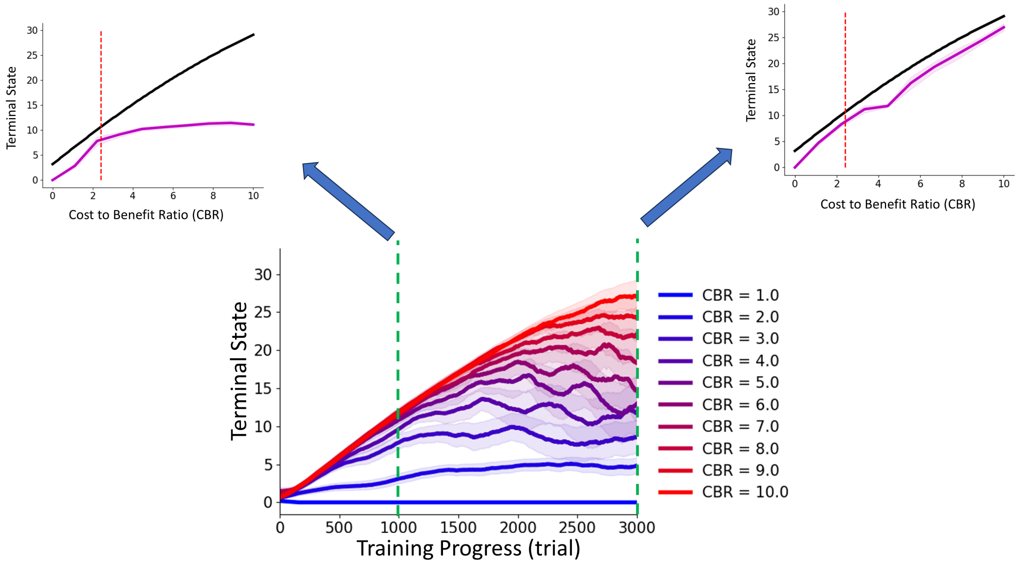

Figure S.7. Sensitivity of the model's optimal terminal state to the number of trials. If we simulate the model for $U = 1000$ the models lose its sensitivity to high values of CBR (inset, left). Yet, if $U = 3000$ the model will be more sensitive to CBR and show a behavior similar to the optimal model.

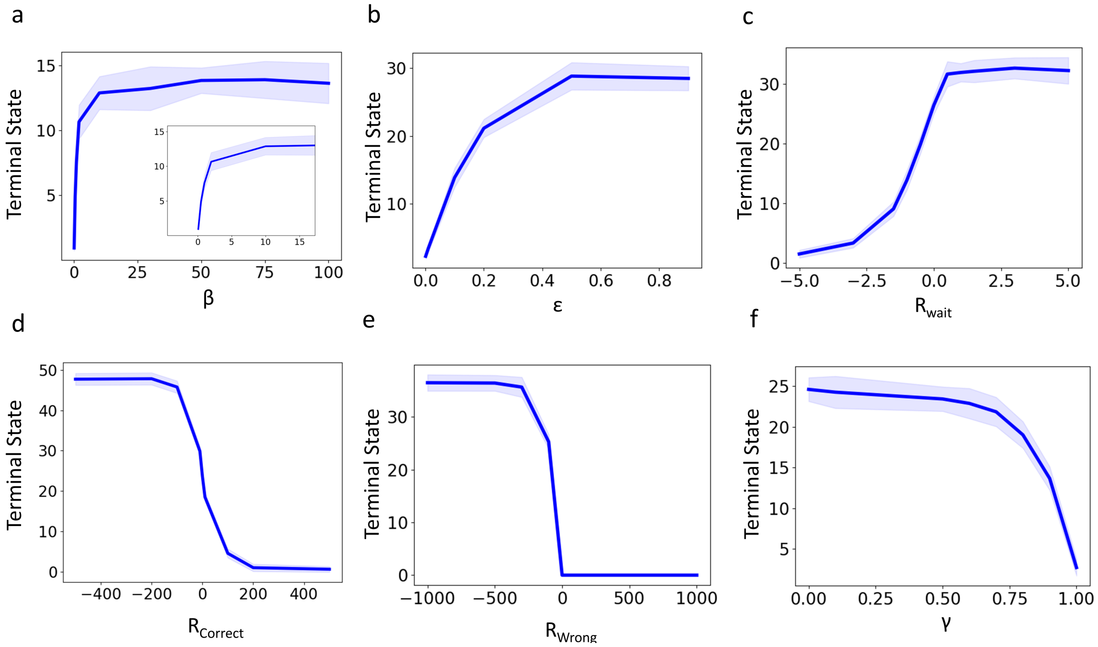

Figure S.8. Systematic study of parameters respect to the terminal state. The shared parameters that have been used in this study are as follows $\mathbf{R} = \{20, --50, -1\}$, $\gamma = 0.9$, $\beta = 50$, $\epsilon = 0.1$, the maximum number of samples per trial = 1000. In each panel, the parameter of the interest (x-axis) was systematically altered while the other parameters were fixed as above. (a) study of β (Eq. (2). The shaded area is STD over simulations (N=30). Inset is the zoomed-in version of (a) for smaller values of β. (b) Same as (a) but for ϵ in Eq. (3). (c) Same as (a) but for reward of waiting. (d) Same as (a) but for the reward of being correct. (e) Same as (a) but for reward of being wrong. (f) Same as (a) but for γ in Eq. (3).

of its behavior (i.e., reaction time and accuracy) as well as at its ”mechanistic” level, for example via examining the model’s Q-table structure and the dynamics of its state transitions.

Building up from our basic model (Eq. (1)-(3) and section 2.3), we present three successful replications and one important failed but instructive replication attempt. Brefily, we replicated the effects of disproportionate training sets ([20], Fig. S.9), reverse pulse ([44], Fig. S.10) and volatility ([52], Fig. S.13). Our basic model could not replicate the effect of differences in RTs for correct vs error trials (Fig. S.11(a)). Motivated by this failure, we introduced a more sophisticated version of the model with an urgency component. This new model replicated the error vs correct RT effects (Fig. S.11(b)) as well as the post-error slowing effect (Fig S.12) as reported in [53].

S.6.1 The impact of disproportionate training set on choice bias and reaction time

To study the role of prior expectations on perceptual decision-making, Hanks *et al* [20] examined human and macaque monkey motion discrimination behaviour under two different schedules differentiated by the frequency of stimulus categories. In *balanced* blocks, the two stimulus categories (e.g., leftward and rightward) were equally (50:50) likely. In the *disproportionate* blocks, one motion direction was 4 times more likely than the other (i.e., 80:20). The results showed that in these latter blocks, after 300-400 trials, choices favored the more frequent category (Fig. S.9(a)-(b) inset) for which RT was also faster.

We compared our basic model’s behaviour (Eq. (1), (2) and (3)) under the two schedules (i.e., 50:50 and 80:20). We began by training two instances of the model in the same balanced schedule (3600 trials). Then, one instance went through another 900 trial of learning in the same balanced schedule. The other instance, however, received 900 trials of the disproportionate schedule. The results did produce the choice (Fig. S.9(a)) and RT bias (Fig. S.9(b)) favoring the more frequent category, replicating the empirically reported findings (Fig. S.9(a)-(b) inset). Examining the evolution of the terminal states in the two instances (50:50 in Fig. S.9.c and 80:20 in Fig. S.9(d)) showed that after the initial balanced schedule, the terminal states of the two instances were positioned in identical positions and as expected, did not change in the subsequent 900 test trials. In the 80:20 schedule, (Fig. S.9(d)) however, the terminal state corresponding to the more probable category drifted slowly closer and closer towards zero. The opposite observation was made for the terminal state corresponding to the less probable category.

This replication is also important for another related reason. Following the standard practice empirical studies of decision making (see Introduction), Hanks and colleagues discarded the first 300-400 trials of the disproportionate blocks in their experimental data. In these initial trials, the behavior was deemed too unstable for the DDM models to cope with. Here, the results in panels a-b of Fig. S.9 by follow [20] and do not include the initial 300 trials of the test phase. But, and here is the important point, the model can also be interrogated during that initial window (the blue areas in Fig. S.9(c)-(d)), thus drawing very specific predictions for the very same discarded period. As such, our model gives unprecedented insights into processes that decision neuroscience, especially in animal studies, has so far discarded as uninterpretable and therefore uninteresting.

S.6.2 Reverse pulse effect

The standard drift diffusion model predicts that if a perceptual stimulus is arranged such that statistically *zero evidence* is presented to the observer, the process of evidence accumulation should prolong and decision be postponed such that reaction times increase without any measurable effect on accuracy. This prediction was tested empirically by the so-called *reverse pulse* paradigm [44]. Specifically, in this paradigm, at some point during the presentation of the random dot motion stimulus, a brief (200 ms) period of *zero evidence* is surreptitiously inserted into the motion stimulus. In the first 100ms of this period, the random motion stimulus follows the statistics dictated by the coherence level of the trial. In the second 100ms of this period, motion stimulus was constructed by *reversing* the first half. In this way the motion evidence in the first and second half should cancel the accumulated evidence from one another. Importantly, this trick was only feasible to test in the low coherence trials otherwise the subjects could clearly notice the motion reversal in the middle of the trial. Behavioral results (Fig. S.10(b)-(c) insets) confirmed the predictions: insertion of the reverse pulse increased the reaction times but did not affect the accuracy.

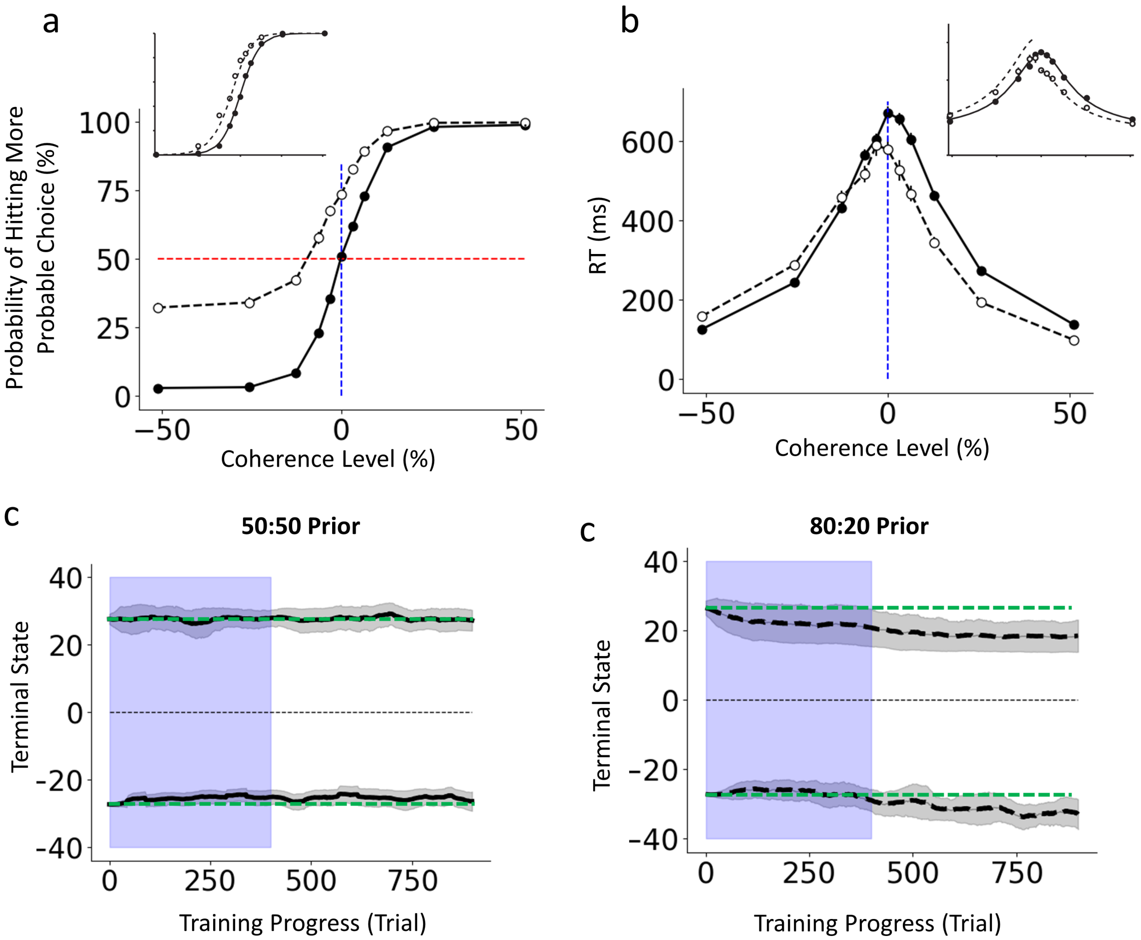

Figure S.9. The impact of disproportionate training schedule on choice bias and reaction time. Choice (a) and RT (b) in different training regimes (Solid line: 50-50 prior, dashed line: 80:20 prior, each condition is simulated with 900 trials). Model behaviour becomes more biased (in both decision and RT) toward more frequent (here: rightward) decision under 80:20 regime. Insets are previous findings from by [20]. Error bars in (a) and (b) are SEM over simulations (N = 30). (c) Evolution of terminal states during 50:50 prior training schedule. The highlighted area is where the data are deemed to be unstable in ref [20]. Grey-shaded areas are SD over simulations (N = 30). the green dashed line shows the terminal state of the model *before* entering balanced or unbalanced conditions. (d) same as (c) but for 80:20. Here, the terminal state is significantly lower than the initial terminal state (green dashed line)

.

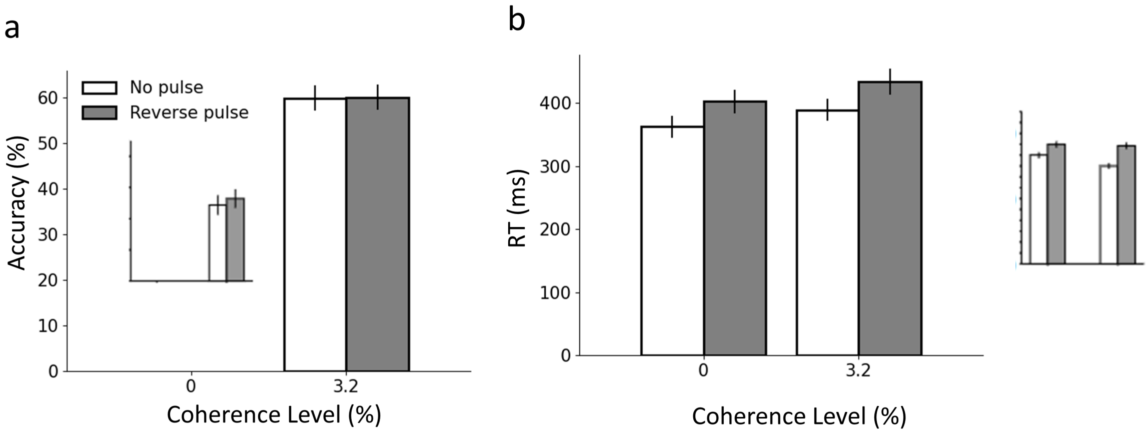

Figure S.10. Replication of the reverse pulse effect. (a) Accuracy and (b) reaction times draw from simulation that compared the basic model performance with (grey) and without (white) reverse pulse. Inset show the previous finding from ref. [44]. Error bars are 95% Confidence Interval across trails (U = 2400).

To replicate this finding, after training two identical instances of our basic model (learning rate=0.1; reward set = 20, -50, -1), we tested one instance with and the other without reverse pulse inserted into their stimulus sequence (Fig. S.10). Except for the reverse pulse window, the sequences of motion stimuli in the two conditions were carefully matched and coherence was restricted to very low values (0% and 3.2%) trials. The trained model (Fig. S.10) replicated the empirically observed behavior (Fig. S.10(b)-(c)) of longer RTs and no change in accuracy.

S.6.3 Reaction times in correct and error trials: the role of urgency signal

A common observation in perceptual decision making is that, controlling for task difficulty, error reaction times are longer than those of correct decisions [26]. This intuitive empirical fact has proven to be a critical challenge for many models of decision dynamics. The canonical form of the drift-diffusion model, for example, cannot account for the longer error reaction times [26, 27, 54]. Our basic model (Eq. (1)-(3)) too is unable to reproduce (Fig. S.11(a)) this widely reported empirical observation.

To address this challenge, previous works have proposed a number of different solutions including collapsing boundaries, urgency signal, and drift rate variability. One idea is inspired by the intuition that in the absence of convincing evidence, we may lower our bar for what counts as acceptable [55]. Under high uncertainty, this idea suggests, the boundary come closer to the starting point as time elapses. The longer the evidence accumulation continues, the lower the decision threshold and thus, the higher the probability of erroneous decisions. In simulations, collapsing boundaries could reproduce the difference between correct vs. error RT [55]. Physiological evidence, however, has not supported this proposition. For example, [53, 3] showed that the maximum firing rate of decision-boundary neurons in macaque lateral intraparietal (LIP) area do not differ between error and correct trials. An alternative that shares some mathematical similarity to collapsing boundaries is the urgency signal. The intuition behind this idea is that the agent prefers to make faster decisions and when the trial takes longer, some sort of internal urgency builds up in the agent, eventually compelling it to commit to one of the choice alternatives even when evidence is not particularly great. This urgency signal has been implemented as an additive term that added to the accumulated evidence. This urgency term monotonically increases with time does not depend on stimulus uncertainty and has been supported by recent neurobiological evidence [56, 54, 24].

Inspired by the idea of the urgency signal [54], we modified our basic model by adding an urgency term to the stimulus evidence, replacing E_t in Eq. (1) by U_t defined such that:

$$U_t = E_t + \text{sgn}(E_t)\rho t \tag{8}$$

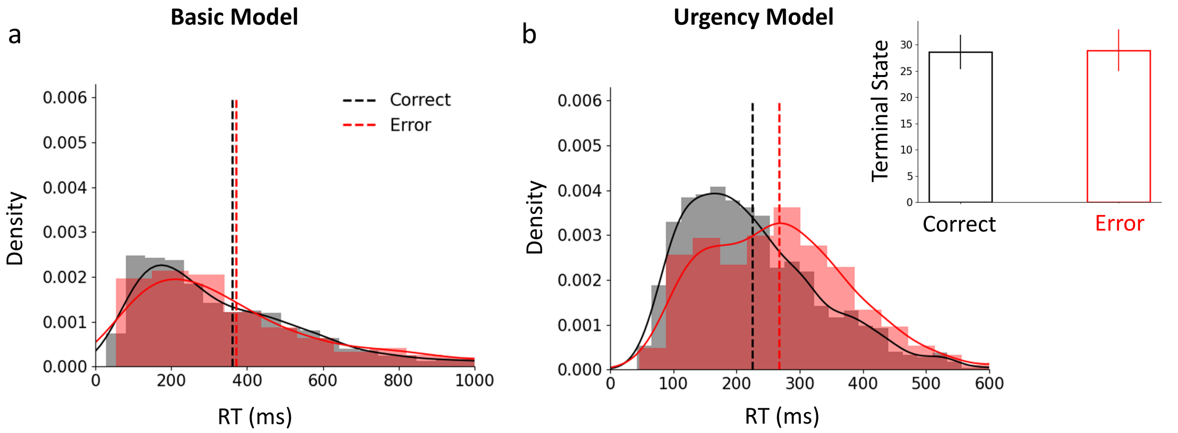

Figure S.11. The urgency effect (a) In the basic model, average error and correct RT are identical. (b) In the urgency model, however, error RTs are longer. The level of coherence is fixed at 12.8% for this simulation. (inset): Average and STD (error bar) of terminal state for correct (black) and error (red) trials. The difference is not significant (t(2399)=1.5, p=0.13).

where $E_t = \mathcal{N}(Kc^u, \sigma^2)$, ρ is the normalization factor (set to 0.005), sgn is the sign function and t is time within a trial [54]. In this formulation, the second term is a time-dependant signal that is added to the stimulus evidence (E_t). Replacing Eq. (1) with Eq. (8), we obtained a new model in which the state variable accumulates both time and evidence. Our simulations showed that this urgency model does reproduce the predicted longer error RT (Fig. S.11(b)). Importantly, comparing the position of terminal states in the Q-tables in correct and error trials we observe that terminal states are similar (Fig. S.11(b) inset). Given that the position of the terminal states corresponds to what DDM literature calls the decision boundary, our simulation results are in line with the previous physiological findings [53, 55, 3] that reported identical level of neural firing for boundary neurons in correct and error trials.

S.6.4 Post Error Slowing

Post Error slowing (PES) is one of the oldest and most well-known behavioural observations in decision sciences [57]. After an error, reaction time in the next trial is slower compared to a trial of the same level of difficulty that follows from a correct outcome. This intuitively understandable and well-known behavioural effect has presented substantial challenges to the computational and neurobiological studies of decision making.

In one study, Purcell et al [53], investigated PES using the same RDK paradigm that we have focused on here and reported that although post-error reaction times were longer than post-correct, accuracy was not different between them. They also examined the monkey LIP's firing rates at the time of response in order to compare the decision threshold in post-error and post-correct trials and did not observe any difference between them. These two sets of findings are problematic because, within the framework of sequential sampling models including DDM, PES is clearly explained as an adjustment of the decision threshold in the current trial based on the previous outcome. This prediction entails that accuracy should also be higher after an error vs after a correct choice. Purcell et al's empirical observations do not fit this prediction, neither at the level of behavior nor neural activity. An alternative suggestion that has been proposed to explain PES is that following an error, sensitivity (i.e., drift rates in DDM) could be modified. However, the empirical observations do not fit this idea either. To account for their findings, Purcell et al. proposed a model in which changes in *urgency* (implemented as collapsing bound) were combined with changes in sensitivity to obtain a model that delivers a change in reaction time but preserves accuracy in post-error trials (see figure 2 of their study [53]).

Earlier, in S.6.3 we introduced urgency in our framework. Here we show that the same formulation, when combined with the outcome from the previous trial provides a more parsimonious account that is consistent with [53] without having to assume any changes in sensitivity. Following [53], we define

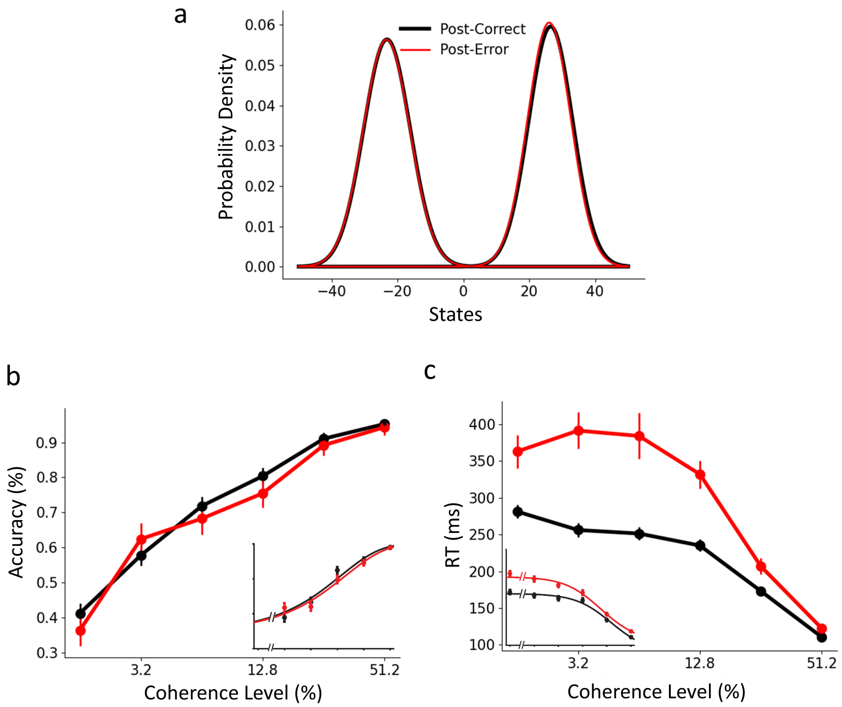

Figure S.12. Replication the Post Error Slowing. (a) Distribution of terminal states in post-error (red) vs post-correct (black) trials. (b) Accuracy of post error vs post correct trials. The accuracy did not change, similar to a previous study. (c) Same as (b) but for RT. Post error trials are significantly slower than the post correct trials; in line with what has been reported before (inset). Error bars are SEM across simulations (U = 2400). Insets show empirical data from [53]

the level of urgency (Eq. (8), also see S.6.3) as a function of outcome in the previous trial as follows:

$$\rho_u = \begin{cases} 0.001 & \text{if } R_{u-1} < -1 \\ 0.003 & \text{if } R_{u-1} > 0 \end{cases} \tag{9}$$

Where R_{u-1} is the reward in trial $u-1$ and ρ_t is the normalization factor, in trial u in equation (8).

The results of the model simulation (number of trials $U = 2400$; learning rate $= 0.1$; reward set $\{20, -50, -1\}$ for correct, error and wait, respectively) are plotted in (Fig. S.12). Data from the second part of the simulation (i.e., trial 1200 to 3600) is shown here. Each trial has been labeled according to its preceding outcome. Several observations are evident. Fig. S.12(a) shows that the terminal states (i.e., decision threshold) do not differ between the post-error and post-correct trials. As would be expected from this fact, at the level of behaviour, no difference is observed in accuracy (Fig. S.12(b)). RTs, however, are longer in post-error trials (Fig. S.12(c)). These findings are in line with previous empirical observations e.g., ref. [53] . The key difference here is that the model proposed here is simpler and more parsimonious involving outcome-dependent modulation of urgency without requiring any modulation of sensitivity.

S.6.5 Impact of evidence volatility on behavior

Studies examining perceptual decision making under uncertainty most often look at the relationship between the first moment, i.e., *mean* of the perceptual evidence (e.g., coherence level in random dot motions, luminance contrast in oriented gratings, vibration amplitude in somatosensory psychophysics) and the behavior [26, 1, 2]. This relationship is indeed captured by our basic model too (3(a)-(b)). Fewer studies have examined the role of the second moment i.e., variance (but see here for a review of second-order perception [58]). A recent study [59] examined human behavior (accuracy and reaction

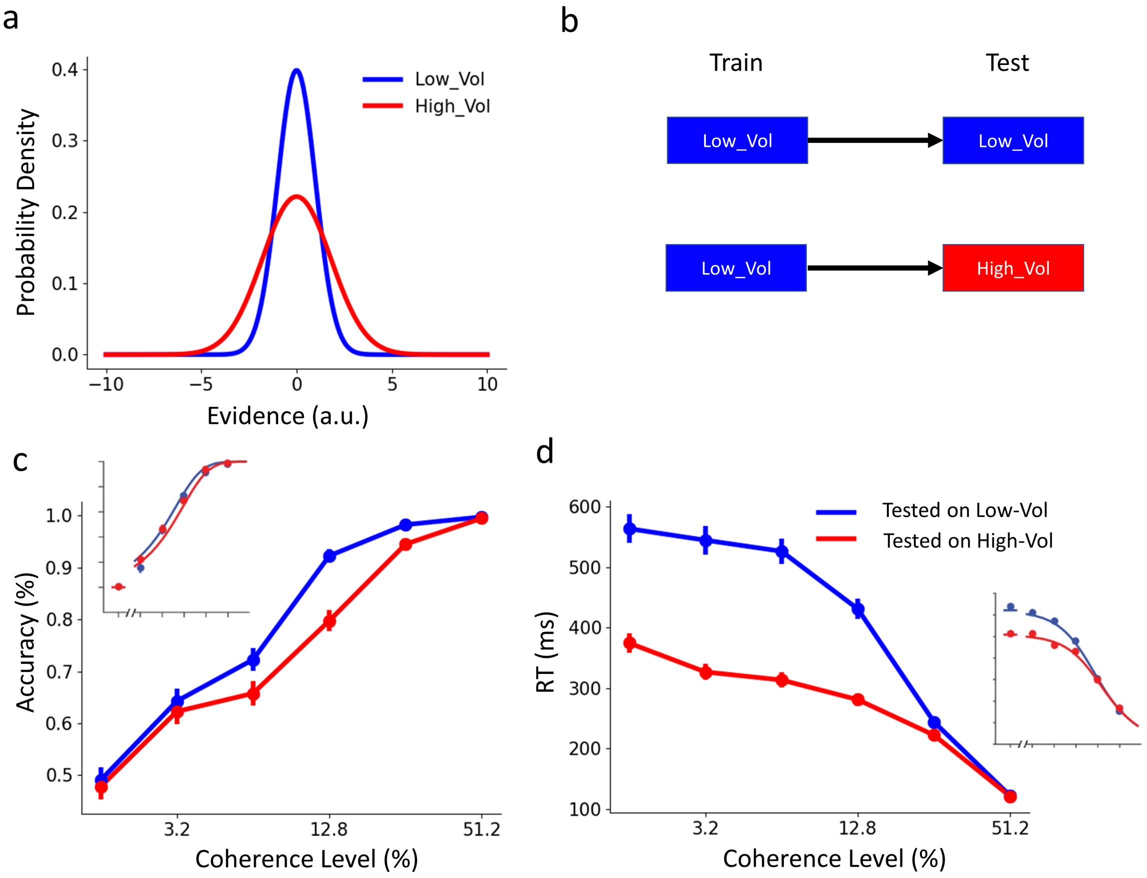

Figure S.13. Replication of the volatility effect. (a) Depiction of the distribution of evidence in two volatility conditions in 0% coherence level. Both conditions have a similar mean (zero) but the variance (volatility) is higher in the high-volatility condition (red) than in the low-volatility condition (blue). (b) The procedure of our simulations. In both volatility conditions, we first trained the model on low-volatility evidence. Then we test the model based on high (red) and low (blue) volatility trials. (c-d) Model behaviors in different volatility conditions. Test accuracy (c) and RT (d) of the model are lower in high volatility condition (red) similar to previous findings [59] (insets). Error bars are SEM across trials (U = 2400).

times) when the mean perceptual evidence was kept constant but its volatility was systematically manipulated. Under high volatility, both accuracy and RT were, reduced (Fig. S.13.c-d inset).

Following [59], we first created two volatility conditions (Fig. S.13(a), low volatility (blue): $\sigma = 1$ in E_t in Eq. (1), high volatility (red): $\sigma = 1.3$). Then, we trained our basic model on low volatility conditions (Fig. S.13(b)) (learning rate = 0.1, U=2400 and tested the trained model under two different conditions of volatility (Fig. S.13(b)). Our simulations showed that under high volatility testing conditions, reaction times and accuracy both decreased (Fig. S.13(c)-(d)) thereby neatly replicating the empirical findings of [59].

5 Discussion

The main contributions of this thesis are twofold. First, we propose a brain-plausible biophysical computational model based on attractor neural networks to explain the neuro-mechanisms underlying social decision-making. Second, we provide a computational framework for understanding and modelling the learning processes involved in making perceptual decisions.

To achieve the first contribution, we employed a fixed duration random dot motion (RDM) task with varying levels of coherence. Subjects participated in two sessions of data collection: (1) Isolated and (2) Social. In the social setting, they were paired with a computer-generated partner whose behaviour was modelled on the subject's behaviour in the isolated session. Subjects were instructed to maximize joint accuracy, which was determined by the decision of the party with higher confidence. Behavioural, eye-tracking, and EEG data were collected from the subjects. The prominent behavioural effect observed and expected was that subjects matched their confidence levels to those of their partners, a phenomenon known as confidence matching (Bang et al., 2017, 2020; Esmaily et al., 2023). We developed an attractor neural network framework (Wong & Wang, 2006) to account for this effect. Through two studies, we demonstrated that a simple confidence-dependent top-down modulation could explain confidence matching. Importantly, we validated and strengthened our model using eye-tracking and EEG data. The model underwent extensive validation checks, including model comparisons and testing different alternative hypotheses.

Second, we address the challenge of understanding the training phase of perceptual decision-making. This effort led to the development of a model aimed at elucidating the mechanisms underlying the training process in perceptual decision-making tasks. By focusing on this training phase, we aimed to uncover the neural and cognitive mechanisms that facilitate the transition from novice to expert decision-makers. This model provides insights into training dynamics and may help optimize training protocols in decision-making studies.

Specifically, we combined reinforcement learning (RL) and the drift diffusion model (DDM) frameworks to simulate perceptual decision-making during training. In our model, the RL component is responsible for learning the decision boundary, which was previously assumed to be a fixed or explicit parameter. By allowing the decision boundary to be learned, we can better explain the decision dynamics during training. Through extensive experiments and analyses, we demonstrated that our model could effectively learn how to make nearly optimal (Watkins & Dayan, 1992) perceptual decisions through time. This may open up new avenues for analysing and utilizing training data that may have previously been considered too messy or useless. Our approach not only enhances the theoretical understanding of decision-making but also has practical applications in improving training methodologies and decision-making strategies across various fields. By addressing these challenges and providing robust solutions, our work may pave the way for future research and applications that can leverage these insights to achieve more effective and efficient training outcomes. In the following sections, we will discuss the considerations and limitations of each project separately.

5.1 Considerations

Here, we provide some challenges and considerations that we addressed in each study. These challenges and considerations are important to highlight as they offer a clearer and more comprehensive understanding of the processes and intricacies each study encountered. We have made concerted efforts to design our experiments, analyses, and computational models to ensure that our findings are reliable and robust. This involved carefully considering the experimental design, the types of data collected, and the ways in which these data could be integrated into our computational models. By addressing these challenges head-on and incorporating a range of considerations into our approach, we aimed to introduce models that not only reflect the complexities of human decision-making and learning but also stand up to rigorous scientific scrutiny.

5.1.1 Project (1)

5.1.1.1 Extensive training

Social neuroscience studies in humans, often involves studies with a large number of participants (N > 30) to ensure the robustness and generalizability of the findings (Bahrami et al., 2010, 2012; Bang et al., 2017; Najar et al., 2020). These studies aim to understand complex social behaviours and interactions, which typically require diverse and extensive data sets to capture the variability in human behaviour and social cognition (Tump et al., 2020, 2022). The larger sample sizes help in achieving statistical power and detecting subtle effects that might be lost in smaller samples. This approach aligns with the field's focus on understanding broad, population-wide phenomena.

In contrast, perceptual decision-making neuroscience focuses on the neural mechanisms underlying decision-making processes by recruiting only few subjects (T. Hanks et al., 2014; Kiani et al., 2014; Purcell & Kiani, 2016a; Vafaei Shooshtari et al., 2019). These studies often involve extensive training of a small number of subjects (N < 10) to ensure consistent and reliable data. The emphasis here is not on learning *per se* but on understanding how decisions are made, based on experimental condition of interest, once the task is well-learned. This approach allows researchers to collect detailed neural data, such as neuroimaging or electrophysiological recordings, from highly trained individuals (T. D. Hanks et al., 2011; Okazawa & Kiani, 2023; Resulaj et al., 2009; Zylberberg et al., 2016).

Many of the computational models employed in individual decision-making studies typically do not account for learning processes (Kiani et al., 2014; Ratcliff et al., 2006; Ratcliff & Starns, 2013; Wang, 2002; Wong & Wang, 2006). Thus, it is advisable to apply these models to subjects who have completed their learning phase, ensuring that no further learning affects their decision-making during the study. To effectively adapt these perceptual models for use in social decision-making research, it is essential to conduct training with a large number of subjects (N > 30). This approach helps to mitigate the influence of learning variables and allows for the assessment of a broad range of subject variability. Addressing this challenge was a critical aspect of the research presented in this thesis, highlighting the complexities involved in adapting individual decision-making models to social contexts.

Through two studies, we recruited more than 30 subjects, of whom 27 successfully passed the training phase. Each individual underwent at least 3,000 trials to ensure that decision-making variables such as reaction time (RT) and accuracy were stable. Notably, we did not impose any instructions or limitations on confidence reporting. Each subject participated in at least five sessions of data collection, which led to some attrition as a few subjects did not show up after the initial sessions. Once we were satisfied with the stability of their behaviour, we began the data collections presented in Project 1. This demanding approach was necessary to ensure that there were no perceptual learning effects in our data. By rigorously training the subjects, we aimed to eliminate the impact of learning variables and achieve a consistent level of decision-making performance. This rigorous preparation ensured that our subsequent analyses could reliably reflect the underlying cognitive processes without being confounded by ongoing learning dynamics.

5.1.1.2 Using EEG and Eye tracker

In the first study, we used eye-tracking and EEG data to validate and strengthen our findings. Specifically, pupil data was used to test whether subjects' beliefs about their decisions are influenced by their partner's confidence. In our attractor neural network model, the partner's confidence altered the firing rate of the model, which may imply that decision evidence, and thus certainty, could also change. Therefore, the model predicts that subjects' beliefs about their decisions should be affected by their partner's confidence. This prediction was confirmed by our analysis of pupil data, which showed changes in pupil size corresponding to the partner's confidence levels.

Similarly, the model also predicts that the ramping of the firing rate, a proxy for evidence accumulation rate, is dependent on the partner's confidence. EEG data, particularly the central-parietal positivity (CPP) component, can measure the rate of evidence accumulation. Our EEG analysis supported the model's prediction, showing that the rate of evidence accumulation was higher when subjects were paired with a high-confidence partner compared to a low-confidence partner. This alignment between EEG data and model predictions further validates our model, indicating that the confidence of a partner influences both the neural mechanisms and subjective beliefs involved in social decision-making.

Importantly, we used pupil and EEG data as complementary measurements to behavioural data. We were particularly interested in measurements relevant to the model but not directly observable through behaviour alone, such as the rate of evidence accumulation and subjective belief about decisions. In our approach, the modelling part was assumed to be the main focus, with EEG and eye-tracking data used to inform and enhance the model. Unlike some studies that aim to find neural correlates to model parameters (Bang et al., 2020; Mahmoodi et al., 2022), Project 1 seek to understand what additional insights neural data can provide to the model that behavioural data alone cannot.

Neural correlates studies are undoubtedly important; model parameters are implicit and varied measurements of behavioural data, and their correlation to neural data is not trivial at all (Gold & Shadlen, 2007; T. Hanks & Summerfield, 2017; Okazawa & Kiani, 2023). We, however, chose to use EEG and eye-tracking as explicit measurements of decision-making processes, verifying and guiding our model with these data. This approach allows us to leverage the strengths of neural measurements to add depth and validation to our computational model, beyond what behavioural data can achieve alone.

5.1.2 Project (2)

The first version of our model was based on zero-step Q-learning (e.g., no temporal difference (TD) term) (Sutton & Barto, 2018) and the action selection method was set to greedy (hardmax) (Sutton & Barto, 2018). The entire analysis and results presented in Project 2 were initially conducted using zero-step Q-learning. Logically, for a new model like this, it made sense to start with the simplest form and then expand it. This incremental approach allows for a clear understanding of the model's basic mechanics before adding complexity. However, without the TD term, conducting optimality analysis was difficult because the optimality theory of Q-learning necessitates the presence of the TD term (Watkins & Dayan, 1992).

To address this, we re-analysed the entire results using a general Q-learning algorithm that included the TD term. Moreover, we changed the action selection policy to SoftMax to allow for exploration (Sutton & Barto, 2018), which may help find optimal solutions. Our initial analysis showed that our Q-learning algorithm drastically failed to follow the normative model when using the original reward set. This observation convinced us to conduct an exhaustive search over the reward landscape to pinpoint a reward set where our model could indeed exhibit behaviour similar to the normative model.

Upon identifying a compatible reward set, we re-analysed everything with this new set of rewards. This was done successfully, and we demonstrated that the model could reproduce all results with the new parameter sets. However, we also aimed to emphasize that even the old reward set could simulate the decision behaviours observed during learning. It remains unclear why the old dataset did not follow the normative model. Potential reasons for this discrepancy include the discretization of the state, the number of trials, lack of sufficient exploration, and the initial values of the Q-table. Yet, the actual reason for this discrepancy is not entirely understood.

To highlight this discrepancy, we chose to use the old reward set for the majority of the analysis (to show it works) and contrasted that with the optimality analysis parameter set. This approach allowed us to clearly demonstrate the differences and provide insights into the conditions under which our model aligns with normative models. By doing so, we underscored the importance of the parameter set, an area we did not fully explore. Future work is needed to study these parameters extensively to identify the range and boundaries that should be used for fitting the model to actual data. This thorough examination will help ensure that the model can be robustly applied across various scenarios, improving its accuracy and reliability in representing real-world decision-making and learning processes.

5.1.2.1 Cost of Waiting

One of the major differences between our RL model and conventional RL models is the presence of a waiting action. While the inclusion of waiting actions is not entirely new (Drugowitsch et al., 2012), our framework may offer a novel procedure for extracting the cost of waiting from data. Intuitively, we can expect that different individuals perceive the cost of waiting differently. By using our model to measure this "hastiness" in individuals, we may provide a pathway for connecting these low-level measurements to high-level psychological traits obtained through questionnaires.

This approach could bridge the gap between quantitative model outputs and qualitative psychological assessments, offering a more comprehensive understanding of decision-making behaviours. Although there is a long and challenging journey ahead to fully grasp these extremely complex phenomena, our model may help advance our understanding by providing a method to quantify and analyse the cost of waiting. This can potentially lead to insights into individual differences in decision-making processes and how they relate to broader psychological traits, paving the way for future interdisciplinary research in this area.

5.2 Future Directions and Limitations:

Here, we provide an overview of the limitations encountered in each study and outline future directions for research building on our findings. These limitations and future directions are important to highlight as they offer a clearer and more comprehensive understanding of the constraints in our research and potential areas for improvement and exploration. Throughout our studies, we meticulously designed our methodologies to ensure the robustness and reliability of our findings. However, it is equally important to acknowledge the inherent limitations that may have impacted our results and to discuss the pathways for advancing our understanding and addressing the gaps identified in our current work.

By discussing these limitations, we aim to provide a balanced perspective on our research, emphasizing the need for cautious interpretation and identifying potential for future work to address these issues. As with any scientific endeavor, there are always new questions and avenues to explore that can further deepen our understanding. The proposed future directions are not only logical extensions of our current work but also represent critical steps toward a more comprehensive and nuanced understanding of these cognitive processes. Embracing these future directions will help to refine existing models and address current limitations. By discussing both limitations and future directions, we aim to encourage continued exploration and innovation in the field, building on our initial findings and pushing the boundaries of what is known.

5.2.1 Project (1)

The confidence recorded in the first study is clearly a combination of the stimulus and the partner's response (both their decision and confidence). Our computational model integrates information about both the stimulus (bottom up) and the partner's response (top down) to compute confidence. However, the measurement of confidence in our model is entirely dependent on neural activity or the decision variable. This approach assumes that the reported confidence of the subject is solely based on their neural activity or decision evidence, without accounting for the subjects' strategic behaviour. For instance, a subject might strategically report high confidence in one trial to influence the collective decision, even if they do not genuinely feel highly confident about their own decision. In this scenario, they might have the same firing rate as they would when reporting low confidence, but their readout of the decision variable differs due to strategic considerations. This indicates that their true certainty based on neural evidence may not always be accurately reflected in their reported confidence.

Although our analysis of pupil data and the model's fit to behavioural data support the model's assumptions to some extent, a comprehensive framework should account for different types of confidence: belief-based confidence and strategic confidence. Belief-based confidence is directly tied to the neural decision evidence, whereas strategic confidence involves a higher-level decision process where subjects might adjust their reported confidence to influence group outcomes or fulfil other objectives. These objectives might be highly context-dependent, with subjects applying different functions to read out neural activities in each context. Our primary focus in Project 1 was to provide a context- and strategy-independent notion of group confidence formation. However, expanding our model to incorporate more context-dependent scenarios could be an interesting avenue for future research.

5.2.1.1 Dynamic Pairs

Project 1 used computer-generated partners to simulate static partners for each subject. This approach provided us with substantial control over designing the partners in a way that could precisely manipulate the variables of interest. However, in real life, partners are not static. Although Bang et al. demonstrated that confidence matching occurs in both static and dynamic settings (Bang et al., 2017), extending our computational framework to include two dynamically interacting individuals is not trivial.

A dynamic partner can adapt and change its behaviour based on the subject's actions, creating a more complex interaction pattern that must be accounted for in the model. This complexity introduces additional variables and dependencies that need to be considered, such as real-time adjustments in confidence levels and strategies based on ongoing feedback. Therefore, while our current model provides valuable insights, future work should aim to develop and validate frameworks that can handle dynamic interactions.

Similar to our current approach, EEG and pupil data could provide insights into the neural mechanisms underlying these dynamic processes. A dynamic social decision-making setting with hyperscanning (Nam et al., 2020), which allows simultaneous recording of neural activity from both interacting individuals, could be an excellent direction for future research. This approach would enable a deeper understanding of how real-time adjustments and mutual influences occur in social decision-making, offering a more comprehensive view of the underlying neural dynamics.

5.2.2 Project (2)

5.2.2.1 The RT Effect

Our models' simulations demonstrate a notable phenomenon: reaction time (RT) tends to increase over time. Initially, decisions are fast, but as the model refines its decisions for accuracy, and longer RTs —an observation contrary to human and rat data. These studies indicate that during initial training, agents often exhibit slower, less precise responses that progressively accelerate and become more accurate.

We have explored several modifications to our model to better align with these observed behaviours. However, a critical question remains: could this behavioural profile be context-dependent? Might instructional cues influence RT profiles? Could external factors like starvation (as seen in monkey studies compared to rat studies) also impact RT during learning? Currently, data is insufficient to provide definitive answers.

To address these gaps, we designed an experiment where subjects perform a perceptual decision-making task with a reward incentive, aiming to maximize their total reward. Crucially, subjects receive no specific instructions beyond the task requirement and can only interact using two keys. The sole communication about the task is: "Maximize your reward."

The key question is whether subjects will opt for rapid, albeit less accurate key presses, or if they will take more time to carefully observe stimuli before responding. Insights gained from this experiment may shed light on the potential impact of instruction on RT.

5.2.2.2 RL and the Attractor Model

In Project 2, we used one of the most widely adopted models of decision-making: the Drift Diffusion Model (DDM). As discussed in Project 1, while the DDM is a highly convenient and interpretable framework, it lacks biological plausibility. This raises an important question: how can our reinforcement learning (RL) framework be adapted to more biologically plausible models of decision-making, such as attractor neural networks?

We have already demonstrated that our RL framework can be seamlessly integrated with non-accumulation-based models. Extending this to the attractor model is both feasible and promising. Similar to other decision-making models we used within our framework, the key variable of interest remains the decision boundary. In the context of an attractor network, the state could be defined as the level of activity across two neural populations—one coding for left decisions and the other for right. As in our model, the state could be encoded as a discrete variable (e.g. positive values indicating dominance of the rightward population, and negative values indicating the leftward population). The available actions would still be "left" and "right" to terminate the trial, and a "wait" action to continue the dynamical evolution of neural activity, effectively allowing evidence accumulation to proceed. The reward function could be defined analogously, reinforcing correct decisions, penalizing incorrect ones, and assigning a small cost to waiting. In this way, the RL agent would learn the dynamics of the decision boundary in a setting that is fundamentally different from DDM, yet more biologically grounded.

A future direction worth exploring is the comparison of how the learned decision boundary evolves in the attractor model versus in the DDM. Such a comparison could yield insights into both the behavioral and neural signatures of decision-making, and inform the development of more unified models of learning and decision processes.

5.2.2.3 Model Fitting

The entire results of Project 2 were derived from model simulations. Our approach was to first understand the model, its dynamics, and theoretical underpinnings, and then move to experimental validation; simply put: theory first, then experiment. However, many open questions cannot be answered solely through model simulations: In a given set up, what would happen to subjects' decision dynamics during learning? These questions require empirical data to answer and cannot be resolved through theoretical analysis alone.

To address these questions, the model needs to undergo a model fitting procedure. This involves defining free parameters, establishing a cost function, and implementing a fitting procedure. Model recovery is essential to ensure the validity of our approach before fitting the model to actual data. We have attempted this procedure, but the nature of our model, which updates at each time step, makes the fitting process more challenging than fitting a simple RL or DDM model alone (Gherman & Philiastides, 2018; Ratcliff & McKoon, 2008). Future work must tackle this problem to make the model's implications clearer. Without fitting, the practical applicability of our model remains uncertain. The ability to fit the model to empirical data is crucial for validating its predictions and understanding its real-world relevance. Solving this fitting challenge will enable us to provide more concrete answers to the open questions and enhance the model's utility in explaining and predicting human decision-making behaviour.

To this end, we designed an experiment in which subjects would go through two separate sessions, spaced weeks apart, with different cost-to-benefit ratios (2 and 3). The core idea is to assess whether we can accurately explain a participant's data from one session using a model fitted on data from the other session. Specifically, we aim to determine how well parameters such as the cost of waiting, learning rate, and exploration coefficient generalize across different reward sets. Our primary focus is on the generalization of the cost of waiting. If we can successfully recover the cost of waiting in each session, it would indicate that this measurement is inherently subject-dependent and orthogonal to the context (reward set). Conversely, if the recovery of the cost of waiting is unsuccessful, it may suggest that the cost of waiting is influenced by the reward set. This study will provide valuable insights into the notion and nature of the cost of waiting, enhancing our understanding of whether this parameter is an intrinsic trait of the subject or contextually driven by the specific reward conditions.

We have already collected data from 10 subjects across two sessions and observed that both accuracy and reaction time (RT) increase with the cost-to-benefit ratio (CBR), consistent with our model's predictions. However, a more formal evaluation of the model and a thorough inspection of the cost of waiting require a proper fitting procedure.

5.2.2.4 Drift Rate During Learning

Our RL model is designed to model the decision boundary. It assumes that the agent already knows how to accumulate evidence but does not know where to place the decision boundary on the accumulated evidence. This assumption might not be entirely correct. Indeed, many studies that model perceptual learning focus on modelling the drift rate (Fontanesi et al., 2019, 2019); this means the agent is exposed to evidence but does not know how to use or accumulate it effectively. There is also neural evidence to support this assumption. For instance, Law et al. (Law & Gold, 2008, 2009) showed that the supposed evidence accumulation area of the brain (LIP) does not show any decodability to different amounts of evidence at the beginning of training, while MT decoding remains well and unchanged by training. This might imply that while evidence is encoded in the brain (MT) consistently, the agent does not initially know how to accumulate it. Therefore, several studies model the drift rate (as a proxy for evidence accumulation) while keeping the decision threshold explicitly defined (Fontanesi et al., 2019).

On the other hand, the same studies that provide evidence for learning in evidence accumulation (Law & Gold, 2008, 2009) may also suggest that the decision threshold increases over time. This aspect has been neglected by previous studies but has been addressed in our research (Masís et al., 2023). To sum up, previous models have typically modelled the drift rate while keeping the decision boundary explicit (either constant or explicitly defined). Our model takes the opposite approach by modelling the decision threshold while keeping the drift rate constant. This approach allows us to explore how the decision threshold evolves with learning, providing insights into how individuals adjust their decision-making criteria over time. By focusing on the decision threshold, we aim to capture a different dimension of the decision-making process that has been overlooked (Masís et al., 2023). This shift in focus helps us understand the balance between the cost of waiting and the benefits of accurate decision-making, as individuals learn to optimize their thresholds to maximize rewards.

Considering both neural and behavioural studies, it is reasonable to assume that an agent might neither know how to accumulate evidence nor where to set the decision boundary initially. Therefore, extending our model to include the evolution of the drift rate is a necessary future step. This extension could help create a more complete picture of perceptual learning by addressing both aspects of the decision-making process.

5.2.2.5 Analytical Understanding

As we discussed earlier, our results in Project 2 are based on simulations. We started with a set of assumptions, ran the model, and gathered the results. Ideally, we would have an analytical understanding of the model, where the decision threshold is explicitly defined as a function of the learning rate, coherence, reward set, and time. While simulations are very helpful, they do not provide the complete picture. It is only through an analytical understanding of our model that we can fully grasp its function, dynamics, and behaviour.

There is one study that attempted to analytically explain how the Q-table for a bandit task evolves over time using path integral analysis (Li & Yeung, 2023). Adapting this methodology to our model may be a promising approach for future research. By applying path integral analysis or similar analytical techniques, we could potentially derive explicit equations that describe the evolution of the decision threshold as a function of various parameters. This would significantly enhance our theoretical understanding. Analytical insights would allow us to better interpret the simulation results and refine the model to capture the complexities of perceptual learning and decision-making more accurately.

5.2.2.6 Decision Making Model of Learning

Our model can learn decision making. But what kind of decision making? The precise mechanisms underlying decision-making remain a topic of debate, both theoretically and empirically. Empirical evidence suggests that boundary neurons do not always follow a simple accumulation process. Instead, they exhibit a range of heterogeneous patterns that challenge the traditional accumulation assumptions of drift diffusion models (Heitz & Schall, 2012). Crucially, we have shown that our model can seamlessly adapt to an *extreme detection* decision-making framework and still produce the canonical behavioral features we expect from a decision-making agent. This flexibility highlights the model's robustness, yet it also underscores a key limitation: the decision-making framework itself—whether based on accumulation or extreme detection—needs to be explicitly defined within our computational architecture. Specifically, the model requires an explicit definition of the transition probability $P(s'|s,a)$ which represents the probability of transitioning to a new state s′ given the current state s and action a. Interestingly, the primary distinction between an accumulation model and an extreme detection model lies in the definition of $P(s'|s,a)$ and nothing more. If we extend our framework to allow $P(s'|s,a)$ to be learned rather than predefined, we could unlock the potential to address more fundamental questions about how decision-making processes are acquired. This extension would allow the model to dynamically infer the most appropriate transition dynamics based on the observed behavior of decision-making agents during learning.

By investigating the properties of $P(s'|s,a)$ within such an adaptive framework, we may gain critical insights into which decision-making model—accumulation or extreme detection—aligns more closely with empirical observations of human and animal behavior. This approach could ultimately help resolve long-standing debates about the true nature of decision-making mechanisms in the brain. Moreover, enabling the model to learn $P(s'|s,a)$ could reveal how decision-making strategies evolve in response to different contexts and uncertainties. This adaptability may offer a deeper understanding of the cognitive flexibility observed in real-world decision-making, where agents must navigate complex and changing environments. Advancing our computational framework to include learning-based representations of $P(s'|s,a)$ holds promise for bridging the gap between theoretical models and empirical data. It may provide a unified explanation for the diversity of decision-making behaviors and shed light on the fundamental principles governing decision-making in the brain.

5.2.2.7 Confidence

In Project 2, we defined methods to extract accuracy and reaction time (RT) from our model. Accuracy is determined by comparing the chosen terminating action of the model with the ground truth experimental direction. RT is defined as the number of wait actions chosen before any terminating action is taken. Note that our model inherently represents decision time. Similar to the Drift Diffusion Model (DDM), RT can be defined as the sum of decision time and non-decision time (usually estimated during fitting).

Our main focus was to model RT and decision-making, similar to traditional decision-making models, while accounting for the learning effect. However, one of the most important and interesting aspects of decision-making is confidence. Confidence in learning is a crucial aspect that has been the subject of several studies (Balsdon et al., 2020; Drugowitsch et al., 2019). How confidence is defined in our model and how it evolves over time could be a very interesting and important subject for future research. One plausible option for defining confidence could be the accumulated value of the Q-value of waiting action during one trial. However, the precise function this value to confidence needs to be articulated, analysed accurately, and extensively. Future research could explore different formulations and their implications, ultimately enhancing our understanding of confidence in the context of learning and decision-making. This could provide deeper insights into how confidence evolves with experience and how it influences decision-making processes, offering a more comprehensive framework for studying perceptual learning and decision-making.

6 Conclusion and Remarks

In this thesis, we presented two studies related to social decision-making and learning. The first study focused on developing a neurobiological model for social decision-making, particularly elucidating the neural mechanisms that may explain how confidence matching occurs in the brain. We employed EEG and eye-tracking data alongside behavioural data to validate our computational model comprehensively. This study examines how humans use information about the confidence of collaborators to guide their own perceptual decision-making and confidence judgments. We addressed this question through a combination of psychophysics, neural and eye data, and computational modelling, resulting in a compelling validation of a framework that can be used to derive and test theory-based predictions about how collaborators use communication to align their confidence and thereby optimize their collective performance.

Studying social decision-making using methods proposed for individual decision-making can be challenging. It requires a large number of subjects (N~30) who undergo extensive training, which is a significant undertaking. In many perceptual decision-making studies, data from the training phase are often discarded because they are considered noisy, messy, and sometimes useless. Our second project aimed to address this blind spot by providing a computational model that attempts to explain decision dynamics during learning.

We developed a perceptual decision-making model embedded within a reinforcement learning (RL) framework that learns how to make decisions over time. Specifically, we created a framework that learns where the decision boundary should be. Through trial and error, the model seeks to find the optimal decision threshold that balances the cost of waiting against external rewards. Our framework allows decision-making and learning variables to be altered, studied, and extended, providing a tool for exploring these processes.

These studies, although conducted in simplistic and controlled scenarios, aim to enhance our understanding of how humans make decisions and learn. Our research offers a pathway for better understanding the intricate processes underlying human decision-making and learning. By advancing these models and methodologies, we hope to contribute to the development of more effective strategies for enhancing decision-making performance in both individual and collaborative settings.

7 General References

Arabadzhiyska, D. H., Garrod, O. G. B., Fouragnan, E., Luca, E. D., Schyns, P. G., & Philiastides, M. G. (2022). A Common Neural Account for Social and Nonsocial Decisions. *Journal of Neuroscience*, *42*(48), 9030–9044. https://doi.org/10.1523/JNEUROSCI.0375-22.2022

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1350–1365. https://doi.org/10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Balsdon, T., Mamassian, P., & Wyart, V. (2021). Separable neural signatures of confidence during perceptual decisions. *eLife*, *10*, e68491. https://doi.org/10.7554/eLife.68491

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1753. https://doi.org/10.1038/s41467-020-15561-w

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., Lau, J. Y. F., Latham, P. E., Bahrami, B., & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 0117. https://doi.org/10.1038/s41562-017-0117

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private–public mappings in human prefrontal cortex. *eLife*, *9*, e56477. https://doi.org/10.7554/eLife.56477

Barrera-Lemarchand, F., Lescano-Charreau, V., Ruiz, J., Cáceres, N., Carrillo, F., & Navajas, J. (2024). Collective Wisdom and the Fermi Method: Improving the Accuracy of Deliberative Groups. *SCT Proceedings in Interdisciplinary Insights and Innovations*, *2*, 257–257. https://doi.org/10.56294/piii2024257

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152. https://doi.org/10.1016/j.neuron.2008.09.021

Britten, K., Shadlen, M., Newsome, W., & Movshon, J. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, *12*(12), 4745–4765. https://doi.org/10.1523/JNEUROSCI.12-12-04745.1992

Burkitt, A. N. (2006). A Review of the Integrate-and-fire Neuron Model: I. Homogeneous Synaptic Input. *Biological Cybernetics*, *95*(1), 1–19. https://doi.org/10.1007/s00422-006-0068-6

Chancel, M., & Ehrsson, H. H. (2020). Which hand is mine? Discriminating body ownership perception in a two-alternative forced-choice task. *Attention, Perception, & Psychophysics*, *82*(8), 4058–4083. https://doi.org/10.3758/s13414-020-02107-x

Colby, C. L., Duhamel, J. R., & Goldberg, M. E. (1996). Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *Journal of Neurophysiology*, *76*(5), 2841–2852. https://doi.org/10.1152/jn.1996.76.5.2841

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Deaner, R. O., Khera, A. V., & Platt, M. L. (2005). Monkeys Pay Per View: Adaptive Valuation of Social Images by Rhesus Macaques. *Current Biology*, *15*(6), 543–548. https://doi.org/10.1016/j.cub.2005.01.044

Drugowitsch, J., Mendonça, A. G., Mainen, Z. F., & Pouget, A. (2019). Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24872–24880. https://doi.org/10.1073/pnas.1906787116

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, *32*(11), 3612–3628. https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F., & Ostojic, S. (2022). The role of population structure in computations through neural dynamics. *Nature Neuroscience*, *25*(6), 783–794. https://doi.org/10.1038/s41593-022-01088-4

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

Esmaily, J., Abharzad, E., Knogler, S., Deroy, O., & Bahrami, B. (2024). *Raising Social Stakes Raises Confidence* (SSRN Scholarly Paper 4844364). https://doi.org/10.2139/ssrn.4844364

Esmaily, J., Zabbah, S., Ebrahimpour, R., & Bahrami, B. (2023). Interpersonal alignment of neural evidence accumulation to social exchange of confidence. *eLife*, *12*, e83722. https://doi.org/10.7554/eLife.83722

Fontanesi, L., Palminteri, S., & Lebreton, M. (2019). Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: A meta-analytical approach using diffusion decision modeling. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(3), 490–502. https://doi.org/10.3758/s13415-019-00723-1

Ganea, D. A. (2021). *Behavioral and pupillometric correlates of perceptual decision making in mice performing a two alternative forced choice task*.

García-Pérez, M. A., & Alcalá-Quintana, R. (2020). Order effects in two-alternative forced-choice tasks invalidate adaptive threshold estimates. *Behavior Research Methods*, *52*(5), 2168–2187. https://doi.org/10.3758/s13428-020-01384-6

Garritsen, O., van Battum, E. Y., Grossouw, L. M., & Pasterkamp, R. J. (2023). Development, wiring and function of dopamine neuron subtypes. *Nature Reviews Neuroscience*, *24*(3), 134–152. https://doi.org/10.1038/s41583-022-00669-3

Gerstner, W., & Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity* (pp. xiv, 480). Cambridge University Press. https://doi.org/10.1017/CBO9780511815706

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press.

Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife*, *7*, e38293. https://doi.org/10.7554/eLife.38293

Gnadt, J. W., & Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Experimental Brain Research*, *70*(1), 216–220. https://doi.org/10.1007/BF00271862

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Haan, E. H. F. de, & Cowey, A. (2011). On the usefulness of 'what' and 'where' pathways in vision. *Trends in Cognitive Sciences*, *15*(10), 460–466. https://doi.org/10.1016/j.tics.2011.08.005

Hagura, N., Esmaily, J., & Bahrami, B. (2023). Does decision confidence reflect effort? *PLOS ONE*, *18*(2), e0278617. https://doi.org/10.1371/journal.pone.0278617

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *Journal of Neuroscience*, *31*(17), 6339–6352. https://doi.org/10.1523/JNEUROSCI.5613-10.2011

Hanks, T., Kiani, R., & Shadlen, M. N. (2014). A neural mechanism of speed-accuracy tradeoff in macaque area LIP. *eLife*, *3*, e02260. https://doi.org/10.7554/eLife.02260

Hanks, T., & Summerfield, C. (2017). Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, *93*, 15–31. https://doi.org/10.1016/j.neuron.2016.12.003

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121. https://doi.org/10.1016/j.tics.2011.12.007

Hautus, M. J., Shepherd, D., & Peng, M. (2011). Decision strategies for the two-alternative forced choice reminder paradigm. *Attention, Perception, & Psychophysics*, *73*(3), 729–737. https://doi.org/10.3758/s13414-010-0076-4

Hegarty, S. V., Sullivan, A. M., & O’Keeffe, G. W. (2013). Midbrain dopaminergic neurons: A review of the molecular circuitry that regulates their development. *Developmental Biology*, *379*(2), 123–138. https://doi.org/10.1016/j.ydbio.2013.04.014

Heitz, R. P., & Schall, J. D. (2012). Neural Mechanisms of Speed-Accuracy Tradeoff. *Neuron*, *76*(3), 616–628. https://doi.org/10.1016/j.neuron.2012.08.030

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500–544. https://doi.org/10.1113/jphysiol.1952.sp004764

Huettel, S. A. (2010). Ten Challenges for Decision Neuroscience. *Frontiers in Neuroscience*, *4*. https://doi.org/10.3389/fnins.2010.00171

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, *14*(6), 1569–1572. IEEE Transactions on Neural Networks. https://doi.org/10.1109/TNN.2003.820440

Katz, L. N., Yates, J. L., Pillow, J. W., & Huk, A. C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature*, *535*(7611), 285–288. https://doi.org/10.1038/nature18617

Kelly, S. P., & O’Connell, R. G. (2013). Internal and External Influences on the Rate of Sensory Evidence Accumulation in the Human Brain. *Journal of Neuroscience*, *33*(50), 19434–19441. https://doi.org/10.1523/JNEUROSCI.3355-13.2013

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science (New York, N.Y.)*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Klein, J. T., Shepherd, S. V., & Platt, M. L. (2009). Social Attention and the Brain. *Current Biology*, *19*(20), R958–R962. https://doi.org/10.1016/j.cub.2009.08.010

Konovalov, A., & Ruff, C. C. (2022). Enhancing models of social and strategic decision making with process tracing and neural data. *Wiley Interdisciplinary Reviews. Cognitive Science*, *13*(1), e1559. https://doi.org/10.1002/wcs.1559

Labutina, N., Polyakov, S., Nemtyreva, L., Shuldishova, A., & Gizatullina, O. (2024). Neural Correlates of Social Decision-Making. *Iranian Journal of Psychiatry*, *19*(1), 148–154. https://doi.org/10.18502/ijps.v19i1.14350

Latimer, K. W., Yates, J. L., Meister, M. L. R., Huk, A. C., & Pillow, J. W. (2015). Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, *349*(6244), 184–187. https://doi.org/10.1126/science.aaa4056

Law, C.-T., & Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nature Neuroscience*, *11*(4), 505–513. https://doi.org/10.1038/nn2070

Law, C.-T., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience*, *12*(5), 655–663. https://doi.org/10.1038/nn.2304

Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife*, *10*, e63282. https://doi.org/10.7554/eLife.63282

Lee, D. G., Daunizeau, J., & Pezzulo, G. (2023). Evidence or Confidence: What Is Really Monitored during a Decision? *Psychonomic Bulletin & Review*, *30*(4), 1360–1379. https://doi.org/10.3758/s13423-023-02255-9

Li, B., & Yeung, C. H. (2023). Understanding the stochastic dynamics of sequential decision-making processes: A path-integral analysis of multi-armed bandits. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *33*(6), 063107. https://doi.org/10.1063/5.0120076

Lo, C.-C., & Wang, X.-J. (2006). Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, *9*(7), 956–963. https://doi.org/10.1038/nn1722

Loughnane, G. M., Newman, D. P., Tamang, S., Kelly, S. P., & O'Connell, R. G. (2018). Antagonistic Interactions Between Microsaccades and Evidence Accumulation Processes During Decision Formation. *The Journal of Neuroscience*, *38*(9), 2163–2176. https://doi.org/10.1523/JNEUROSCI.2340-17.2018

Luft, C. D. B., Zioga, I., Giannopoulos, A., Di Bona, G., Binetti, N., Civilini, A., Latora, V., & Mareschal, I. (2022). Social synchronization of brain activity increases during eye-contact. *Communications Biology*, *5*(1), 1–15. https://doi.org/10.1038/s42003-022-03352-6

Mahmoodi, A., Nili, H., Bang, D., Mehring, C., & Bahrami, B. (2022). Distinct neurocomputational mechanisms support informational and socially normative conformity. *PLOS Biology*, *20*(3), e3001565. https://doi.org/10.1371/journal.pbio.3001565

Masís, J., Chapman, T., Rhee, J. Y., Cox, D. D., & Saxe, A. M. (2023). Strategically managing learning during perceptual decision making. *eLife*, *12*, e64978. https://doi.org/10.7554/eLife.64978

Mastrogiuseppe, F., & Ostojic, S. (2018). Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron*, *99*(3), 609-623.e29. https://doi.org/10.1016/j.neuron.2018.07.003

Miranda, B., Malalasekera, W. M. N., Behrens, T. E., Dayan, P., & Kennerley, S. W. (2020). Combined model-free and model-sensitive reinforcement learning in non-human primates. *PLOS Computational Biology*, *16*(6), e1007944. https://doi.org/10.1371/journal.pcbi.1007944

Mishkin, M., & Ungerleider, L. G. (1983). *Object vision and spatial vision: two cortical pathways*.

Mojzisch, A., & Krug, K. (2008). Cells, circuits, and choices: Social influences on perceptual decision making. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 498–508. https://doi.org/10.3758/CABN.8.4.498

Moore, M., Katsumi, Y., Dolcos, S., & Dolcos, F. (2021). Electrophysiological Correlates of Social Decision-making: An EEG Investigation of a Modified Ultimatum Game. *Journal of Cognitive Neuroscience*, *34*(1), 54–78. https://doi.org/10.1162/jocn_a_01782

Morris, T. (2022). *An evaluation of the gap detection ability of mice using two-alternative forced choice tasks*. https://scholarsbank.uoregon.edu/xmlui/handle/1794/27384

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., & Malach, R. (2005). Coupling Between Neuronal Firing, Field Potentials, and fMRI in Human Auditory Cortex. *Science*, *309*(5736), 951–954. https://doi.org/10.1126/science.1110913

Najar, A., Bonnet, E., Bahrami, B., & Palminteri, S. (2020). The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLOS Biology*, *18*(12), e3001028. https://doi.org/10.1371/journal.pbio.3001028

Nam, C. S., Choo, S., Huang, J., & Park, J. (2020). Brain-to-Brain Neural Synchrony During Social Interactions: A Systematic Review on Hyperscanning Studies. *Applied Sciences*, *10*(19), Article 19. https://doi.org/10.3390/app10196669

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126–132. https://doi.org/10.1038/s41562-017-0273-4

O’Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735. https://doi.org/10.1038/nn.3248

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K., & Kelly, S. P. (2018). Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. *Trends in Neurosciences*, *41*(11), 838–852. https://doi.org/10.1016/j.tins.2018.06.005

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

Okazawa, G., & Kiani, R. (2023). Neural Mechanisms That Make Perceptual Decisions Flexible. *Annual Review of Physiology*, *85*(Volume 85, 2023), 191–215. https://doi.org/10.1146/annurev-physiol-031722-024731

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1. https://doi.org/10.1167/5.5.1

Pisauro, M. A., Fouragnan, E., Retzler, C., & Philiastides, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI. *Nature Communications*, *8*(1), 15808. https://doi.org/10.1038/ncomms15808

Platt, M. L., & Glimcher, P. W. (1997). Responses of intraparietal neurons to saccadic targets and visual distractors. *Journal of Neurophysiology*, *78*(3), 1574–1589. https://doi.org/10.1152/jn.1997.78.3.1574

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Purcell, B. A., & Kiani, R. (2016a). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*, *113*(31), E4531–E4540. https://doi.org/10.1073/pnas.1524685113

Purcell, B. A., & Kiani, R. (2016b). Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. *Neuron*, *89*(3), 658–671. https://doi.org/10.1016/j.neuron.2015.12.027

Ratcliff, R. (1978). *A Theory of Memory Retrieval*. 50.

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences*, *106*(16), 6539–6544. https://doi.org/10.1073/pnas.0812589106

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83. https://doi.org/10.1037/a0014086

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging, practice, and perceptual tasks: A diffusion model analysis. *Psychology and Aging*, *21*(2), 353–371. https://doi.org/10.1037/0882-7974.21.2.353

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266. https://doi.org/10.1038/nature08275

Rilling, J. K., & Sanfey, A. G. (2011). The Neuroscience of Social Decision-Making. *Annual Review of Psychology*, *62*(Volume 62, 2011), 23–48. https://doi.org/10.1146/annurev.psych.121208.131647

Roitman, J. D., & Shadlen, M. N. (2002). Response of Neurons in the Lateral Intraparietal Area during a Combined Visual Discrimination Reaction Time Task. *The Journal of Neuroscience*, *22*(21), 9475–9489. https://doi.org/10.1523/JNEUROSCI.22-21-09475.2002

Rojas, J.-C., Marín-Morales, J., Ausín Azofra, J. M., & Contero, M. (2020). Recognizing Decision-Making Using Eye Movement: A Case Study With Children. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.570470

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science (New York, N.Y.)*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Shadlen, M. N., & Kiani, R. (2013). Decision Making as a Window on Cognition. *Neuron*, *80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony Unbound: A Critical Evaluation of the Temporal Binding Hypothesis. *Neuron*, *24*(1), 67–77. https://doi.org/10.1016/S0896-6273(00)80822-3

Shadlen, M. N., & Newsome, W. T. (2001). Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936. https://doi.org/10.1152/jn.2001.86.4.1916

Shen, J., Liu, N., Li, D., & Zhang, B. (2022). Behavioral Analysis of EEG Signals in Loss-Gain Decision-Making Experiments. *Behavioural Neurology*, *2022*, 3070608. https://doi.org/10.1155/2022/3070608

Shen, Y., & Zhou, B. (2021). *Closed-Form Factorization of Latent Semantics in GANs* (arXiv:2007.06600). arXiv. https://doi.org/10.48550/arXiv.2007.06600

Shepherd, S. V., Deaner, R. O., & Platt, M. L. (2006). Social status gates social attention in monkeys. *Current Biology*, *16*(4), R119–R120. https://doi.org/10.1016/j.cub.2006.02.013

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: Empirical tests of theoretical predictions. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1865–1897. https://doi.org/10.1037/a0016926

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between integration and non-integration strategies in perceptual decision making. *eLife*, *9*, e55365. https://doi.org/10.7554/eLife.55365

Subramanian, A., Chitlangia, S., & Baths, V. (2022). Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, *145*, 271–287. https://doi.org/10.1016/j.neunet.2021.10.003

Summerfield, C., & Miller, K. (2023). Computational and systems neuroscience: The next 20 years. *PLOS Biology*, *21*(9), e3002306. https://doi.org/10.1371/journal.pbio.3002306

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

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. *Advances in Neural Information Processing Systems*, *12*. https://papers.nips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html

Tremblay, S., Sharika, K. M., & Platt, M. L. (2017). Social Decision-Making and the Brain: A Comparative Perspective. *Trends in Cognitive Sciences*, *21*(4), 265–276. https://doi.org/10.1016/j.tics.2017.01.007

Tump, A. N., Deffner, D., Pleskac, T. J., Romanczuk, P., & M. Kurvers, R. H. J. (2024). A Cognitive Computational Approach to Social and Collective Decision-Making. *Perspectives on Psychological Science*, *19*(2), 538–551. https://doi.org/10.1177/17456916231186964

Tump, A. N., Pleskac, T. J., & Kurvers, R. H. J. M. (2020). Wise or mad crowds? The cognitive mechanisms underlying information cascades. *Science Advances*, *6*(29), eabb0266. https://doi.org/10.1126/sciadv.abb0266

Tump, A. N., Wolf, M., Romanczuk, P., & Kurvers, R. H. J. M. (2022). Avoiding costly mistakes in groups: The evolution of error management in collective decision making. *PLOS Computational Biology*, *18*(8), e1010442. https://doi.org/10.1371/journal.pcbi.1010442

Turner, W., Angdias, R., Feuerriegel, D., Chong, T. T.-J., Hester, R., & Bode, S. (2021). Perceptual decision confidence is sensitive to forgone physical effort expenditure. *Cognition*, *207*, 104525. https://doi.org/10.1016/j.cognition.2020.104525

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, *8*(1), 14637. https://doi.org/10.1038/ncomms14637

Vafaei Shooshtari, S., Esmaily Sadrabadi, J., Azizi, Z., & Ebrahimpour, R. (2019). Confidence Representation of Perceptual Decision by EEG and Eye Data in a Random Dot Motion Task. *Neuroscience*, *406*, 510–527. https://doi.org/10.1016/j.neuroscience.2019.03.031

Valsangiacomo, F. (2023). Pupil's Decision-making Ability in the Context of Sustainable Development: Construction of a Typology of Decision-making Processes. *Journal of Education for Sustainable Development*, *17*(1), 78–100. https://doi.org/10.1177/09734082231183344

van Kempen, J., Loughnane, G. M., Newman, D. P., Kelly, S. P., Thiele, A., O'Connell, R. G., & Bellgrove, M. A. (2019). Behavioural and neural signatures of perceptual decision-making are modulated by pupil-linked arousal. *eLife*, *8*, e42541. https://doi.org/10.7554/eLife.42541

Wang, X.-J. (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron*, *36*(5), 955–968. https://doi.org/10.1016/S0896-6273(02)01092-9

Wang, X.-J. (2008). Decision Making in Recurrent Neuronal Circuits. *Neuron*, *60*(2), 215–234. https://doi.org/10.1016/j.neuron.2008.09.034

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292. https://doi.org/10.1007/BF00992698

Wei, Z., & Wang, X.-J. (2015). Confidence estimation as a stochastic process in a neurodynamical system of decision making. *Journal of Neurophysiology*, *114*(1), 99–113. https://doi.org/10.1152/jn.00793.2014

Wimmer, K., Compte, A., Roxin, A., Peixoto, D., Renart, A., & De La Rocha, J. (2015). Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nature Communications*, *6*(1), 6177.

Wimmer, K., Ramon, M., Pasternak, T., & Compte, A. (2016). Transitions between Multiband Oscillatory Patterns Characterize Memory-Guided Perceptual Decisions in Prefrontal Circuits. *Journal of Neuroscience*, *36*(2), 489–505. https://doi.org/10.1523/JNEUROSCI.3678-15.2016

Wong, K.-F., & Wang, X.-J. (2006). A Recurrent Network Mechanism of Time Integration in Perceptual Decisions. *The Journal of Neuroscience*, *26*(4), 1314–1328. https://doi.org/10.1523/JNEUROSCI.3733-05.2006

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*. https://doi.org/10.3389/fnint.2012.00079

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). *The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision*. 31.