In search of uncertainty:
representations of uncertainty across various levels of perception and cognition

Ádám Ferdinánd Koblinger

Central European University
Department of Cognitive Science

In partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Cognitive Science

Supervisor: József Fiser

Secondary supervisor: Máté Lengyel

Budapest, Hungary
2024

Declaration of Authorship

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgement is made in the form of bibliographical reference.

The present work includes studies that appeared in the following publications:

1. The review in Chapter 1 was published in:

   Koblinger, Á., Fiser, J., and Lengyel, M. (2021). Representations of uncertainty: where art thou? Current Opinion in Behavioral Sciences, 38:150–162

   Author contributions: The review paper was written with equal contributions from all three authors.

2. An earlier version of the of the study in Chapter 2 with a preliminary set of data and with a different hypothesis was published in:

   Lengyel, M., Koblinger, Á., Popović, M., and Fiser, J. (2015). On the role of time in perceptual decision making. arXiv preprint arXiv:1502.03135

   The experimental paradigm and the basic measurements and controls in the preprint and the present work were the same therefore, their descriptions were also the same.

   Author contributions: M. Lengyel and J. Fiser were involved in the conceptualization. M. Lengyel formalized the models. All authors were involved in designing the study. Á. Koblinger recorded and anaylzed the experimental data. Á. Koblinger, M. Lengyel and J. Fiser were involved in interpreting the results. M. Lengyel and J. Fiser wrote the paper.

3. The data from the first 6 experiments in Chapter 3 was also part of József Arató's dissertation:

   Arató, J. (2018). Active learning as a link between environmental statistics and the development of internal representations. PhD thesis, Central European University

4. The data from Chapter 4 was also part of Theoklitos Amvrosiadis' dissertation:

   Amvrosiadis, T. (2023). Representation of perceptual uncertainty in mouse primary visual cortex. PhD thesis, The University of Edinburgh

Other parts of the dissertation contain research that was performed collaboratively, and will potentially be submitted for publication:

1. Chapter 3 will be submitted for publication with József Arató and József Fiser.

   Contributions: Á. Koblinger and J. Arató were both involved in data collection. J. Arató designed the first six experiments. Á. Koblinger modified the experimental design for the last two experiments, conceptualized the theoretical framework and performed the data analyzes. (Both the theoretical model and the data analysis appear for the first time in the current thesis.). Á. Koblinger and J. Fiser wrote the text.

2. The study in Chapter 4 was a collaborative effort involving Ádám Koblinger, Theoklitos Amvrosiadis, Nathalie Rochefort and Máté Lengyel.

   Contributions: All contributors were involved in the conceptualization and the design of the experiment. Á. Koblinger performed the behavioural analysis, implemented the neural network decoder, validated it on synthetic data, and assisted in neural data processing (including $\Delta F/F$ computation, data structuring and neuron matching across different recording sessions). T. Amvrosiadis conducted the experiments in the Rochefort lab at the University of Edinburgh, preprocessed the neural data (motion correction, decontamination, ROI detection and neuron matching), made improvements to the neural network decoder, and fitted it to mouse behaviour.

Ádám Ferdinánd Koblinger

Abstract

Uncertainty is an inevitable companion of life in our complex world. It is increasingly believed that to act effectively under such circumstances, humans and animals rely on approximately probabilistic computation and an internal model that becomes attuned to environmental regularities and can efficiently complement the incomplete sensory observations with insights distilled from past experiences. However, for efficient behavior, the internal model must maintain veridical information about its own uncertainty. While there is increasing evidence that humans and animals are aware of the uncertainty associated with their decisions, the extent of their uncertainty representations is unclear. Specifically, it is unknown whether they represent uncertainty in a task-dependent manner, solely at the level of decisions, or in a fully Bayesian manner, representing uncertainty just about every aspects of their internal model. In this dissertation, I address this gap by first arguing for the preeminence of fully Bayesian models in terms of their generalization abilities over the alternative candidates. Then, I present three studies that clarify different characteristics of human and animal internal models relevant to assessing the degree to which uncertainty is encoded in biological internal models.

Chapter 1 provides a normative justification for the use of fully Bayesian representations, demonstrating their superior data and memory efficiency compared to task-dependent representations. I critically review the literature, highlighting the lack of conclusive evidence on the extent of the brain's uncertainty representation, and propose experimental paradigms to address this gap, setting the stage for the subsequent chapters.

In Chapter 2, I present experimental evidence based on a novel behavioural paradigm that human internal models meet one of the fundamental prerequisites for fully Bayesian models: they simultaneously represent uncertainties about multiple internal variables. Moreover, I found that explicit uncertainty reports about a variable (e. g., orientation) are based on gradually emerging probabilistic perceptual representation of that variable rather than on other, related variables (e.g., contrast) that can serve as proxies for the sensory reliability.

In Chapter 3, I use another novel behavioural paradigm and ideal observer analysis to demonstrate that humans automatically employ complex internal models with multiple variables and parameters, even in simple decision-situations where such complexity may not seem necessary. These complex internal models are then updated in a Bayesian manner when changes in task statistics are encountered, providing further support for the fully Bayesian brain hypothesis.

In Chapter 4, I propose a novel hybrid experimental paradigm combining neural and behavioral approaches to identify the neural traces of the potential perceptual uncertainty representation that are distinct from the representation of uncertainty directly related to the decision. Next, I demonstrate the practical application of this method to behavioral data from mice performing an orientation estimation task together with neural activity from their primary visual cortex (V1) recorded with calcium imaging. This data provides preliminary evidence that mouse V1 represents perceptual, rather than decision uncertainties, which is encoded in the temporal activity patterns within a trial, rather than in the spatial activity patterns across the population.

Together, these results shed a more focused light on the hitherto unexplored extent to which uncertainty is encoded in the brain and provide a consequential support to the proposal that complex brains use complex, approximately probabilistic processes and a broad representation of uncertainties to cope with challenges of their complex and uncertain environment.

Acknowledgements

The path to a doctorate is never walked alone, and I'm certainly no exception. I'm grateful to everyone who supported me along they way, and I would like to highlight those who helped me the most.

First, I want to thank my supervisors, József Fiser and Máté Lengyel.

József always received my ideas with a critical yet open mind. He helped me a lot in organising my thoughts, implementing my ideas and putting my results into a consumable form. He also made sure that I never lost sight of the broader perspective in my work.

From Máté, I learned just how important attention to detail is when it comes to the quality of scientific work. Under his guidance, I learnt how to transform abstract concepts into testable mathematical models and how to rigorously validate those models. I am also thankful to him for giving me the opportunity to visit his lab at the University of Cambridge.

I am very grateful to both of my supervisors for their immense support that allowed me to attend prestigious summer schools and conferences.

I would also like to thank all the current and former members of the CEU Vision Lab community for their great company and for making the lab such a pleasant place to work. I would especially like to thank those who assisted me with the experiments and data collection: Márton Nagy, József Arató, Dávid Magas, and Benjámin Márkus.

I'm also grateful to the ever supportive community of the Cognitive Science Department. I owe special thanks to Réka Finta and Ildikó Varga for always ensuring the administrative side of things ran smoothly.

I would like to thank Nathalie Rochefort and Theoklitos Amvrosiadis for giving me a glimpse into the world of neuroscience.

Last but certainly not least, I would like to thank my family for their unwavering support. Without them I wouldn't be able to do this!

Contents

List of Figures

Chapter 1

Normative justification for a broad representation of uncertainty

Perception is often described as probabilistic inference requiring an internal representation of uncertainty. However, it is unknown whether uncertainty is represented in a task-dependent manner, solely at the level of decisions, or in a fully Bayesian manner, across the entire perceptual pathway. To address this question, we[1] first codify and evaluate the possible strategies the brain might use to represent uncertainty, and argue for the advantages of fully Bayesian representations. In such representations, uncertainty information is explicitly represented at all stages of processing, including early perceptual areas, allowing for flexible and efficient computation in a wide variety of situations. Next, we critically review neural and behavioral evidence about the representation of uncertainty in the brain agreeing with fully Bayesian representations. We argue that sufficient behavioral evidence for fully Bayesian representations is lacking and suggest experimental approaches for demonstrating the existence of multivariate posterior distributions along the perceptual pathway.

[1]This chapter has been published, and therefore, I use plural pronouns.

1.1 Introduction

In order to efficiently interact with our environment, we need to evaluate the potential consequences of our decisions. Critically, these decisions are usually based on information that is limited and ambiguous in several ways. For example, when we choose where to look for our bicycle at the parking station at the end of the day, we are coping with occlusions by other similar bikes, low visibility, and incomplete memories about where our bike was left when we came to work. Therefore, in general, we cannot know with certainty the values of those variables that are relevant to our decisions, and optimal decision making requires that we take this uncertainty into account.

Indeed, a large body of evidence indicates that humans, and other animals, make decisions by representing their uncertainty (Knill and Pouget, 2004; Griffiths et al., 2010; Bach and Dolan, 2012; Ma and Jazayeri, 2014). However, it remains unclear how general a computational strategy it is for the brain to represent and compute with uncertainty in complex environments characterised by many interacting variables (i.e. arguably just about any real-life scenario). In particular, it is largely unknown whether the brain represents uncertainty "opportunistically", only about the variables that are relevant for the decision at hand, or "constitutively", about many variables simultaneously, including ones that are not directly relevant for the current decision making situation. In the context of our example above, the opportunistic strategy would only represent uncertainty about the single high-level decision variable (the location of the bike relative to where we stand). In contrast, the constitutive strategy would represent uncertainties about several perceptual and other variables that feed into the decision process, such as the reliability of perceived color in the darkness, the ambiguity of shape information given partial occlusions within the crowd of bikes, and the precision of our memories about the layout of the parking station.

The distinction between opportunistic and constitutive representations of uncertainty has not been explicitly articulated before and is therefore the main focus of this review. We begin by building on the classical framework of Bayesian decision theory (Jaynes, 1996) to formalise the distinction between these representational strategies as task-dependent and fully

Bayesian recognition models, evaluate their respective theoretical advantages and disadvantages, as well as the empirical evidence that may be interpreted as supporting them. We also explore different hybrid solutions between these two extreme representational strategies, and discuss the consequences of each of them for the neural representation of uncertainty. We then argue that current evidence is insufficient to clearly distinguish between these different strategies, and propose experimental approaches that are appropriate for identifying their behavioral signatures.

1.2 Representing uncertainty: the basics

In order to understand the distinction between opportunistic and constitutive representations of uncertainty, we first take a step back, and briefly discuss why representing uncertainty at all is relevant for decision making in the first place. The mathematical framework of Bayesian decision theory provides an answer to this question (Box 1), (Jaynes, 1996). In this framework, the problem of optimal decision making – and the role of uncertainty in it –, can be formalised by defining a relationship between a handful of key variables in the observer's internal model of the decision making situation (Fig. 1.1, 1st column): the decision variable (in our running example: expressing the location of the bicycle relative to where we stand now), the utility (time spent searching for the bike), the action (turning right or left, or going straight), and the observation (current visual input as well as the memory traces stored from the time when we left the bike at the station). The specific decision making task is ultimately defined by the utility function (also called the reward or loss function) that determines how the utility obtained depends on the decision variable and the action taken. The internal model that defines the relationships between all these ingredients is also called a *generative model* because it describes the observer's beliefs about how the world generates observations and utilities (the latter contingent upon their own actions).

Critically, as the decision variable itself is not observed directly (it is a *latent variable*), its value can only be inferred (based on the observations), and this inference usually carries uncertainty. This uncertainty is formalised by the Bayesian posterior as the probability that the

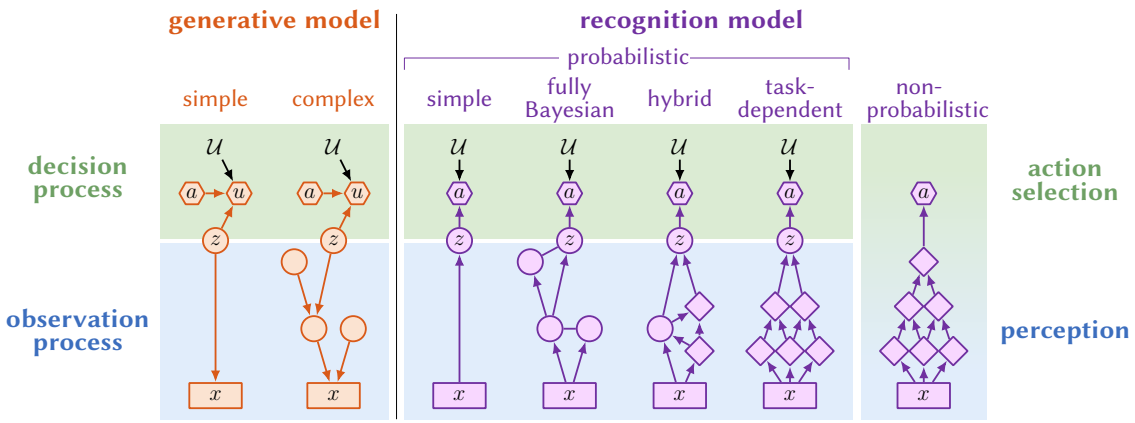


Figure 1.1. Taxonomy of generative and recognition models in decision making. Blue and green backgrounds correspond to the two components of Bayesian decision theory: the observation and decision processes of the generative model (orange), and the perception and action selection modules of the recognition model (purple), respectively. Note that 'perception' here is broadly construed to include all cognitive processes (e.g. sensory perception or memory) that have access to information ('observations') that is relevant for the decision making task. Rectangles indicate observed variables (x), circles indicate latent variables (including the decision variable, z) which are part of the generative model and are probabilistically computed in the recognition model, diamonds indicate non-probabilistically computed internal variables of the recognition models, hexagons indicate variables specific to the decision process: the action (a) and the utility obtained (u). The utility function ($\mathcal{U}$) is shown without a bounding box to indicate that it is a parameter that is constant across trials or time steps, while other quantities change over time or trials. Left: generative models. All generative models describe how z is related to x, and how it (exclusively) determines the u obtained for a given a (as parameterized by $\mathcal{U}$). **Simple generative models** only have a single latent variable, z. **Complex generative models** have multiple latent variables beside z. Right: recognition models. All recognition models compute action a from observations x. **Probabilistic recognition models** compute a posterior over z given x (Eq. 1.1), which they then combine with $\mathcal{U}$ to compute a (Eqs. 1.2a and 1.2b). The **non-probabilistic recognition model** computes a directly from x, without computing a posterior over z, and without explicitly representing $\mathcal{U}$. Probabilistic recognition models are further subdivided based on what other variables are computed probabilistically while computing the posterior over z: **simple models** do not have any other variables, complex models do, with **fully Bayesian**, **hybrid**, and **task-dependent models** probabilistically computing all variables, a subset of them, or none of them, respectively.

decision variable might take any particular value given the information in the observations (Eq. 1.1). Thus, computing the optimal action under a given internal model requires 1. computing the Bayesian posterior over the decision variable, and then 2. computing the expected utility of each available action under this posterior (Eq. 1.2a). The optimal action is simply the one that yields the highest expected utility (Eq. 1.2b). Straightforward this description may sound, the corresponding computation is anything but: in fact, both steps have prohibitive

computational costs in the general case. Borrowing (and slightly extending) the standard terminology from machine learning (Dayan and Hinton, 1996), we call the algorithmic architectures that implement (at least approximately) optimal decision making *recognition models*. Although, in order to rigorously formalise the concept of uncertainty, we derived the definition of a recognition model based on a generative model, it is only the recognition model that needs to be actually implemented in the brain for decision making (though see Hinton et al., 1995, for potential uses and signatures of generative models also being implemented). A key question is then what recognition models the brain employs and specifically which, if any, of their internal variables are computed probabilistically, i.e. such that uncertainty about their values is represented.

The conceptually most straightforward recognition model is broken down into two discrete steps that directly correspond to the observation and decision process components of the generative model: computing the posterior, and choosing the action (Fig. 1.1, 3rd column). Indeed, these two steps have been suggested to correspond more broadly to the computations underlying perceptual (and memory; Hemmer and Steyvers, 2009) processes, and action selection, respectively (Yuille and Bülthoff, 1996; Körding and Wolpert, 2006). In this case, the representation of uncertainty is a given: the perceptual module computes a (potentially approximate) posterior probability distribution over the decision variable. Thus, such recognition models are *probabilistic*. However, note that, ultimately, the optimal action is just a function of the observations (Eq. 1.2b). This function could be implemented directly without having to ever compute explicitly the posterior over the decision variable. In this case, even if the computation of the optimal action is broken down to several internal steps and quantities, none of these need to correspond to the decision variable as such, or its Bayesian posterior. Thus, the resulting recognition model is *non-probabilistic* (Fig. 1.1, 7th column).

If optimal actions can be computed non-probabilistically, why bother with the costly computation of a posterior over the decision variable at all? Indeed, some of the most successful deep learning architectures of today do not use probabilistic representations (LeCun et al., 2015). Yet, there is widespread evidence for uncertainty about decision variables being represented in the brain (Fiser et al., 2010; Vilares et al., 2012; Walker et al., 2020). This suggests

Box 1 - Bayesian decision theory

According to the agent's internal model (Fig. 1.1, 1st column), there is a decision variable, z, of which the value determines how the utility, u, depends on the different actions they might choose, a (decision process). The exact form of this dependence is determined by the utility function $u = \mathcal{U}(a, z)$. However, z is not directly accessible to the agent, hence it is a *latent variable*. Instead, the agent makes *observations*, x, that refer to all the information that are available to them at the time of making the decision (observation process). Importantly, the agent must also have some knowledge about the (potentially noisy and ambiguous) relationship between z and x (Fig. 1.1, 1st column, arrow connecting z and x).

Given these ingredients, Bayesian decision theory proceeds in two steps to compute the optimal choice of a. First, the value of z needs to be inferred based on x. As z cannot be determined with certainty, Bayes' rule is used to compute a posterior distribution that expresses the probability with which z might take any particular value given the information in x:

$$\mathcal{P}(z|x) = \frac{\mathcal{P}(x|z)\ \mathcal{P}(z)}{\int \mathcal{P}(x|z')\ \mathcal{P}(z')\ \mathrm{d}z'} \tag{1.1}$$

where $\mathcal{P}(x|z)$ (the *likelihood*) expresses the probability of observing x when z takes on a particular value, and $\mathcal{P}(z)$ (the *prior*) expresses the overall frequency with which z is believed to take on any particular value. Second, the utility expected from choosing each action is computed by considering all possible settings of z, computing the corresponding utilities and then taking the average of these utilities weighted by the corresponding posterior probabilities:

$$\bar{\mathcal{U}}(a, x) = \int \mathcal{U}(a, z)\ \mathcal{P}(z|x)\ \mathrm{d}z \tag{1.2a}$$

and the action with the maximal expected utility is chosen

$$a(x) = \underset{a'}{\operatorname{argmax}}\ \bar{\mathcal{U}}(a', x) \tag{1.2b}$$

a normative reason for such probabilistic representations. Here we briefly review four such reasons, each of which implies that the additional computational cost of probabilistic representations can be offset by their increased data- and memory-efficiency (i.e. that they can perform well with less learning and require less memory):

Task-flexibility. The modular architecture of probabilistic recognition models endows them with great flexibility when a new task is encountered (Houlsby et al., 2013). In this case, the 'perception' module of the recognition model can be kept the same, and only 'action selection' needs to be adapted to the new task by incorporating the corresponding new utility function into it. For example, if one day a car blocks your way to the right at the parking station, then the action of turning right suddenly yields very low utility no matter where your bike is parked. Nevertheless, with a probabilistic recognition model, we can keep using the same perceptual module to compute the relative location of the bike as you have always done. In contrast, an altogether new non-probabilistic recognition model would need to be constructed (or learned) whenever the utility function changes, precisely because non-probabilistic recognition models allow no neat splitting between perception and action selection.

Information fusion. When information from multiple observations needs to be fused, probabilistic recognition models for each observation can be combined efficiently. This is because the posterior (or, more precisely, the likelihood; Box 1) over the decision variable they compute is the 'sufficient statistic' of the corresponding observation. Information fusion is not only relevant for classical cases of (multi)sensory cue combination, when the different observations correspond to different sensory modalities (Ernst and Banks, 2002; Alais and Burr, 2004) or cues (Moreno-Bote et al., 2011) (e.g. depth cues for judging the distance of our bike) or memory (Körding and Wolpert, 2004) (as when combining our current percept of a bike that looks like ours in the distance and our memory of where we left our bike), but also for accumulating evidence across successive observations over time (Brunton et al., 2013) (successive views of the parking station as we walk towards our bike). Optimal sensory cue combination can be achieved by combining the individual probabilistic recognition models appropriate for each sensory (or memory) observation (Ma et al., 2006), and optimal evidence accumulation

can be performed by applying the same probabilistic recognition model recursively (Beck et al., 2008) (also known as Bayesian filtering (Chen et al., 2003)). In contrast, to achieve the same performance with non-probabilistic recognition models, different models need to be constructed (or trained) for each combination of sensory cues that may need to be combined in sensory cue combination. For evidence accumulation, they would need to have access to the entire relevant history of observations at once.

Active sensing. Information gathering is often an active rather than a passive process: we can control our sensors (e.g. move our eyes) so that we receive information that is most useful for solving a task (finding our bike). The Bayesian posterior over the decision variable, as computed by probabilistic recognition models, can be directly used to form an objective for active sensing (Yang et al., 2016a). For example, active sensing can be achieved by actions whose predicted sensory consequences would leave the least amount of uncertainty about the decision variable (quantified by the entropy of the posterior), or by actions that would result in a posterior under which the expected utility (using the optimal action) is the highest (Yang et al., 2016b). Importantly, for computing the possible sensory consequences of these actions in both cases, the same generative model of decision making can be extended to include the effects of active-sensing-related actions on observations (Fig. 1.1). Such a recycling of resources makes active sensing data-efficient. In contrast, information gathering in non-probabilistic recognition models implies an extension of the space of possible actions (e.g. in the case of our bike search, our actions become tuples consisting of which way to look *and* then turn) requiring to learn an entirely new recognition model for each active sensing action. Even worse, the available supervision for information gathering actions in this case is very sparse and indirect - we make many eye movements while searching for our bike, these eye movements by themselves are not rewarded, and each of them contributes only very indirectly to our ultimate success or failure to find the bike. In sum, achieving efficient active sensing by this method is just altogether impractical.

Learning. In most cases, the recognition model can be improved with experience, such as learning to distinguish our bike from similar bikes often parked at the same station. However, this can be a non-trivial task without receiving ground-truth information. For example, if we

accept occasionally a colleague's offer of a lift home while hesitating at the station about the direction to find our bike, we learn neither the true value of the decision variable (where the bike was), nor the correctness of the action we would have chosen (which way to go). Thus, there is no obvious way to learn from those occasions – as long as one uses a non-probabilistic recognition model. In contrast, by using a probabilistic recognition model, one can still take advantage of such 'unlabelled' data for improving the recognition model by averaging over the possible adjustments that would be appropriate for different values of the decision variable. Importantly, to weight appropriately the different terms in this average, the uncertainty about the decision variable (as quantified by its likelihood or the posterior, Box 1) needs to be taken into account. In other words, probabilistic recognition models are particularly well-suited for un- or semi-supervised forms of learning.

There are also ways to interpolate between probabilistic and non-probabilistic models in an attempt to combine the best of both worlds: the data- and memory-efficiency of probabilistic models on one hand, and the computational efficiency and direct performance guarantees on the relevant decision tasks of non-probabilistic models on the other (Box 2).

1.3 Representing uncertainty in a complex world

Although the simplest two-step generative model described above succinctly summarizes the essential elements of Bayesian decision making (Fig. 1.1, 1st column), it belies the real complexity of the natural environment. In most cases, our observations are generated by a complex mesh of interactions between a large number of latent variables, of which the decision variable is but one (Fig. 1.1, 2nd column). For example, our perceived view of the parking station is jointly determined by a number of features characterizing each individual bike and car (their make, model, color, accessories, locations, etc.) as well as the lighting conditions and our viewing angle among other aspects.

In complex environments, of which the internal generative model includes multiple latent variables, probabilistic recognition models can be further subdivided based on how many

Box 2 - Between probabilistic and non-probabilistic recognition models

Amortized inference

Definition: Computational resources are shared across subsequent uses of the recognition model. In the context of Bayesian inference, this can be accomplished by treating latent variables as if they were observed, substituting a previously inferred value for them.

Machine learning application: Once a non-probabilistic deep neural network has been extensively trained on some tasks, it can be reused almost entirely (bar the final one or few layers) in novel but related tasks, such that it can achieve state-of-the-art performance already after only minimal training on those tasks (Mathis et al., 2018).

Behavioural evidence: "Certainty-equivalent" form of amortisation has been shown to account for suboptimalities in high-level human probabilistic reasoning (Gershman and Goodman, 2014) as well as lower level "conditioned perception" effects, whereby an earlier perceptual discrimination decision biases later perceptual estimates of the same variable (Stocker and Simoncelli, 2008).

Loss-calibrated inference

Definition: Rather than using general-purpose algorithms to speed up probabilistic inference, knowledge about the utility (or loss) function specific to the current task is used for setting up the particular approximate inference algorithm to be used.

Machine learning application: Loss-calibration increases the expected utility of decisions during variational inference (Lacoste-Julien et al., 2011).

Behavioural evidence: A particular form of loss-calibration, called "utility-weighted inference", has been shown to account for a number of widely observed, seemingly irrational cognitive biases in human decision making, including the over-representation of extreme event probabilities (Lieder et al., 2018).

of the latent variables they represent probabilistically. As mentioned, the conceptually most straightforward, but computationally most ambitious, recognition model is based on a direct, *fully Bayesian* inversion of the generative model (Fig. 1.1, 4th column). In fully Bayesian recognition models, the (joint) posterior over all latent variables of the generative model implies a constitutive representation of uncertainty because uncertainty is represented for variables irrespective of whether they are directly relevant for the current decision making task (i.e. whether they are the designated decision variable). At the other – but still probabilistic – extreme, the recognition model only computes a posterior over the decision variable, but not over any of the other variables of the generative model (Fig. 1.1, 6th column). Of course, there may still be multiple internal variables storing interim results of the computations of recognition, but those need not correspond to the latent variables of the generative model, and even if they do, no posterior needs to be computed over them at all (e.g. even when representing the color of each bike in a visual scene, only the single best estimate of each color is represented, rather than its full posterior.) Thus, in this case, the recognition model remains *task-dependent* as it represents uncertainty opportunistically only for the decision variable relevant for the current task. Finally, there also exists a continuum of intermediate, *hybrid* models between these two extremes: these models compute the posterior over, and thus represent uncertainty about only a subset of latent variables but not all of them (Fig. 1.1, 5th column).

Note that the benefits of probabilistic recognition models enumerated in the previous section already apply to even the least ambitious of these models, the task-dependent recognition model. Meanwhile, the computational complexity of inference in a fully Bayesian recognition model grows exponentially with the number of latent variables in the general case. Thus, extending our previous discussion, we can ask why bother with the costly computation of a joint posterior over all (or a large number of) the latent variables? Indeed, the training of deep neural networks, for example for image classification, usually requires uncertainty to be correctly represented only in the decision variable (image class label) – if at all –, and can still lead to human- or even super human-level performance (Krizhevsky et al., 2012). However, there are indications that the brain may be closer to the fully Bayesian model and represents uncertainty about sensory as well other variables, and not just the ones related directly to the

decision (Bach and Dolan, 2012). The normative reasons for following this strategy are based on the same general trade-off between computational costs and data- and memory-efficiency that we argued underlay the benefits of probabilistic recognition models:

Task-flexibility. Just as different tasks may differ in their utility function over a particular latent variable, they may also differ in the identity of the latent variables (or subsets of latent variables) upon which the utility function depends. For example, it is only in the context of the specific task of locating our bike that out of all the latent variables characterising the parking station, it is the location of our bike that happens to play the special role of the decision variable. Once we find our bike and cycle through the station, suddenly other variables (the location of the exit, the predicted trajectory of an approaching car, etc) would subsume the status of the decision variable for choosing our actions (which way to turn the handle bar). In such cases, when using task-dependent recognition models, separate models need to be constructed for each of these tasks. A fully Bayesian recognition model is much more economical in that perception can proceed unchanged, as it constitutively computes the posterior distribution over all the latent variables, and only action selection needs to be modified to reflect the new utility function.

Information fusion. As we argued above, efficient information fusion across different observations (as in sensory cue combination or evidence accumulation) requires probabilistic representations. Importantly, this efficiency of probabilistic recognition models is only guaranteed as long as they represent uncertainty about a sufficiently large set of variables. This is because the computations underlying cue combination and evidence accumulation can only be performed efficiently (by a simple combination of individual recognition models for cue combination, and by simple recursive operations, one observation at a time, for evidence accumulation) if different observations are statistically independent given the latent variables about which uncertainty is represented (Clark and Yuille, 2013). However, in complex environments, observations are rarely dependent only on the single latent variable that happens to determine the utility of our actions (i.e. the decision variable). For example, two subsequent views of the parking station (the observations) across which we need to fuse information to better infer the location of the bike (the decision variable), will not be statistically indepen-

dent given the location of our bike, because there are a number of other latent factors which influence both views (the location of other bikes, lighting conditions, etc.). In turn, conditioning on all these latent variables can make the observations independent. (Sidenote: this is closely related to the problem of what constitutes an appropriate representation of 'state' in sequential decision making tasks, which are an extension of the non-sequential decision making tasks we are considering here, and which can be formalised as Markov decision processes in the realm of reinforcement learning; Sutton and Barto, 2018.) Thus, in these cases, only recognition models that are sufficiently close to being fully Bayesian will be able to fuse information efficiently. In contrast, task-dependent recognition models, in which only the decision variable is represented probabilistically, inherit the same problems that we argued non-probabilistic recognition models have: information fusion requires a separate recognition model for each combination of cues, or access to a history of observations.

The advantage of inferring multiple latent variables jointly is particularly well exposed in tasks that require "explaining away": a special form of credit assignment in updating beliefs when an additional observation leads to drastic changes in the posterior. Such updating occurs when the additional observation reveals that previous observations that have been attributed to a particular cause should in fact be credited to an entirely different cause. A classical example for this is when an object that initially looks convex based on its shading suddenly looks concave when we receive evidence that it is actually lit from below not above (Adams et al., 2004). This can be implemented naturally in a recognition model that jointly infers the shape of the object and the direction of light.

Active sensing. Generally, active sensing is a sequential process: it takes multiple adjustments of our sensors to collect adequate information before making a decision. For example, when looking for our bike at the parking station, we move our gaze to several locations to reduce our uncertainty about the location of our bike before we decide which way we turn. Merging information across consecutive observations is precisely the problem of information fusion we discussed above. Therefore, the very same arguments explain why fully Bayesian representations are useful for active sensing in complex environments.

Learning. Formally, learning is 'just' another type of information fusion, albeit on a slower time scale than what we considered so far. Instead of fusing information across multiple observations within a single trial to update the decision variable or other latent variables (as in evidence accumulation), learning requires fusing information across multiple trials to update the recognition model itself. Therefore, the same arguments that justify representing uncertainty about latent variables in a recognition model also apply to representing uncertainty about the recognition model itself (its parameters, structure, or form; Kemp and Tenenbaum, 2008): efficient online learning requires probabilistic representations. As in all these cases uncertainty needs to be represented about quantities (parameters, etc) in the recognition model other than the decision variable, such recognition models are not task-dependent any more by our definition, and are closer to the fully Bayesian end of the spectrum. For example, in neural networks, a probabilistic representations of synaptic weights (the parameters of the recognition model) has been shown to be advantageous as it allows optimal adaptation of learning rates on novel tasks (Aitchison and Latham, 2014) and helps avoiding catastrophic forgetting across multiple tasks (Huszár, 2018).

Moreover, just as representing uncertainty about the decision variable (and other variables, as we saw above) can allow efficient information gathering in active sensing, representing uncertainty about the recognition model can allow efficient information gathering for active learning, i.e. to choose inputs that are expected to improve the future performance of the recognition model most (Yang et al., 2016b). On the first occasion when we search for our bike at the parking station, we might choose a direction that leads to a longer expected searching time than the optimal choice, and use the extra time to familiarise ourselves with the station so that to improve our recognition model and thus our future search performance. Representing uncertainty about the recognition model can help us focus our exploration of the station where we know least about it. Eventually, once we have little uncertainty left about it, we can decide to stop exploring altogether.

recognition model	implementation	behavioural data	neural data
task-dependent	DDM	psychometric & chronometric curves of perceptual decisions (Gold and Shadlen, 2007; Kiani and Shadlen, 2009)	decision-related ramping activity of LIP single cells (Kiani and Shadlen, 2009; Shadlen and Newsome, 2001; Gold and Shadlen, 2007)
	PPC	psychometric & chronometric curves of perceptual decisions (Beck et al., 2008); cue combination (qualitatively) (Ma et al., 2006)	Poisson-like variability of cortical neurons (Ma et al., 2006); single cell activity in LIP (Beck et al., 2008)
	DNN	object recognition and categorisation performance (LeCun et al., 2015; Yamins and DiCarlo, 2016)	feature selectivity along the hierarchy of visual cortex (Kriegeskorte, 2015)
hybrid	CRP	human categorization (Griffiths et al., 2007)	–
fully Bayesian	belief propagation	bistable perception (Leptourgos et al., 2020), hallucinations (Jardri et al., 2017)	tight balance between excitation and inhibition (Deneve, 2005)
	extended PPC	–	anatomy and physiology of the olfactory bulb (Grabska-Barwińska et al., 2017)
	DDC	–	dopaminergic or hippocampal activity (Vértes and Sahani, 2019)
	sampling	cue combination (Moreno-Bote et al., 2011); multistable perception (Moreno-Bote et al., 2011; Gershman et al., 2012)	various static (Berkes et al., 2011; Orbán et al., 2016; Haefner et al., 2016) and dynamic (Echeveste et al., 2020) activity patterns of the early visual cortex

Table 1.1. Compatibility of specific implementations of probabilistic recognition models with behavioural and neural data. Details of the specific models are discussed in the main text. As we also note there, implementations appropriate for fully Bayesian recognition models could also implement task-dependent recognition models (but not vice versa). Thus, data listed here as supporting such implementations (e.g. multistable perception for sampling) does not necessarily provide support for fully Bayesian recognition models *per se*. Experiments providing support for fully Bayesian recognition models are discussed later (see also Table 1.2).

1.4 Implementing probabilistic recognition models in the brain

Whether probabilistic recognition models in perception are task-dependent, fully Bayesian, or hybrid, has important implications for how they might be implemented in the brain. For example, at the broadest level, fully-Bayesian recognition models must maintain a globally coherent representation of their latent variables' joint posterior. This requires that information about higher level cognitive variables should have effects on inferences about lower-level variables, i.e. potentially strong top-down influences on sensory cortical areas (Lee and Mumford, 2003). There is a large swathe of experimental data on such top-down interactions (Gilbert and Li, 2013). More specifically, recent studies provided evidence for trial-specific priors, cued by auditory stimuli, affecting both overt decisions and early visual cortical responses in a visual perceptual decision making task (Kok et al., 2013; Aitken et al., 2020). While task-dependent models may not be formally incompatible with such top-down influences, they also do not make any specific prediction about them.

In the following, for a finer level of distinction, we review previous specific suggestions for how the brain might represent uncertainty, and group these representations by the kind of recognition models they may be able to implement. In Table 1.1, we also provide pointers to some of the key empirical data that have been suggested to support them. We note, however, that most of these data only provide circumstantial evidence for the corresponding representations as yet. Thus, more work will be necessary that directly contrasts the predictions of different representations and compares them to experimental data (Echeveste et al., 2020).

There have been three influential proposals for how task-dependent recognition models might be implemented in the brain (Fig. 1.1, 6th column): the drift diffusion model (DDM; Gold and Shadlen, 2007), probabilistic population codes (PPCs; Ma et al., 2006), and deep neural networks (DNNs; Kriegeskorte, 2015; Yamins and DiCarlo, 2016). The DDM is a psychological process-level model of decision making, of which the behavioral signatures and neural underpinning have been extensively investigated (Table 1.1). According to the DDM, decisions are based on gradually accumulating noisy evidence obtained from sensory inputs (Kiani and

Shadlen, 2009) or memory traces (Shadlen and Shohamy, 2016). Optimal decision making requires both the accumulated evidence and the time elapsed since the beginning of accumulation in such tasks (Kiani and Shadlen, 2009), such that these two quantities together form the (sufficient statistic of the) 'decision variable', z, in our formulation (Fig. 1.1). Importantly, the evidence accumulated by the DDM is always about a specific decision variable that is relevant for the current task (e.g. saccade left or right). Therefore, the DDM is a *bona fide* task-dependent probabilistic recognition model.

According to probabilistic population codes (PPCs), neural populations encode probabilistic information in a format that is both easy to read out by downstream areas and allows simple, biologically plausible neural operations (e.g. linear summation of inputs) to implement probabilistically optimal processing of such information (Ma et al., 2006; Beck et al., 2008). As some PPCs were specifically designed to capture hierarchical probabilistic computations for example in a cue combination task (Ma et al., 2006), they might superficially appear to be fully Bayesian. Indeed, the neural architecture of such PPCs typically consists of several (feed-forward connected) layers, each encoding probabilistic information. Nevertheless, in our classification, they implement task-dependent recognition models because, at least in their originally proposed form (Ma et al., 2006), all layers encode probabilistic information about the same single decision variable (its likelihood based on different subsets of observed variables).

Recently, the most popular recognition models have been deep neural networks (DNNs). DNNs are typically trained on a given task in an end-to-end fashion by providing (a typically large set of) example input-output pairs from which these architectures can learn to generalise and generate the correct output to novel inputs. As such, DNNs are often used – and certainly construed – as fundamentally non-probabilistic recognition models that generate their output (the equivalent of a in our terminology) without performing any probabilistic computations on the way. Nevertheless, when trained with the appropriate (cross-entropy based) loss function, routinely used for example in image classification tasks, neural activities in the layer before a (the last hidden layer) come to essentially encode the posterior distribution of the decision variable z. As this probabilistic representation only emerges at this final

stage, these DNNs are task-dependent probabilistic recognition models. In addition, even the probabilistic representation of z can be poorly calibrated in these models, as has been demonstrated, for example, with so-called 'adversarial samples' (Goodfellow et al., 2014). We posit that this lack of proper calibration in the last hidden layer might be the consequence of the missing representations of uncertainty in the intermediate layers. Indeed, it has been suggested that endowing DNNs with more fully Bayesian probabilistic representations (e.g. by dropout) might improve their calibration of uncertainty on adversarial samples (Smith and Gal, 2018).

Regarding fully Bayesian recognition models (Fig. 1.1, 4th column), there have been two different classes of neural representations suggested. In principle, each of these neural representations is also able to support task-dependent recognition models, but the differences between them are best exposed when applied to fully Bayesian recognition models. In the first such representation, neural activities represent parameters or sufficient statistics of the posterior over *all* relevant latent variables. One recent example of this class is distributed distributional codes (DDCs; Vértes and Sahani, 2018). DDCs have been shown to have a number of computationally appealing properties, in particular when the recognition model corresponds to a complex hierarchical generative model (i.e. the very setting which motivates fully Bayesian recognition models in the first place, Fig. 1.1, 2nd column) and needs to be learned in an unsupervised way from experience.

Other examples of parametric neural representations do not attempt to represent the full joint posterior over all latent variables and instead use a cruder factorised approximation, in which only the marginal posteriors over individual latent variables are represented (Raju and Pitkow, 2016; Deneve, 2005; Grabska-Barwińska et al., 2017). Although such "marginally" fully Bayesian recognition models lose all information about posterior correlations between latent variables, this simplification also greatly reduces the complexity of neural dynamics required to implement them based on methods borrowed from machine learning, such as belief propagation (Raju and Pitkow, 2016; Deneve, 2005) or more general variational approximation schemes (Grabska-Barwińska et al., 2017). Some models in this class can be seen as extensions of PPCs to the (marginally) fully Bayesian case (Raju and Pitkow, 2016; Grabska-

Barwinska et al., 2013), inheriting some of their appealing properties, albeit with substantially more complex neural dynamics.

The other class of fully Bayesian recognition models uses a sampling-based representation of uncertainty (Hoyer and Hyvärinen, 2003; Fiser et al., 2010). In these models, neural responses represent the latent variables themselves such that the distribution of neural responses generated by the network's dynamics over some time period represents the joint posterior distribution of the recognition model. As such, these models again approximate the full joint posterior over all latent variables. There is substantial converging behavioral and neural evidence for sampling-based fully Bayesian recognition models at least in the early visual cortex (Table 1.1). Nevertheless, it remains to be seen whether such models apply to other perceptual domains and brain areas. This will be particularly interesting in settings in which inference over dynamically changing variables needs to be performed, as the time requirements of sampling may produce unique testable predictions in these domains (Lengyel et al., 2015).

There is one specific domain where hybrid recognition models (Fig. 1.1, 5th column) have – by necessity – been proposed in cognitive science. When the generative model includes an infinite number of latent variables, Bayesian inference necessarily needs to focus on a finite subset of these to tractably compute the posterior. This is the realm of non-parametric Bayesian inference in machine learning (Orbanz and Teh, 2010). Such non-parametric Bayesian models have been suggested to underlie a number of cognitive processes (Austerweil et al., 2015). For example, a non-parametric Bayesian inference algorithm (called the "Chinese restaurant process", CRP) can be used to infer which out of a potentially infinite number of categories does each item in a training set belongs to, and what the common characteristics of items in each category are (Aldous, 1985). Nevertheless, a systematic exploration of hybrid recognition models has not been pursued in cognitive neuroscience. This could be an interesting avenue for future research because these models have the potential to achieve a useful balance between performance and efficiency.

recognition model	normative advantages			
	task-flexibility	information fusion	active sensing	learning
probabilistic vs. non-probabilistic	Whiteley and Sahani, 2008; Qamar et al., 2013	Körding and Wolpert, 2004; Lengyel et al., 2015; Drugowitsch et al., 2016; Van den Berg et al., 2017	Paulun et al., 2015; Yang et al., 2016a	Zylberberg et al., 2018; Behrens et al., 2007
fully Bayesian vs. task-dependent	Denison et al., 2018; Lengyel et al., 2015	—	—	—

Table 1.2. Behavioral evidence for probabilistic and fully Bayesian recognition models. The studies providing behavioural support that the brain's recognition model is probabilistic (top row) or fully-Bayesian (bottom row) are organized according to the normative advantage of the probabilistic or fully Bayesian recognition model they verified behaviorally (columns). Below we summarize the evidence for probabilistic recognition models only, the existing evidence (or lack thereof) for fully-Bayesian recognition models is discussed in the main text. **Task-flexibility.** Humans (or monkeys) were found to exhibit a high degree of task-flexibility in tasks requiring generalization either to new utility functions (Whiteley and Sahani, 2008) or to new stimuli giving rise to posterior distributions that are qualitatively different from those previously experienced in the task (Qamar et al., 2013). **Information fusion.** Studies of information fusion showed that humans can near optimally combine two sources of information. Notably, in several classical cue-combination studies, participants had plenty of everyday experience with combining the information from the two tested sensory modalities to enhance the absolute accuracy of their decisions (Ernst and Banks, 2002; Alais and Burr, 2004). Thus, these tasks required only modest generalization, and as such could be solved with non-probabilistic recognition models. Other studies showed optimal information fusion even when participants needed to combine two sources of information information that they had never combined before (Körding and Wolpert, 2004), or a sequence of observations while the number and informativeness of observations was varied (Drugowitsch et al., 2016). In these cases, having a separate (non-probabilistic) recognition model for each task condition would be infeasible. Moreover, humans provided reliable uncertainty reports about their own performance when the difficulty of trials was modulated by multiple task parameters (e.g. the number and contrast of items in a scene) (Lengyel et al., 2015), suggesting that uncertainty reports were based on a unified representation of uncertainty (i.e. a single posterior distribution) rather than heuristic estimates corresponding to the different task parameters. Furthermore, in line with the unified uncertainty representation, the reported uncertainties also reliably predicted stimulus-independent fluctuations in performance over and above those controlled by the experimentally defined cues (Van den Berg et al., 2017; Koblinger et al., 2019, COSYNE, conference). **Active sensing.** Eye movements are almost never rewarded directly as such, but typically depend on participants' inferences about the currently viewed stimulus. Thus, they offer an ideal test bed for assessing behavioral signatures of probabilistic representations. For example, while performing the same visual search task in widely different lighting conditions, humans near-optimally adjusted their eye-movements to the changed lighting conditions, suggesting that they could efficiently generalize their eye-movement strategies across a wide range of posteriors (Paulun et al., 2015). In a visual pattern categorization task, eye movements were also shown to be optimized for information search (Yang et al., 2016a). Critically, this eye movement strategy correctly took into account the constantly evolving posterior distribution that a probabilistic recognition model computed over pattern category (the decision variable) based on (the growing set of) previous fixations in a trial (observations). **Learning.** As non-probabilistic recognition models provide no principled basis for unsupervised learning, appropriate stimulus reliability-dependent updating of a recognition model can be taken as a hallmark of the recognition model being probabilistic. Such optimal updating was reported in a perceptual discrimination task without feedback, in which human participants used their uncertainty estimates about the stimulus (the decision variable) to correctly update their estimate of the base rate of the stimulus (a parameter of the recognition model) (Zylberberg et al., 2018). Similarly, in an economic decision task (Behrens et al., 2007), participants near-optimally adjusted the learning speed of a dynamically fluctuating reward rate (decision variable), despite a lack of direct feedback about reward rates.

1.5 Behavioural evidence for fully Bayesian recognition models

The top row of Table 1.2 summarizes the four normative advantages of probabilistic recognition models over non-probabilistic ones that we discussed above and show how these advantages translate into experimental designs that distinguish between these classes. Our approach is based on the superior generalization properties (high data and memory efficiency) of probabilistic recognition models. The critical insight is that while non-probabilistic recog-

nition models can also learn to solve any decision making task, they can only do so after sufficiently long training. Thus, a proper test of probabilistic recognition models must create a situation in which the data- (or memory)-inefficient strategy of non-probabilistic models would be pushed to its limits. The strongest experimental tests capitalize on the extreme data efficiency of probabilistic models allowing one-shot generalization, previously investigated under the rubric of "Bayesian transfer" (Maloney and Mamassian, 2009). In addition, Table 1.2, top lists other, more subtle experimental tests that can still provide supporting evidence, based on the data and memory efficiency of probabilistic recognition models. While there have been previous proposals for the criteria that such experiments must meet (Maloney and Mamassian, 2009; Ma and Jazayeri, 2014), our approach based on normative principles allowed us to extend these proposals to other kinds of experiments that had not been considered in this context before.

Analogously to the probabilistic vs. non-probabilistic distinction, we can also use the normative advantages of fully Bayesian vs. task-dependent recognition models to suggest experimental strategies for distinguishing these in behavioral measurements (Table 1.2, bottom row). As a disclaimer, we note that the experimenter can never have perfect knowledge about which latent variables constitute the internal generative model of the subject, and therefore, behavioral tests are insufficient to distinguish between fully Bayesian and hybrid recognition models. Nevertheless, by demonstrating that participants represent the uncertainty of latent variables other then the decision variable, one may be able to exclude task-dependent recognition models.

In general, there is a basic experimental criterion that needs to be met regardless of the specific normative advantage we aim to utilize: we need to use complex stimuli that are characterized by multiple latent variables that differ in the level of (un)certainty with which they can be inferred. Without this across-variable diversity in uncertainty, one cannot exclude the possibility that participants summarize the uncertainty of the whole stimulus in a single value.

Task-flexibility. Given the constitutive uncertainty representation of a fully Bayesian recognition model, it can rapidly switch between utility functions that treat different latent variables as the decision variables. This can be tested in a sequential manner by making each latent variable the decision variable, one-by-one, by changing the utility function of the task across the trials. There is, however, a caveat of this method, that once the identity of the decision variable is revealed, a task-dependent recognition model is sufficient to solve the task. Nevertheless, this problem can be eliminated by revealing the identity of the decision variable only after the stimulus presentation, a typical strategy in multi-item working memory tasks (Ma et al., 2014). Multi-item working memory tasks represent a special case of this approach in which separate latent variables correspond to distinct items in a multi-element visual scene, and the identity of the queried item is only revealed once the stimulus disappears (typically after a delay period). Despite the widespread use of working memory experiments, only a small fraction of them is appropriate for identifying probabilistic recognition models at all (Van den Berg et al., 2017; Denison et al., 2018; Lengyel et al., 2015), and even within this smaller set of studies, we are only aware of two which were appropriate for distinguishing fully Bayesian from task-dependent recognition models (Denison et al., 2018; Lengyel et al., 2015). In these studies, the uncertainties associated with different items within the same scene were systematically varied by their contrast (Lengyel et al., 2015) or by an extraneous attentional cue (Denison et al., 2018). The representation of uncertainty about the queried item was assessed directly from participants' uncertainty reports (Lengyel et al., 2015; Denison et al., 2018), or indirectly from their categorization decisions (with non-trivial category boundaries, thus requiring an appropriate representation of uncertainty about stimulus orientation; Denison et al., 2018). Both studies provided evidence for participants' simultaneous representation of probabilistic information about multiple items in a scene.

The main drawback of the method used in typical working memory experiments is that it tests the (probabilistic) representation of latent variables (items in scene) one-by-one. Therefore, it can only provide evidence for the representation of the marginal posterior distributions of individual latent variables, not a full joint posterior over all of them. Testing the representation of joint posteriors would require complex utility functions that depend on more than a single latent variable. Spatial tasks, in which latent variables correspond to different spatial

dimensions rather than different items, seem a natural choice for this. Indeed, humans have been shown to be able to integrate their uncertainty with complex utility functions in such tasks (Maloney et al., 2007). We suggest that adapting this approach to the study of fully Bayesian recognition models is a promising avenue for future research.

Information fusion. An important advantage of fully Bayesian recognition models over task-dependent ones is that they can efficiently fuse information across observations that are not independent given the decision variable. In order to experimentally test this, there need to be decision variable-independent correlations among observations due to additional latent variables in the task's generative model. In the domain of motor control, it has been argued that not only the state of the environment (decision variable) but also that of the body (additional latent variable) determine motor errors (observations), (Berniker and Kording, 2008). Critically, just as the state of the environment changes continually so does the state of the body, thus creating correlations across the sequence of motor errors that we experience, requiring a joint probabilistic representation of environmental and body state for optimal behavior. Indeed, a fully Bayesian recognition model jointly inferring both latent variables successfully accounted for behavior in a variety of motor adaptation experiments (Berniker and Kording, 2008, 2011). Nevertheless, without explicitly investigating how well alternative models might be able to fit the data, the possibility of task-dependent recognition models cannot be fully excluded.

Interestingly, this kind of information fusion has not been exploited more generally to test for fully Bayesian recognition models. We suggest that future experiments could investigate information fusion in paradigms in which the decision variable (e.g. the color of a different object on each trial) needs to be estimated (e.g. based on observations of reflected light from the surface of the object) in the presence of non-decision ("nuisance") latent variables that have predictable temporal correlation structure across trials (mimicking e.g. slowly changing lighting conditions across the day). A key manipulation would be providing extra information about the nuisance variable (e.g. by revealing the color of the lighting source) on some trials. Solving such tasks successfully requires both dynamical inference over the nuisance variable and explaining away between the nuisance and decision variables, which taken together

implies performing joint inference over both variables – something that a task-dependent recognition model only inferring the decision variable could not achieve.

Active sensing. Just as in the case of passive information fusion, there can also be decision variable-independent co-variation across actively selected observations, which in turn can depend on nuisance variables. If these correlations modulate the informativeness of observations about the decision variable, then the active control of the sensors (presumably optimizing information about the decision variable) can benefit from representing the uncertainty of these nuisance variables. Although, once again, we are not aware of using such an active sensing approach to studying fully Bayesian recognition models, we suggest that it could be a fruitful future research direction. For example, the visual pattern categorization experiment used to study active sensing by (Yang et al., 2016a; described in Table 1.2), could be extended such that correlations across observations (pixels of an image) not only depend on the decision variable (stripy vs. patchy pattern, respectively defining the fall-off of spatial correlations between pixels to be longer in one direction than the other, or to be isotropic) but also on a nuisance variable (e.g. wavelength or spatial scale) that is not directly relevant for the task, but still influences correlations among observations. In this case, active sensing eye movements can benefit from inferring this nuisance variable (together with the decision variable), and this benefit should lead to behaviorally identifiable signatures.

Learning. As we saw previously, when learning needs to proceed unsupervised, the optimal adjustment of model parameters depends on the posterior distribution of the decision variable. The more accurately the recognition model approximates this posterior distribution, the more efficient learning will be. Thus, the efficient information fusion of fully Bayesian recognition models that improves inferences about the decision variable by also representing uncertainty about other latent variables (see above) should also improve unsupervised learning. Once again, this advantage of fully Bayesian recognition models has not been used to design specific experiments, although it could potentially reveal deep connections between the representation of uncertainty and learning.

1.6 Conclusion

In contrast to distinguishing probabilistic from non-probabilistic recognition models (Table 1.2, top), there is a notable paucity of behavioural experiments studying the fully Bayesian vs. task-dependent distinction (Table 1.2, bottom). This is not surprising given that this distinction has so far attracted little attention even at a conceptual level. The goal of this review was precisely to fill this gap.

First, we discussed a spectrum of possible recognition models with different uncertainty representations that can all compute optimal decisions in a given task, albeit at very different computational, data and memory costs. We argued that in this respect, fully Bayesian recognition models stand out with their superior data and memory efficiency, which allows for efficient generalization across a wide range of tasks and observation conditions. Given the parsimony of the hypothesis, it seems appealing to assume that general-purpose human and animal brains implement fully Bayesian recognition models in order to flexibly use the limited amount of experience they may have with any one task. Although there are neurally plausible implementations of fully Bayesian recognition models that can explain a number of neurophysiological observations, the available evidence is not conclusive and, ultimately, behavioral evidence will also be necessary to establish whether the recognition models the brain implements are task-dependent or closer to being fully Bayesian. Therefore, in this chapter, we organized the normative benefits of probabilistic and, more specifically, fully Bayesian recognition models from the perspective of the key cognitive advantages they offer (task-flexibility, information fusion, active sensing, learning). This allowed us to establish a set of experimental criteria that is suitable for distinguishing task-dependent and fully Bayesian (or hybrid) recognition models.

In the following chapters, guided by these experimental criteria, I will explore the extent to which the brain's recognition model represents probabilities. In Chapter 2, building on the concept of task-flexibility, I will examine whether the simultaneous representation of multiple marginal posterior distributions observed in working memory is already present in perception, and if so, what processes shape these representations. In Chapter 3, I will test the

probabilistic nature of implicit learning by investigating human decision-making in complex dynamic situations with multiple internal variables of the task. In Chapter 4, I approach the question from a different direction by introducing a novel hybrid approach and combining behavioral and neural data analysis of mice to search for neural traces of perceptual posterior distributions, distinct from the decision variable's posterior. Finally, in Chapter 5, I will conclude the findings regarding how the brain's recognition model manages uncertainty and propose potential future directions for further investigation.

Chapter 2

Uncertainty representations beyond the decision variable

I found that explicit uncertainty reports consistently reflected the manipulations of difficulty across the three stimulus features and remained predictive of accuracy even when the mean effect of stimulus features was factored out. Furthermore, the explicit uncertainty reports became increasingly predictive of estimation accuracy as presentation time became longer. I analyzed the data within the framework of sequential sampling models (SSMs). The modelling allowed me to distinguish between whether uncertainty is directly encoded in the perceptual representation or inferred from proxies provided by the stimulus features. I confirmed the generality of my method using two separate versions of SSMs – noisy evidence accumulators and probabilistic samplers. My experimental findings were in line with the behavior of SSMs that utilize genuine probabilistic perceptual representations rather than cognitive proxies for assessing uncertainty. Notably, as contrast in our trials was manipulated at the level of individual items, my results imply that perceptual uncertainty during such tasks is encoded at the item level, offering a more nuanced representation of uncertainties beyond a mere summary statistic of overall uncertainty.

2.1 Introduction

As discussed in detail in the introductory chapter, representing all latent variables of interest to the brain with their uncertainty is particularly useful for data and memory efficiency. However, the growing evidence supporting the existence of probabilistic representations in the brain concerns almost exclusively the decision variables of the task. The technical difficulty of measuring the representation of other types of variables (and parameters) of the internal representation has left the nature of these other representations largely unexplored. The present chapter aims to fill this gap.

Most studies testing the hypothesis of probabilistic decision making – whether focusing on the optimal computation (Körding and Wolpert, 2004; Zylberberg et al., 2018) or utilization (Whiteley and Sahani, 2008; Qamar et al., 2013) of the decision variable – have leveraged the direct measurability of the decision variable's representation. In contrast, directly measuring the representation of variables other than the decision variable (e.g. perceptual variables) is typically not feasible and this makes the testing of these representations complicated. However, there are exceptions, for example those working memorys (WM) experiments that use visual stimuli with multiple items, each with a unique visual feature (perceptual variable), and only reveal some time after the stimulus offset which item was the "target" whose visual feature (decision variable) needs to be recalled. In these tasks, the identity of the decision variable is unknown to participants during the formation of perceptual representation. By converting one of the perceptual variables into the decision variable only after the stimulus presentation is over, it becomes possible to test whether this variable was encoded probabilistically from the outset before it become the decision variable. In this study, I took inspiration from these WM experiments and developed a perceptual (orientation) estimation task with exactly such a structure. My task contained stimuli with multiple oriented items in each trial, and the target item's identity was revealed only after the presentation of the stimulus was completed. The participants task was then to report their best estimate of the target item's orientation (decision variable) and their subjective uncertainty about this estimate. To avoid confusion, here I note that although mine is an estimation task, throughout the chapter I will sometimes refer to the orientation of the target item as the decision variable, and con-

sequently I will refer to the estimation process as decision making, in order to be congruent with the terminology of the rest of the thesis.

To test the core hypothesis that uncertainty is represented at the level of individual items, it was necessary to modulate task difficulty at this granular level, thereby making the individual items' impact on uncertainty identifiable. I achieved this by independently adjusting the contrast of each item within the same display. Similar item-specific stimulus manipulations have been used in earlier working memory (WM) studies. These studies successfully demonstrated that both humans (Denison et al., 2018; Yoo et al., 2021) and animals (Devkar et al., 2017) take into account item-specific WM uncertainty in their decisions. However, my study focused on perceptual rather than WM representations and this led me to diverge from the traditional WM paradigm in two key ways. First, I completely eliminated the delay between the stimulus offset and the exposition of decision variable (the target item's identity), which is typical in WM tasks (≈ 1 sec or more). By testing the nature of the representation right after the end of stimulus presentation, I potentially reduced the involvement of WM in the decision process. Second, to gain a deeper understanding of the nature of the subject's perceptual representation, I gathered information on its temporal evolution by manipulating the stimulus presentation time. This manipulation allowed me to test whether the formation of the perceptual representation is instantaneous or requires time, and if the latter is true, exactly what processes take up time. In my study, I opted for using static stimuli – for which all relevant information is present at the moment of stimulus onset – rather than dynamic stimuli, e.g. random dot motion patterns. This is because collecting information about relevant dynamic variables such as speed or direction of motion, by definition, requires time due to the physical constraints of the stimulus and this could be confounded with time requirements of the underlying general mechanisms of perceptual processing.

Although my paradigm is suitable for testing the perceptual representation at the level of individual items, to draw meaningful conclusions about the item-level uncertainty representations, it is crucial to first confirm that the perceptual representation is, indeed, probabilistic. For this, I must rule out the possibility that uncertainty reports are based on the value of other nuisance variables that affect the quality of representation rather than on the percep-

tual representation itself. For example, the observed value of the target's contrast provides a reliable *proxy* for the accuracy of orientation representation, which can potentially be used to provide sensible uncertainty reports without having a truly probabilistic perceptual representation of orientation. There have been two distinct approaches in the literature to address such confounds. The first approach builds on the idea that a proper representation of uncertainty should act as a common currency between distinct stimulus manipulations, thereby naturally using a common scale to express the expected effects of these manipulations on accuracy (De Gardelle and Mamassian, 2014; de Gardelle et al., 2016). In the absence of intrinsic uncertainty representations, it would require considerable cognitive effort to learn the appropriate mapping between individual proxies and their combinations to the quality of the representations, which might even be intractable for the large number of latent variables needed to describe everyday decision situations. Hence, demonstrating that equal-sized changes in accuracy, caused by two or more distinct stimulus manipulations, result in equal-sized changes in reported uncertainties – thus putting these manipulations on a common scale – would bolster the assumption that uncertainty is indeed properly represented. The second approach posits that a proper uncertainty representation should also reflect variability in the quality of representation that is independent of the stimulus – variability that is present even if the same stimulus is presented multiple times (Ma, 2012; Honig et al., 2020). This variability might originate from the fluctuation of internal processing, e.g. due to attentional fluctuations, in which case this approach is essentially equivalent to the common scale approach, by extending it to include not only the effect of stimulus manipulations but also those of internal states to which the experimenter may have no access (at least when attention is not experimentally modulated as in Denison et al., 2018). To be comprehensive, I employed both approaches here. On the one hand, I used a total of three nuisance parameters to control task difficulty: contrast, presentation time, and set size (number of items simultaneously presented in a display). On the other hand, each stimulus was presented multiple times to the participants to assess stimulus-independent fluctuations in behaviour.

Importantly, manipulating presentation time allowed me to fit process-level computational models to the potentially time-dependent behavioural performance of the subjects. These models, part of the sequential sampling family (Forstmann et al., 2016), explained perceptual

processing in terms of the gradual accumulation of temporally fluctuating evidence samples. The models varied along two dimensions: the type of information represented by the samples and the extent to which uncertainty judgements were aided by proxies. The samples could either provide noisy information about a point estimate (e.g. drift diffusion model; Ratcliff and McKoon, 2008) or act as probabilistic samples representing the histograms of an entire posterior distribution (Fiser et al., 2010). These samples were then evaluated by ideal observers, which relied on proxies to varying extent to infer the uncertainty of the target orientation estimate. At one end of the spectrum, the observer neglected the samples altogether, basing its uncertainty judgments entirely on proxies (Fig. 2.1, first pink arrow), assuming a perfect mapping of proxies to uncertainties. In this scenario, the recognition model might still be probabilistic based on our definition (depending on how well it generalizes to novel stimuli), as it provides meaningful uncertainty reports, but the probabilistic computation occurs only at the decision phase and thus it is entirely task-dependent, with no uncertainty evaluated (only proxies encoded) at the perceptual phase. Importantly, such a model lacks information about potential proxy-independent fluctuations in the quality of perceptual representations, and is potentially very fast compared to the sampling-based models, as only point estimates of the proxies need to be encoded. At the other end of the spectrum, the ideal observer bases its uncertainty judgments entirely on the samples (Fig. 2.1, second pink arrow). In this case, the samples carry information about the quality of perceptual representation, which, if item-specific, makes the recognition model (at least up to a certain extent) task-independent. This perceptual representation is inherently time-dependent – more samples lead to a more accurate representation – and as I will show, it improves both the estimation accuracy and the reliability of uncertainty judgments over time. Finally, there is an in-between model that primarily relies on samples but takes advantage of proxies to enhance the quality of the ideal observer's inference, particularly in situations with a small number of samples (Fig. 2.1, second and third pink arrows jointly). Initially, this model leans more heavily on proxies, gradually shifting to rely more on the samples as more time passes. Importantly, this model is again task-independent, although it is optimized through clever "proxy-based" priors.

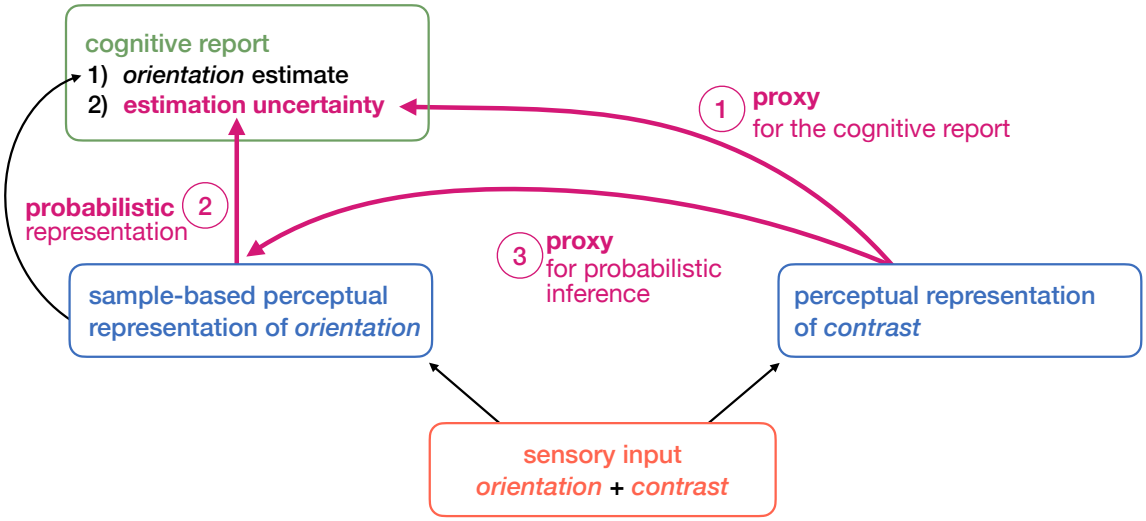


Figure 2.1. Potential processes behind the cognitive reports. The sensory input (orange box) contains information about both the decision variable (orientation) and the nuance variables (e.g., contrast), which contribute to forming the perceptual representations (blue boxes). (Although orientation and contrast representations are depicted separately for clarity, they may not be as distinct in reality. This distinction only serves to illustrate the role of proxies here.) In terms of cognitive reports (green box), the orientation estimate is derived from the (sample-based) perceptual representation of orientation, but the uncertainty judgement is not necessarily so (pink arrows). Uncertainty judgements may rely entirely on proxies (first arrow), or alternatively, it could be inferred from the perceptual representation of orientation (second arrow), provided it contains probabilistic information. This inference could be enhanced by the use of proxies (third arrow).

I conducted extensive simulations with all model variants and compared the synthetic behavioral patterns to the experimentally measured human behavior to reveal the nature of human uncertainty representations.

2.2 Sequential sampling models

To understand how uncertainty representations are formed and what information they rely upon, I turned to Sequential Sampling Models (SSMs), the dominant process-level models of perceptual decision making (Gold and Shadlen, 2007; Forstmann et al., 2016). These models have been successful in explaining the time dependence of decision accuracy and subjective confidence reports (Kiani and Shadlen, 2009; Kiani et al., 2014) and, by considering normative factors, they could elucidate the distribution of reaction times in a wide range of conditions (Drugowitsch et al., 2012). Moreover, recently these models have also been proposed

as suitable models for working memory, capturing its capacity limits in tasks similar to my experiment (Schneegans et al., 2020).

In order to keep my model-based analysis general, I tested two fundamentally different classes of SSMs that attribute sample variability to different causes: either to the presence of noise both in the external sources and in the internal sensory processes (Ratcliff and McKoon, 2008; Shadlen and Kiani, 2013) , or to the stochastic nature of the approximate Bayesian computations that the brain is assumed to perform (Fiser et al., 2010; Orbán et al., 2016; Pitkow, 2016).

These models enabled me to generate specific predictions regarding how certainty correlates with the accuracy of perceptual estimates. I tested three variants of each model, differentiated by the extent to which they rely on proxies for certainty computation. By comparing the predictions of these models to human behaviour, I was able to assess whether humans employ probabilistic perceptual representations that contain information about their own reliability, or if they merely rely on proxies to estimate the reliability of their non-probabilistic perceptual representations.

2.2.1 Noise model (classical evidence accumulation)

In traditional SSMs, samples carry noisy pieces of evidence about the decision variable. By accumulating multiple independent samples, the noise-related uncertainty can be effectively reduced and eventually eliminated over time. In my model, to be consistent with the experimental paradigm, samples carry information about the perceptual variables, but I only model the particular variable that later becomes the decision variable.

At equal time intervals, N independent evidence samples ($\tilde{x}_n$, $n \in \{1, 2, ..., N\}$) about the true stimulus orientation (x) are drawn from a circular normal (von Mises) noise distribution centered on the ground truth (Fig. 2.2A, left side). This distribution is characterized by its circular precision (ρ_{S}) that varies from trial to trial expressing the natural variability of

uncertainty and the effect of stimulus features:

$$\tilde{x}_n \sim \mathrm{vM}(x, \rho_\mathrm{S}) \tag{2.1}$$

(Note that I use an unconventional notation here, parameterising the von Mises distribution by its precision rather than its concentration, see the conversion in Eq. A.9). These samples constitute the noisy sensory representation. A hypothetical downstream *ideal evidence accumulator* (IEA), acting as the ideal observer of the evidence samples, accumulates these samples and, based on the generative model of the samples (Fig. 2.2B, left side, without gray part), computes the posterior distribution of the target orientation (Fig. 2.2C, left side):

$$\mathcal{P}(x \mid \tilde{x}_{1:N}) \propto \mathcal{P}(x) \prod_{n=1}^{N} \mathcal{P}(\tilde{x}_n \mid x) \tag{2.2}$$

$$= \mathcal{P}(x) \int \mathrm{d}\rho_\mathrm{S}\, \mathcal{P}(\rho_\mathrm{S}) \prod_{n=1}^{N} \mathrm{vM}(\tilde{x}_n; x, \rho_\mathrm{S}) \tag{2.3}$$

(For the full expression, including all the potential variables of the generative model, including stimulus strength, see the Appendix, Section A.2.1.) I refer to this distribution as the (posterior) predictive distribution of the IEA to avoid confusion with the perceptual posterior that will be introduced later at the description of the Signal model. I assume that this predictive distribution forms the basis of explicit uncertainty reports.

Importantly, in the Noise model, all the IEA's uncertainty stems from the noisiness of the samples. Since this uncertainty can, in principle, be completely eliminated over time, it is a *reducible* form of uncertainty.

2.2.2 Signal model (probabilistic sampler)

Sampling-based probabilistic models can be seen as special versions of SSMs (Lengyel et al., 2015). In the probabilistic sampling model, it is the perceptual posterior that is stochastically represented by the samples, in such a way that the histogram of samples approximate the

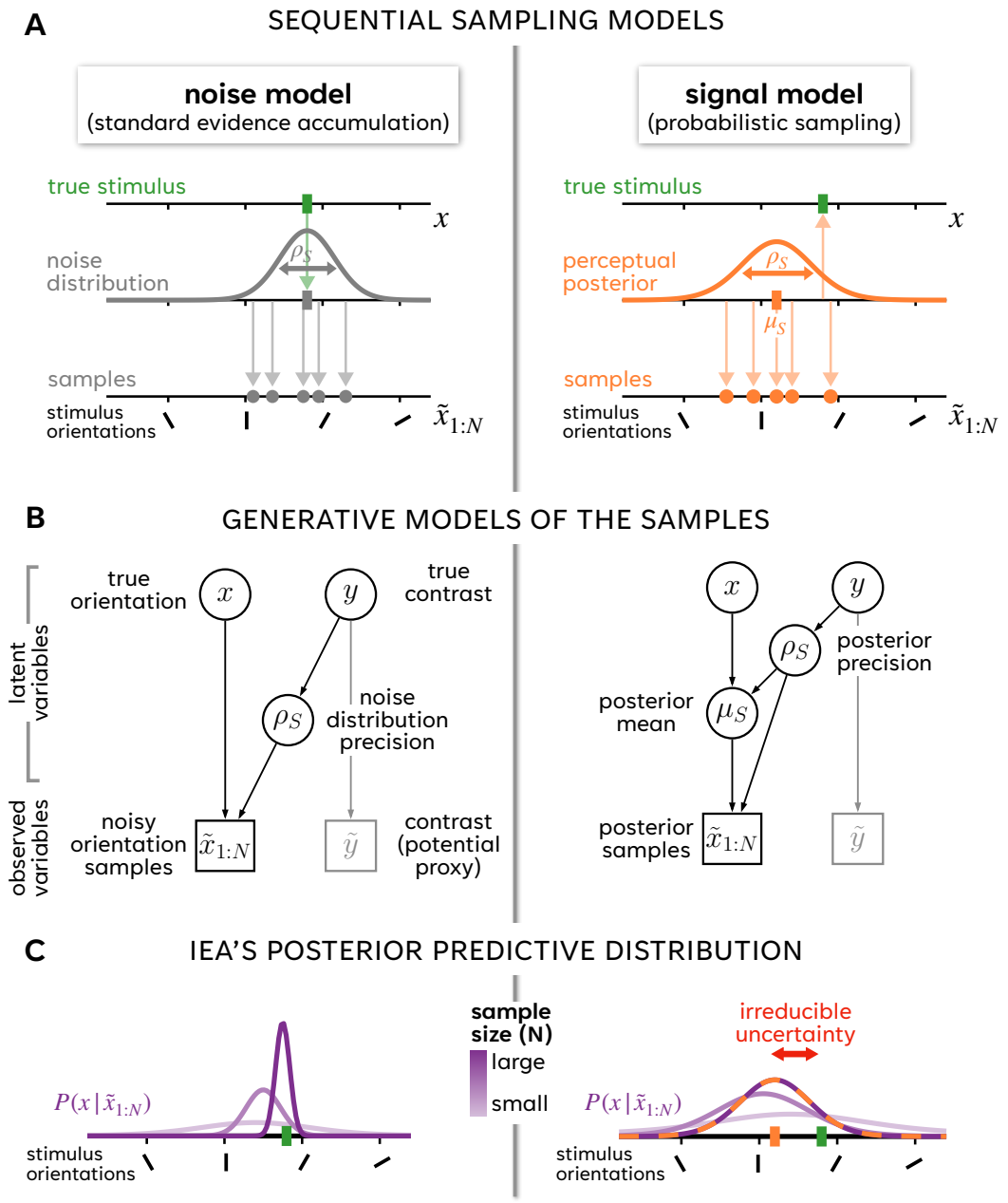


Figure 2.2. Sequential sampling models of perception. A. Graphical comparison of the noise model (gray) and the signal model (orange). In the noise model, the sampling distribution is centered on the true target orientation of the stimulus (green), while in the signal model, the true target orientation acts as if it was also sampled from the perceptual posterior. **B**. Graphical representation of the sample generating processes. (For the signal model, only the variables that differ from the noise model are labeled.) Gray arrows indicate the potential presence of proxies – additional observations of nuisance variables beyond the orientation samples—that assist in inferring the posterior precision. **C**. The (posterior) predictive distributions computed by the downstream ideal evidence accumulators. In both models, the predictive distributions become narrower with increasing sample size. In the noise model, the distribution converges to a Dirac-delta distribution (infinitely narrow distribution) centered on the ground truth, eliminating all uncertainty in the orientation estimate. In the signal model, it converges to the finite precision perceptual posterior, indicating an irreducible part of the uncertainty that cannot be eliminated by further sample accumulation.

actual (and generally intractable) posterior. Thus, sample variability serves as a signal – hence the name –, encoding the precision of the posterior, rather than being mere noise.

In our model, the perceptual posterior of the stimulus orientation is a von Mises distribution parameterized by its mean (μ_{S}) and precision (ρ_{S}), both of which vary from trial to trial, expressing the natural variability of uncertainty and the effect of stimulus features (Fig. 2.2A, right side). At equal time intervals, an independent probabilistic sample ($\tilde{x}_n$) is drawn from this posterior:

$$\tilde{x}_n \sim \mathrm{vM}(\mu_{\mathrm{S}}, \rho_{\mathrm{S}}) \tag{2.4}$$

Importantly, due to the nature of probabilistic representations, the true stimulus orientation acts as if it was just another sample drawn from this posterior (Fig. 2.2A, upward pointing orange arrow):

$$x \sim \mathrm{vM}(\mu_{\mathrm{S}}, \rho_{\mathrm{S}}) \tag{2.5}$$

To put the Noise and Signal models on equal footing, here again, an IEA, which knows the generative model of the samples (Fig. 2.2B, right side), is assumed to compute the (posterior) predictive distribution of target orientation given the samples (Fig. 2.2C, right side, without gray part):

$$\mathcal{P}(x \mid \tilde{x}_{1:N}) \propto \mathcal{P}(x) \prod_{n=1}^{N} \mathcal{P}(\tilde{x}_n \mid x) \tag{2.6}$$

$$= \mathcal{P}(x) \int \mathrm{d}\rho_{\mathrm{S}}\, \mathcal{P}(\rho_{\mathrm{S}}) \int \mathrm{d}\mu_{\mathrm{S}}\, \mathrm{vM}(\mu_{\mathrm{S}}; x, \rho_{\mathrm{S}}) \prod_{n=1}^{N} \mathrm{vM}(\tilde{x}_n; \mu_{\mathrm{S}}, \rho_{\mathrm{S}}) \tag{2.7}$$

(For the full expression, see the Appendix, Section A.2.1.) This distribution forms again the basis of decisions.

In contrast to the Noise model, just one part of the total uncertainty is reducible. This reducible uncertainty arises because a finite number of samples cannot capture perfectly an underlying distribution. The remaining uncertainty, controlled by ρ_{S}, is *irreducible* (Fig. 2.2C,

right side), and its presence is precisely what makes probabilistic perceptual representations desirable.

2.2.3 Comparing the two SSM models

The main difference between the two SSM approaches can be summarized neatly using the Bayesian encoding-decoding terminology (Zemel et al., 1998; Lange et al., 2023). In both models, the IEA acts as a Bayesian 'decoder', performing optimal perceptual inference based on the true generative model of the observed samples. However, in the Signal model, the samples themselves 'encode' Bayesian posteriors, whereas in the Noise model, the samples encode only a point estimate, albeit noisily.

2.2.4 Proxies for inference

In both the Noise and Signal models, I assume that presentation time controls the number of samples (N) accumulated during stimulus presentation, while contrast and set size determine ρ_{S}, the precision of the sampling distribution (at least in expectation). I refer to the combined effect of the last two factors as stimulus strength ($y \in [0, 1]$). The IEAs are assumed to fully observe the sample number (N) but not necessarily stimulus strength.

In both models, ρ_{S} defines the reliability of the individual samples about the ground truth target orientation ($x^{\star}$). In a given trial, t, $\rho_{\mathrm{S}}^{(t)}$ is inferred based on the empirical sample distribution ($\tilde{x}_{1:N}^{(t)}$) and the (potentially noisily observed) stimulus strength ($\tilde{y}^{(t)}$). This knowledge is incorporated in the computation of the posterior predictive distribution (Eq. 2.3 and Eq. 2.7). The predictive distribution, now including $\tilde{y}$, can be expressed as follows:

$$\mathcal{P}\Big(x \mid \tilde{x}_{1:N}^{(t)}, \tilde{y}^{(t)}\Big) = \int \mathrm{d}\rho_{\mathrm{S}}^{(t)}\, \mathcal{P}\Big(x \mid \tilde{x}_{1:N}^{(t)}, \rho_{\mathrm{S}}^{(t)}\Big)\; \mathcal{P}\Big(\rho_{\mathrm{S}}^{(t)} \mid \tilde{x}_{1:N}^{(t)}, \tilde{y}^{(t)}\Big) \tag{2.8}$$

$$\propto \int \mathrm{d}\rho_{\mathrm{S}}^{(t)}\, \mathcal{P}\Big(x \mid \tilde{x}_{1:N}^{(t)}, \rho_{\mathrm{S}}^{(t)}\Big)\; \mathcal{P}\Big(\tilde{x}_{1:N}^{(t)} \mid \rho_{\mathrm{S}}^{(t)}\Big)\; \mathcal{P}\Big(\rho_{\mathrm{S}}^{(t)} \mid \tilde{y}^{(t)}\Big) \tag{2.9}$$

This formulation shows that, in the large sample limit, $\tilde{y}^{(t)}$ has minimal influence on the predictive distribution because $\tilde{x}^{(t)}_{1:N}$ already provides strong evidence about $\rho_{\mathrm{S}}^{(t)}$. However, in the low sample limit, $\tilde{y}^{(t)}$ becomes more critical. In the extreme case of having only a single sample, and when ρ_{S} is independent of x, $\tilde{y}^{(t)}$ serves as the only trial-specific information about $\rho_{\mathrm{S}}^{(t)}$. Interestingly, in this case, if $y^{(t)}$ wasn't observed at all, there would be no variability in the shape (just the position) of $\mathcal{P}\left(x \mid \tilde{x}^{(t)}_{1:N}, \tilde{y}^{(t)}\right) = \mathcal{P}\left(x \mid x_1^{(t)}\right)$ across trials. Intuitively, the IEA's uncertainty should be reflected in the shape of the posterior predictive distribution, therefore if its shape remains constant, the model's uncertainty will also stay constant across trials, even though its estimation accuracy is still be modulated by $y^{(t)}$. Thus, when the sample size is small, $\tilde{y}^{(t)}$ serves as a useful proxy for estimating the y dependence of the IEA's accuracy.

I developed three different base versions of the IEA model, each corresponding to varying degrees of reliance on proxies:

1. **Proxy-only (no inference):** This model relies solely on proxies, not accounting for the variability of samples to estimate ρ (or the IEAs predictive distribution). In this case, only the stimulus-specific prior mean is taken into account for making certainty reports (Fig. 2.3F leftmost image)

2. **IEA without proxies:** This model does not observe the stimulus strength variable (y) at all (Fig. 2.2B models without the gray arrow), so it cannot use it as a proxy for ρ_{S}, and the IEA needs to rely on a stimulus-agnostic ρ_{S} prior (Fig. 2.3C leftmost image).

3. **IEA with proxies:** This model fully observes stimulus strength (Fig. 2.2B models with the gray arrow, which, in this case, expresses an identity relation between y and $\tilde{y}$). This knowledge puts stimulus-specific constraints on the shape of the ρ_{S} prior (Fig. 2.3E leftmost image).

Besides the three base models, there could be a range of models between version 1 and 2 in which stimulus strength is imperfectly observed (or observed with noise). In these cases, $\tilde{y}$ is modelled as being drawn from a noise distribution (see, Methods) that depends on the

true stimulus strength and a single 'observedness' parameter, λ, which interpolates smoothly between the fully unobserved ($\lambda = 0$, version 1) and fully observed ($\lambda = 1$, version 2) stimulus strength scenarios.

2.2.5 SSM predictions

Model simulation

I simulated the behavior of each variant of the Noise and Signal models at different presentation times, i.e. samples sizes (N), and stimulus strengths ($y \in [0, 1]$).

Generative model of the samples: In each simulated trial, the true target orientation, $x^\star$, was fixed to 0° (unbeknown to the model, thus without loss of generality). The precision of the sampling distribution, ρ_S, was randomly drawn from a Beta distribution centered on the stimulus strength, y:

$$\rho_\mathrm{S} \sim \mathrm{Beta}\left(y \cdot K\left(y\right), \left(1 - y\right) \cdot K(y)\right) \tag{2.10}$$

such that the variance (v) of the distribution was kept fixed across all stimulus strengths:

$$K(y) = \frac{y\left(1 - y\right)}{v} - 1 \tag{2.11}$$

In the Noise model, the mean of the sampling distribution, μ_S, was set to equal $x^\star$ (i.e., 0°). In the Signal model, μ_S was drawn from a circular normal distribution centered on $x^\star$ with precision ρ_S:

$$\mu_\mathrm{S} \sim \mathrm{vM}(0, \rho_\mathrm{S}) \tag{2.12}$$

Next, N independent samples were drawn from the sampling distribution:

$$\tilde{x}_{1:N} \sim \mathrm{vM}(\mu_\mathrm{S}, \rho_\mathrm{S}) \tag{2.13}$$

and a noisy observation of the stimulus strength was generated:

$$\tilde{y} \sim \mathrm{Beta}\left(\mu(y,\lambda), \sigma^2(\lambda)\right) \tag{2.14}$$

where

$$\mu(y,\lambda) = \lambda\, y + \frac{1-\lambda}{2} \qquad \sigma^2(\lambda) = \frac{1}{12}(1-\lambda)^2 \tag{2.15}$$

The λ parameter governs the observedness of y. When $\lambda = 1$, the observation is noiseless ($\tilde{y} = y$), meaning that the stimulus strength is perfectly observed. When $\lambda = 0$, the stimulus strength is not observed ($\tilde{y} \sim \mathrm{Unif}(0,1)$), so $\tilde{y}$ contains no information about y. Intermediate values of λ smoothly interpolate between these two extremes, such that the variance is y independent.

Behavioural reports: The orientation estimate, μ, was always the sample mean:

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2.16}$$

which equaled the mean of the IEAs' posterior predictive distribution (see Appendix, Section A.2.1), due to the flat orientation prior (matching the experimental distribution).

The certainty judgement depended on the model variant:

1. For the IEA models, with or without proxy, it was the expected cosine error under the IEAs' predictive distribution:

$$\rho = \int \mathrm{d}x \, \cos(x - \mu) \, \mathcal{P}(x|\tilde{x}_{1:N}, \tilde{y}) \tag{2.17}$$

   The solution for this integral, across different models, is provided in Appendix Section A.2.1.

2. For the proxy-only variants, it was assumed that the model knew (e.g. from previous experience) the expected cosine error's (ρ) dependence on proxies ($\tilde{y}$) and the sample number (N):

$$\rho = \int \mathrm{d}e \, \cos(e) \, \mathcal{P}(e \mid \tilde{y}, N) \tag{2.18}$$

   where

$$e = x^{\star} - \mu \tag{2.19}$$

   and used this ρ to make a certainty judgment.

To increase the predictions' generality, I introduced a potential noise to the uncertainty reports, which could deteriorate the accuracy of uncertainty estimates. Noisy certainty reports were sampled from a Beta distribution centered on the noiseless certainties:

$$\rho_{\text{noisy}} \sim \text{Beta}(\rho \, \zeta, \, \rho \, (1 - \zeta)) \tag{2.20}$$

(now, parameterized by its mean and variance) where ζ is the parameter that defines the magnitude of noise. Later on, I denote the noise on a given trial with ϵ, which is simply $\rho_{\text{noisy}} - \rho$.

Evaluation of the model performance

I used the accuracy metric to characterize the performance of the model and later of the participants (Fig. 2.3B), which is defined as the average cosine distance between the true stimulus orientation, $x^\star$, and the IAE's estimate, μ, computed on some subset of trials ($\mathcal{T}$):

$$\text{accuracy} := \frac{1}{\mathcal{T}} \sum_{t \in \mathcal{T}} \cos(\mu_t - x_t^\star) \tag{2.21}$$

Data analysis

To find qualitative differences between the model variants, I examined the relationship between their reported certainty and their measured accuracy in the simulated data. Since the IEAs are ideal observers of the orientation samples and the (potential) proxies, in the noiseless case, their certainty reports would accurately predict their accuracy at the reported certainty level, and we call them *well-calibrated*. However, in the presence of noise, the IEAs won't appear well-calibrated anymore based on their noisy certainty reports. One way to address this issue is by grouping the trials according to the variables that define their stimulus properties, which would average out the noise in the certainty reports at the group level. Critically, however, the well-calibrated property of the ideal observer does not automatically manifest once trials are grouped in such a way. Instead, the form of the certainty-accuracy relationship depends on both the chosen model variant and the specific method used for grouping the trials (Drugowitsch et al., 2014). Specifically, the certainty-accuracy relationship will only appear well-calibrated when the variables according to which trials are grouped are also available to (i.e. observable by) the ideal observer, thus providing a diagnostic for identifying which variables are and are not available to the ideal observer. I will exploit this dependency to distinguish between the different model variants. Specifically, I will examine how the certainty-accuracy relationship apparently deviates from being well-calibrated when different grouping strategies are used.

In the following analysis, I will use three distinct grouping methods. Trials will be grouped based on (1) the subjective certainty (ρ) the model expresses (Fig. 2.3C, second column), (2) the objective strength of the stimulus (y) (Fig. 2.3C, third column), or (3) the subjective certainty the model expresses, but with the effect of the stimulus marginalized out (Fig. 2.3C, fourth column; see Methods). For all three methods, trials of different presentation times will be analysed separately (Fig. 2.3C, second to fourth column).

Before I go into the details of the model prediction, I give a brief summary of the main results: As shown in Fig. 2.3 C-F and Fig. 2.4, the model variants using proxies to different extents can be distinguished from each other based on the qualitative differences observed in the three types of trial grouping, but these might not necessarily suffice to separate the Noise and Signal models.

Prediction of the Signal models

First, I discuss the Signal model's predictions in detail, then I apply the same line of arguments to the Noise model to highlight the similarities and differences between the two.

In the Signal model, the sampling distribution's precision (ρ_{S}) on a given trial determines the magnitude of irreducible uncertainly corresponding to the discrepancy between the sampling distribution's mean (μ_{S}) and the ground truth ($x^{\star}$). To make accurate uncertainty judgements, the model must infer ρ_{S} based on the available samples and proxies (Eq. 2.9). The IEA ultimately computes ρ, instead of ρ_{S}, as the ultimate predictor of accuracy by also taking into account the reducible part of its uncertainty (due to the finite number of samples).

I consider first the without-proxy variant of the IEA model, which does not observe the stimulus strength, and thus relies on a generic prior of ρ_{S} in every trial, regardless of stimulus strength (Fig. 2.3C, first column, upper row). In the very short presentation time limit, when only a single sample is available, the best estimate of ρ_{S} for each trial is simply the generic prior's mean, and thus all variation in the certainty reports, ρ, is due to noise. This makes certainty-based grouping effectively equivalent to random grouping (with respect to accuracy), resulting in constant accuracy across certainty bins. Therefore, the slope of the accuracy-certainty regression is zero (Fig. 2.3C, second column, upper row, bright purple line).

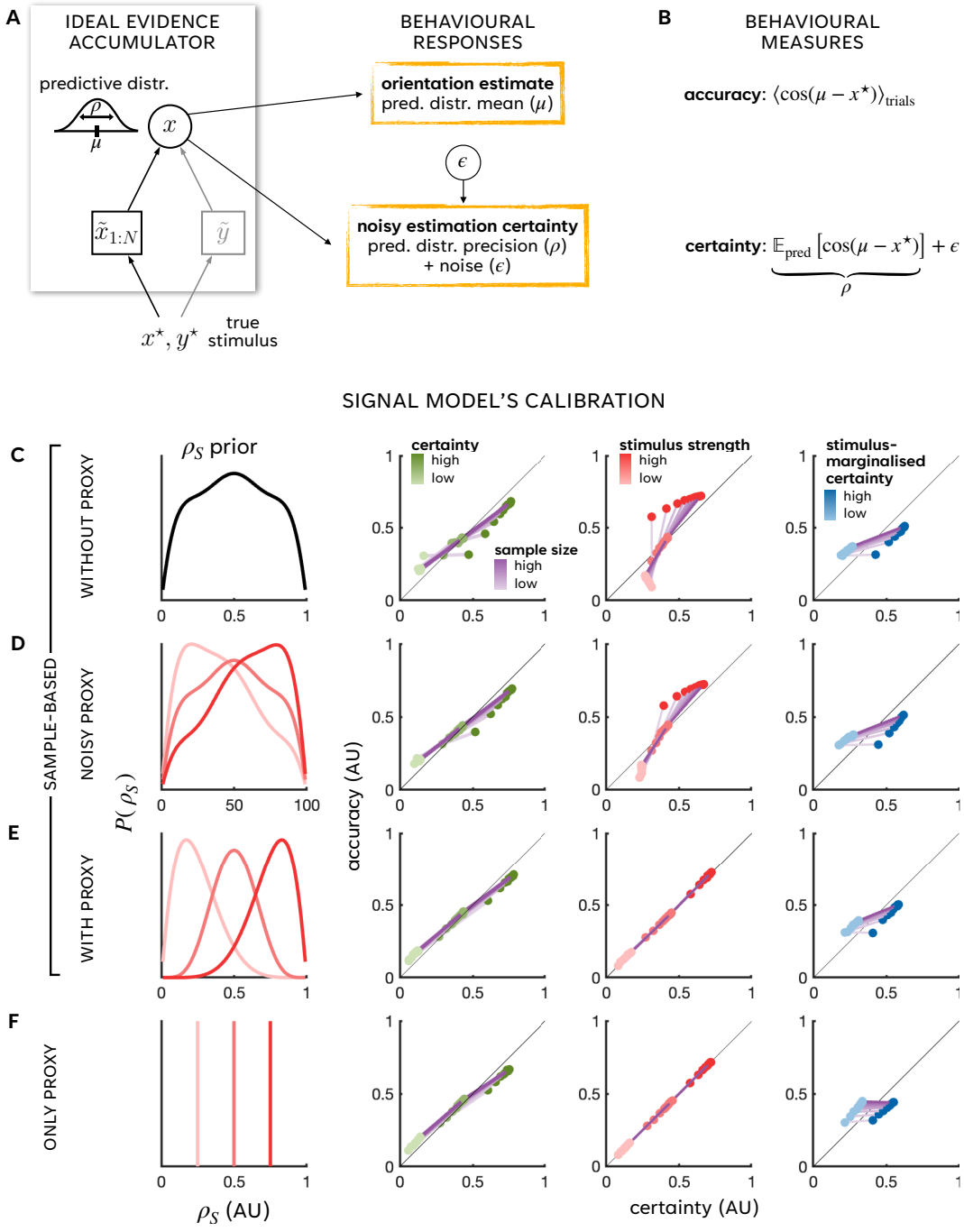


Figure 2.3. Behavioural predictions of the Sampling model's Ideal Evidence Accumulators (IEAs). A. IEAs compute the predictive distribution of stimulus orientation and report its mean (μ) and noisy circular precision ($\rho + \epsilon$) as the orientation estimate and its uncertainty, respectively. **B**. Behavioral measures used for analysis are accuracy, defined as the average cosine error of estimates across a set of trials, and uncertainty, defined as the expected cosine error (plus potential noise) on a given trial according to the IEA. **C-F**. Behavioural predictions of the model variants differing in their reliance on proxies. 1^{st} column: Priors of the sampling distribution's precision. 2^{nd} to 4^{th}: Accuracy vs. certainty as a function of sample size (purple shades), averaged according to either the certainty reports (green shades), stimulus strength (e.g., contrast, red shades), or the stimulus-marginalized certainty reports (blue shades).

This slope will increase with presentation time (Fig. 2.3C, second column, upper row, purple gradient), as accumulated samples gradually reveal the true shape of the sampling distribution in the current trial, and it will asymptotically converge to a limit-slope, slightly below the 45° slope of the well-calibrated line. This limit slope is determined by the conditional distribution of ρ_{S} given the stimuli (Fig. 2.3C, first column, middle row, red lines) and the magnitude of certainty noise controlled by ζ.

Crucially, the certainty-accuracy relationship looks very different if we choose to group the data according to the objective stimulus strength, y, instead of the subjective certainty reports, ρ. In this case, the stimulus feature used to bin data strongly influences accuracy, but this variant of the model has no access to it. In the limit, when responses are based on a single sample, stimulus strength strongly modulates the informativeness of this single sample about the ground truth, but certainty reports do not reflect this modulation at all (a single sample carries no information about its own informativeness). This implies that the accuracy-certainty slope is at the other extreme, being 90°. This slope is not much affected by the certainty noise, as random noise largely averages out across the trials within a stimulus strength bin. Again, as more samples are collected, and consequently the trial-specific ρ_{S} (which is affected by the stimulus strength) is getting better estimated by the observer, the slope gradually regresses back to the well-calibrated line. (There is no limit line in this case, as the effect of noise largely averages out.) The slopes exceeding 45° (at small sample sizes) reflect a regression to the generic prior mean effect, which diminishes as additional samples provide more evidence. This leads to the overestimation of easy trials and the underestimation of hard trials, a phenomenon known as the 'hard-easy effect' (Drugowitsch et al., 2014; Khalvati et al., 2021). An important prediction of the sequential sampling framework is that the strength of the hard-easy effect varies over time.

In the IEA model with proxies (that are perfectly observed), the accuracy-certainty regression shows quite different patterns. I explore this variant under conditions where the stimulus-induced variability of ρ_{S} is comparable to the magnitude of stimulus independent fluctuations, and both are significantly larger then the magnitude of certainty noise. In the IEA plus proxy model, the ρ_{S} prior on any given trial is conditioned on the observation of the

nuisance variable (Fig. 2.3E, first column). Therefore, when we bin the data according to the certainty reports, the slope is already very close to the limit-line with just a single sample available, and there is no significant improvement over time (Fig. 2.3E, second column). Interestingly, the proxies also provide sufficient information for an exact estimate of the accuracy within a stimulus strength condition, therefore the accuracy-certainty points are on the well-calibrated line irrespective of the sample size when grouped by stimulus strength (Fig. 2.3E, third column).

If we now consider a model where the stimulus strength is only partially observed ($\lambda = 0.5$ in the simulation), then its ρ_{S} prior interpolates between the priors of the IEA model with no proxies and the IEA model that relies on noiseless proxies (Fig. 2.3D, first column). As a consequence, this model's behavior will also interpolate between the former model's behavior (Fig. 2.3D, second and third column). In this case, the slopes deviate from the 45° well-calibrated line when the sample size is small, but not to the same extent as in models that do not use proxies.

Up to this point, none of the arguments above utilized the fact that the stimulus-conditioned ρ_{S} prior still has a finite width, therefore the same arguments and consequently the same accuracy-certainty patterns hold for the proxy-only model, which has only access to the prior means (Fig. 2.3F, second-third column). This is because the information conveyed by the samples were obscured by the information contained in the proxies. To differentiate between the two models that use proxies, we factored out the effect of the stimulus. We divided the trials into low and high certainty groups (median split) within each stimulus condition and compared the across-condition averages between these groups. If samples provide additional information about ρ_{S} beyond the proxies, certainty should still predict accuracy even after factoring out the stimulus effect, resulting in a slope greater than 0 for sample sizes greater than 1 (Fig. 2.3C-E, fourth column). Otherwise, if only the proxies are used, the slope would be zero regardless of presentation time (Fig. 2.3F, fourth column).

Prediction of the Noise models

As we saw in the foregoing section, analyzing the slopes of the certainty-accuracy lines provides a way to distinguish between the different variants of the Signal model. We can follow very similar arguments for the Noise model. There, the lack of irreducible uncertainty means that the quantity that predicts accuracy is the reducible uncertainty that a finite amount of evidence accumulation leaves behind. In order to correctly estimate this, the observer needs to know the precision of the sampling distribution (ρ_{S}) and from this point the arguments are essentially the same as for the Signal model. The main difference between the Signal and the Noise models' predictions (Fig. 2.4A and B) is that in the Noise model there is a more pronounced improvement in accuracy (and, consequently, in certainty) with the progression of time.

Incorporating attentional lapses into the model

When the same task is carried out by humans, it is conceivable that in some proportion of the trials, no samples are collected at all – or at least not from the target item's location. On these *lapse* trials both accuracy and certainty are minimal. To explore how such lapses affect the qualitative model predictions, we included lapse trials to the model. In the simulations, orientation estimates on lapse trials were randomly selected from the range $[0°, 180°]$, and certainty was set to zero. We modeled the probability of a lapse as being proportional to the product of presentation time and stimulus strength, indicating that more challenging trials are more likely to result in lapses.

Interestingly, the inclusion of lapse trials did not change most of our qualitative predictions (Fig. 2.4C and D). The only significant change was that the slope was greater than zero in the stimulus-marginalized grouping condition, even for the proxy-only model. This effect was due to the binary nature of behavior introduced by the lapses (lapse vs no-lapse). However, it can be easily tested whether the slopes are real by excluding the zero-certainty trials from the analysis and checking if the slopes remain positive.

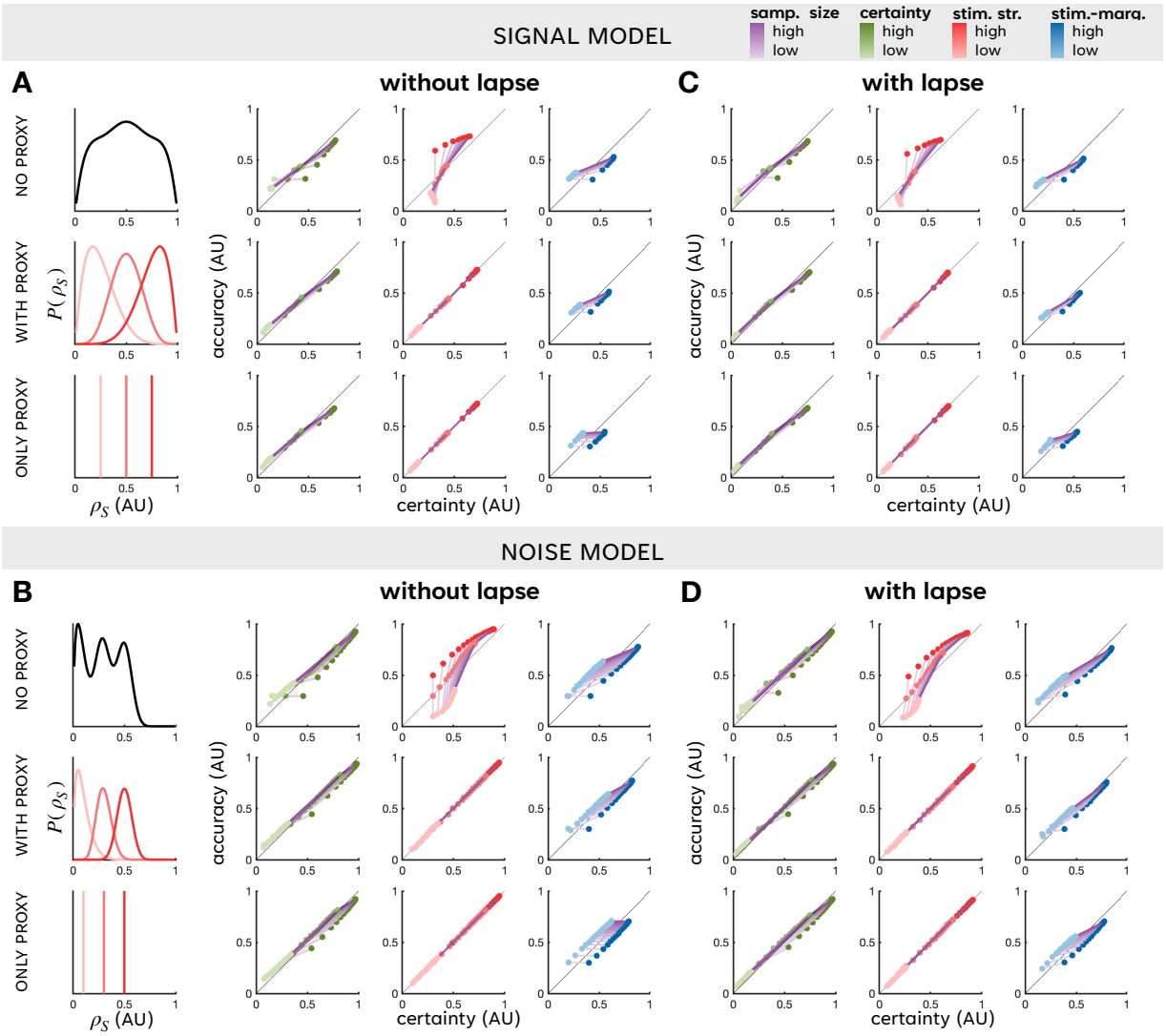


Figure 2.4. The comparison of different Sequential Sampling Model variants. A-B. Comparison of the Signal (**A**) and Noise (**B**) models' predictions when there are no lapse trials. Noise model predicts a more pronounced shift of the accuracy-certainty points towards the upper right corner (indicative of a the reducibility of uncertainty), compared to the Signal model. **C-D.** When incorporating lapse trials, the accuracy-certainty slopes in the stimulus-marginalized grouping condition are no longer zero.

2.3 Certainty-accuracy relationship in an orientation estimation task

2.3.1 Experimental paradigm

In order to asses the extent and algorithmic realization of human uncertainty representations, I developed a new experimental paradigm. The basis of my paradigm was a standard

orientation estimation task, implemented in two slightly different versions. I will first detail one version of the task (fig. 2.5A), and summarize the differences between the two versions at the end.

In each trial, subjects first saw a blank screen with a fixation dot. After maintaining fixation continuously for 1100 msec, which was verified by an eyetracker, a display appeared with a variable number ("set size": 1–6, randomly chosen) of 1-degree-long line segments equidistant (with a randomly chosen rotation) around a circle (extending 7° of visual angle in diameter). The line segments' contrast levels were sampled randomly (without replacement within a display) from the set $\{0, 0, 30, 40, 50, 60, 70, 80, 90, 100\%\}$ (zero contrast appeared twice as frequently as the other contrasts), and orientations uniformly from $0^\circ - 180^\circ$. The display appeared for one of nine possible durations ("presentation time"): 50, 75, 100, 133, 167, 200, 300, 400, or 600 msec. If participants broke fixation during stimulus presentation, the trial was omitted from the later analyses. After the display disappeared, a mask of random noise appeared with a small red circle identifying the position of one of the segments in the preceding display. The subject's task was to report as quickly as they could their estimate of the orientation of the segment in the cued position simultaneously together with the subjective assessment of uncertainty in their estimate by drawing a single line on a tablet with a stylus (fig. 2.5B-C). The orientation of the line indicated the estimated orientation of the line segment, while the length of the line corresponded to subjective uncertainty (a longer line indicated less certainty). After the subject responded, the mask and cue disappeared and a small segment appeared at the tested location with the orientation chosen by the subject and with a gray wedge around the line segment with a width (subtended angle) corresponding to the reported uncertainty. (The true orientation of the segment was not displayed.) This feedback display appeared for 500 msec, after which a new trial began.

The wedge provided a natural way to express circular uncertainty, representing a range within which the target line has a "high probability" of falling. To concertize "high probability" and to enhance the quality of subjective uncertainty estimation, a scoring function (Eq. 2.28) was used to assess subjects' performance on each trial. Subjects were instructed that their goal was to maximize their score which was calculated by combining the accuracy and certainty

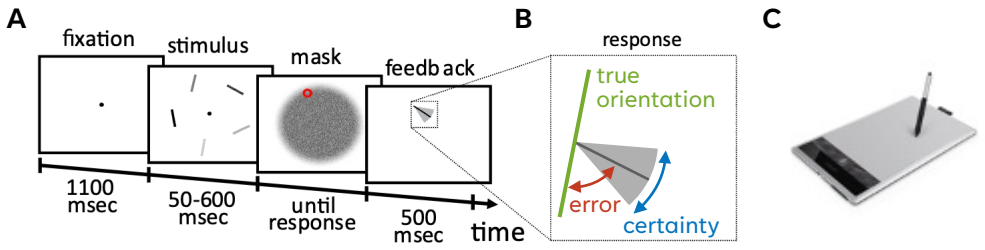


Figure 2.5. Experimental design. A. In each trial, the subject made an orientation estimation judgement and provided information about the orientation and their subjective uncertainty by drawing a single stroke on a tablet. See text for details. **B**. Estimation error between the true and reported orientations, and level of certainty were the dependent variables in the experiment. **C**. Wacom Bamboo tablets were used to record the subjects' responses.

of their response. As a scoring function, I used the log probability of the true stimulus orientation under a circular Gaussian (von Mises) distribution defined by the subject's response (segment orientation – mean, wedge width – precision). This scoring function can be maximised if the subject's uncertainty report reflects their true subjective uncertainty which in turn is predictive of the errors they are making (Jaynes, 1996). To prevent subjects from developing simple feedback-based strategies while keeping them alert, subjects received only grouped feedback after every 10 trials in the form of an average score. Subjects completed 3-4 sessions of 900 trials across multiple days. To familiarize themselves with the procedure and to facilitate the precision of mapping from uncertainty to line length, prior to each test session subjects had a practice session with 50 trials during which they received feedback after every trial including the true orientation of the cued segment. Practice trials were twice as long as the test trials and data from these trials was not included in the analyses.

In the other version of the task (not shown), Gabor patches were used as items instead of line segments, the set size options were limited to 3 or 6, contrast levels were sampled from $\{0, 5, 8, 14, 22, 37, 61, 100\%\}$ and presentation times were chosen from $\{33, 50, 83, 133, 200, 600\,\mathrm{msec}\}$. Unlike the other version, the Gabor patch variant included only a single practice session prior to the first test session, featuring a restrictive stimulus set to minimize the potential for learning an optimal heuristic strategy (presentation times: $\{83, 200, 600\}$; contrasts: $\{0, 14, 22, 37, 61, 100\%\}$). This time, I added another practice task that I designed to teach participants how to accurately produce wedges of varying sizes with a single stroke. Aside from these differences, the two experimental variants were identical.

2.4 Results

In the experiment using line segments, we collected data from a total of $N = 6$ subjects, five of whom were naive while the last one was informed about the goal of the experiment. We found no difference in performance between the naive and informed subjects confirming that the paradigm measured direct reactions of the subjects without much cognitive influence.

For the version of the experiment using Gabor patches, data were again collected from a total of $N = 6$ subjects, all naive about the purpose of the experiment. However, two of them were excluded from subsequent analysis based on post hoc considerations (see Methods). These exclusions were due to the virtual absence of certainty-error correlations in their data.

2.4.1 Basic measurements and controls

First, I checked whether in my paradigm I measured the relevant aspects of human performance. In order to measure the typical pattern of trial-by-trial error and uncertainty, the stimuli must cover the entire space of orientation, the subject's perception needs to follow the true stimuli, and response movements need to be ballistic. Fig. 2.6 confirms the uniform stimulus orientation distribution and the ballistic response movements, but especially for the Gabor patch version, responses were biased away from the vertical ($90°$) orientation (Fig. 2.6A), but this bias decreased as confidence increased (Fig. A.2). Nevertheless, on average, subjects' judgement closely followed the true orientation of the target line segment (Fig. 2.6B), therefore I neglected this repulsive effect in the later analyses. Moreover, their stroke was a straight line (Fig. 2.6C) with an average deviation from the straight line between the starting and endpoints below $3.6 \pm 1.6\%$ of the length of the stroke for both stimulus type. In addition, I calculated the time profile of the strokes and found that subjects' mean duration of drawing was 390 msec with a standard deviation of 270 msec for the line segment version, and 330 msec with a standard deviation of 220 msec for the line segment version. This suggests that subjects drew the line segments with a fast, single stroke without much fine-tuning, explicit cognitive deliberation, or modulation by different aspects of the task.

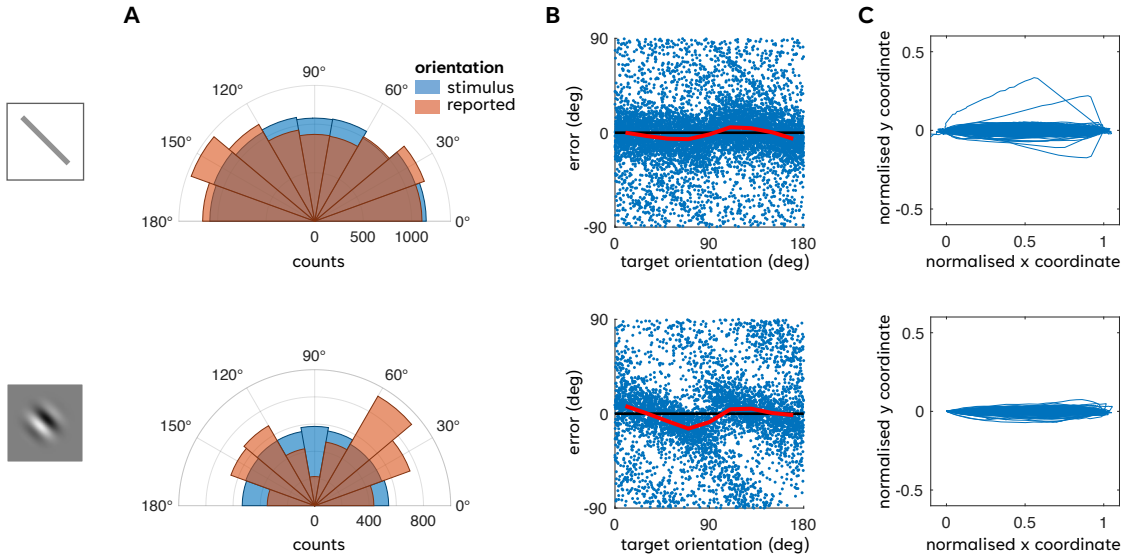


Figure 2.6. Control measures. Experiments with both stimuli (upper raw - line sigment, lower row - Gabor patch) give veridical trial-by-trial information about subjects' error and subjective uncertainty. **A**. The distributions of the test items' true (blue) and reported (orange) orientation (blue). **B**. Trial-by trial correspondence between the line segments' true orientation and the signed error of reported orientation across all subjects and trials. Red line indicates the circular average error as a function of target orientation. **C**. Trajectories of strokes for all subjects normalized (rotated and scaled) such that they go from (0,0) to (1,0).

2.4.2 The representation of error and uncertainty

Next, I tested whether subjects' uncertainty reports were predictive of their estimation errors. Fig. 2.7 shows the results for each subject from both experiments with trials binned by reported certainty and the resulting error histograms fitted with a circular Gaussian. Despite individual variations, each subject showed the same general relation of increasing certainty corresponding to steadily decreasing error in their performance. This suggests that subjects had a reliable representation of the quality of their perceptual information and faithfully reported this through their stroke. Thus, my experimental paradigm and response method successfully captured subjects' trial-by-trial error and certainty.

2.4.3 Stimulus-dependence of accuracy and certainty

Subsequently, I analysed the impact of the applied stimulus manipulations on performance. Specifically, I tested how they affect the certainty-accuracy relationship. For this analysis, it

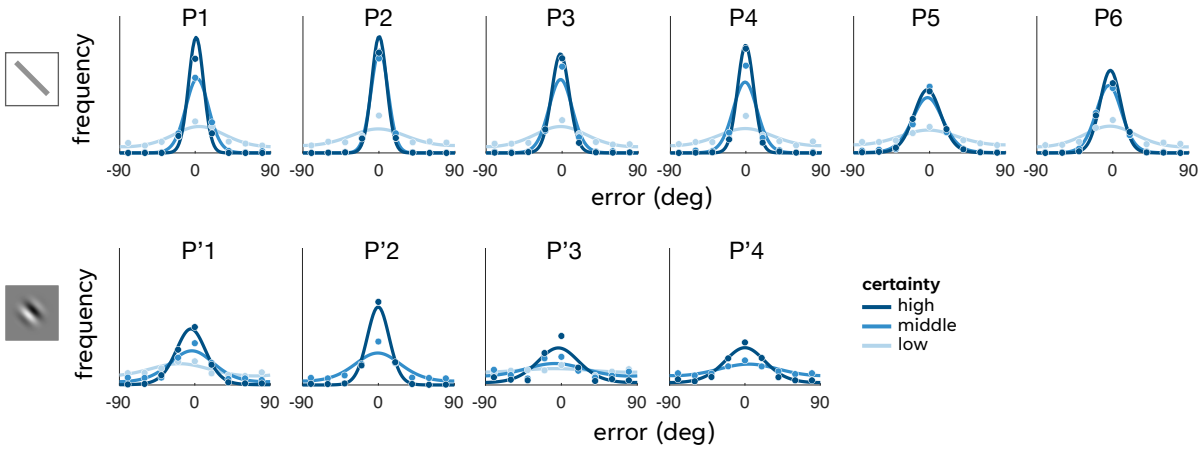


Figure 2.7. Relation between error and subjective uncertainty. Subjects' orientation estimation error changed according to their subjective certainty. **A**. Circular gaussian fit of orientation estimation error histograms (dots) corresponding to different levels of reported certainty for stimuli containing line segments. **B**. Same as A., but for stimuli containing Gabor patches.

was essential to correctly interpret the participants' certainty reports. However, despite the applied scoring, there is no guarantee that participants could perfectly learn the intended certainty scale. To minimize errors arising from discrepancies between the intended and actual scales, I monotonically rescaled the reported certainties (Eq. 2.28). I used the mapping that minimized the mismatch between the average accuracy and certainty of long presentation time trials for which even the no-proxy model anticipated well-calibration.

First, I tested the effect of presentation time (Fig. 2.8A), while averaging trials across contrast and set size conditions to marginalize out the influence of these stimulus manipulations. Regardless of the type of items used in the experiment (lines vs. Gabors), the reported certainty closely matched the accuracy at all presentation times. Initially, both measures improved monotonically with increasing presentation time – justifying the choice of SSMs – until they reached a saturation point. I approximated the time of saturation by identifying the smallest presentation time at which accuracy and certainty were no longer significantly lower than the time-averaged accuracy and certainty at longer presentation times. To factor out the effect of inter-subject variability from this analysis, I first z-scored the measures within each subject. In the experiment with line segments, the (approximate) saturation point for accuracy was at 133 msec (one-tailed paired t-test: $t(5) = -0.09$, $p = 0.46$; compared to 100 msec, $t(5) = -2.1$, $p = 0.04$) and for certainty was at 100 msec ($t(5) = -1.2, p = 0.14$; com-

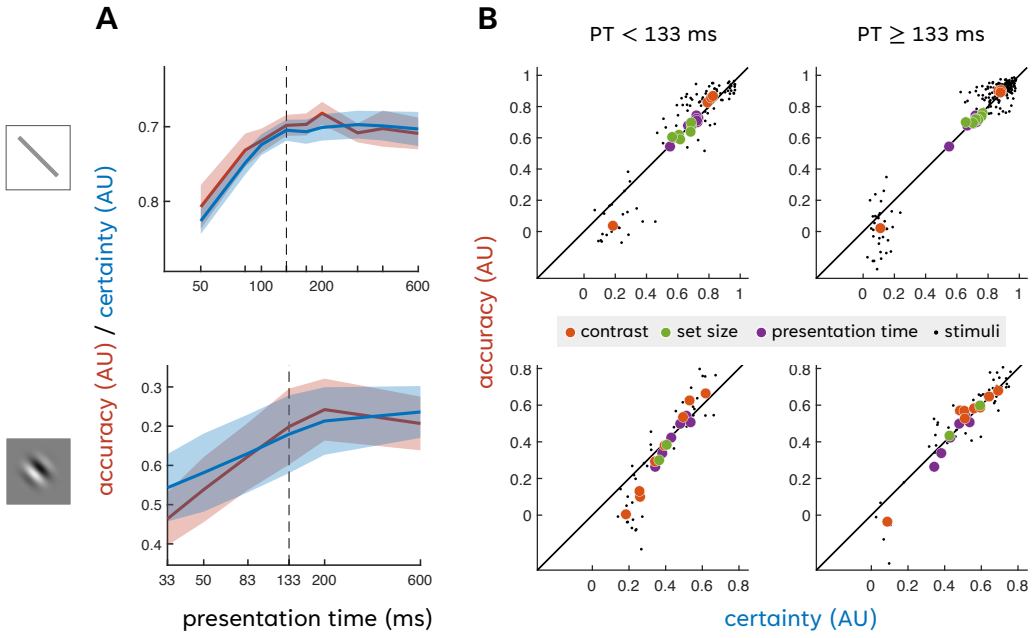


Figure 2.8. Stimulus dependency of accuracy and certainty. A. Certainty closely matches accuracy and both improves with the length of presentation time (log scale) until it reaches an asymptote at around the saturation-point (dashed black line). **B**. Average certainty from the repeated presentation of stimuli closely matches the accuracy measured on the corresponding stimuli (small black dots), with a better match observed after the saturation-point compared to before (right vs. left column, respectively). This relationship also holds true for the group averages when the grouping is based on the different stimulus manipulations (larger color dots).

pared to 83 msec: $t(5) = -6.6$, $p < 0.001$). In the Gabor experiment, the saturation point of accuracy was also at 133 msec ($t(3) = -0.89$, $p = 0.22$; compared to 83 msec: $t(3) = -3.27$, $p = 0.023$) and of certainty was at 83 msec ($t(3) = -2.3$, $p = 0.052$; compared to 50 msec: $t(3) = -10.9$, $p < 0.001$). Based on these results, I conclude that both accuracy and certainty reached their asymptote at around 133 msec (or maybe a little earlier), therefore I set the saturation point for 133 msec for the later analysis.

Next, I compared the different stimulus manipulations based on their effect on the certainty calibration. For this I first computed accuracy and certainty separately for each individual stimulus (Fig. 2.8B, black dots), where a stimulus was defined by a feature-combination triplet (presentation time, contrast, set size). I then evaluated the marginal influence of each feature type by grouping the stimuli according to the feature being analyzed and averaging accuracy and certainty across the other features. According to the models, if participants are nearly ideal observers, accuracy and certainty should align with the $45°$ line of well-calibration,

regardless of the grouping method, provided that either the feature in question was observed or the asymptotic region has reached. To test this, I analyzed the effects of contrast and set size separately for presentation times below and above the saturation point. I note here, that no such separation is needed for presentation time if, as assumed, it is observed.

The results shows that regardless of the type of stimuli and the the grouping method, the accuracy-certainty points are close to the 45° well-calibrated line even for short presentation times, and the closeness improves further after the saturation point (Fig. 2.8B). This result already suggests that certainty acts as a common currency across very different difficulty manipulations, but time is needed to fine-tune the calibration. In the next section, I provide a more detailed analysis of this time dependence.

2.4.4 Time-dependence of certainty calibration

To determine which SSM variant aligns best with behavior, I conducted tests as outlined in Section 2.2.5.

Fig. 2.9 illustrates how the calibration of certainty changes over time under the three different grouping conditions: certainty-based, stimulus-based, and stimulus-marginalized certainty-based grouping. In the behavioral experiments, unlike in the simulations, stimulus strength was influenced by the complex interaction of two stimulus features (contrast and set size). This interaction prevented me from establishing a priori the order of the stimuli based on their strength, and thus I could not apply the same colour scheme in the empirical plots that was used for the synthetic plots. Instead, I color-coded the certainty-accuracy points based on presentation time. To assess the participants' level of calibration, I calculated the best fit lines for each participant's certainty-accuracy points. This line minimized the total squared distance of the certainty-accuracy points from their orthogonal projections. I show here the average best-fitting lines across participants for the shortest presentation times below the saturation point, as well as for the asymptotic data pooled across presentation times at and above the saturation point (Fig. 2.9 A and C). In addition, I plot separately the slopes of the best fitting lines (Fig. 2.9 B and D).

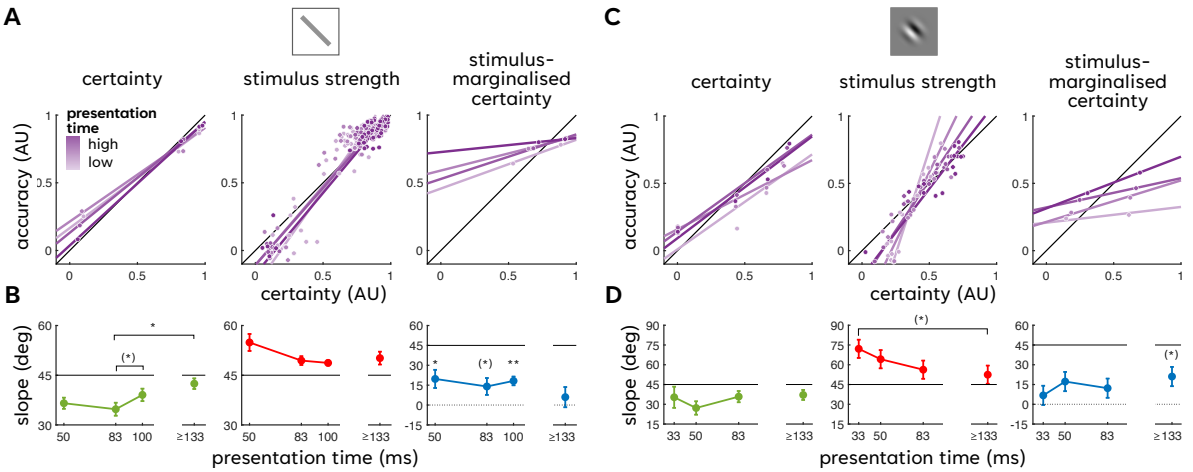


Figure 2.9. Calibration of certainty as a function of time. **A**. and **c**. Best fitting lines to the accuracy-certainty points computed using different grouping methods at various presentation times (purple shading). **B**. and **D** Across-subject averages and standard errors of the best fitting lines' slope. For the certainty- and stimulus-based grouping (green and red lines, respectively), stars indicate significance levels from post-hoc multiple-comparisons following repeated measures of ANOVA (**: 0.01; *: < 0.05; (*): < 0.01). For stimulus-marginalized certainty grouping, stars indicate the deviation of individual points from 0, measured by two-tailed t-tests (same significance notations as before).

The empirical results are in qualitative agreement with the predictions of the IEA models that may noisily observe the proxies or may not observe them at all. There is a general tendency for the slope to increase over time when data is grouped based on the explicit certainty reports (Fig. 2.9 B and D, green lines) and to decrease over time when grouped based on the stimulus features (Fig. 2.9 B and D, red lines), gradually approaching the well-calibrated line in both cases. I tested the significance of these trends using repeated measures ANOVA, and found significance only in a subset of cases. Specifically, for the certainty-based grouping, the slope increase was significant only for the line stimuli ($F(3,15) = 5.21$, $p = 0.012$), while for the same stimuli the decrease for stimulus-based grouping was just not significant ($F(3,15) = 3.22$, $p = 0.053$). For the Gabor stimuli, the stimulus-based grouping showed a significant decrease, with a strong level of significance ($F(3,9) = 13.84$, $p = 0.001$). The lack of significance might be due to the low number of participants. The slope's relative proximity to the 45^0 line, even at very short presentation times, suggests that participants either have access to multiple samples by that time or they are relying on proxies that they perceive with some level of noise.

Furthermore, when I divided the data into high- and low-certainty groups after accounting for the effect of the stimulus, I consistently obtained fits with positive slopes (Fig. 2.9 B and D, blue lines). When pooling the data across all presentation times (not shown), accuracy in the high-certainty group (lines: M = 0.92, SD = 0.02; Gabors: M = 0.67, SD = 0.24) was significantly higher than in the low-certainty group (lines: M = 0.7, SD = 0.08; Gabors: M = 0.25, SD = 0.14), as measured by a one-tailed paired t-test (lines: t(5) = 7.67, $p < 0.001$; Gabors: t(3) = 7.96, $p = 0.002$). Even after excluding trials with 0 certainty and 0 contrast (where 0 certainty is justified), the difference between the high-certainty (lines: M = 0.96, SD = 0.01; Gabors: M = 0.77, SD = 0.23) and low-certainty groups (lines: M = 0.86, SD = 0.04; Gabors: M = 0.57, SD = 0.33) remained significant (lines: $t(5) = 8.32$, $p < 0.001$; Gabors: t(3) = 3.96, $p = 0.026$). These results indicate that participants were sensitive to fluctuations in the quality of their perceptual representations beyond what could be explained by visible stimulus features or the observability of the item.

2.4.5 Presentation time-dependence of accuracy, certainty and reaction time

Up until this point, I have examined the effects of different difficulty manipulations collectively. Now, to reveal their individual impacts on accuracy and certainty, I plot these quantities as functions of presentation time, with the data grouped by either contrast or set size (Fig. 2.10). Additionally, I analyze reaction times in a similar fashion to see how the experimentally identified decision process (IEA with partially observed proxies) is influenced by the proxies.

Overall, both accuracy and certainty increase, while reaction time decreases with increasing presentation time and stimulus strength (higher contrast and smaller set size). However, there are two notable exceptions: First, for zero-contrast stimuli, certainty actually decreases with longer presentation times. Second, in the experiment using Gabor stimuli, after an initial decrease, reaction time starts to increase again as presentation time lengthens.

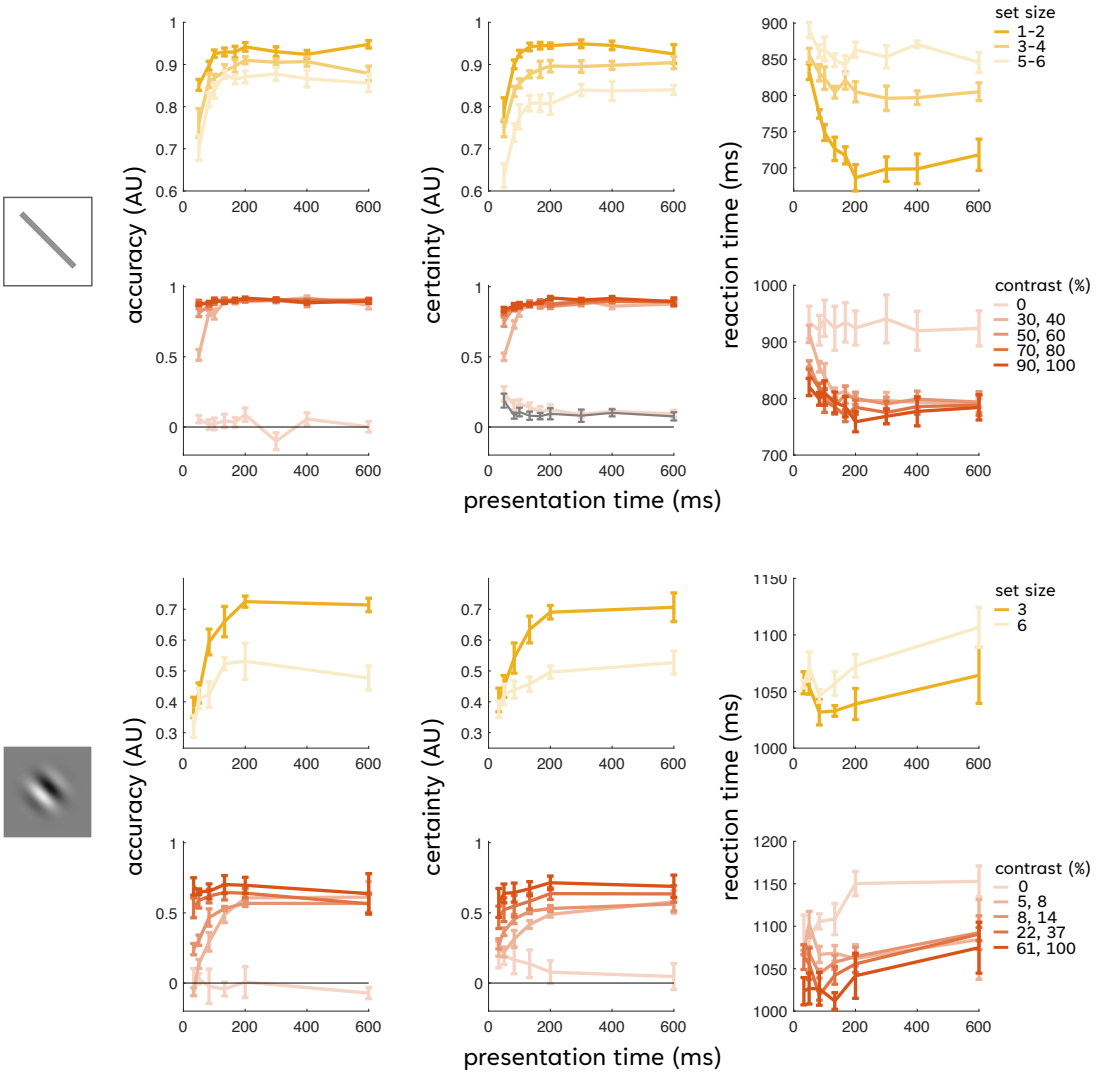


Figure 2.10. Presentation time dependence of accuracy, certainty and reaction time The figure shows the average accuracy, certainty, and reaction time across participants as a function of presentation time, along with the standard errors. Data is grouped either by set size (yellow) or by contrast (red), with 0 contrast data omitted in the set size grouping. The gray line corresponds to the certainty on empty stimuli.

The decrease in certainty of the zero-contrast trials with longer presentation times can can be naturally explained with SSMs. In this framework, accurate certainty estimation takes time. At short presentation times (or small sample sizes), the ρ_S prior dominates certainty estimation. This is a regression-to-prior-mean effect, the strength of which is modulated by the observedness of stimulus strength. As presentation time increases (more samples are collected), this regression effect is getting weaker. In the case of zero-contrast trials, there is increasing evidence that the sampling distribution is broad (in this case, samples are from the uniform distribution), which consequently reduces the IEA's certainty. Importantly, this

argument is only valid if we assume that samples are taken even from those locations where the item's contrast is zero (or nothing has been displayed). Therefore, if this argument is true, it has the important implication that the representation of local features is automatic in the brain, which fits well with the fully Bayesian view. One might consider an alternative explanation, that this effect is due to swap-errors, i.e. the incorrect reporting of a non-target item as in Bays (2016). If the probability of these errors were time dependent, this would seem to explain the error in question. However, since the effect is present even when the stimulus consists of a single zero-contrast line (effectively being an empty stimulus), and thus there is nothing to swap, the swap error-based explanation of the effect can be rejected.

The stimulus dependence of reaction times suggests that not only perception but also decision making requires time, despite that all sensory evidence have been presented before the onset of decision making process. One possible explanation of this is that the (sampling-based) perceptual representation is only accessible to the decision mechanism through a secondary sampling process, which again takes time. Alternatively, it is possible that it is the process of perceptual sampling that continues after the stimulus offset. In either case, the termination of sampling process(es) are likely governed by a subjective cost-benefit trade-off – balancing the accuracy gained from collecting more samples against the cost of both the sampling and the time investment (opportunity cost) (Drugowitsch et al., 2012). An interesting avenue for future research is to distinguish between these alternatives, with a particular focus on the somewhat unexpected asymptotic increase in reaction times, that seemingly contradicts the predictions of the standard normative models (Drugowitsch et al., 2012).

Finally, there is another interesting pattern in my data. While the set size-driven differences in accuracy and certainty persist even asymptotically, differences driven by contrast diminish over time (except for the zero-contrast condition), at least for the line stimuli. These asymptotic behavioral differences might imply that the two manipulations introduce different types of uncertainty. Set size appears to induce irreducible uncertainty, meaning that representing more items comes at the cost of sacrificing the precision of the individual item's representation, while contrast induces reducible uncertainty, likely related to the signal-to-noise ratio of the representation or maybe to the delayed onset of sampling. Crucially, however, explicit

reports seem to reliably track both types of uncertainty manipulations. Constructing an SSM that can accurately track its own uncertainty in the presence of both reducible and irreducible uncertainties presents an intriguing challenge for future research.

2.5 Discussion

In this study, utilizing a novel perceptual estimation paradigm with multi-item stimuli, I demonstrated that the human brain relies on probabilistic perceptual representations that simultaneously encode variable-specific uncertainty information about multiple internal variables (items). This was an important step in the progress of my Thesis, as the capacity for variable-specific uncertainty representation is a prerequisite for fully Bayesian representations.

In addition to the main result, my work provides new ways to investigate the nature of uncertainty representations. For example, by experimentally separating perceptual processing from decision making and by manipulating the stimulus presentation time, I could gain new insights to the temporality of perceptual decision making and could show that both the formation of perceptual representation and the process of decision making requires time. Previous approaches either ignored the time aspect of perceptual decision making altogether (Maloney and Mamassian, 2009) or did not distinguish the two subprocesses (Kiani et al., 2008; Kiani and Shadlen, 2009; Kiani et al., 2014). My purely behavioural approach also complements research on perceptual decision making that also seeks to dissociate the two processes by focusing primarily on neural measures (Wyart et al., 2012; Mostert et al., 2015).

I was also able "to look under the hood", that is, to gain information about the sources upon which explicit uncertainty reports rely – whether they are based on genuinely probabilistic representations or on "proxy" variables that directly affect the quality of these representations (Barthelmé and Mamassian, 2010; De Gardelle and Mamassian, 2014; Meyniel et al., 2015b; de Gardelle et al., 2016; Adler and Ma, 2018). By developing the first formalization of the concept of a "proxy" in the sequential sampling context, I could investigate how they

could enter to the perceptual decision making at different stages of the process. While handling proxies at the cognitive level may seem intractable under realistic conditions – given the numerous factors influencing uncertainty at the same time – I could demonstrate that using proxies as informed priors for perceptual inference, particularly in the low-sample limit, may still be a viable strategy for improving the quality of inference. By showing that the calibration of certainty reports were notably high already at the shortest presentation times, I also identified the traces of such behavior in my experimental data.

My task design also comes with certain limitations. For instance, latent variables can only be measured one at a time, which restricts the approach to examining marginal distributions rather than joint distributions. To overcome this, one would need to use stimuli with correlated item features and potentially complex utility functions that take more than one item as input. When complex utility functions are used, the ability to measure the (joint) perceptual probabilities through explicit reports would likely be lost, as expressing complex multivariate functions with simple experimental methods (like making a stroke) is challenging. However, model-based approaches (Denison et al., 2018; Yoo et al., 2021) could still be used to reconstruct the multivariate perceptual distribution underlying behavioral judgments.

A further limitation of my method is that, although the decision variable is not known during the accumulation of sensory evidence, the possible options are limited to a narrow set (1-6 item directions). As a result, uncertainty might only be represented for this limited set of variables. While this strategy is sufficient for efficiently solving the experimental task, it falls short in real-world scenarios, where complex scenes, involving many latent variables, must be processed, and the future decision situations are often unpredictable. To better approximate these real-world conditions, one could increase the "richness" of latent variables (both in number and type) that constitute the set of potential decision variables, making the identity of upcoming decision variable less predictable. In our experimental paradigm, I could simulate this scenario by incorporating contrast and set size estimation tasks in addition to orientation estimation, and by randomly alternating between these tasks across trials. However, modifying the task to this extent – where participants must first figure out the type

of task to answer before preparing the response – risks disrupting participants' spontaneity (i.e., answering without deliberation) and may lead them to adopt other cognitive strategies.

Finally, although our goal was to assess the quality of perceptual representation rather than that of working memory, it is questionable how well this can be achieved in a paradigm where decisions are based on prior observations, even if those observations are made just before the decision. This concern is directly tied to the issue of how distinct perceptual and working memory representations are (Bays et al., 2024). Specifically, whether working memory simply maintains the perceptual representation (even if the code is dynamic) or rather transcribes it into a qualitatively different form. Especially in the latter case, a critical question is how much of the uncertainty measured in our task is attributable to perceptual uncertainty versus to the process of storing information in working memory. The influence of working memory might dominate in two scenarios. First, when the limitations of working memory are greater than the limitations of perceptual representation. Second, when the perceptual representation does not encode uncertainty at all (or at least if working memory is insensitive to this uncertainty). Under both of these conditions, working memory becomes sensitive mostly to its own uncertainty limitations and it contaminates the uncertainty measured under the label of "perceptual uncertainty". To minimize this confound, we omitted the gap between stimulus offset and target cue onset and introduced a noise mask immediately after stimulus offset, which has been proposed to curtail ongoing activity related to the previous perceptual stimulus (Tomić and Bays, 2024). However, to ensure we capture the momentary perceptual experience, we could no longer hide the decision variable during perception. In this case, the representation of other variables than the decision variable could only be investigated indirectly through the advantages of probabilistic representations enlisted in Chapter 1.

While the current experimental paradigm has its limitations, it also offers several promising avenues for further exploration. For example, it offers the possibility to distinguish Noise and Signal models – a topic I've begun to explore, though without reaching a definitive conclusion so far. The two models differ in how they explain sensory variability: one attributes it to the presence of noise, while the other attributes it to the nature of probabilistic code. Importantly, in both cases, downstream computations have the potential to explicitly evaluate the uncer-

tainty arising from the two processes, provided they implement ideal evidence accumulators. My initial results in this exploration, based on the line stimulus experiment, indicated that the Signal model provides a better qualitative and quantitative fit to the data than both the vanilla Noise model and its extended version (Koblinger et al., 2019, COSYNE, conference). In the extended Noise model, the sampling distribution can be biased, but unlike in the Signal model, this bias is independent of variability. However, later I incorporated the possibility of early termination of the sampling process into the models, a phenomenon that has been documented in animal perceptual decision making studies (Kiani et al., 2008), and the qualitative differences between the two models types diminished, and their quantitative distinctions become less pronounced. Whether the two models can still be distinguished quantitatively with more refined experiments (such as the one using Gabor stimuli) and improved data analysis remains an open question for future research.

Another potential avenue for future research stems from the fact that our paradigm inherently separates perceptual and decision-making processes, allowing them to be studied independently. Just as analyzing presentation time provided insights into the perceptual process, examining reaction times can offer valuable information about the decision making process. Staying with the SSM models, and building on normative assumptions on when to terminate the sampling process and commit to a decision (Drugowitsch et al., 2012) we may be able to answer what kind of samples – i.e. probabilistic samples (Fischer and Whitney, 2014), noisy memory samples (Shushruth et al., 2022), or ongoing sensory samples that persists after stimulus offset (Tomić and Bays, 2024) – take time to process during the decision process.

In conclusion, this study advances our understanding of the scope and temporal-nature of probabilistic perceptual representations, and it lays the groundwork for future research that seeks to clarify the content of momentary perceptual activity (probabilistic vs noisy information) and the complexities of the decision making processes. Despite the simplicity of the paradigm used, it demonstrates considerable potential for further exploration into the internal representations of a probabilistic brain.

2.6 Methods

2.6.1 Inclusion Criteria

To determine which subjects to include in the analysis, the raw certainty reports (ρ_c) were linearly transformed so as to maximize the log-likelihood of signed errors (e) under the transformed certainty reports:

$$\rho_{\mathrm{tf}} = b_0 + b_1\,\rho_c \tag{2.22}$$

such that

$$\{b_0, b_1\} = \underset{a_0,a_1}{\operatorname{argmax}} \prod_{n=1}^{N} \mathrm{vM}(e_n; 0, \rho_{\mathrm{tf}}(a_0, a_1)) \tag{2.23}$$

Only subjects for whom b_1 was significantly positive were included in the further analysis.

2.6.2 Scoring Function

The scoring function was the log probability of the true stimulus orientation ($x^\star$) under a circular Gaussian (von Mises) distribution defined by the subject's response (segment orientation – mean (μ), wedge width (w) – concentration (κ)):

$$\mathrm{score} = \kappa(w)\,\cos(x^\star - \mu) - \ln I_0\left(\kappa(w)\right) \tag{2.24}$$

where $w/2$ was defined as the distance at which the distribution decreases to half its peak. This gives the following equation for the concentration:

$$\kappa = \frac{\ln 2}{1 - \cos\frac{w}{2}} \tag{2.25}$$

We applied a correction to κ to ensure that the maximal w expresses maximal uncertainty ($\kappa = 0$):

$$\kappa' = \kappa - \frac{\ln 2}{2} \tag{2.26}$$

From the concentration, certainty (circular precision, ρ) can readily be computed:

$$\rho = \frac{I_1(\kappa')}{I_0(\kappa')} \tag{2.27}$$

2.6.3 Rescaling Certainty Reports

The scoring scheme was designed to teach subjects the ideal mapping (which maximizes the obtained cumulative score) between wedge width (w) and the predictive posterior's concentration (κ, related to the precision of internal representation). However, it is not guaranteed that subjects accurately learned this mapping during the short training session. To account for this possibility, we estimated the actual w to κ mapping that was used by the subjects through function fitting. We still assumed that w is related to the width of the posterior, but allowed for the possibility that at a distance of $w/2$, the posterior distribution decreases not necessarily to half its peak value, but to another proportion, $1/C$:

$$\kappa' = \frac{\ln C}{1 - \cos\frac{w}{2}} - \frac{\ln C}{2} \tag{2.28}$$

with C being the free parameter.

This type of correction only makes sense if the subjects are well-calibrated. Therefore, we fitted the model to data grouped by stimulus, since this method is the least sensitive to noise and we only analyzed trials with presentation times above 200 msec, as performance has already saturated at this point, and according to the IEA models, the regression-to-prior-mean effect is minimal.

2.6.4 Simulating the Ideal Evidence Accumulator

For each model variant, I used three stimulus strengths and simulated a total of 30,000 trials (10,000 per stimulus strength condition). In each trial, I simulated 10 samples for the Signal model ($\tilde{x}_{1:10}$) and 20 samples for the Noise model ($\tilde{x}_{1:20}$). I computed the model responses for each cumulative subsets of the 10 or 20 samples ($\tilde{x}_1$, $\tilde{x}_{1:2}$,...). The derivations of the relevant computations are in Chapter A.

2.6.5 Stimulus-marginalized certainty

To evaluate how well certainty predicts accuracy while controlling for the effect of stimulus strength, I grouped trials by stimulus. For each stimulus, I performed a median split on the trials based on the reported certainty, dividing them into a low-certainty and a high-certainty group. I then calculated the accuracy and average certainty within each of these groups. This process was repeated for every stimulus. Finally, I averaged the accuracy and certainty across all stimuli.

Chapter 3

On the complexity of internal models

We[1] investigated the rules of human perceptual decision making in complex dynamic situations when changes in the external conditions could be explained by multiple, equally feasible adjustments of the internal model rather than by one possible interpretation. Using hierarchical Bayesian modeling and a novel behavioral paradigm, we identified through response biases the internal representations observers used during their decision and found that in such situations observers' interpretation is strongly modulated by the specific dynamics of the input sequence. We show that this behavior could be captured by assuming that observers rely on representations with detailed dynamics of each parameter of their internal model and use this information to readjust their model to properly account for changes in the input sequence. These results are compatible with a fully Bayesian view of perceptual decision making, in which uncertainty at various levels of the external input is optimally accounted for.

[1]This chapter is being prepared for publication. It therefore uses plural pronouns.

3.1 Introduction

Making decisions is one of the most fundamental cognitive act performed by humans and animals that includes just about every type of behavior (Mellers et al., 1998; Kahneman, 2013; Newell et al., 2022). Importantly, everyday decision making always occurs in context and this context changes perpetually. These changes could be continuously evolving small alterations (e.g. feeling an increasing discomfort), (Newell and Shanks, 2014) or occasionally occurring major modifications through distinct events (e.g. receiving the news about the crash of the stock market), (Resulaj et al., 2009). Experimental investigations of decision making imitate this evolving context by repetitive sequence of trials (Goldstone, 1998; Heilbron and Meyniel, 2019; Lengyel and Fiser, 2019; Lee et al., 2020) and the effect of event-based changes by volatility-based adaptation studies (Nassar et al., 2010; Gallistel et al., 2014). These simplified studies identified sequential adaptation effects influencing the current decision due to the stimuli and decisions of the preceding trials (Fischer and Whitney, 2014) and volatility-specific adaptation of the single learning parameter of the decision process (Behrens et al., 2007; Glaze et al., 2015; Piray and Daw, 2020). However, these studies do not explore the feasible situation when the decision process can be influenced by multiple adaptable parameters and the problem is not only how much "the" parameter needs to be changed but also "which" parameter(s) need to be adjusted.

To address this gap, in the present study, we explored how well a multi-parameter model can capture human behavior in a classical sequential 2-AFC decision making paradigm when the context and its change during decision making are more reminiscent to natural conditions. We designed our experiment so that multiple equally adequate but contradicting interpretations of the circumstances could generate different "contexts" that the decision could be based on during the trials. We also varied the dynamics of the changes across contexts on a wide range to explore the effect of these different types of changes on decisions.

Based on our experiments, we obtained three fundamental results. First, we show that under these conditions, human decision making behavior cannot be explained by the presently available decision making models as they are fundamentally controlled by other aspects than

the known serial effects and overall volatility of the input. Second, through detailed computational analysis and modeling, we show that explaining our results requires a decision making process that automatically maintains a sophisticated internal model of multiple aspects of the complex setup with corresponding parameters and with their unique dynamics well beyond merely capturing a general volatility measure of the observed situation. Third, we generate a number of counter-intuitive predictions based on our model and show by testing that human behavior confirm them.

Our findings clarify the scope of earlier decision making results based on simple experimental setups and suggest that the general mechanism of human sequential decision making involves an implicit and automatic inferential process that uses a complex internal representation of the current situation and a probabilistic explaining-away-based mechanism to select between multiple plausible interpretations of the current observation based on all parameters of the internal representation.

3.2 Unexpected pattern of human decision making results after detecting a change in context

To test human decision making behavior in changing context, first we ran two different variants of the classical sequential 2AFC categorization task. The general design of these two and all the other experiments in the present study followed the simple structure of a 2AFC decision task, in which participants decide which of the two possible shapes is hidden under the variable amount of Gaussian noise on a given trial (Figure 1A). In all experiments, the following three general features of the setup were fixed across trials except for changing at one or two change-points (CP) during the entire course of the experiment: 1) the temporal noise structure (how the noise level of the next stimulus was selected), 2) the appearance probability (AP, in what fraction of the trials the more frequent shape appeared), and 3) whether there was a feedback or not at the end of the trial (Fig. 3.1B). The changes occurring at the CP were either abrupt from one trial to the next one or gradually introduced in the course of 80 trials called the transition period (TP). Importantly, feedback was provided to the

participants about the correctness of their choice only up to the change point, and they were not informed about the presence and nature of changes in the features to avoid influencing them how to set up and update their internal model.

In Experiments 1-2, we varied the experimental parameters abruptly vs. gradually at a single CP located after the 200th trial in the 500-long sequence to test whether these variations altered the participants' decision behavior (Fig. 3.1C inset). The first 200 trials (Training) used identical parameter settings in both experiments. In particular, a) the Gaussian noise added to the shape image on each trial followed an adaptive staircase to identify the noise level at which the participants' performance was around chance level, b) the two shapes appeared equally often across trials (AP =50%), and c) the participants received feedback about the correctness of their answers to bring their assumptions about the task statistics closer to a common baseline. In both experiments, all the parameters changed at the single CP after the 200th trial. Specifically, feedback after each trial was not provided any more and the noise level of each trial was selected randomly from a uniform distribution between the Min ("No noise") and the Max value identified by the adaptive staircase during the training. Most importantly, APs also changed to the same new level (AP =65%) either instantaneously (Exp 1) or gradually in an 80-trial-long TP (Exp 2) so that one of the shapes became more frequent across the trials.

By the 80th trial after the CP, all changes were completed and the last 220 trials (Test) the conditions across the two experiment were again identical. Response biases based on these last 220 trials of the Test were assessed to identify the internal models the participants used in response to the introduced changes. For both of these two and all subsequent experiments, two biases were computed for the first and second half of the Test trials, respectively, to quantify the persistence of biases. The biases were the offset parameters of the best-fit psychometric curves to the data. Note that participants performed exactly the same task in all the trials during the Training, the TP and the Test periods of all the experiments, only the underlying structure of the task changed: Training and TP set up the context and the Test section provided the measure of human decision making.

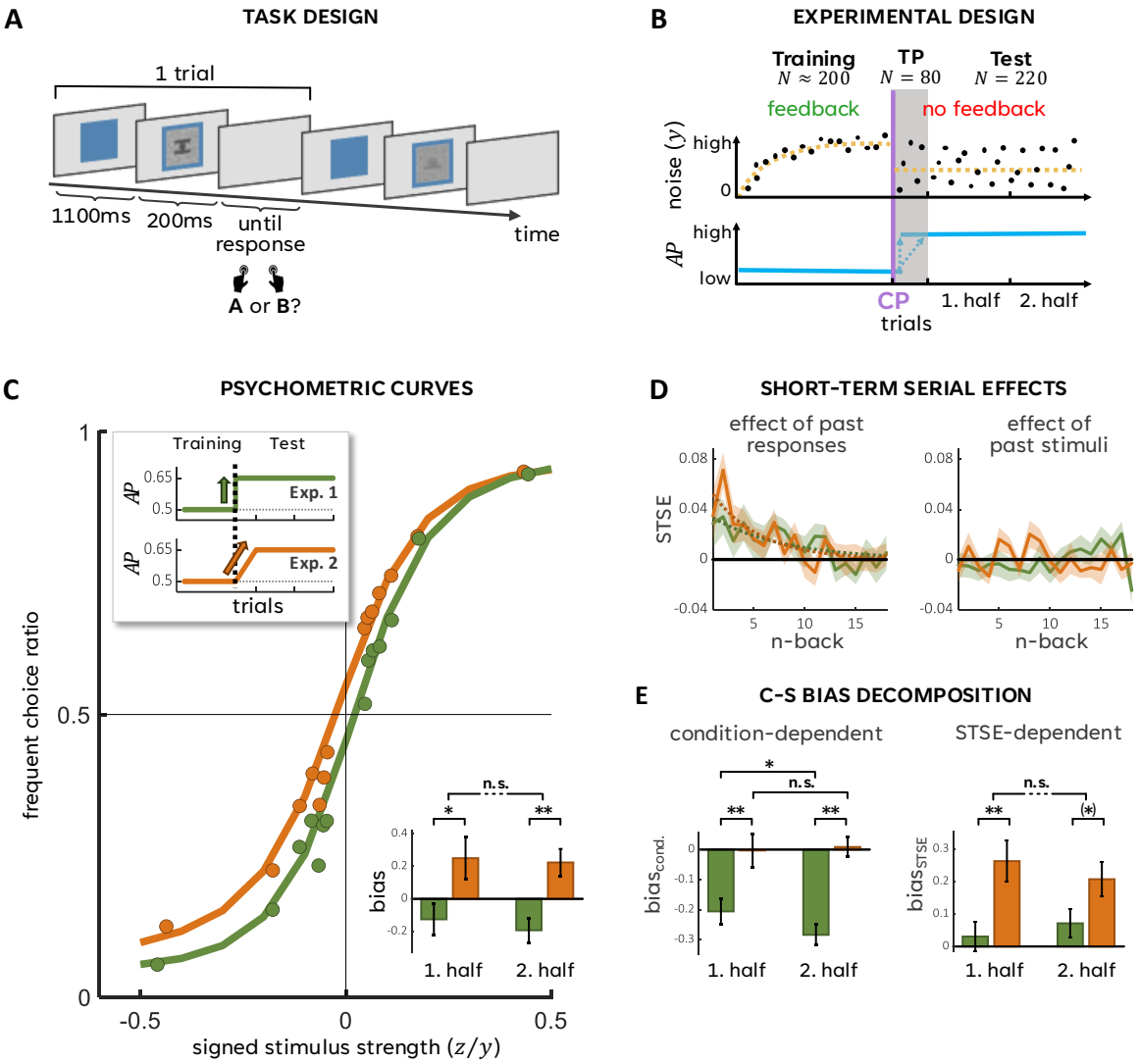


Figure 3.1. Diverging change induced behaviour. A. Temporal trial structure. **B** Three task features change at the CP, separating the Training and Test phases: 1) Feedback is restricted to the Training phase. 2) Training employs a 2up-2down adaptive staircase method to determine the maximal value of random uniform noise in the Test phase (black dots, with a temporal average represented by the yellow dotted line). 3) The AP of shapes (blue line) changes either abruptly or gradually over 80 trials (dotted blue arrows). **C**. The measured ratio of frequent choices (dots) in relation to stimulus strength (inverse variance of Gaussian pixel noise), with best-fitting sigmoid curves (lines) for the two experimental conditions (orange and green) shown in the inset. The two psychometric curves are shifted in the opposite direction and this bias is persistent throughout the experiment (bar plot). **D**. The impact of stimulus and decision from n trials ago on the current decision (dotted lines - best exponential fits). **E**. The biases in figure C are the sum of an experimental conditioned-dependent bias term and the average short-term serial effects of past decisions.

We conducted Experiments 1 and 2 with two separate groups of participants ($N_1 = 18$, $N_2 = 19$) and obtained two counter-intuitive results compared to common sense expectations. First,

according to conventional accounts of decision making in changing environments (Behrens et al., 2007; Nassar et al., 2010), the decision making behavior in these two experiments including trials with high noise level, should be identical from about 100 trials on in the Test session after AP stabilizes at 65%. Second, observers in both experiments should choose, especially in trials with high noise, the shape that has appeared 65% of the time in the preceding dozens and hundreds of trials, reflecting the accumulated prior knowledge about the shape APs. We found that the observers' behavior followed a completely different pattern. Specifically, the responses in Exp 1 showed the polar opposite bias from that in Exp 2. Observers in Exp 2 followed the expected pattern of choosing the overall more frequent shape ($M = 0.24 \pm 0.1$, $t_{18} = 2.518$, $p = 0.021$, $\mathrm{Cohen's}\, d = 0.021$), whereas observers in Exp 1 preferred to chose the shape that they saw half as often in the preceding trials ($M = -0.18 \pm 0.07$, $t_{17} = -2.49$, $p = 0.023$, $\mathrm{Cohen's}\, d = 0.023$). The difference between the biases of Exp1 and Exp2 was significant in both the first ($t_{35} = -2.59$, $p = 0.014$, $\mathrm{Cohen's}\, d = 0.83$) and in the second halves ($t_{35} = -4$, $p < 0.001$, $\mathrm{Cohen's}\, d = 1.29$) of the Test period (Fig. 3.1C). In addition, we found no significant difference between the biases measured in the first and second half of the same experiments (Exp1: $t_{17} = 0.946$, $p = 0.357$, $\mathrm{Cohen's}\, d = 0.201$; Exp2: $t_{18} = 0.202$, $p = 0.842$, $\mathrm{Cohen's}\, d = 0.052$), indicating that much of the internal model adjustment responsible for the biases occurred at the CP, without major revisions later. In sum, neither the similarity of the results in Exps 1 and 2 nor the preference for the more likely shape in Exp 1 was confirmed and this pattern remained the same across the 220 trials of the Test session.

One set of the mechanisms potentially capable to explain multiple-trial-based higher-order phenomena in sequential perceptual decision making has been intensively explored under the label of Short-Term Serial Effects (STSE). STSE incorporates a combination of negative biases due to short-term stimulus adaptation and positive biases due to short-term attractive serial dependence induced by past stimuli and decisions and other post-perceptual processes (Fischer and Whitney, 2014; Fritsche et al., 2017). However all these processes reportedly act based on the past 5-10 trials at a time scale of 10-15 sec making them an unlikely candidate for explaining our results. Nevertheless, to assess the contribution of STSE in our paradigm, we calculated the STSE bias, defined as the change in the probability of giving a 'frequent'

response depending on whether the stimulus or response n^{th} trial back was 'frequent' rather then 'rare' (Methods). We found that prior decisions, but not stimuli, had a strong influence on current decision (Fig. 3.1D) and this effect diminishes roughly exponentially over time (dotted lines) with the time scale compatible with previous reports (Fritsche et al., 2017). More importantly, these STSE were highly similar across these two (Fig. 3.1D) and all following experiments in this study (Supplementary, Fig. B.1) and thus, they could not explain our results of diverging and long-lasting biases. Since in this study we focus on the long-term biases, in the rest of this paper, we always remove the biases of STSE and present only the purely long-term ones (Fig. 3.1E). We note that due to interactions between long- and short-term effects, the removal of the biases due to STSE does not amount to the same simple subtraction across all the experiments. Instead, we included an independent predictor in the psychometric function to account for the STSEs, effectively factorizing out the STSE from the bias term (Section 3.7.3).

3.3 Computational analysis of complex human decision making

To find an explanation to our puzzling results, we start this section with forming a hypothesis about human decision making. Next we formalize our experimental paradigm in a probabilistic model based on our hypothesis and conduct a thorough computational analysis of the model. Finally, we evaluate how the model explains our results and derive a number of concrete and testable hypotheses for our subsequent empirical investigation which we will test in the subsequent section.

3.3.1 Human decision making is based on choosing between competing dynamic interpretations

Our starting point is the earlier assertion that capturing the real complexity of the world necessitates a complex internal model endowed with multiple parameters (Koblinger et al., 2021). The main hypothesis of the present study is that the fundamental challenge that

humans resolve during decision making in a complex dynamic environment after detecting an alteration in the surrounding context is interpreting the detected alteration as a change in one rather than some other subset of parameters of their complex internal model. Importantly, the alternative interpretations could – in various combination – describe the input equally well but promote very different decisions. To illustrate with an example, if someone is driving on a road full of potholes in foggy weather and she sees no more potholes for a while, she could reason that either the road ahead has already been resurfaced or that the fog has increased. Both options are perfect descriptions of the sensory evidence, but they support very different decisions to make about what to do next when a darker spot suddenly appears on the road: drive around the spot if it is interpreted as a pothole or go through it, if the road is assumed to be fine and the fog thinner at that spot. We posit that the largest fraction of human decision making cases under natural conditions represent samples of such a situation.

In such complex cases, the changes in the characteristics of the observed stimuli do not reveal univocally the changes of the latent model parameters and inferring the current state of the parameters becomes impossible without additional knowledge, for example knowing the dynamic properties of the parameters. In our example, an easy interpretation of the driving situation becomes impossible based exclusively on the apparent frequency of spots on the road since this frequency is not directly related to either the rules of construction or the accumulation of fog. On the other hand, if the observer has access to the dynamics of the two alternative explanations and those are different (e. g. when a road gets repaired, it tends to be done for a large segment, whereas fog accumulation on the road seems to vary quickly within some dozen meters), this difference can be used for making the right decision. In principle, such a solution would require our perceptual system monitoring and storing information of the dynamics of individual internal parameters and automatically use this information in the inference process during sensory decision making. Our second hypothesis is that humans follow such a dynamical context-based decision making process and given the contextual richness of our experimental paradigm, this interpretation process is responsible for our intriguing results.

Presently, there is no empirical evidence in the literature suggesting that the brain involuntarily maintains and uses a detailed internal representation of each internal parameter and their dynamics instead of simply encoding a limited number of decision parameters and using only the global volatility of the environment. This is because almost all previous investigations of sequential decision making processes used a task design with the strong simplifying feature that there was only a single parameter in the corresponding internal model that could change, e.g. the mean of the stimulus distribution (Behrens et al., 2007; Nassar et al., 2010; Gallistel et al., 2014; Glaze et al., 2015; Zylberberg et al., 2018), and this fact was obvious to the observers. Therefore, the underlying task can be framed as a tracking problem, in which successful behaviour requires the accurate detection of changes in a single parameter embedded in noise followed by a proper adjustment of the parameter. Some of these models also used information about the dynamics of the relevant parameter to improve change-detection efficiency (Behrens et al., 2007; Glaze et al., 2015; Piray and Daw, 2020). However, due to having a single dynamic parameter in these studies, the dynamics of this parameter and of the observed variables become directly linked thereby greatly simplifying the tracking problem since monitoring the overall volatility of the observed variable was sufficient to estimate the dynamics of the relevant parameter. Such a modelling setup captures well the structure of the typical behavioral and neurophysiological decision making experiments, but it fits only a small fraction of real life situations. In contrast, as demonstrated below, our experimental setup is suitable to investigate complex decision making since human behavior after the CP in this setup can be captured only by a model that possesses two characteristics. First, it includes two latent parameters that, in different combinations, can give two or more equally good description of the statistics of observations. Second, these parameters can in principle have distinct dynamics, so this information could be utilized to select the appropriate interpretation when facing a change.

To formalize our main hypothesis, we describe our experiment in a hierarchical Bayesian observer model (Fig. 3.2A). In our model, the observed stimulus x_t of each trial t is generated by the combination of two latent variables, the shape identity z_t and the magnitude of noise y_t superimposed on the shape image. To make a decision, the ideal observer infers the identity of shape, i. e. the decision variable z_t, by combining its prior assumptions about the appearance

probability (AP) of the two shapes, with the likelihood of shapes given the observations

$$P(z|x) \propto P(x|z) \cdot P(z; AP) \tag{3.1}$$

Importantly, the likelihood function:

$$P(x|z) = \int P(x|y, z) \cdot P(y|z)\, \mathrm{d}y \tag{3.2}$$

could be distorted if one of the shapes, on average, tends to be either more noisy (related to the second term of the integral in Eq. 3.2) or just less detectable in equal amount of noise (related to the first term of the integral in Eq. 3.2). This potential imbalance of noisiness is formalized in our model by a prior on the shape of the likelihood function by introducing the second main parameter of our model, normalized relative visibility (RV), that is scaled to be compatible with the first main parameter AP (Fig. 3.2A). Specifically, we parameterized RV so that arbitrary change in AP could be replaced by a change in RV and at the AP = 0.5 RV = 0.5 condition the model is unbiased (see Methods). Notably, when RV deviates from the unbiased condition, the nuisance variable y_t, which originally is not related directly to the decision, becomes informative about the shape identity z_t, thus allowing y_t to indirectly influence the decision via explaining away. Under this condition, when feedback is not provided about the correctness of the choice in a trial, the ambiguities of the two latent variables propagate one level up in the hierarchy and effectively couple the inference making on the priors AP and RV thereby eliciting an explaining away situation at the level of the parameters. This setup provides the first necessary conditions of a model suited for investigating complex decision making, having the potential for multiple competing interpretations of the observed stimuli.

To satisfy the second condition, we made our model dynamic by assuring that both prior parameters could change at any point in time during the experiment. This was implemented by two additional parameters in the model D_{AP} and D_{RV}, called hyperpriors, that encoded the dynamics of the parameters (Fig. 3.2A). Due to the dynamic variability of the setup, D_{AP} and D_{RV} hyperpriors must also be involved in the inference of the two latent prior parameters AP and RV. This experimental design and the corresponding generative model represent a

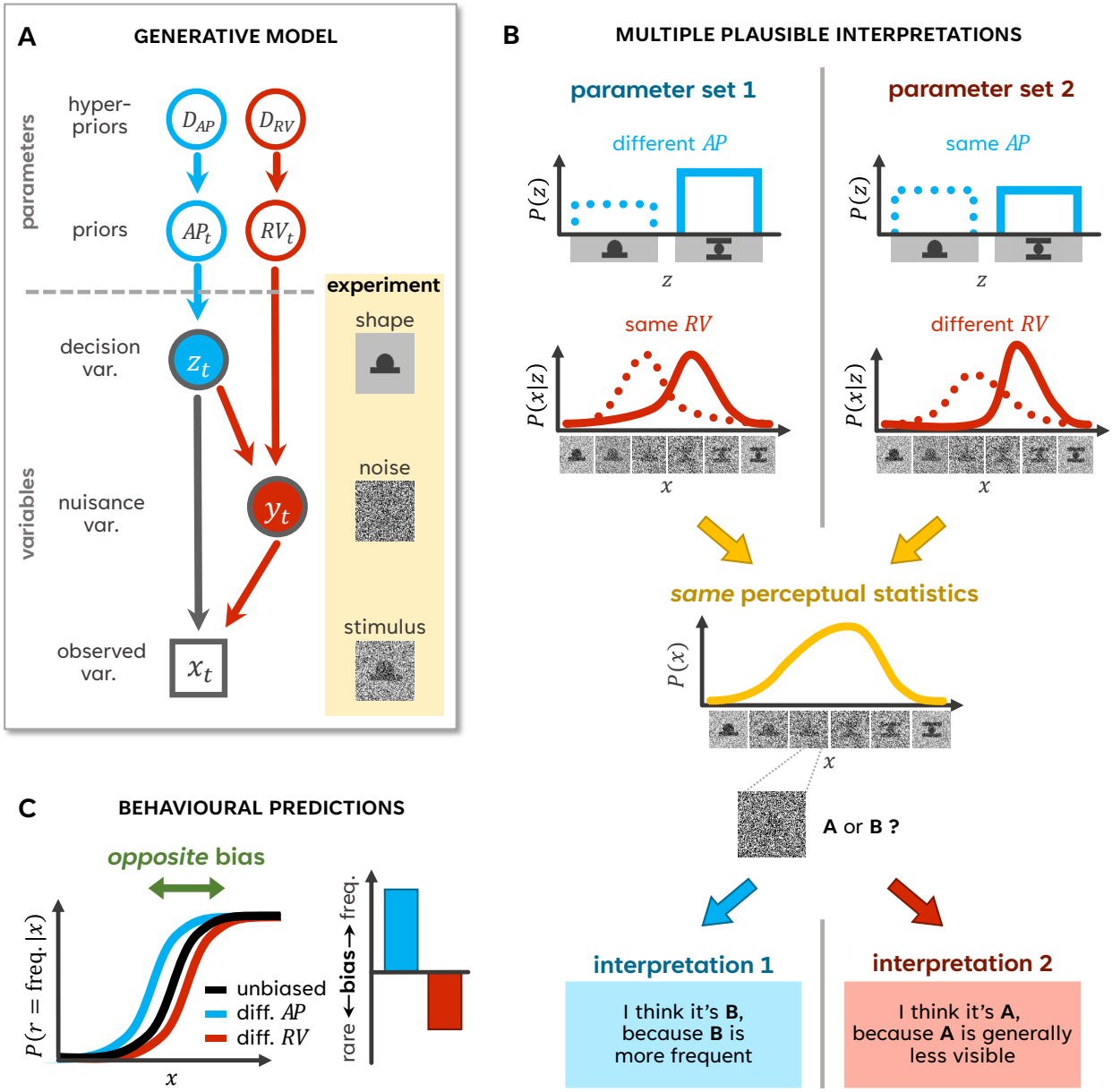


Figure 3.2. Modeling equally plausible interpretations. A. Variables and parameters of the complex generative model. The two-way branching structure induces the explaining away effect at the level of prior parameters. **B.** Illustration of how two distinct parameter sets can produce the same perceptual statistics, with each interpretation leading to different decision strategies. **C.** Biases of the psychometric curve corresponding to the two interpretations in B.

minimal setup in a 2-AFC classification paradigm required to investigate decision making based on choosing between competing interpretations.

3.3.2 Treating complex decisions with competing interpretations within a static model

To illustrate how our model can capture our experimental results by interpreting the same observations differently with different parameter combinations, we use the simplified static version of our model, momentarily omitting parameters D_{AP} & D_{RV} for the sake of clarity. Building on the intuition of our example with the patchy road in foggy weather, the key insight of our approach can be captured as follows. Without feedback in the Test trials following the CP, observers have no direct information about the AP and RV parameters, therefore, they have to infer them from the observed noisy stimulus distribution. However, the same stimulus distribution can be elicited by different parameter settings. For example, if shape B seems to be noticeably more frequent in the sequence than shape A after the CP, this may actually be true (Fig. 3.2B, parameter set 1). Alternatively, the two shapes may appear equally often but shape A generally embedded in more noise compared to shape B (Fig. 3.2B, parameter set 2), and this fact of A being less noticeable creates the illusion of appearing less frequently. While both interpretations are equally valid descriptions of the observed statistics (Fig. 3.2B, yellow distribution), they lead to very different decisions, especially on the noisiest trials. In trials where the shape is completely covered by noise, the first alternative assuming different APs suggests to chose shape B as it is overall more frequent, while according to the second alternative based on different RVs, the observer would imply that the noise covers shape A since A tends to be more noisy in general. In short, the two alternative interpretations induce opposite biases in the observer's decision and this difference in biases can be experimentally tested by measuring the horizontal shift in the psychometric curve. (Fig. 3.2C). We used this measure in our experiments to asses the generative models that observers use in their decisions (Methods).

To obtain specific model-based predictions, we formalize the above rationale within a probabilistic framework. In the case of a two-dimensional likelihood function over AP and RV, the maximum values of the function are located in a continuous region and we confirmed by simulations that in our experimental setup, these values are grouped along a straight ridge (Fig. 3.3A). The area under the ridge includes two special points and the correspond-

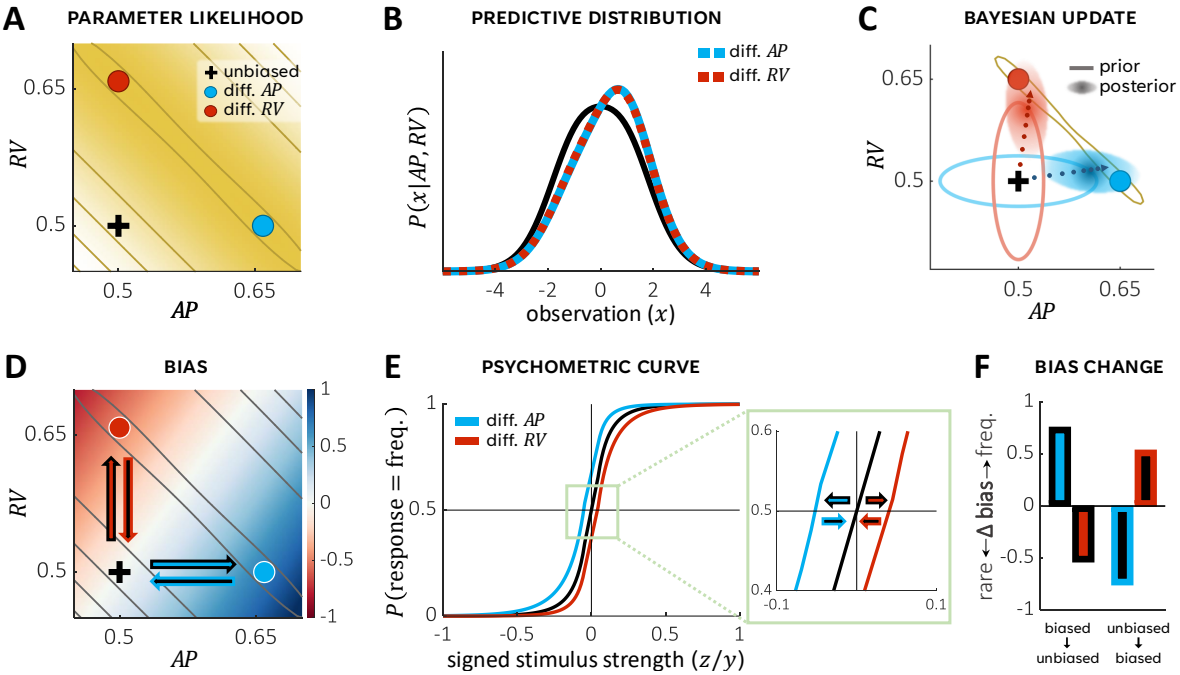


Figure 3.3. Behaviour of the static model. A. Heatmap and contour plot of the two dimensional parameter likelihood function in the $AP - RV$ space, with the observations generated by the parameter combination indicated by the red dot. There are two special points on the maximum likelihood ridge: one where AP is unbiased ($= 0.5$) and the bias of the perceptual statistics is fully explained by the RV parameter (red dot), and another where the reverse is true (blue dot). **B.** Predictive distributions of two models corresponding to the special points in A. **C.** Simulated trajectories of the MAP parameter estimates across trials (dotted arrowa) after a sudden change in the stimulus-generating parameters from the black cross to the red dot. Initial conditions are compared based on whether the RV parameter (red ellipse and trajectory) or the AP parameter (blue ellipse and trajectory) is more uncertain a priori. **D.** Heatmap and contour plot of the decision bias in the two dimensional parameter space. We show potential parameter trajectories (arrows) between the unbiased parameter combination and the special parameter combinations. This trajectories could belong to hypothetical update processes. **E.** Simulated psychometric functions corresponding to the special parameter combinations. The shifts in the psychometric functions (inset arrows) align with the parameter trajectories shown in panel D. **F.** Bias difference between the endpoints of the trajectories in D and E (bar colors correspond to the arrow colors in D and E).

ing models, one with different appearance probabilities ($AP \neq 0.5$) but the same relative visibility ($RV = 0.5$) for shapes A and B and another one with identical appearance probabilities ($AP = 0.5$) but biased relative visibility ($RV \neq 0.5$), (Fig. 3.3A, blue and red dots, respectively). These models correspond to the two parameter sets in Fig. 3.2B. Importantly, although each point around the ridge with approximately equal likelihood values represent different models with distinct parameter settings, these models predict roughly the same

stimulus distribution thus providing an almost equally good explanation of the sensory input as confirmed by simulations using the two special models (Fig. 3.3B).

While our model so far satisfies the requirement of providing multiple alternative internal representations of the sensory input stream, it raises a new question: How to choose between these equally suitable models? For example, when the observer having an unbiased model ($AP = 0.5$ and $RV = 0.5$) detects a discrepancy between the perceived stimulus distribution and the one predicted by her model, how does she adjust the parameters of her model to choose between different new models that offer equally plausible interpretations of the discrepancy? Following probability theory, the model parameters should be updated in proportion to their prior uncertainty. After accumulating sufficient evidence, the chosen model will eventually converge to the one with maximum likelihood parameter setting, if any, or settle in one of the equally good options, but the exact trajectory of the convergence is determined by the priors used in the starting model. In particular, the initial part of this trajectory will be strongly attracted towards models that differ from the starting model in the less certain parameter, that is models positioned along a trajectory more aligned with the RV axis for the starting model with higher RV uncertainty and conversely, models positioned along a trajectory more aligned with the AP axis for the starting model with higher AP uncertainty (Fig. 3.3C). We confirmed by simulations that this difference between the initial part of the two trajectories is manifested by markedly different decision biases of the psychometric curves depending on the parameter priors (Fig. 3.3D). Establishing a direct mapping between this probabilistic framework and the logical reasoning about biases presented in the previous section has an important consequence: if human decision making is, indeed, probabilistic and it is controlled by the relative uncertainty in the parameters of the two alternative interpretations, a change in this relative uncertainty should make the observer alter her interpretation as well. We use this feature to link our model to our experimental results.

3.3.3 Treating competing interpretations based on a dynamic model

One way to manipulate uncertainties for changing the observer's interpretation is to adjust the dynamics of the system: the faster a parameter tends to change, the more uncertain the

observer is about its future value. This dynamics-related uncertainty contributes to parameter uncertainty and hence, to the interpretation of the stimulus statistics in the same way as other uncertainties do in the static case, but only if the observer has knowledge about the dynamics of the parameters. As mentioned above, in simple models, this information can be easily estimated by using a simplified proxy, namely the volatility of the environment due to the direct link between the dynamics of the single decision variable defining the problem and the dynamics of the observed variables. However, this strategy does not generalize to more complex internal models with several differently behaving latent parameters since, the observed volatility only reflects the intertwined aggregate dynamics of the interacting parameters, not the individual dynamics separately. To disentangle the dynamics in such situations, a more detailed characterization of parameter dynamics is required.

Such a more detailed characterization can be given by using the 2D-space specified by the axes of change frequency and change magnitude (Fig. 3.4A). Change of each parameter in this space may occur (1) infrequently but with a large magnitude (change-point process), (2) frequently but with a small magnitude (diffusion process),(3) not at all along either axis (static) or (4) frequently with a large magnitude along both axes (unpredictable). Each of these conditions specifies a transition probability function with a distinct and parametrizable shape expressed by the hyperpriors (Fig. 3.2B, $D_{AP}\,\&D_{ND}$). In principle, both hyperpriors of our model can be located in any of these four regions, but if one of the them is either unpredictable or static, then the task itself is either unpredictable or equivalent to the simple case with one dynamic prior parameter. Therefore, we focus on the interesting situation when the two hyperpriors define dynamics that are sufficiently different from each other along the diffusion process vs. change-point process axis (Fig. 3.4A). In this case, the characteristics of the observed changes do provide a strong cue for interpretation: a prior parameter with a more change-point (CP) type dynamics can explain a large and rapid observed change, while a prior with a more diffusion process-like dynamics is more suitable to describe a slow drift of the observed statistics. As a consequence, while a complex model with multiple hyperpriors defining latent variables with different dynamics can provide more than one equally adequate interpretations of the steady-state behavior of the observed variables through dif-

ferent interactions, these models can be clearly distinguished in their adequacy based on their dynamics.

To illustrate such equally adequate interactions more formally, we ran simulations with different dynamic versions of our Bayesian model described in Fig. 3.2A. We investigated how simple and complex dynamic versions of the model starting from a highly confident initial state with unbiased parameters (i.e. both AP and RV being at 50%) would adapt to a novel and biased statistics defined by AP changing to 75% (i.e. shape A appears in 75% of the trials), when this transition from 50% to 75% bias is introduced either abruptly or gradually over the coarse of 120 trials. We compared three models: two simple models with one dynamic parameter, where the transition of single AP parameter followed either a drift diffusion or a change-point dynamics (Fig. 3.4B orange and green, respectively), and one complex model with two dynamical parameters, a drift diffusion type AP dynamics and a change-point type RV dynamics (Fig. 3.4C burgundy).

Our simulations with the simple models revealed that the resulting representation of the dynamic parameter, defined by the posterior over AP (Fig. 3.4B, left panels), is largely independent of both the parameter dynamics (orange and green) and the actual speed of change (light and dark colours). This illustrates that with one dynamic parameter, there exists only one interpretation of the novel statistics, in our case the true steady-state AP after the 120-trial-long transition period. Consequently, the response biases are almost identical in all conditions, and these biases persist long after the transition period (Fig. 3.4B, right panels). In contrast, the complex model arrives at different representations depending on the actual speed of change. When the observed change is sudden, model states with parameter combinations that require more change from the current state by the parameter that has CP dynamics (here, RV) have relatively higher posterior values compared to combos that would change more the parameter with diffusion dynamics (here, AP) (Fig. 3.4C, left panel). Conversely, in response to slow changes, the complex model prefers combos for explaining the new condition that change more the parameter with diffusion dynamics (AP) (Fig. 3.4C, middle panel).

Crucially, this leads to very different (even opposite) response biases in the two cases, and this difference persists long after the change occurred (Fig. 3.4C, right panel). This persistence indicates that both interpretations fit the observed data equally well, which eliminates the need for the observer to revise the initially selected interpretation in the close future. We verified by simulation that both the two simple and the complex models arrived and remained at a state after the change, in which their posterior predictive distributions of the observations were identical and veridical (Fig. 3.4D). For completeness, we tested all other possible scenarios of parameter and dynamics combinations of simple and complex models and found results similar to the ones obtained by using the above setups with the only difference that the generated biases had the opposite sign (nor shown).

This analysis provides an explanation to our initial results. If humans use specific information about the dynamics of the latent parameters of their internal model according to the probabilistic computation specified above, introducing the same parameter change either abruptly or gradually in our experiments should prompt the observers to choose different interpretations of the input according to the predictions of our complex dynamic internal model, and their selected interpretation should be manifested by opposite observed biases in their responses (Fig. 3.3D, inset). Moreover, we predict that a targeted experimental manipulation of the observers' prior experience about the dynamics of the two parameters (D_{AP} &D_{RV}) that alters the relative uncertainty of those parameters should lead to observers changing their interpretation of the same input sequence and selecting a different adjustment of their internal model with a bias opposite to those measured without the targeted manipulation.

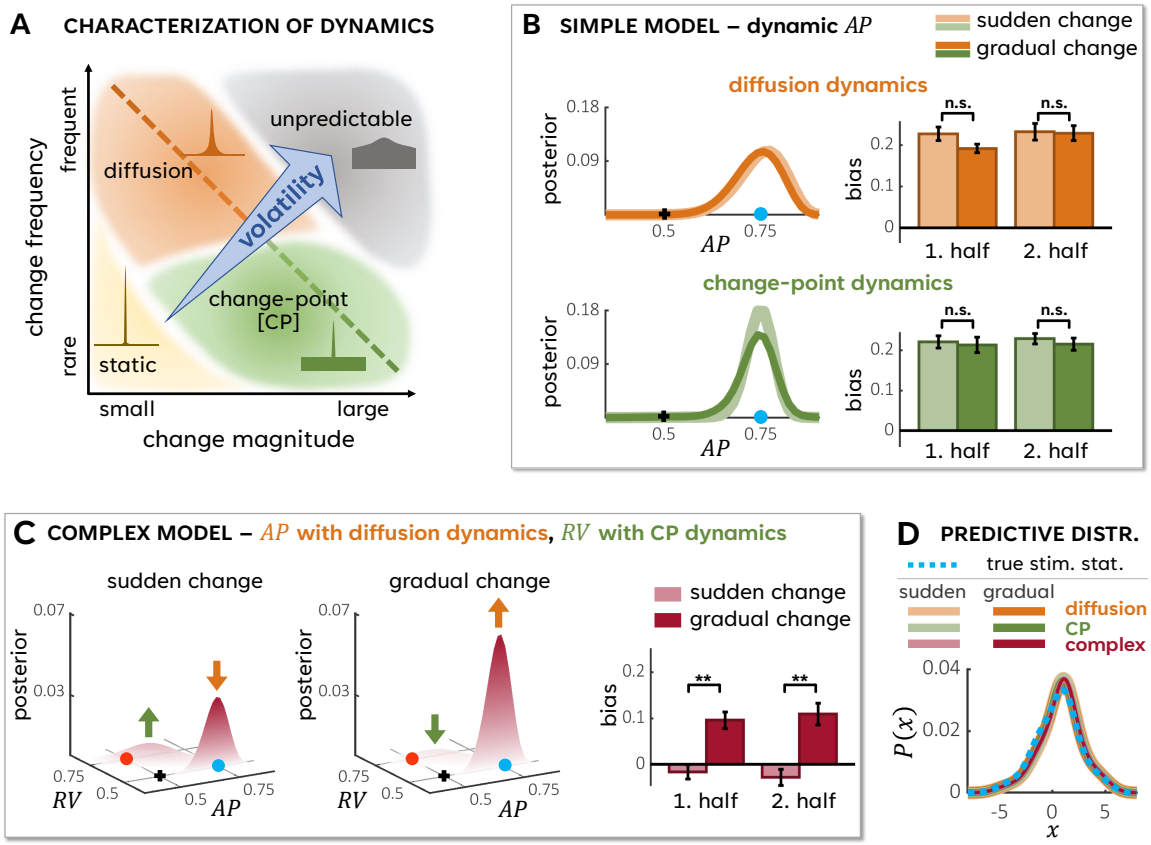


Figure 3.4. Characterization of parameter dynamics and the behaviour of the dynamic model. A. Volatility can be decomposed according to the *frequency* and expected *magnitude* of individual changes. In the resulting 2D space of dynamics, four characteristic regions can be distinguished based on the corresponding transition probability distributions [TPD]: 1) narrow TPDs define diffusion processes (orange), 2) however, if the TPD is extremely narrow the environment is approximately static (yellow), 3) the mixtures of narrow and wide TPDs define change-point processes (green), 4) and if the TPD is too wide, the environment is unpredictable (gray). **B.** Simple models with either drift (orange) or CP (green) type AP dynamics end up with similar AP posteriors by the end of the TP (left column) and produce roughly equal behavioural biases (right column) short and long (1. and 2. half) after the TP, regardless of whether the change was sudden (light color) or gradual (dark color). The marginal posterior distributions are concentrated on the true $AP = 75\%$ value (blue dot) in either case. **C.** A complex dynamic model, with diffusion type AP and CP type RV dynamics, will end up with different joint parameter posterior distributions depending on whether the change is sudden (left panel) or gradual (middle panel). As a consequence, significantly different long-term biases emerge in the two cases (right panel). **D.** For all models, the posterior predictive distribution (solid lines) is approximately the same as the true stimulus distribution (dotted blue line).

.

3.4 Evidence of humans choosing between competing dynamic interpretations during decision making

3.4.1 Assuming internal selection between interpretations captures unexpected human decision making behavior

The two predictions derived from the probabilistic model in the previous section not only allow a direct comparison between outcomes of the model's simulation and the results of Experiments 1&2, but it also provides a further surprising prediction for a new Experiment 3 (Fig. 3.5). Experiment 3 is identical to Exps 1&2 in all respect except that two changes were concatenated at the CP, first an instantaneous increase of AP from 50-50% to 65-35% and then immediately a gradual decrease of AP back to chance level over 80 trials. This manipulation resulted in a 500-long trial sequence, in which apart from a brief 80-trial-long spike at the CP all trials were presented in the unbiased AP condition.

The simple and complex models described in Section 3.3 make qualitatively distinct predictions and provide confirmatory simulation results about the expected relative response biases measured during the Test of Exps. 1-3 (Fig. 3.5B). If the participants use the simple model, the bias after the CP is ultimately determined by the steady-state statistics of the Test and not influenced by the type of transition. In this case, Exp1 & Exp2 that have identically biased steady-state statistics should generate equal biases during the Test that significantly differ from zero either positively or negatively depending on whether the AP or the RV parameter is dynamic, respectively (Fig. 3.5B, upper row, green and orange bars). The same model predicts that the bias in Exp3 should be zero given the balanced steady-state statistics (Fig. 3.5B, upper row, yellow bar).

In contrast, if participants use a complex model, their behavior would show a completely different pattern matching the two counter-intuitive features found empirically in Exps 1&2. First, the temporal characteristics of the transition at the CP should have a strong influence on the interpretation leading to a significant difference between the biases of Exp1 & Exp2 despite their identical steady-state observed statistics. The sign of the difference would be

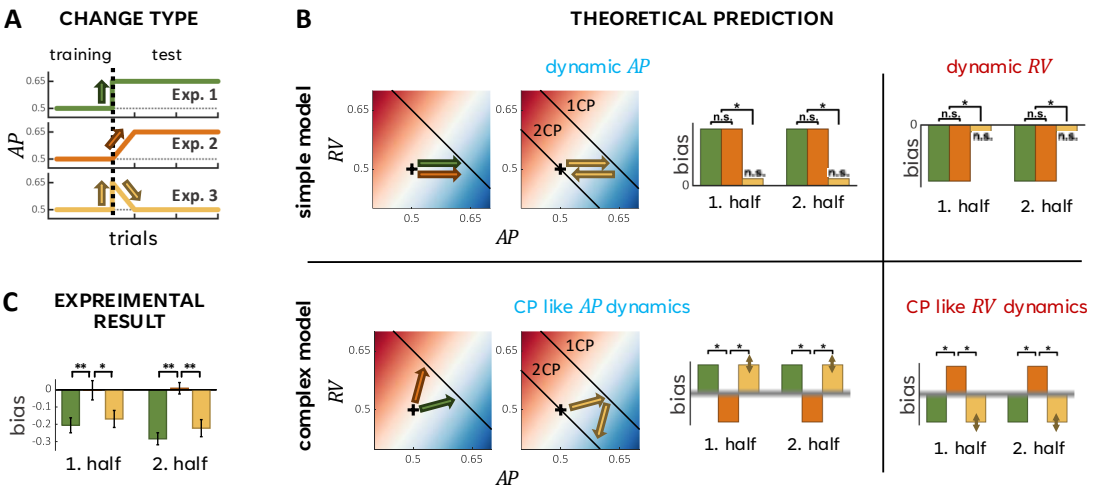


Figure 3.5. Experimental dissociation of simple and complex models. A. Experimental conditions. **B**. Qualitative predictions of the simple (upper) and complex (lower) models. Hypothetical trajectories in the parameter space (arrows within the heatmaps on the left) and the long-lasting biases (bar plot) are shown for the experimental conditions when AP is the dynamic parameter (upper left) or it has the more CP-like dynamics along the diffusion-CP axis compared to the RV parameter (lower left). Biases are also shown for the reverse parameter allocations, when the role of AP and RV are swapped. **C**. Experimentally measured biases during the first and second half of the test phase.

determined by the relative position of the two parameters' dynamics along the diffusion - CP axis. Specifically, if AP has the more CP-like dynamics, a more positive bias should follow the instantaneous (Exp 1) than than the gradual change (Exp 2), while the opposite pattern should hold if RV has the more CP-like dynamics (Fig. 3.5B, lower row, green and orange bars). The second predictions is that either in Exp 1 or Exp 2 (the one with negative bias), the participants should systematically choose the shape appearing significantly less frequently in the preceding tens and hundreds of trials.

The third counter-intuitive predictions of the model is that the bias in Exp3 should not become zero even long after the AP perturbation is completed and the sequence returned to the balanced steady-state statistics (Fig. 3.5C, lower panel, yellow bar). This occurs because the internal parameter with a more CP-like dynamics is mostly involved in explaining the initial sudden increase of the observed AP and less so in the subsequent gradual return to the balanced state, while the opposite is true for the other internal parameter with more diffusion-like dynamics. This two-step adjustment in the $AP - RV$ parameter space results in a return to a different segment of the ML ridge that crosses the coordinates of the neutral

bias model, settling on a model in which the bias will be positive or negative depending on which of the RV and AP parameters is assumed to have more CP-like dynamics. This leads to a bias in Exp3 that differs significantly from that of Exp2 and has the same direction as the bias of Exp1 (Fig. 3.5B, lower panel, yellow vs. green).

To test these predictions, we have run Experiments 3 with a new group of participants (N_3=20) and found that participants' bias in Exp 3 as well as the combined pattern of the average biases across the three experiments was highly compatible with the bias pattern predicted by the complex internal model having more CP-like RV dynamics without any adjustment of parameters between experiments. The difference between Exp1 and Exp2 was significant in both the first ($t_{35} = -2.85$, $p = 0.007$, $\mathrm{Cohen's}\, d = 0.92$) and the second halves ($t_{35} = -6.09$, $p < 0.001$, $\mathrm{Cohen's\, d} = 1.96$) of the Test period (now, with STSEs factored out from the bias). The same was true for the difference between Exp2 and Exp3 in both the first ($t_{37} = 2.24$, $p = 0.031$, $\mathrm{Cohen's}\, d = 0.7$) and the second halves ($t_{35} = 3.84$, $p < 0.001$, $\mathrm{Cohen's\, d} = 1.21$) of the Test period. Thus the qualitative pattern of results remained the same again across the two halves of the Test period also in line with the predictions of the complex model. We found no significant difference between the biases measured in the first and second half of Exp2 and Exp3 (Exp2: $t_{18} = -0.21$, $p = 0.836$, $\mathrm{Cohen's}\, d = 0.055$; Exp3: $t_{19} = 1.13$, $p = 0.272$, $\mathrm{Cohen's}\, d = 0.232$), but the negative counter-intuitive bias of Exp1 became slightly, but significantly stronger in the second half of the Test phase ($t_{17} = 2.582$, $p = 0.019$, $\mathrm{Cohen's}\, d = 0.446$). The relative stability of the biases indicates that much of the internal parameter adjustment occurred at the CP, without major revisions later, since all interpretations described the post-CP steady-state equally well.

3.4.2 Snapshot models of local steady state statistics cannot describe human decision making

In Exp1 and Exp3 that produced similar biases, the APs in the first 20 trials of the Test phase were the same, while they were quite different form the AP in Exp2 that produced a different bias. This rises the possibility that after noticing a CP, observers start *tabula rasa* and the bias

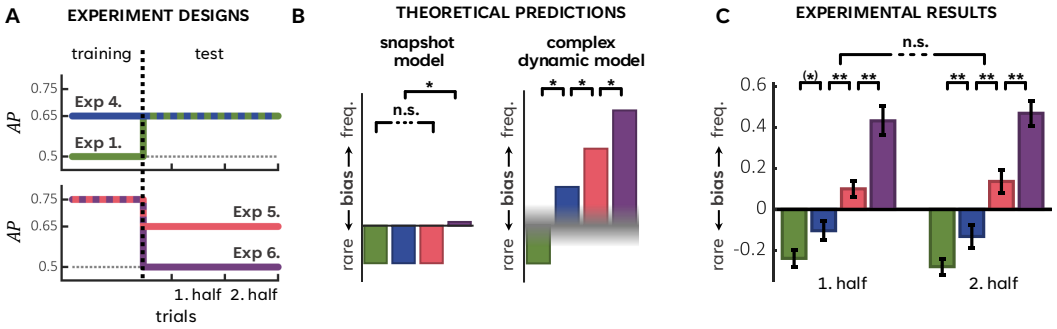


Figure 3.6. Testing the snapshot model. **A**. Experimental conditions. **B**. Qualitative predictions of the snapshot model (left) and the complex dynamic model (right). **C**. Experimental results.

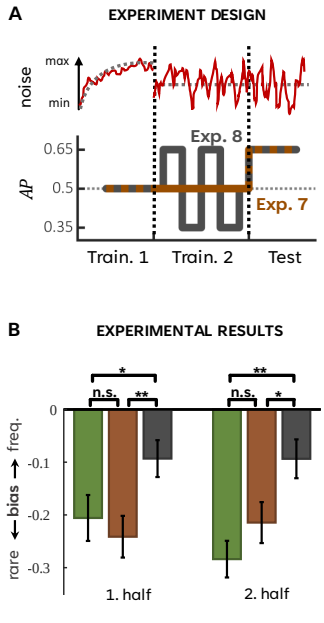
in their behavior is determined by the statistics of the stimulus sequence provided immediately after the CP rather than by the full set of characteristics (i.e. starting and end point and dynamics) of the change. To compare the predictions of this alternative *snapshot model* and the complex model, three additional experiments were run. Specifically, in Exp4 ($N_4 = 18$), AP was set to 65% in both the training and the test phase, completely eliminating the AP change during the experiment. In Exp5 & Exp6 ($N_5 = 16$ & $N_6 = 19$), the AP was 75% during the training, and reduced to either 65% or 50% during the test phase, respectively, reversing the direction of change compared to Exp1. The snapshot model would predict the same bias in Exp4 and Exp5 as in Exp1, and a near-to-zero bias in Exp6 as in Exp2 (Fig. 3.6B, left side). In contrast, the previously identified complex dynamic internal model with more CP-like RV dynamics makes specific predictions about the ordering of the biases measured in Exp1 and in the three control experiments (Fig. 3.6B, right side). Conceptually, since according to this model the sudden AP change in Exp1 elicited a negative bias, the complete elimination of the change in Exp4 should reduce the negativity of this bias. In addition, reversing the direction of change in Exp5 & Exp6 should further shift the bias to the positive direction. Interestingly, since the sudden change in Exp6 is larger then in Exp5, the model should predict more positive bias in Exp6 than in Exp5, despite that the actual AP being larger in the test phase of Exp5 (Fig. 3.6C). Notice that with the complex model, specific predictions can be made only about the relative magnitude and ordering of the biases and not of their exact sign since that would require more specific knowledge of the model parameters.

The behavioral results of the three experiments showed that the participants' biases clearly followed the ordering predicted by the complex dynamic model (Fig. 3.6B-C), and there were no significant differences across the two halves of the Test phase in any of the new control experiments (Fig. 3.6C, see the statistics in Fig. 3.7D). All the predicted differences between biases were significant ($p < 0.05$) in both the first and the second half of the Test, except for one condition (Exp1-Exp4, first half), which was not significant but approached significance (see details in Fig. 3.7C). These results confirm that the observed behavioral biases are not determined by the local statistics observed immediately after the CP, but by the combination of the perceived direction, magnitude, and speed of the change, as predicted by the complex model.

3.4.3 Changing the history of observed dynamics can strongly influence the choice of the implicitly applied internal model

The preceding results confirm that during even the simplest perceptual decision, humans make up their mind in a complex manner relying on multiple internal variables including z and y, and their prior parameters, AP and RV that reflect summary statistics of z and y based on earlier experiences. These results also suggest that observers selected their internal models idiosyncratically by implicitly assuming more or less CP-like dynamics to RV and AP. A relevant question is whether observers' assumptions about the dynamics of AP and RV, captured by the hyperparameters (D_{AP}, D_{RV}) in our model, is predetermined or, similarly to AP and RV, they can change flexibly based on previous experience.

To address this question, we investigated which local changes at the CP could be primarily responsible for the observers assumption about the dynamics of AP and RV. First, we examined whether the abrupt change in the local noise distribution around the CP alone could cause the RV parameter appearing as having more CP-like dynamics. In Exp7 ($N_7 = 33$), we made one modification compared to Exp1 by introducing a 100 trial long second training phase (Training2) between the original training phase (Training1) and the Test (Fig. 3.7A). Training1 and Training2 were identical with one difference: instead of using the adaptive staircase method, noise in Training2 was drawn from the same $P(y)$ marginal distribution as



C

first sample	second sample	phase	t	df	p	Cohn's d
Exp1	Exp2	FH	-2.85	35	0.007	0.92
		SH	-6.09		0.000	1.96
Exp2	Exp3	FH	2.24	37	0.031	0.70
		SH	3.84		0.000	1.21
Exp1	Exp4	FH	-1.90	34	0.066	0.62
		SH	-2.88		0.004	0.94
Exp4	Exp5	FH	-3.37	32	0.002	1.13
		SH	-3.34		0.002	1.12
Exp5	Exp6	FH	-3.58	33	0.001	1.19
		SH	-3.89		0.000	1.29
Exp1	Exp7	FH	0.57	49	0.572	0.16
		SH	-1.18		0.243	0.34
Exp1	Exp8	FH	-2.20	51	0.032	0.63
		SH	-3.33		0.002	0.95
Exp7	Exp8	FH	-2.82	66	0.006	0.68
		SH	-2.26		0.027	0.54

D

sample	t	df	p	Cohen's d
Exp1	2.58	17	0.019	0.45
Exp2	-0.21	18	0.836	0.06
Exp3	1.13	19	0.272	0.23
Exp4	0.26	17	0.801	0.07
Exp5	-0.75	15	0.464	0.19
Exp6	-1.41	18	0.177	0.20
Exp7	-0.56	32	0.579	0.12
Exp8	0.01	34	0.991	0.00

Figure 3.7. Training the dynamic hyperpriors. A. The experimental conditions were extended with a second training phase, featuring the same noise statistics as the test phase. **B**. Experimental results of the new experiments compared to the results of Exp 1. **C**. Detailed statistics of the two-sample t-tests comparing participants' biases across different experiments. Tests were conducted separately for comparing the first halves (FH) and second halves (SH) of the Test phases. **D**. Detailed statistics of the paired t-tests comparing participants' biases across the two halves of the Test phase within each experiments.

in the Test thereby eliminating the change of the marginal noise distribution at the CP. Had the abrupt change $P(y)$ been the main reason why RV seemed to have more CP-dynamics, our change in Exp7 should reduce the negativity of bias found in Exp1 (Fig. 3.7B). However, we found the bias in Exp7 not being significantly different from the bias in Exp1 (Fig. 3.7C, Ref Table) eliminating the change in local noise distribution at the CP as the main factor determining the choice of parameter dynamics.

Next, we tested whether longer-term observed statistics could control the choice of parameter dynamics. In Exp8 ($N_8 = 35$), we introduced one further change to Exp7: the experimental AP alternated between 70% and 30% in 20 trial long blocks during the middle 80 trials of Training2 (Fig. 3.7A). With the introduction of frequent and abrupt changes in AP during Training2 with feedback still provided, the observers perceived the AP dynamics being more volatile and CP-like. If this alternation in the level of AP experienced before the Test session influences the observers' internal description of AP dynamics as more CP-like, the AP parameter should be involved more in the interpretation of CP-type changes, which in

turn should decrease the negativity of the bias found in Exps 1&7. Indeed, observers' bias in Exp8 was significantly less negative than in Exp7 confirming the cardinal role of previously collected information on dynamics of AP in selecting the internal model after a CP model (Fig. 3.7B, Table). In summary, the combined results of the eight experiments together suggest that observer using a complex internal model that can simultaneously capture diverse parameter dynamics. When more than one equally fitting interpretations of an observed change is available, observers incorporate in their decision the dynamic characteristics of the change as well as high-level and specific knowledge about the dynamics of the parameters of the internal model to determine which interpretation will be selected for explaining the new condition.

3.5 Reaction time-based confirmation of the internal model's complexity

In the previous sections, we used the qualitative predictions of an abstract Bayesian model to argue that human decision making adapts to dynamic changes in the input by relying on a complex internal model. In this section, we further validate this result by linking the prior parameters of the abstract Bayesian model (AP and RV) to the parameters of a process-level model – the bounded evidence accumulator (BEA), (Forstmann et al., 2016; Ratcliff and McKoon, 2008) – to explain idiosyncratic reaction time (RT) patterns in the data that are beyond the explanatory scope of the original Bayesian model. By showing that the complexity of the internal model is critical in explaining the diversity of reaction time patterns across the experiments, we provide strong confirmation of our initial findings.

According to the BEA framework, 2AFC (perceptual) decisions are driven by the accumulation of evidence about the relative probability of the two options until a decision threshold in favour of one option is reached. This modeling framework is well suited to explaining how different factors, such as prior beliefs and stimulus strength, affect reaction time. The BEA models have two key parameters: (1) the starting point of the accumulation process and (2)

the bias of the accumulation rate towards one option. Importantly, these two parameters can be conceptually linked to the AP and RV parameters of the complex Bayesian model.

The key observation to link the abstract model to the process-level model is that the AP and RV parameters have conceptually very different roles in the trial-by-trial perceptual inference when it is modelled as Bayesian inference. The AP parameter defines the prior term of Bayes' rule expressing the beliefs of the participants prior to the observations and, therefore, is independent of the stimulus. In contrast, the RV parameter is related to the likelihood term that defines the statistical relationship between the decision options and the observed stimulus. As a result, the two types of parameters influence the process of decision formation and, thus, the within-trial time-courses of stimulus-dependent vs. -independent parameters biases qualitatively differently (White and Poldrack, 2014). Specifically, biases related to stimulus-independent parameters show up much stronger when the decision is fast compared to when it is slow (Fig. 3.8A, top, left). In contrast, biases related to stimulus-dependent parameters are equally present in both fast and slow decisions (Fig. 3.8A, top, right). These effects are usually described with BEA models (White and Poldrack, 2014), by assuming that the starting point of the accumulation process is influenced by the stimulus-independent parameters (such as AP; Fig. 3.8A, bottom, left) while the rate of accumulation is biased by the stimulus-dependent parameters (such as RV; Fig. 3.8A, bottom, left).

These observations indicate that after establishing the monotonic relationship between the parameters of the abstract Bayesian model and the parameters of the BEA model, the abstract parameters should capture the experimentally measured time dependence of the behavioural biases within trials.

We tested this prediction in two steps. First, we independently estimated the best-fitting AP-RV combinations of the abstract model for each participant (Fig. 3.8B), and validated through simulation that the model accurately captured the data (Fig. 3.8C). In the second step, we freeze the per-participants AP and RV parameter estimates and use them as regressors for the BEA's parameters to fit the response-times-augmented data assuming a linear relationship between the abstract and the process-level parameters. During this step, short-term response patterns were also allowed to affect the reaction times in addition to the AP and

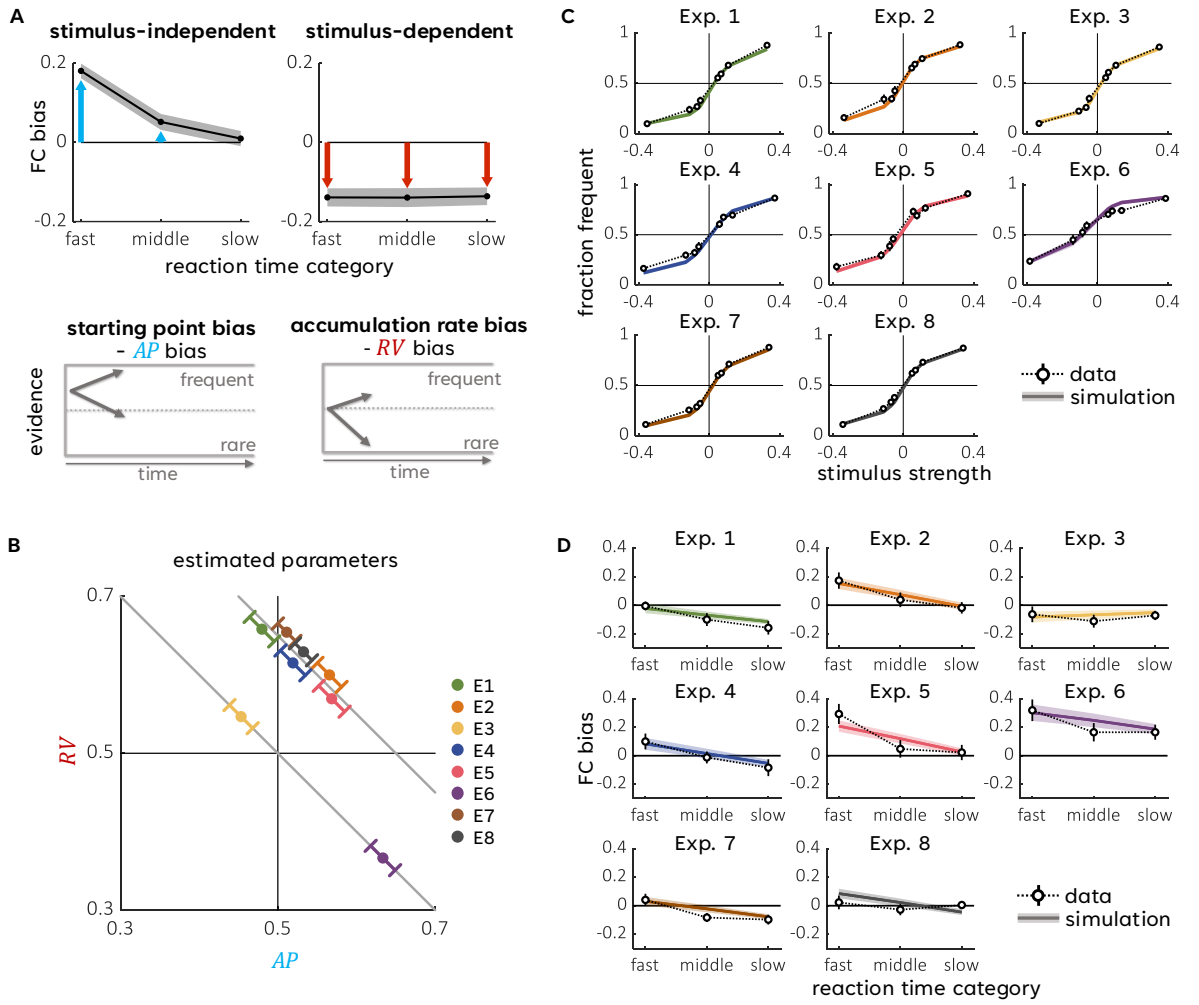


Figure 3.8. Reaction time-based validation of the complex model. A. Top row: Typical RT-dependent bias patterns associated with parameters affecting stimulus-dependent and stimulus-independent aspects of the decision making process. The figures were simulated using the BEA model. The **FC bias** is the difference between the fraction of correct responses measured for the frequent and rare objects. Bottom row: Illustration of the key BEA model parameters and their link to the abstract model parameters. **B.** Best-fit parameters of the complex model (mean $\pm$ sem across participants), with the parameters constrained to the maximum likelihood ridge corresponding to the steady-state experimental statistics. **C.** Fraction of frequent responses (mean $\pm$ sem) as a function of stimulus strength computed either from the experimental data (black circles) or from the synthetic data (colored lines) simulated by the EA model with best fitting parameters. **D.** FC biases (mean $\pm$ sem; see the definition in panel A) as a function of reaction time category (equale percentile split), computed either from the experimental data (black circles) or from synthetic data simulated by the BEA model with the best fitting parameters (coloured lines).

RV parameters. Importantly, all the remaining parameters in this step were shared across all the participants and experiments, since the goal was to find a general mapping between the abstract and the process-level parameters that holds for each participant irrespective of which experiment they participated in. We found that the process-level model with its parameters

being tied to the corresponding abstract Bayesian model's parameters could reliably capture the reaction time-dependence of the behavioural biases (Fig. 3.8D). Furthermore, Bayesian model comparison showed that the simple models, which scaled only one parameter of the BEA model, or a BEA model that relied exclusively on short-term response history could not fit the reaction time patterns as well as the complex model did (Section B.4). In summary, this independent verification method confirmed that a complex model with two adjustable parameters was, indeed, necessary to describe the full complexity of the behavioural data.

3.6 Discussion

In this chapter, we explored the simplest forms of complex decision situations where the precise underlying structure remained undisclosed to the observer despite an extended period of familiarization. We found that in these situations, participants exhibited seemingly counter-intuitive behaviour that could not be adequately captured by standard, simplistic models of sequential decision making that attribute all changes to a single hidden cause. Specifically, we observed that in some cases participants developed long-term biases against the observed statistics, and even short-term perturbations of otherwise balanced statistics were able to induce persistent behavioral biases. Moreover, participants' performance was notably influenced by the dynamics of the perturbation suggesting that this dynamics is exquisitely encoded in the their internal representation. Our results contribute two important novel insights to the field of complex decision making: 1) In ambiguous decision situations, regardless of the task's simplicity, participants automatically resort to their complex hierarchical internal models with multiple dynamic parameters. 2) When encountering changes in the task statistics, participants adjust those parameters of their complex internal model the dynamics of which are best aligned with the change.

In order to fully grasp the true nature of the decision situation, we developed a complex Hierarchical Bayesian Model (HBM) and demonstrated that this model could qualitatively capture all the idiosyncrasies of our data. This HBM was complex in the sense that it implicitly represented multiple alternative interpretations of the observations through differ-

ent parameter settings. When the need for revise the current interpretation emerged due to changes in the conditions, this model relied on higher-order statistics encoded by the dynamics of the parameters to select between the alternatives to explain the data. In this way, our complex HBM could naturally exhibits a fundamental phenomenon of complex probabilistic models at the level of parameters, known as "explaining away". Explaining away refers to the phenomenon that the system maintains multiple possible interpretations of the observations (weighted by their probabilities) until an internal need or external information is received that can resolve the ambiguity.

The idea that complex HBMs are suitable models of human cognition has been around before (Griffiths et al., 2008; Orbanz and Teh, 2010; Austerweil et al., 2015; Whittington et al., 2020). For example, HBMs have previously been successfully used to elucidate a remarkable human capacity: the ability to efficiently build abstract knowledge from limited evidence (Tenenbaum et al., 2011). For this purpose, the key property of HBMs was their ability to dynamically grow based on simple rules, enabling them to effectively discover complex structures underlying raw observations (Kemp and Tenenbaum, 2008). However, computational explorations of how to tune the parameters of such an existing HBM and relating the obtained results to human behavior has been restricted mostly to very simple chain or tree structures (Behrens et al., 2007; Heilbron and Meyniel, 2019; Piray and Daw, 2020). For example, while Heilbron and Meyniel (2019) emphasised the importance of deviating from the chain-structure when investigating the hierarchical nature of human cognition, they still used a simple tree structure that did not allow for competing interpretations of a given input. As a result, very few studies could focus even theoretically on the important effect of explaining away on parameter learning in complex probabilistic networks let alone relating this effect to behavioral data.

The current study aimed at filling this gap by exploring the consequences on decision-making based on a complex and dynamic internal model in a changing environment. Our findings revealed a nuanced interaction between the speed of environmental changes and the idiosyncratic parameter dynamics of the complex internal model. Importantly, our results concerning the mechanism of perceptual decision making are conceptually consistent with previous

findings in the field of motor-learning. In that domain, employing a complex model representing both internal (body-related) and external (environment-related) parameters could explain the change speed-dependent arbitration between explaining motor errors with these two types of parameters (Berniker and Kording, 2011). The consistency of results from these two distinct domains implies that the same universal mechanism may underlie the adaptation of both motor responses and abstract decisions.

The explaining away phenomenon has mainly been discussed in the context of momentary perceptual inference, (Adams et al., 2004) even if the information triggering the phenomenon was obtained over a longer period of time (Sotiropoulos et al., 2011). In our study, we specifically investigated how explaining away impact parameter learning, when not only the individual observations, but also the statistical regularities across these observations have multiple interpretations. Our finding that humans automatically employ complex HBMs when faced with such complexity indicates that explaining away is a fundamental aspect of cognition that is also present in learning situations.

The prior work by (Courville, 2006; Dayan and Kakade, 2000) also investigated explaining away at the level of parameters by using the framework of Bayesian causal modelling and focusing on the classical conditioning phenomenon known as backward blocking. However, in classical conditioning, the two latent parameters representing the causal strength of the two conditioning stimuli on the Unconditioned Stimulus are direct memory representations of the sensory input. In our case, the internal parameters created a latent space of possible and equally good solutions for interpretations and the explaining away effect operated on implicit equivalence classes of possible interpretations rather than on directly represented sensory items. Therefore, a significant contribution of our study is providing a demonstration that humans utilize complex models even when the actual task does not explicitly calls for this, and showing that explaining away emerges even when the competing effect of the two parameters could be distinguished only indirectly by relying on more nuanced second-order statistics such as the dynamics of observations.

Explaining away effects can be described by the concept of contextual modulation. For example, in a recent study, Heald and colleagues showed that observers could quickly switch

between different previously stored contexts in a sensory-motor learning task (Heald et al., 2021). In particular, Heald and colleagues investigated whether, and if so then how the brain could choose between gradually learning a new interpretation (either by slowly adapting the current parameter combination or by abandoning the present combination and find a new combination from tabula rasa) vs. reactivating an already stored parameter setting representing an earlier context. In this framework, context is defined at the level of parameters by a vicinity of a particular set of these parameters and the input dynamics provide a cue to decide whether to stay in the current context or to change (i.e. weather to adjust or completely reset a single parameter while storing the old one). In contrast, in our case, the dynamics of the parameters set up the context at a higher level by determining which parameters should be used to explain the observed changes (i.e. which parameters to adjust among the many) without specifying what values these parameters need to take for the explanation. Both of these two complementary processes are likely to be present and cooperate in human decision making in a format where the larger decision situation modelled in our work determines the context in which the more specific mechanism modelled by Heald and colleagues completes the adjustment of the parameters.

In my work, I focused on the decision behavior after an individual Change Point (CP) and identified a new effect that influenced this behavior across a couple of hundred trials after the CP. Being an effect that depends on multiple trials, a potential candidate to explain our effect is short-term serial effects (STSEs) identified in previous studies and also present in our sequential decision-making experiments. In the literature, STSEs have been explained in various ways and through different assumed mechanisms including sensory adaptation, hot-hand, balancing between local and global statistics, Bayesian adaptation (Cicchini et al., 2024) and resource rational computation (Prat-Carrabin et al., 2024). In these proposals, STSEs have been linked both to attractive biases of previous decisions, (e.g. hot hand bias) and to repulsive effects of previous stimuli (due to e. g. adaptation) (Bosch et al., 2020). In our experiments, we observed a strong attractive bias, which decreased over time but remained constant across all experimental conditions and was independent of the long-term effects induced by CPs (see Supplementy, Section B.1). Thus, this attractive STSE lasting for 5-7 trials we found can explain neither the long-term speed-dependent biases nor the non-intuitive negative bias

that we observed in Exp1. or the between-experiment variations of biases. Our results are not compatible with another proposed explanation based on a balancing mechanism either that was hypothesised to explain short-to-intermediate term effects in a tilt estimation task by bringing the statistics observed in the recent past closer to the statistics of long-term past (Chopin and Mamassian, 2012), because that mechanism evokes a repulsive STSE in contrast to what we found in our experiments. In sum, none of the proposed mechanisms the literature can provide a viable alternative to our complex HBM to capture the effects reported in this chapter. It should be noted that, in contrast, our complex HBM can in principle, capture the emerging short-term effects, since the observed waning attractive STSE can be induced by the Bayesian update of the drifting AP parameter.

3.7 Methods

3.7.1 Stimuli and Procedure

Two shapes out of a set of 11 were randomly selected for each participant to serve as the discrimination stimuli. On each trial, the stimulus of size of 204*204 pixels ($\approx$ 4.7 visual angle) was presented centrally (circa 4.7 visual angles) on an iMac 27" (2560*1440) using Psychophysics Matlab toolbox. Participants watched a screen in a dimly lit room at a viewing distance of 60 cm and used the left/right buttons of the keyboard to provide responses. Instructions emphasized accuracy over reaction times but did ask for timely responses as there was an upper limit of four seconds to respond. The instructions did not mention stimulus probabilities nor changes in task structure. Trials were presented on a grey background display within a blue "box", a 256*256 pixels large blue square, spanning approximately 5.7 visual angles. On each trial, a shape embedded in Gaussian noise was presented for 200 ms, while the thin frame of the box (12 pixels wide) remained visible. After the stimulus disappeared, the center of the box turned white until the participant responded. After the response it reverted to blue until the next stimulus. The interval from the response to the next stimulus (RSI) was sampled randomly from a normal distribution with mean=1100 ms and SD=100 ms. During the training block, negative feedback was given after each mistake (in the form

of red exclamation marks) and no feedback after correct responses. During the test block there was only feedback if participants made a mistake on the 1/8th lowest noise trials in order to maintain attention (performance was over 90% in these trials, totaling to approximately 1% of test trials with feedback). At each trial, varying levels of Gaussian noise were added to the grayscale stimulus (Fig 4.1B). The training started at a low noise level, and the variance of the Gaussian noise was gradually increased with a "2up/2down" adaptive staircase procedure. The training lasted for 180-200 trials, to have an estimate of discrimination threshold. After 15-30 seconds of break, a 300-trial long test phase followed, where the Gaussian noise was sampled uniformly-randomly between low noise and the threshold reached during training.

3.7.2 Participants

189 Hungarian students (18-30 year old) participated in 8 experiments (between 18 and 34 in each experiment) and received monetary compensation. The participants gave informed consent before the start of the experiment and were unaware as to the purpose of the study.

Subject inclusion criteria

We used weak inclusion criteria to exclude from further analysis participants who were likely inattentive to the task or incorrectly used the response button. Participants data need to met two inclusion criteria: 1) There had to be a significant positive Pearson correlation ($p < 0.05$) between the responses, r, and the signed stimulus strength $y \cdot z$. 2) The Pearson correlation between the stimulus strength y and the correctness of responses $r = z$ had to be positive for both $z = 1$ and $z = -1$. Altogether 11 subject out of 189 couldn't meet the criteria, the average inclusion rate was 0.94 ± 0.05 ($\pm$std) in the eight experiments, and the lowest inclusion rate was 0.86 in Exp1.

3.7.3 Psychophysical analysis

We estimated the behavioral biases of participants by employing hierarchical logistic regression, simultaneously fitting all participants' data from all experimental conditions using the

numpyro probabilistic programming language (Phan et al., 2019). This allowed us to separate long-term behavioural biases (which depend on experimental condition) from short-term serial effects (which are the same in all experiments).

In the model, the behavioural bias of participant p within the i^{th} half of the Test phase was defined as the constant coefficient ($\beta^{p,i}_{bias}$) of the logistic function (S) fitted to the participant's responses:

$$P^{p,i}(r_t = 1|y, z) = (1 - \lambda^p) \cdot S\Big(\beta^{p,i}_{bias} + \beta^p_{stim} \cdot \frac{z_t}{y_t} + \beta_{past} \cdot \mathrm{R_t}\Big) + \frac{\lambda^p}{2} \tag{3.3}$$

This equation establishes a relationship between the probability of a "frequent" response ($r_t = 1$) in trial t and two predictors: the signed strength of the current stimulus (z_t/y_t), and the weighted sum of the past 30 responses ($\mathrm{R_t}$). The later term accounts for short-term serial effects, which were consistently observed in the data regardless of experimental manipulations (see Section B.1). However, when computing the overall bias for Fig. 3.1C, the β_{past} coefficient was set to 0. Otherwise, the impact of past responses was incorporated using exponential discounting with a decay parameter, τ:

$$\mathrm{R_t} = \sum_{n=1}^{30} \frac{e^{-\tau \cdot n}}{\sum_{n=1}^{30} e^{-\tau \cdot n}} \cdot r_{t-n}. \tag{3.4}$$

To account for potential attentional lapses, on λ_p proportion of the trials responses were selected randomly with equal probability instead of relying on the logistic function S.

Considering that short-term effects exhibited similar patterns across different experimental conditions, we assumed that the two corresponding parameters, $\{\beta_{past}, \tau\}$, were shared among all participants across all experimental conditions. The rest of the parameters were independent across participants $\{\beta^p_{stim}, \lambda_p\}$ and in case of the bias parameter,$\beta^{p,i}_{bias}$, also for the two halves of Test phase. To express the assumption that the bias is primarily determined by the experimental condition, we incorporated an experimental condition-dependent prior for the bias parameter:

$$\beta^p_{bias} \sim \mathrm{Normal}(\beta^e_{bias}, \sigma) \tag{3.5}$$

whose standard deviation (σ) was fitted to the data.

3.7.4 Ideal observer analysis

Ideal observer model

We modelled the participants' decisions with the ideal observer model illustrated in Fig. 3.2B. In each trial t, the ideal observer computes the posterior distribution over stimulus categories ($z_t \in \{\pm 1\}$) based on the noisy observation in the current trial (x_t) and the prior beliefs parameterized by AP_t & RV_t. Since the prior parameters are latent parameters in the hierarchical model, their values also need to be inferred based on all the previous observations, $x_{1:t-1}$. When computing the posterior of z_t, the prior parameters are integrated out as follows:

$$P(z_t|x_{1:t}) = \iint \mathrm{d}AP_t \, \mathrm{d}RV_t P(z_t|x_t, AP_t, RV_t) \cdot P(AP_t, RV_t|x_{1:t-1}) \tag{3.6}$$

This posterior distribution forms the basis of the ideal observer's decisions ($r_t \in \{\pm 1\}$).

Learning (in the dynamic model)

We treated the ideal observer as a Bayesian learner, optimally updating its prior parameters with the noisy observations according to its internal model. We put the assumption into the internal model that the prior parameters, AP_t & RV_t, change over time independently of each other, following separate first-order Markov processes characterized by the transition probabilities, $P(Q_t|Q_{t-1})$, $Q \in \{AP, RV\}$. With the Markov assumption, the posterior distribution of the prior parameters are computed iteratively:

$$\begin{aligned} &P(AP_t, RV_t|x_{1:t}) \propto P(x_t|AP_t, RV_t) \cdot P(AP_t, RV_t|x_{1:t-1}) = \\ &= P(x_t|AP_t, RV_t) \iint \mathrm{d}AP_{t-1} \, \mathrm{d}RV_{t-1} P(AP_t|AP_{t-1}) \cdot P(RV_t|RV_{t-1}) \cdot P(AP_{t-1}, RV_{t-1}|x_{1:t-1}) \end{aligned} \tag{3.7}$$

The transition probabilities were defined as the mixtures of a beta and a Dirac-delta distributions:

$$Q_t \sim \nu \cdot \mathrm{Beta}(\mathrm{mean} = Q_{t-1}, \mathrm{std} = s) + (1 - \nu) \cdot \delta_{Dirac}(Q_{t-1}) \tag{3.8}$$

The specific type of parameter dynamics determined the parameters of the mixture distribution. For drift diffusion dynamics, ν was fixed to 1 and s was small (relative to its theoretical maximum). For CP dynamics, s was large and ν was much smaller than 1 (see Supplementary). In in the graphical model of Fig. 3.2B, the abstract D_{AP} and D_{RV} hyperpriors stands for the parameters of the mixture distributions.

Noisy observations

We characterize the stimulus with a single real number, μ_x (in this part, we omit the trial index, t, for simplicity), whose sign is determined by the stimulus identity, z, and whose absolute magnitude increases monotonically with stimulus strength, $s = 1/y$ (defined as the inverse standard deviation of the Gaussian pixel noise, y):

$$\mu_x(y, z) = z \cdot \left(\mu_{\mathrm{amp}} \cdot \left(\frac{s(y) - s_{\min}}{s_{\max} - s_{\min}} \right)^{\alpha} + \mu_{\min} \right) \tag{3.9}$$

Here, $s_{\min} = \min(s)$, $s_{\max} = \max(s)$, and μ_{amp}, $\mu_{\min}$ and α are three scalar parameters that scale the stimulus strength.

To capture the imperfections of the sensory system (e.g. internal noise), we model the sensory observation, x, as a random sample drawn from a unit standard deviation normal distribution centered on the stimulus value, μ_x:

$$x \sim \mathrm{Normal}(\mu_x(y, z), 1) \tag{3.10}$$

Note, that we don't loose generality by setting the standard deviation to one, since the signal-to-noise ratio can be arbitrarily adjusted by appropriately setting the parameters μ_{amp} & $\mu_{\min}$.

While we, as experimenters, need to represent separately the functions $P(y)$ and $P(x|y, z)$ to leverage our knowledge about the noise when fitting the model parameters, the participants, who do not have access to y, can suffice with representing only the integral of the two func-

tions, $P(x|z)$, implicitly marginalizing out the unknown y. We parameterize this likelihood with RV in such a way that

$$P(x) = \sum_{z=0}^{1} P(x|z; RV = 0.5) \cdot P(z; AP = q) = \sum_{z=0}^{1} P(x|z; RV = q) \cdot P(z; AP = 0.5) \tag{3.11}$$

where

$$P(x|z; RV = 0.5) = \int \mathrm{d}y P^{\star}(x|y) \cdot P^{\star}(y) \tag{3.12}$$

and $P^{\star}(y)$ is the experimental noise distribution and $P^{\star}(x|y)$ is the estimated noise likelihood from the model fitting. The $P(x|z; RV = q)$ distributions are numerically computed following the method discussed in the Supplementary Information (Section B.4.1).

While $P^{\star}(y)$ was continuously changing during the training due to the adaptive staircase method, moreover underwent an abrupt change at the CP, we could not be sure that participants could learn its Test distribution perfectly. Therefore we allowed it to deviate from the uniform distribution. Instead of the uniform $P^{\star}(y)$ distribution we used $P^{\star}(s)$

With the above parametrisation $\{AP = q, RV = 0.5\}$ and $\{AP = 0.5, RV = q\}$ generate the same $P(x)$, and roughly the same is true for each combination where $AP + RV = q + 0.5$ and $q \in [0.5, AP]$.

Response model

Ideally, participants would always choose the option with the higher posterior probability (maximum a posteriori, or MAP decision), because this strategy would maximize the rate of correct responses. This is the strategy that we used for simulating the responses of the dynamic model. However, when fitting the static model we allowed two deviations from the ideal response strategy: First, to model the potential attentional lapses of the participants, on λ proportion of the trial the model choose the responses randomly instead of relying on the posterior. 2) To account for the short-term serial effects that were present in the data, but were independent from the experimental conditions (see Supplementary), we transformed

the log posterior ratio, forming the basis of decision, by adding a past-dependent term (R_t) to it:

$$P(r_t = 1|x_t) = (1 - \lambda) \cdot \frac{1}{1 + \exp\left(\beta(P(z_t = 1|x_t) - P(z_t = 0|x_t) + \kappa\, R_t)\right)} + \lambda \cdot 0.5 \quad (3.13)$$

R_t was the term that we used for fitting the psychometric curve defined in Eq. 3.4, and β was set to 100 making it virtually equivalent to choosing the maximum of the transformed log posterior ratio.

Model fitting (of the static model)

We employed hierarchical Bayesian inference (using `numpyro` probabilistic programming language; Phan et al., 2019) to estimate the parameters of the static model. This hierarchical model was fitted to data from all participants across all experimental conditions at once. The parameters AP, RV and λ were estimated individually for each participants, while all the remaining parameters, $\{\mu_{\text{amp}}, \mu_{\text{min}}, \alpha, \tau, \kappa, \beta\}$, were shared across participants and conditions. Sharing the parameters allowed us to independently estimate the long-term biases associated with AP and RV from the STSE, and made the inference more robust overall. We searched for the $\{AP, RV\}$ parameter combinations on the maximum-likelihood parameter ridge (to ensure model identifiability) defined by the linear equation $AP_{GT} = AP + RV - .5$, where AP_{GT} is the ground truth appearance probability of the experiment.

When fitting the reaction time data, using both the simple and complex models, we allowed a simple one-parameter transformation to the AP-RV parameter pairs, applying the same transformation for all participants across all sessions (see Section B.5). We justified this because there was an experimental parameter that participants might have misjudged, and we could not reliably infer their beliefs about it either. The one-parameter transformation captured the effect of potential misjudgment well, but still kept the constraints imposed by the AP-RV parameter estimates on the RT fit strong.

3.7.5 Bounded evidence accumulator model

To capture reaction times, we employed the bounded evidence accumulator (BEA) model with constant decision boundaries. This model posits that noisy evidence is accumulated over time, and a decision is made when the accumulated evidence reaches one of the decision boundaries at 0 or a, corresponding to the two possible choices, $r = -1$ or $r = -1$, respectively. The accumulated evidence at trial t and time τ is defined by the integral

$$e_t(\tau) = w_t + \int_0^\tau (v_t + \epsilon_\tau)\,\mathrm{d}\tau \tag{3.14}$$

Here, w_t is the starting point of accumulation, v_t is the accumulation rate, and ϵ_τ is the momentary noise drawn from a standard normal distribution independently in each time bin.

The parameters w_t and v_t varied from trial to trial based on the signed strength of the current stimulus ($\frac{z_t}{y_t}$) and the past decisions (R_t, Eq. 3.4). Specifically:

$$w_t = S(w_{bias} + \kappa_w \cdot R_t) \tag{3.15}$$

$$v_t = a \cdot (v_{bias} + v_{amp} \cdot y_t \cdot z_t + \kappa_v \cdot R_t) \tag{3.16}$$

In line with (White and Poldrack, 2014), the bias of starting point only depends on AP and the bias of accumulation rate only depends on RV. We assume linear relationship between the abstract Bayesian parameters and the bias terms:

$$w_{bias} = w \cdot (AP - 0.5) + 0.5 \tag{3.17}$$

$$v_{bias} = v \cdot (RV - 0.5) \tag{3.18}$$

The distribution of accumulation termination times were computed from w_t, v_t and a using the numerical method described in (Navarro and Fuss, 2009) (see Supplementary). Reaction

times were modelled as the accumulation termination time plus a non-decision time $t_{nd} \sim \mathrm{LogNorm}(\log(t_0), \sigma_{t_0})$.

We fitted the BEA model simultaneously to all participants in all experiments using the BADS optimizer in Matlab. All free parameters of the model were identical for all participants, so all the variations across participants were attributed to the estimates of AP and RV, which were obtained from the independent static Bayesian model fits.

Chapter 4

Identifying uncertainty representations in early visual cortex

In this chapter, I move beyond purely behavioural analysis and explore the neural traces of posterior representations. In a fully Bayesian brain, uncertainty about all latent variables of the internal model would be represented in the neural activity, potentially encoded in (partially) distinct neural populations. Here, I introduce a novel data-driven approach – developed in collaboration[1] – that was designed to test which latent variables' posterior distributions (if any) are encoded in specific populations. This approach combines model-based analyses of behavioral data and population decoding analyses of simultaneously recorded neural data. I also present proof-of-concept results from applying this analysis approach to mouse primary visual cortex (V1) calcium imaging data recorded during a perceptual decision making task. In line with the topic of the thesis, we[2] aimed to identify the traces of behaviourally estimated perceptual posteriors in the neural activity, distinct from those of decision posteriors. After the initial validation of the method on synthetic data, we found preliminary evidence for the representation of such perceptual posteriors in mouse V1. Upon extensive validation of the method and further confirmation of the results, this work could offer the first neural evidence supporting the fully Bayesian brain hypothesis.

[1] For details regarding the distribution of contributions, please refer to the Declaration of Authorship at the beginning of the Thesis.

[2] In this chapter, where I want to emphasise the joint contribution, I will use the first-person plural.

4.1 Introduction to the approach

The ultimate product of the brain is behaviour; thus, neural activity is best understood in light of the behaviour it brings about (Krakauer et al., 2017). Guided by this principle, we developed a novel two-staged approach to search for the neural representations of uncertainty and tested it on mouse neurophysiological data obtained from a direction discrimination task (Fig. 4.1, top).

The first stage of this approach focuses solely on the behaviour of the animal, temporarily setting neural data aside. In this initial step, a Bayesian ideal observer – fitted to the trial-by-trial behaviour of individual subjects – computes the posterior distributions for two latent variables on each trial: a low-level perceptual variable (the orientation of the stimulus; Fig. 4.1, bottom right) and a high-level decision variable (which of the two responses is correct; not shown). In the second stage, these 'behavioural' posteriors are used as targets for different population decoder algorithms that are applied to neural responses recorded in the corresponding trials. Acknowledging the existence of competing hypotheses that differs on how neural activity might represent probability distributions, two distinct decoders were developed, corresponding to the dominant theories of probabilistic representations: (1) one that uses the within-trial temporal patterns of responses (Fig. 4.1, bottom left), as suggested by the neural sampling hypothesis (Fiser et al., 2010), and (2) another that uses the spatial pattern of responses across the population (Fig. 4.1, bottom middle), as suggested by theories e.g. probabilistic population codes (Ma et al., 2006) and distributed distributional codes (Vértes and Sahani, 2018). By quantifying the decodability of posteriors, this approach allows testing of which variables' uncertainties are represented in neural activity, and how.

A key advantage of this method, compared to previous approaches (Walker et al., 2020), is that it use no neural data for estimating the subjects' posteriors. This avoids the circularity of reasoning that the 'target' perceptual posteriors would be inferred from the same neural data in which their existence is meant to be demonstrated. Instead, we inferred the internal representation solely from behaviour, assuming this representation is fully Bayesian, and only

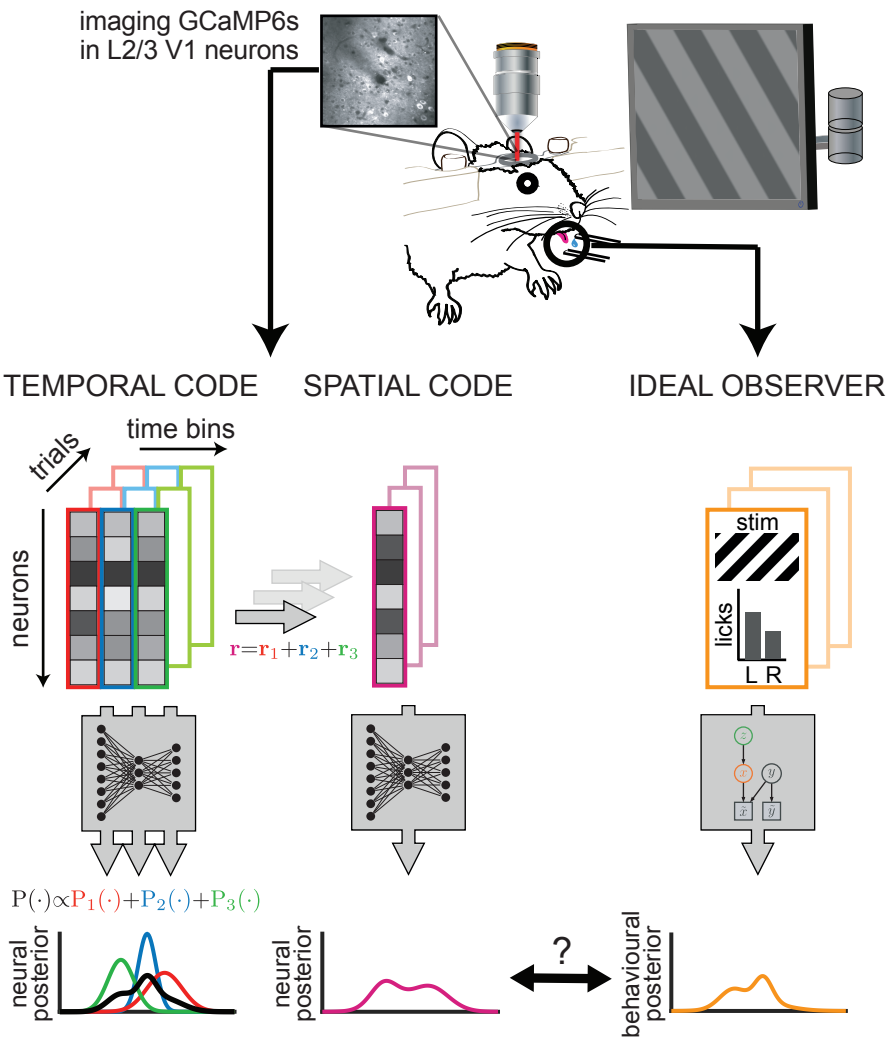


Figure 4.1. A data-driven approach for revealing the neural representations of uncertainty in mouse V1. This approach combines model-based analyses of behavioural data and population decoding analyses of simultaneously recorded neural data. **Experimental data:** Alongside the mouse's licking behaviour, the activity of layer 2/3 pyramidal cells in mouse V1 was recorded using calcium imaging during a 2AFC perceptual decision making task. **Behavioural analysis:** Based on the stimulus-response pair data (orange box), an ideal observer provided trial-by-trial estimates for the posterior distributions of two latent variables: a low level perceptual variable (orange distribution), and a high level decision variable (not shown). **Neural analysis:** The behavioural posteriors were used as targets for two population decoders (feedforward neural networks) differing in the assumption whether posterior distributions are represented by temporal (red, blue, green) or a spatial (purple) code (see details in Section 4.4).

in the subsequent step, leaving the raw behavioural data behind, we assessed whether the neural activity is consistent with the inferred representation.

Given that our primary goal was to identify the representations of perceptual posteriors, we applied this novel approach to data obtained from mouse primary cortex (V1), which area is a prime candidate for perceptual representations. Specifically, we recorded the activity of layer

2/3 pyramidal cells in V1. This recording region was chosen for two main reasons: First, it is rich in direction-selective neurons (Niell and Stryker, 2008), aligning well with the demands of our experimental task. Second, as one would expect from a candidate area for posterior inference – a computation involving the integration of priors (expectations of the internal model) with likelihoods (sensory evidence), with the results then being forwarded for downstream computations – layer 2/3 receives both bottom-up (via layer 4) and top-down inputs (directly and via layer 5), and it projects to higher-order (visual) areas (Lee and Mumford, 2003; Harris and Mrsic-Flogel, 2013).

By applying our data-driven approach to a specifically tailored data set, we created the first opportunity to identify the neural traces of the putative fully Bayesian brain model.

4.2 Experimental paradigm

Here, I summarize the most important details of the experimental paradigm essential for interpreting the results. For a more detailed description see Amvrosiadis (2023).

4.2.1 Two-alternative forced-choice (2AFC) visual discrimination task

In order to estimate the posterior distributions that are potentially represented in the neural activity of the mice, we developed a motion direction discrimination task. In each trial, the animal was presented with a moving grating, based on which it made a decision whether the movement direction was closer to 45° (left) or 135° (right) by licking one of two reward spouts that were placed on either side of the animal's mouth (Fig. 4.2A).

Experimental protocol: For this study, we analysed data from three animals, each undergoing several recording sessions over multiple days (one session per day). During these sessions, the visual stimuli was presented on a computer screen in the right visual field of the head-fixed animals, while their neural activity was measured with a dual two-photon laser microscope. The mice performed the task for water reward. To maintain the animals' moti-

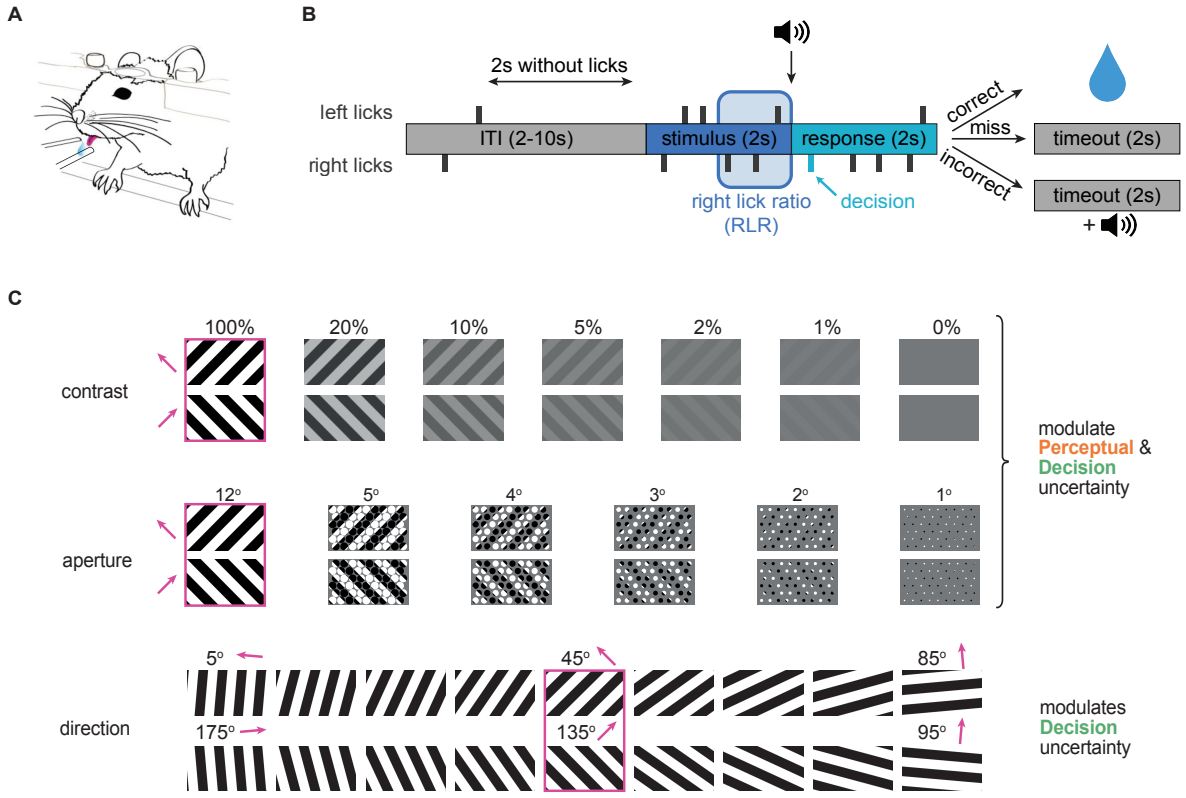


Figure 4.2. 2AFC direction discrimination task. A. Arrangement of the reward spouts in relation to the head-fixed mouse. **B.** Temporal structure of the trials and the role of licks at different trial stages. **C.** The stimulus set used in the experiment. Magenta arrows illustrate motion direction (note the unconventional notation that the angles increase counterclockwise) and magenta boxes highlight the 'target' stimulus pair.

vation, their general access to water was restricted, keeping them at 85-90% of their baseline body weight.

Trial structure: Fig. 4.2B illustrates the trial structure. Each trial started with a gray screen, during which the animal was required to withhold liking. After either successfully withholding licking for 2 seconds or a maximum wait time of 10 seconds (whichever came first), a moving grating was presented for 2 seconds. Mice were free to lick during stimulus presentation and the right lick ratio (RLR) during the second half of this period was used as an indicator of their decision certainty. Immediately following the stimulus period, the screen turned gray and a 2 sec long response period began, cued by an auditory stimulus (Pure tone, 500Hz, 500ms). The animal's choice was determined by its first lick during this period. A correct choice (licking into the direction congruent with the motion) was immediately rewarded with a 10 μl water droplet. In contrast, an incorrect lick triggered an aversive auditory tone

lasting 100 ms and resulted in a 2 sec timeout. If the animal did not lick at all during this period, the trial was classified as a "miss," and a 2 sec timeout was also imposed.

Stimulus manipulations: Fig. 4.2C presents all the stimuli used in the experiment, grouped according to the applied uncertainty manipulations. The difficulty of the trials was manipulated in three distinct ways, with only one manipulation type used per session. Starting with a pair of full contrast, unoccluded 'target' gratings moving either to the $45°$ and $135°$ directions (Fig. 4.2C, magenta frames), we introduced uncertainty by adjusting either 1) the contrast of the gratings in 7 steps, $\{0, 1, 2, 5, 10, 25, 100\%\}$; 2) the aperture of a grid mask (a 6x9 grid of apertures) partially occluding the grating in 6 steps, $\{1, 2, 3, 4, 5, 12°\}$ ($12°$ being the unoccluded grating); or 3) the movement direction in 18 steps, $\{5, 15, 25, ...175°\}$ ($90°$ being the decision boundary).

Importantly, while the contrast and aperture manipulations were designed to influence both perceptual uncertainty and, indirectly, decision uncertainty, the direction manipulations were intended to affect solely decision uncertainty, but leave perceptual uncertainty unchanged – i.e., perceptual uncertainty should remain invariant to direction manipulation (Walker et al., 2023). This partial decoupling of the two types of uncertainty was a crucial design feature for testing the fully Bayesian brain hypothesis, as it allowed for the potential experimental dissociation between perceptual and decision uncertainties.

We introduced two distinct perceptual uncertainty manipulations to test whether the perceptual posterior decoded from neural activity specifically represents perceptual uncertainty rather than the variables that contribute to it (contrast and aperture) – a crucial feature of proper uncertainty representations (Walker et al., 2023). However, we have not yet conducted this test.

Training: Prior to the recording sessions, the mice underwent several habituation and training sessions to familiarize them with the experimental setup, reduce their default side biases, and achieve a discrimination accuracy over 70%. During training, only the two 'target' stimuli (full contrast, full aperture, $45°$ or $135°$) were presented to prevent the mice from learning simpler decision strategies that wouldn't involve the representation of uncertainty.

4.3 Behavioural analysis

4.3.1 An ideal observer-based approach

Behavioural responses were analyzed using an ideal observer-based approach. This formalizes a generative model (GM) that describes the stochastic process by which stimuli are generated in the experiment (Fig. 4.3 A). Here, in each trial, the GM first samples the trial type determining the correct decision (left or right lick; Fig. 4.3A, z) , followed by the "true" stimulus orientation and "true" stimulus strength (Fig. 4.3A, y). Lastly, it generates noisy observations based on the true value of orientation and strength (Fig. 4.3A, $\tilde{x}$ and $\tilde{y}$). The formal mathematical description of the GM is in the Methods (Section 4.6.1).

By the direct inversion of this GM, on each trial, the ideal observer (Fig. 4.3B, inset) infers Bayesian posterior distributions (see the details in Section C.1) over two relevant latent variables (here: the orientation of the stimulus, and whether licking left or right is the correct response) based on the (potentially noisily perceived) stimulus (here: a drifting grating characterized by its direction and stimulus strength), and predicts the behavioral response (here: the rate of licking left vs. right) based on these inferences (Fig. 4.3B). The model assumes that the same licking strategy is maintained during both the stimulus period – used to compute the right lick ratios (RLR) – and the first lick of the response period, which determines the choice.

This model is fitted to stimulus-response pair data as inputs, and provides data-driven trial-by-trial estimates of the posteriors that the subject computes as outputs (Fig. 4.1, right).

4.3.2 Biases of the ideal observer

We allowed the generative model to be biased and suboptimal in various ways, which was necessary to account for complex patterns in the licking behavior that we will discuss in the result section. These biases were controlled by a total of 10 parameters, visually illustrated in Fig. 4.4A. The mathematical formalization is again in the Methods (Section 4.6.1).

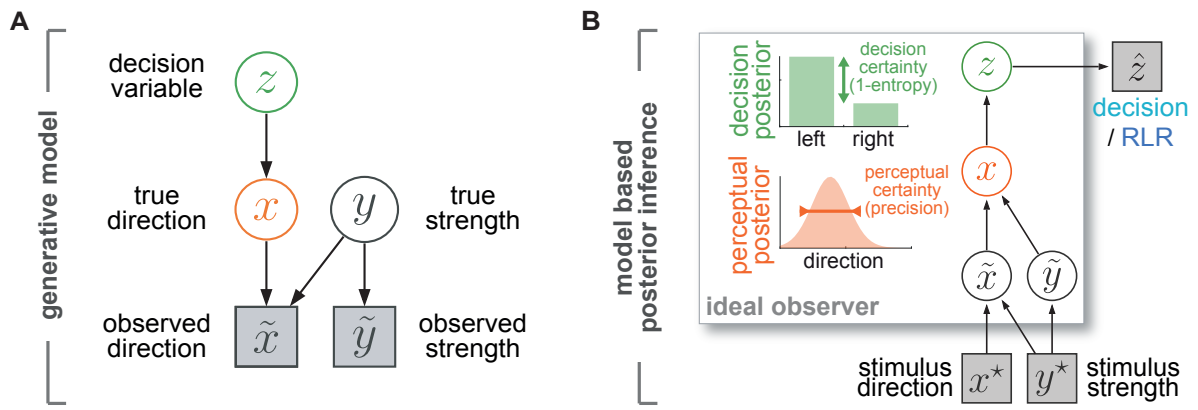


Figure 4.3. The ideal observer model. A. The graphical model of the stochastic generative process responsible for producing the stimuli and observations. **B**. Generative model-based inference of the perceptual (orange) and decision posteriors (green). The ideal observer's inference is an intermediate step in the process of lick generation ($\hat{z}$) based on the stimulus features ($x^\star, y^\star$).

Specifically, the sensory precision of the generative model (defined as the circular precision of the sensory likelihood) could vary significantly with the stimulus direction, being more precise around a specific direction than elsewhere (Fig. 4.4A, left). The relationship between the physical features that determine stimulus reliability (contrast and aperture) and the model's stimulus strength variable, y, could be non-linear and the overall observedness of y could range from perfect to zero (Fig. 4.4A, middle). Finally, the mapping between the decision posterior and the response could follow a highly asymmetric function with respect to the signed log posterior odds (Fig. 4.4A, right).

4.3.3 Results of the behavioural analysis

We analyzed the licking responses of three mouse subjects (in total 3053, 2252, and 985 trials, pooled across all their sessions), fitting the model to their licking behavior during the stimulus presentation phase of the trials. We only used data from the second half of the stimulus presentation, as by this time both the licking behaviour and the neural signal had stabilised. Prior to this, licking was dominated by the animal's default behavior that was triggered by the stimulus onset but was independent of the content of the stimulus and the neural signal exhibited strong stimulus-induced transients.

Inferred model parameters: For two out of three animals, the estimated model parameters indicated a strong direction dependence in their sensory precision (Fig. 4.4B, first panel), with much higher precision around the right target direction ($135°$) compared to other directions. Only one animal observed the stimulus strength to some extent (Fig. 4.4B, third panel), and all animals' appeared to exhibit response biases (Fig. 4.4B, third panel).

The identified strong sensory bias may seem surprising at first, but it might be justifiable for an observer with limited sensory resources (Lieder and Griffiths, 2020). In the presence of such limitations, optimizing resource allocation based on task demands can significantly improve performance, but potentially elicit strong biases (Wei and Stocker, 2015). One of our candidate assumptions is that, rather than distinguishing between the two motion categories ($45°$ vs. $135°$), the animals may have been performing a detection task – deciding whether the direction matched the rightward 'target' ($135°$) or not – and optimizing their perception based on the demands of this task. However, whether this strategy would account for the observed biases is a question for future work.

Validation of the fitting method: To verify the reliability of the parameter fitting method, we assessed the recoverability of the model parameters using synthetic data. First, a synthetic responses were generated for the actual experimental stimulus sets, either with random parameter settings or using the best-fitting parameters from the animal data. Then, the model parameters were estimated by fitting the model to the synthetic data in the same way as for the experimental data. Based on this test, parameter recovery was not always reliable when using random parameters (Fig. 4.4C, light grey dots), but the best fitting parameters of the animals were always recovered with high precision (Fig. 4.4C, red dots).

Beyond assessing the quality of parameter recovery, we developed an alternative method to evaluate the reliability of the model fitting by focusing on the feature most relevant to our purposes: how well the posteriors estimated from the recovered parameters (see Section C.3) matched the ground-truth synthetic posteriors. We made a pointwise comparison of the estimated and ground-truth posterior distributions (Fig. 4.4D, grayscale heatmap), and to better illustrate the trends, we also compared the average data binned according to the ground truth posterior values (Fig. 4.4D, colored dots). To quantify the alignment between the estimated

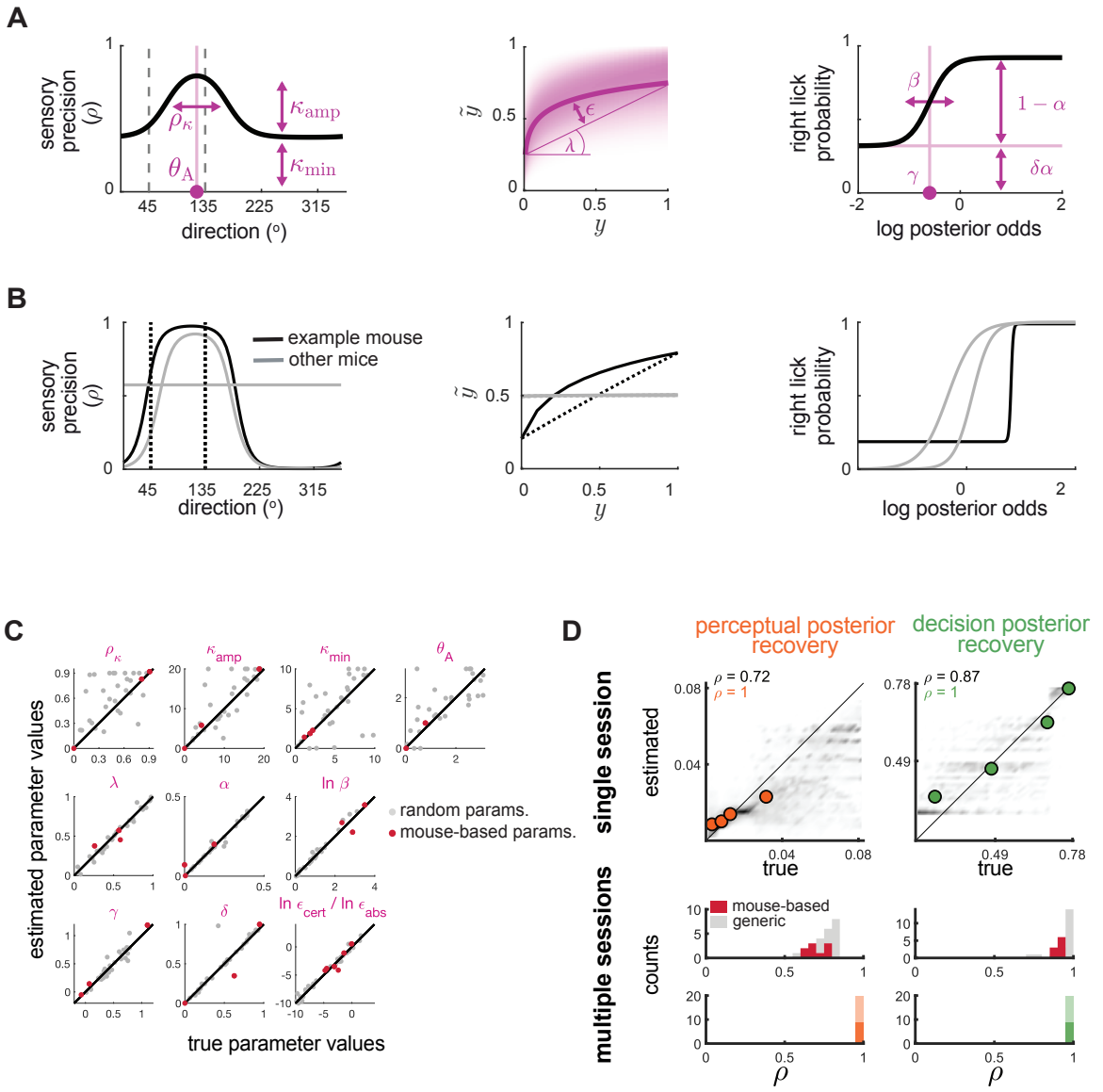


Figure 4.4. Behavioural model parameters and their recoverability. A. Visual representation of the model parameters. Left: sensory precision ($\rho = I_1(\kappa)/I_0(\kappa)$, see Methods) as a function of direction; middle: distribution of observed stimulus strength ($\tilde{y}$) given the ground truth stimulus strength (y) and its observedness (λ); right: licking model. **B**. Illustration of the best fitting models for the experimental data from three mice, one of them being the example mouse in Fig. 4.5B (black line). **C**. Comparison between the estimated model parameters and the true model parameters that were used to generate the synthetic data. True parameters were either randomly sampled (gray dots) or were the best-fitting model parameters for the experimental data (red dots). **D**. Upper row: Pointwise comparison of the estimated and true posteriors within a single session (gray heatmap), and their averages binned by the true posterior values (coloured dots). Lower rows: Histograms of the Pearson correlations of the heatmaps (upper row) and scatter plots (lower row) across multiple simulation sessions, corresponding to the parameter values in panel C.

and ground-truth posteriors, we calculated the Pearson correlation of the pointwise value pairs (Fig. 4.4D, gray-red histograms) as well as the correlation of their averages (Fig. 4.4D, colored histograms). The correlation between the true and estimated posteriors were gener-

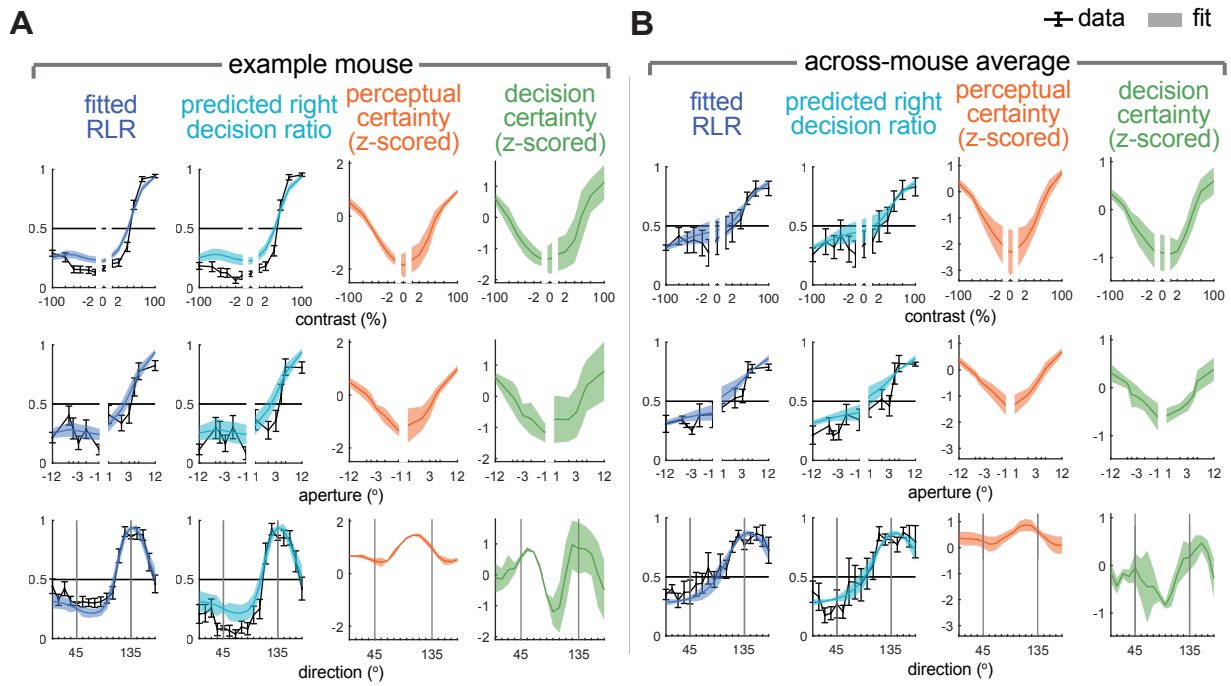
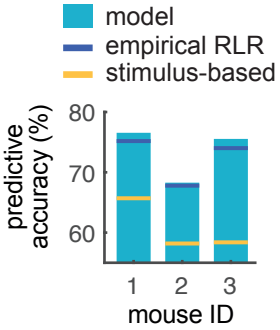


Figure 4.5. Behavioural fits. A. The average behavior across animals in response to distinct stimulus manipulations (black lines), along with the model predictions from the best-fitting models (shaded areas). The z-scored certainty estimates are purely model-based, so no black lines are shown for these plots. Error bars and shaded shading represent the across-animal standard error of the mean. **B**. Same as in A, but for the example mouse. Here, the error bars and the shading represent the across trials standard errors.

ally high, despite the inclusion of data with poorly recovered parameters ((Fig. 4.4D). These results show that the posterior estimates are generally reliable (at least provided that the model assumptions are close to the true generative model of the animal).

Fit quality: By fitting the model solely to licking during the stimulus presentation (Fig. 4.5A, far left), we were able to predict, in a cross-validated way, licking during the response phase of trials (Fig. 4.5A middle left).

Importantly, incorporating sensory bias into the model allowed it to explain complex patterns of the psychometric curves (Fig. 4.5B), which were particularly apparent for one of the mice (the example mouse in Fig. 4.4B). For example, there was a counterintuitive increase in the RLR of this mouse as the contrast of leftward-moving stimuli increased that was at least qualitatively reflected in the model behaviour as well (Fig. 4.5B, upper far left) . According to the fitted model, this animal accurately perceived rightward stimuli, but not so much leftward stimuli, sometimes even perceiving them as rightward (Fig. 4.4B, left). Since the animal par-



	model	empirical RLR	stimulus-based
M1	76.6	75.3	65.7
M2	68.3	67.9	58.2
M3	75.5	74.3	58.4

Figure 4.6. Ideal observer's predictive accuracy. Across period predictive accuracy of the model (light blue bars), the empirical RLR (dark blue line) and the stimulus-based approach (yellow line). The table shows the numerical values (in percentages) corresponding to the bar plot.

tially observed the contrast (Fig. 4.4B, middle), when it mistakenly perceived a high-contrast leftward stimulus as rightward, the observed high contrast boosted its confidence in this incorrect perception, leading to an increased proportion of rightward lickings. The occasional misperceptions of high-contrast leftward stimuli (related to the strong sensory biases) were enough to create the unexpected contrast pattern.

Stimulus dependence of uncertainty: After fitting the model, internal estimates of perceptual certainty (the precision of the x posterior), and decision certainty (the negative entropy of the z posterior) varied with stimulus attributes as expected: contrast modulated both perceptual and decision certainty (Fig. 4.5A-B, right, top and middle), while orientation mainly modulated decision certainty (Fig. 4.5A-B, right, bottom). Based on this result, dissociating between the two types of uncertainty seems plausible.

Quantification of the predictive performance: We predicted the animals decision during the response period based on the subjective distributions we inferred from their behaviour during the stimulus period. To quantify the predictive accuracy of this model-based approach, we computed the *probabilistic fraction correct* metric proposed by Houlsby et al. (2013). This measure is the geometric mean of the predictive probability assigned by the model to the subject's actual response in each trial. To justify the validity of our model-based approach, we compared its predictive performance to the maximal achievable performance of alternative approaches that do not infer posteriors, but rely solely on either the empirical RLR or the stimuli alone for making predictions.

The predictive accuracy of the model-based approach exceeds by a large margin the accuracy of the stimulus-based approach and showed a slight improvement over the empirical RLR-based approach for all animals (Fig. 4.6). Although this advantage is small, it supports the use of model-based inference for estimating subjective distributions.

4.4 Neural analysis

Neural responses were analyzed using a population decoding approach. On each trial, the input to the decoder is the set of time-resolved responses of a population of neurons (Fig. 4.1B, left, top). These responses are analyzed in two complementary ways. For evaluating a *temporal code*, we apply a neural network decoder for each time bin in a trial separately (Fig. 4.1B, left, top; red, blue, and green time windows). Although we use a neural network-based decoder for flexibility, critically, it is the same decoder that we apply to each time window (Fig. 4.1B, left, middle). Thus, neural responses in each time window are decoded separately into individual distributions (Fig. 4.1B, left, bottom; red, blue, and green distributions) and are averaged to yield a final "neural" posterior (Fig. 4.1B, left, bottom; black distribution – this is a smoothed generalization of a simple Monte Carlo representation, in which individual distributions would effectively be deltas). For evaluating a *spatial code*, we simply swap the decoding and averaging steps: we first average (or sum) responses for each cell within a trial (Fig. 4.1B, center, top; magenta frame) and then decode this purely spatial pattern of activities directly into a single neural posterior (Fig. 4.1B, center, bottom; magenta distribution). Note that neural decoders for temporal and spatial codes map between the same kind of input (a single vector of neural activities) and final output (a neural posterior) for temporal and spatial codes, and are thus chosen to be identical in their architecture and complexity (number of parameters) for a fair comparison. Finally, in either case, the decoder is trained to match the neural posterior to the ideal observer's behavioral posterior trial-by-trial, by minimizing the average discrepancy between the two distributions, as specified by a summary statistic-based loss-function.

Optimisation loss

At first, we used the average KL divergence across trials (t) to quantify the discrepancy between the behavioural ($\mathrm{P}^{\mathrm{behav}}$) and neural posteriors ($\mathrm{P}^{\mathrm{neur}}$):

$$\mathcal{L} = \frac{1}{T} \sum_{t} \mathrm{KL}[\mathrm{P}_t^{\mathrm{neur}} \,||\, \mathrm{P}_t^{\mathrm{behav}}] \tag{4.1}$$

However, this approach was later replaced as it does not align well with the characteristics of the temporal code.

Constructing a loss-function that is also suitable for a temporal code is not straightforward. This is because a temporal code may, in the limit, imply a sampling-based code, in which each momentary component distribution is a delta distribution, such that the full distribution is a mixture of deltas. However, standard information theoretic measures of mismatch between two distributions (in our case: the behavioural and neural posteriors), such as the Kullback-Leibler divergence and other measures derived from it, give degenerate results when the optimised distribution is a mixture of deltas. Therefore, we opted for a 'projection pursuit' approach instead that was based on matching a set of summary statistics between the two distributions. These linear projection-based summary statistics (i.e. integrals of specific, potentially nonlinear functions over the distributions) could be readily computed even for sample-based representations (essentially yielding Monte Carlo estimates of the integrals). Furthermore, arguably, any posterior represented in the brain ultimately serves the purpose of such integrals being computed over it, when being marginalised out for computing expected losses for decision making, or sufficient statistics for learning. Thus, a summary statistic-based loss may also be considered more 'ecological'.

One complication with using a summary statistic-based loss, compared to using an information theoretic loss, is that it requires the specification of the particular set of statistics that are to be matched, and this choice will inevitably entail some degree of arbitrariness. Here, in order to make minimal assumptions, we chose an 'unsupervised' approach. We reasoned that those statistics would be most important for matching the neural to the behavioural posteriors that best discriminate among the behavioural posteriors themselves. We computed the set of orthogonal functions that best discriminated the behavioural posteriors in terms of maximising the average squared difference between their corresponding summary statistics

(i.e. the integral of the function over the different posteriors). We achieved this by performing principal component analysis on (the discretized version of) the posteriors. (In other words, we computed the eigenvectors of the covariance matrix of $\mathrm{P}^{\mathrm{behav}}$.) As a result, each principal component (PC_i) corresponded to a function whose summary statistic could be used to discriminate between behavioural posteriors. Intuitively, the larger its associated eigenvalue (α_i) was, the better the function that it represented discriminated behavioural posteriors by its corresponding summary statistic – and the more relevant it was assumed to be from the perspective of the task. We thus used a loss-function for optimizing the decoders that penalized the discrepancy between the decoded neural posterior ($\mathrm{P}^{\mathrm{neur}}$) and the target ($\mathrm{P}^{\mathrm{behav}}$) along each PC_i commensurate with its behavioural relevance as measured by α_i.

This loss-function consists of two terms expressing two independent objectives:

$$\mathcal{L} = \frac{1}{T} \sum_t \sum_i \alpha_i \left(\langle \mathrm{P}_t^{\mathrm{neur}}, \mathrm{PC}_i \rangle - \left\langle \mathrm{P}_t^{\mathrm{behav}}, \mathrm{PC}_i \right\rangle \right)^2 + \lambda \mathrm{H}_m \left[\mathrm{P}^{\mathrm{neur}} \right] \tag{4.2}$$

The first *mismatch* term penalises the average discrepancy between $\mathrm{P}_t^{\mathrm{neur}}$ and the corresponding $\mathrm{P}_t^{\mathrm{behav}}$ across trials ($t \in T$), such that the discrepancy is the average squared difference between the projections of the posteriors onto the PC_is, weighted by the α_is.

The second *entropy penalty* term penalises the average entropy of the momentary distributions, thereby incentivising the momentary distribution to be narrow, akin to the point-like Monte Carlo samples. To ensure fairness, this penalty term is also applied to the spatial code.

Normalized mismatch

We evaluate the combination of a latent variable and a neural code by computing the resulting mismatch between the behavioral posteriors for the given latent variable and the neural posteriors for the given neural code after optimization. As the posteriors of different candidate latent variables may have fundamentally different complexities (or even supports), for a fair comparison, we normalize the mismatch by the average mismatch between every

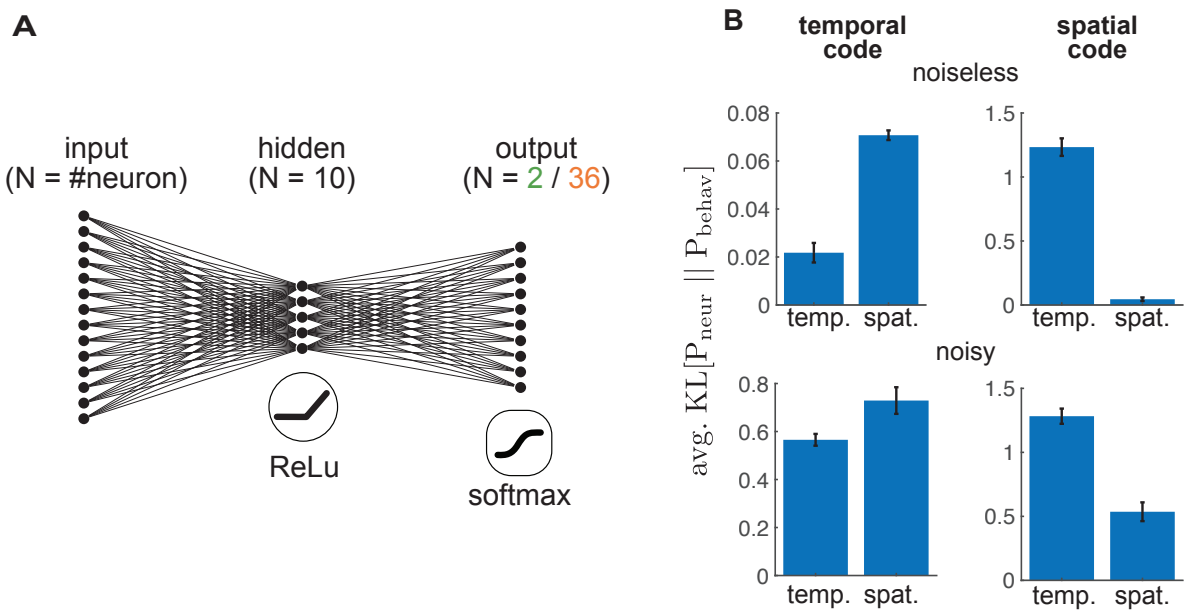


Figure 4.7. Neural network decoder. A. Illustration of the feedforward neural network decoder. **B**. Validation of the discriminative power of the neural network decoder trained with KL-loss. Bar plots show the average KL divergence between the neural and behavioural posteriors averaged across five runs ($\pm$ std).

combinations of the behavioural posteriors:

$$\mathcal{M} = \frac{\frac{1}{T}\sum_{t}\sum_{i}\alpha_i\left(\langle \mathrm{P}_t^{\mathrm{neur}}, \mathrm{PC}_i\rangle - \langle \mathrm{P}_t^{\mathrm{behav}}, \mathrm{PC}_i\rangle\right)^2}{\frac{1}{T^2}\sum_{t,t'}\sum_{i}\alpha_i\left(\langle \mathrm{P}_t^{\mathrm{neur}}, \mathrm{PC}_i\rangle - \langle \mathrm{P}_{t'}^{\mathrm{behav}}, \mathrm{PC}_i\rangle\right)^2} \tag{4.3}$$

Neural decoder details:

We opted for artificial neural networks as our decoder due to their flexibility. Specifically, we used feedforward neural networks (Fig. 4.7A) with a single hidden layer of 10 fully connected ReLU neurons and a Softmax output layer (36 categories for decoding x posteriors, and 2 for z posteriors). The networks were trained running the Adam optimization algorithm from 5 initial conditions, for 40 epochs each, and we used a 1:1 train:test split.

(Preliminary) validation of the neural decoder:

We aimed to evaluate each latent variable - neural code combinations to identify which one of them explains best the neural activity of the mouse V1. To prove that our method is indeed capable of this, the first step would be its validation on synthetic data. However, up until now,

I have only validated the KL-based loss and I have only tested how well it can discriminate between the temporal and spatial codes. The summary statistic-based loss, which was actually used for data analysis, and whether the method can discriminate between perceptual and decision posteriors, have yet to be validated.

For the validation, I generated spike data from 50 Poisson neurons with circular Gaussian shaped tuning functions ($e^{\cos(x-mu)-1}$) uniformly spacing the $[0°, 360°]$ direction range, with their amplitudes linearly scaled by the stimulus strength ($y \in [0, 1]$). The 'sensory' input to these neurons reflected the statistics of the experimental stimulus set.

In each trial, I generated spikes for a unit time, and divided the spikes into three equal time bins (as in the analysis of animal experiment). I then decoded the posterior distribution of the input direction using the optimal decoder (Abbott and Dayan, 2001, p. 16-23), both from the averaged activity and from the time-resolved activity of the neurons.

To simulate different coding schemes, I set the 'behavioral' target posterior in two ways: (1) as the posterior distribution decoded from the cumulative activity, mimicking a spatial code, or (2) as the sum of the distributions decoded from the time-resolved activities, mimicking a temporal code. Finally, I tested how well the neural network decoder could recognize the target distributions from either the time-resolved or cumulative synthetic activities. I included a version of models with added Dirichlet noise (symmetric, $\alpha = 3$) to assess the robustness of the method.

This limited validation demonstrated that the neural network-based decoding approach has a potential to distinguish between temporal and spatial neural codes with high accuracy, tough there is a slight bias against temporal codes (Fig. 4.7B). This validation, needs significant improvements in the future, for which I provide suggestions in the discussion section.

4.4.1 Results of the neural analysis

Data

The inputs to the decoder were dF/F responses of L2/3 neurons in V1, recorded using GCaMP6s imaging (Henschke et al., 2020). We used the data recorded in the second half of the stimulus presentation, which coincided with the time window used for the behavioural RLR computation. Omitting the first half of stimulus presentation from the analyses had two additional benefits: it excluded the stimulus-evoked transient activity, and since it is reasonable to assume that processing motion (even if stationary) requires at least a brief evidence accumulation phase, this phase was likely omitted as well, preventing it from being confounded with the temporal uncertainty code (see Chapter 2, Noise vs. Signal model). For the temporal code, trials were divided into 3 time bins (in the animal experiment, the autocorrelation of the signal allowed for at most three independent samples per trial).

To compare perceptual and decision posteriors, we needed neurons that were recorded in at least two different session types corresponding to the two types of uncertainty. This required matching the identified neurons across sessions, a process that was largely manual. To support and verify this manual procedure, I developed an automated validation method (see Section 4.6.3). However, no neurons were recorded across all three session types, therefore we excluded aperture modulation from the neural analyses. To maximize the number of neurons analyzed, we used data from the single direction-contrast session pair for each mouse that had the highest number of matched neurons. After matching, we were left with 277, 289, and 229 trials and 59, 57, and 132 neurons for the three tested mice, respectively.

Results

We first performed the principle component analysis (PCA) of the behavioural posteriors (Fig. 4.8A), then we fitted all together eight neural decoders (2 latent variables and 4 coding schemes including 2 controls) to the training set and compared their performance on the held-out test set. Fig. 4.8B show category probabilities under the neural (y-axes) vs. behav-

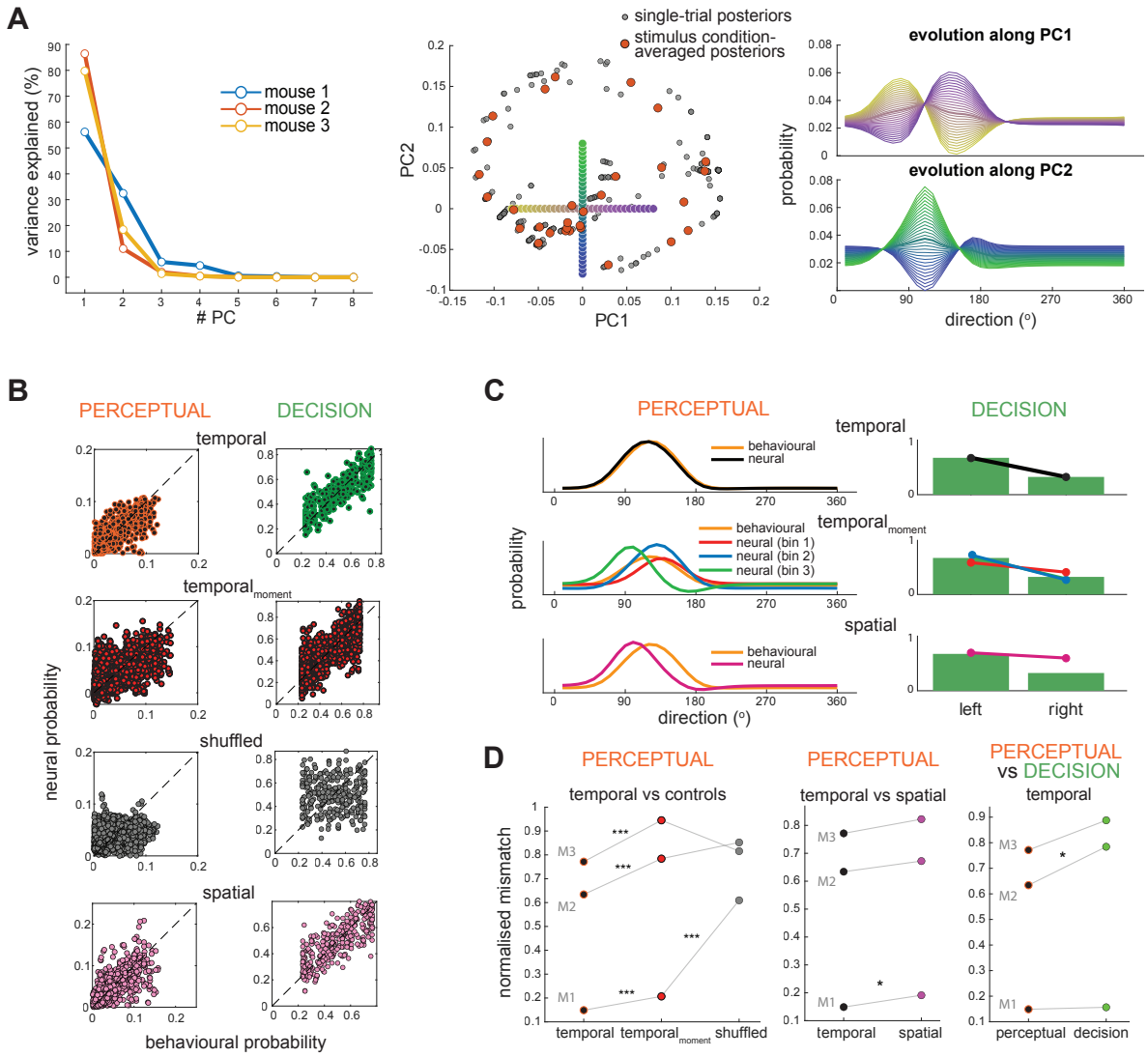


Figure 4.8. Results of the neural analysis. A. Principal component analysis of the behaviourally inferred perceptual posteriors. Left: The eigenspectrum of the stimulus-conditioned average posteriors. Middle: Mouse 1's stimulus-conditioned average posteriors (red dots) and single-trial posteriors (gray dots) in the space spanned by the first two principal components. Right: evolution of the perceptual posterior along the first two principal components. **B**. Pointwise comparison of the behavioural-neural posterior pairs. **C**. Example behavioural (orange and green) and inferred neural (colors as in main figure) posterior pairs for each comparison in panel B. **D**. Normalized mismatch between the behavioural and neural posteriors across different latent variables and coding schemes using the best fitting λ values for each coding scheme ($\lambda = 0$ for the spatial code and $\lambda = 0.05$ for the temporal code; see Fig. C.2).

ioral posteriors for all the posteriors of the test set and Fig. 4.8C show example neural and behavioral posteriors from representative trials.

As for the main test, we compared four models: encoding by temporal / spatial code of perceptual / decision posteriors. The model using temporal code of perceptual uncertainty was consistently the best across all animals (Fig. 4.8D, middle and right), albeit the difference

was significant in only two animals, and even for them, it was significant for only one model comparison each (see Appendix, Fig. C.1).

As controls, we checked that the temporal model, for which averaging neural posteriors across time bins was essential, was better than using the underlying neural posteriors of individual bins (Fig. 4.8D, left, black vs. red), or when its neural posteriors were optimized to match trial-shuffled behavioral posteriors (Fig. 4.8D, left, black vs. gray), (see Appendix for significance values, Fig. C.1).

4.5 Discussion

In this chapter, I presented a novel data-driven approach that we developed to investigate whether and how uncertainty is represented in specific neural populations. We implemented the necessary data analyses tools, and demonstrated the usage of our approach by applying it to experimental data from a mouse decision making task with calcium imaging. This experiment aimed at addressing the central quest of this thesis: detecting the traces of perceptual posterior representations, distinct from the decision posteriors that determine behavior. In doing so, it also tackled an inevitably related question – what coding scheme the brain uses to represent probabilities (spatial vs. temporal). While we found preliminary evidence suggesting that the mouse primary visual cortex may use temporal codes to represent perceptual uncertainty, this evidence is currently very limited. Further synthetic validation, conceptual refinement, and potentially improved experimental data are required to reach a more definitive conclusion.

A crucial step moving forward is to rigorously validate the neural decoder on synthetic data, ensuring that it can differentiate between coding schemes and latent variables. This validation process would ideally be built on biologically inspired, image-computable probabilistic models of perception, corresponding to the two candidate coding schemes, as they likely provide the closest approximations to the actual neurons of the brain. For example, for the temporal code the Gaussian scale mixture model would be a strong candidate, as it has been successfully used in the past to explain various stimulus-induced static and dynamic patterns

of neural activity (Orbán et al., 2016; Echeveste et al., 2020), (see Chapter 1). This validation process should also include testing different loss functions to determine which one is the most effective at distinguishing between alternative coding schemes, which one encourages most the generation of point-like samples, and which normalization method is the most suitable for comparing the posteriors of different latent variables.

In addition to testing how well our decoder can distinguish between different probabilistic coding schemes, it is crucial to evaluate whether it can detect when the inspected code is non-probabilistic. In this context, it is particularly interesting to compare its performance with previous approaches that, similar to ours, also combined model-based behavioral data analysis and neural population decoding but in the opposite order (Walker et al., 2020). Walker et al. recorded population activity from the visual cortex of monkeys during an orientation categorization task, where the animals' uncertainty about orientation influenced their behavior. First, they decoded orientation information from neural responses with a well-calibrated representation of uncertainty, without access to behavior. Then, using the decoder's output, they could predict stimulus-independent fluctuations in uncertainty that were evident in the animals' behavior. Walker et al. interpreted this as identifying a neural code for uncertainty. However, it is possible that they simply decoded the quality of noisy perceptual representations optimally – something the monkey's brain might also accomplish through downstream processing (see Chapter 2, Noise model and the ideal evidence accumulator) – without actually identifying the probabilistic perceptual representation itself. Comparing how the two methods perform in this regard could offer valuable insights into the methodologies used to study probabilistic coding.

Another important issue that cannot be ignored during model validation is the potential discrepancy between the abstract variables that intuitively characterise the task (e.g., orientation) and the variables that the brain (or at least the tested neural population) actually represents (Zemel et al., 1998; Lange et al., 2023; Lengyel et al., 2024). For example, L2/3 pyramidal cells in mouse V1 might be more sensitive to the direction of local image patches than the overall stimulus direction. This raises the question of what implications it has for our model selection method if the variable of the behavioural posterior differs from the variable

of the neural posterior (even if the latter can be derived from the former). In such cases, local uncertainties may either average out at the global level or the reverse may occur. For example, in a stimulus composed of many local patches moving in different directions – such as the one that was used in Hénaff et al. (2020) – the local image patches may be unambiguous, but the global direction could be uncertain due to their diversity (this relates to the concept of multiplicity discussed in Sahani and Dayan, 2003). In the latter case, even if the brain is using a temporal code, which would be evident in the representation of the uncertain global direction, when measuring the local directions, it might appear as though it is using a spatial code. What might offer a remedy for this problem are the well-documented feedforward and feedback connections in the cortex, which are believed to play a critical role in hierarchical probabilistic inference, carrying top-down prior expectations and bottom-up evidence between the internal variables represented at different hierarchical levels (Lee and Mumford, 2003; Haefner et al., 2016). These connections could carry the imprint of the uncertainty code between local and global variables. Future work should address this issue and validate the approach with these considerations in mind.

Finally, there is room for improvement in the experimental paradigm as well. Although the experiment was designed specifically to measure probability distributions, its implementation in this pilot study was not ideal in several respects. On one hand, it introduced strong behavioral biases, likely originating at the perceptual level (as suggested by our models). Not only did this increase the complexity of behavioral modeling, but more importantly, it impaired the credibility of the model-based estimates of perceptual posteriors, as their explanation required strong modelling assumptions. On the other hand, due to the species of the tested animals, the task's complexity had to be kept low. As a result, even though uncertainty was explicitly manipulated through different stimulus features, accounting for this uncertainty in the choices was not necessary to solve the task optimally. This limits the task's ability to detect the use of probabilistic representations from behavior alone. We had hoped that the animals' uncertainty would spontaneously manifest in measurable behavioral characteristics, but apart from the weak modulation of licking ratio during the stimulus period, we found no behavioral measure that could be attributed to uncertainty independently of other factors (such as contrast and aperture). Nevertheless, the decision posteriors we estimated

from the spontaneous right lick ratios and the stimuli that elicited them were marginally better predictors of choices during the held-out response period than the right lick ratios directly fitted to the choices, and much better predictors than the stimuli alone. This supports the assumption that the decoded posteriors are indeed represented by the mice, but to draw more definitve conclusions, a task would be required where optimal performance depends on the accurate representation of the reliability of the choices. However, so far, evidence of animals learning such tasks has only been found in other species, e.g. rats (Lak et al., 2014) and monkeys (Walker et al., 2020). How far this approach can be taken with mice is an open question, which we have already begun exploring in a follow-up experiment.

In summary, our achievements thus far amounts to the design and early implementation of a promising approach that still requires further fine-tuning and testing to realize its true potential. Once completed, however, this approach should not only provide the first neural evidence regarding the extent of uncertainty representations but also be widely applicable across a broad range of paradigms.

4.6 Methods

4.6.1 Ideal observer details:

Stimulus generation: By definition, the generative model of an ideal observer exactly matches the actual processes that generate the stimuli in the experiment. However, for mathematical and practical conveniences, we allowed the generative model that we used for data analysis to deviate slightly from the exact experimental conditions.

In both the model and the experiment, the binary trial type is randomly sampled with equal probability:

$$z \sim \text{Bernoulli}(0.5) \tag{4.4}$$

However, in the model, conditioned on the trial type, stimulus orientation is sampled from a continuous uniform distribution:

$$x \sim \begin{cases} \text{Uniform}(90^{\text{o}}, 270^{\text{o}}), & \text{if } z = 1 \\ \text{Uniform}(-90^{\text{o}}, 90^{\text{o}}), & \text{if } z = 0 \end{cases} \tag{4.5}$$

This contrasts with the discrete direction values used in the experiment. The reason for this modification was to ensure that the perceptual posterior is a continuous function.

Another difference between the model and the experiment is related to the distribution of stimulus strengths. In the experiment, high stimulus strengths (high contrast and aperture) were over-represented to keep the animals' motivated. In contrast, the model samples stimulus strength from a uniform distribution:

$$y \sim \text{Uniform}(0, 1) \tag{4.6}$$

This simplification was chosen because, in preliminary synthetic tests, the shape of the actual distribution was unreliable to recover, and adding this flexibility to the model did not seem to visibly improve the fit to the mouse data either (not shown).

However, even if there were no discrepancies between the theoretical model and the experiment, we cannot expect the internal model of a real mouse to be aligned perfectly with the experimental setup. The experiment involved multiple training and test sessions, each with different stimulus statistics. The mouse developed its internal representation based on a combination of these experiences, and it is unlikely that it could rapidly adjust this generic representation for each individual session. Taking the representation's inertia into consideration, a single generic GM, which accounted for all the potential manipulations, was used to fit the data across all sessions, regardless of the specific manipulation applied.

Observations: In addition to the structure of the experiment, the generative model also contains a model of the sensory observations. According to the model, the observed orientation is sampled from a circular Gaussian (von Mises) distribution centered on the true grating

direction:

$$\tilde{x} \sim \mathrm{vonMises}\left(x, \kappa\right) \tag{4.7}$$

whose concentration (κ) is linearly scaled by the stimulus strength, y, and is potentially direction dependent (Fig. 4.4A, left panel):

$$\kappa = y\, \kappa_0 \left(x; x^\star, \rho_\kappa, \kappa_{\mathrm{amp}}, \kappa_{\mathrm{min}}\right) \tag{4.8}$$

so that

$$\kappa_0 = \kappa_{\mathrm{min}} + \kappa_{\mathrm{amp}}\, e^{\kappa_f(\rho_\kappa)\, \cos(x - x^\star)} \tag{4.9}$$

where $\kappa_f()$ function maps the von Mises distribution's precision parameter to its concentration parameter (inverse of Eq. A.9). Assuming direction dependence for the κ_0 parameter was crucial for capturing the asymmetries of the psychometric functions shown in the result section.

An important part of the observation model is that it defines a mapping between the physical contrast and aperture and the model's stimulus strength variable, y. This mapping is a priori unknown, so assumptions are needed. We assume a monotonic scaling between the normalized contrast (c) or aperture parameters (a) (a multiplicative normalization was used to set their maximum to 1) and the y variable (Fig. 4.4A, middle panel) in the following form:

$$y = \frac{\ln(c + \epsilon_{\mathrm{cont}}) - \ln(\epsilon_{\mathrm{cont}})}{\ln(1 + \epsilon_{\mathrm{cont}}) - \ln(\epsilon_{\mathrm{cont}})} \tag{4.10}$$

The above equation is for the contrast case, but the same formula is applicable to the aperture too, if we replace c with a and ϵ_{cont} with $\epsilon_{\mathrm{apert}}$. The observation of stimulus strength was modeld exactly as in Chapter 2 (Eq. 2.14), and was parameterized with the λ observedness parameter (Fig. 4.4A, middle panel).

Response model: Finally, the ideal observer's response (here: its licking behavior) is ideally determined by its decision posterior. In the current experiment, the animal's choice was its first lick during the response period. However, in reality, the animals already began licking

shortly after the stimulus onset, and by the second half of stimulus presentation, the ratio of right to left licks become highly predictive of their upcoming choice (Fig. 4.5, comparison of the light and dark blue columns). Therefore, we modelled the direction of each lick (we did not model the occurrence of licks, just their direction) as an independent sample from a Bernoulli distribution, with a fixed parameter (q_{RL}: right lick probability) that remained unchained between the stimulus and response periods, and which depended on the decision posterior of the observer.

If behaviour were optimal, always the option with the higher posterior probability were chosen (MAP, maximum a posteriori choice), meaning that:

$$q_{RL} = \underset{z}{\operatorname{argmax}}\, P(z|\tilde{x}, \tilde{y}) \tag{4.11}$$

However, to account for the suboptimalities of real mice, we modeled the response probabilities as a monotonic function of the log posterior odds (Fig. 4.4A, right panel):

$$q_{RL} \sim (1-\alpha)\, S\left(\tilde{x}, \tilde{y}; \beta, \gamma\right) + \alpha\, \delta \tag{4.12}$$

where

$$S(\tilde{x}, \tilde{y}; \beta, \gamma) = \text{Sigmoid}\left(\beta\, \left(\ln \frac{P(z=1|\tilde{x}, \tilde{y})}{P(z=0|\tilde{x}, \tilde{y})} - \gamma\right)\right) \tag{4.13}$$

Here, α is a lapse rate, indicating that in α proportion of the trials, the animal chooses the right option with a fixed probability, δ, irrespective of its decision posterior. On the other $(1-\alpha)$ proportion of the trials, the response probability is a sigmoid function of the log posterior odds. The β parameter interpolates between completely random responses ($\beta = 0$) and a deterministic strategy ($\beta = \infty$), while γ is a response bias, shifting the sigmoid in a similar way to how unequal prior probabilities would – though, unlike priors, it does not affect perceptual inference.

4.6.2 Quantification of the predictive performance

The predictive performance was quantified with the probabilistic fraction correct metric (Houlsby et al., 2013):

$$f_{\text{prob}} = \prod_{i=1}^{n} \mathcal{P}_i(d_i)^{\frac{1}{n}} = e^{\frac{1}{n}\sum_{i=1}^{n} \ln \mathcal{P}_i(d_i)} \tag{4.14}$$

Here, n is the total number of trials, d_i is the subject's actual decision on trial i, and $\mathcal{P}_i(d_i)$ is the predictive probability assigned by the model to the subject's decision on the i^{th} trial.

RLR-based approach

If each lick's direction is sampled independently from the same Bernoulli distribution during both the stimulus and response period, then the empirical RLR is the maximum likelihood estimator of the Bernoulli parameter. Thus, RLR appeares to be a reasonable approximation for the predictive probability. However, the approximation's quality dependents heavily on the number of licks during the stimulus presentation from which RLR is calculated. In the extreme case of a single lick, the RLR's value is binary (either 0 or 1). This becomes problematic when there is at least one trial where the single lick during the stimulus presentation is in the opposite direction of the upcoming decision, as this would zero out the probabilistic fraction correct metric. To avoid this extreme behavior, we compress the RLR-based approximation as follows:

$$\mathcal{P}_i(d_i) \approx (1-\epsilon) \cdot RLR_i + \frac{\epsilon}{2} \tag{4.15}$$

where ϵ is set to the value that maximizes the predictive performance of the RLR-based approach in three-fold cross-validation.

Stimulus-based approach

Let $n_{k,j}$ be the number of trials in which response k was given to stimulus j and let $n_j = \sum_k n_{k,j}$ be the total number of trials with stimulus j. The stimulus-based approach achieves maximal predictive performance, if it assigns the following predictive probability to the trials

based on their stimulus-response pair ($\{d, s\}$, respectively):

$$\mathcal{P}_i(d_i = k|s_i = j) \approx \frac{n_{k,j}}{n_j} \tag{4.16}$$

4.6.3 Neural data processing

Neural recording

We recorded the activity of layer 2/3 pyramidal cells in the mouse primary visual cortex using two-photon calcium imaging (GCaMP6s).

Neural pre-processing

I received the decontaminated fluorescence signals (F) measured in previously identified region of interests (ROIs) corresponding to individual neurons. The neural analysis was based on the time-dependent $\phi = \frac{\Delta F}{F}$ value, which I calculated for each neuron, n, and trial, i, using the following formula:

$$\phi_{n,i}(t) = \frac{F_{n,i}(t) - F^0_{n,i}}{F^0_{n,i}} \tag{4.17}$$

Here, the baseline fluorescence $F^0_{n,i}$ was calculated for each neuron as the average activity within the 2 s time interval before the stimulus presentations measured in five consecutive trials centered on the current trial. This 5-trial averaging balanced the need to filter out global trends while preserving local fluctuations in neural activity.

Matching cells across imaging sessions

To track neurons across multiple imaging sessions, ROIs from different sessions that belong to the same neuron must be matched. Initially, the matching procedure was based on the visual comparison of ROIs (Amvrosiadis, 2023). However, this manual approach was prone to errors, which became apparent during the visual inspection of the resulting matchings. To address this, I developed a graph theory-based algorithm to systematically identify and visualize erroneous matches, enabling their efficient review and correction.

The algorithm assigns a node in a graph to each identified ROI in such a way that the ROIs from different sessions are placed in separate rows. Then it draws edges between the nodes that correspond to manually matched ROI pairs. A set of connected nodes (i.e. nodes that are linked by a continuous path of edges) defines a matching, or in technical terms, connected component (?? A). Each matching corresponds to exactly one physical neuron.

A proper matching has to meet two conditions:

- Transitivity: If two nodes are both connected to a common third node, there must also be an edge directly connecting them, as all nodes belong to the same neuron. This requirement is equivalent to stating that a proper matching forms a complete subgraph, in which every node is connected to every other node. (??B left side illustrates transitivity violation.)
- Congruence: Each matching should contain only a single node from a given row (imaging session). Otherwise, two visually distinct ROIs within one imaging session would belong to the same physical neuron. (??B left side illustrates an incongruent matching.)

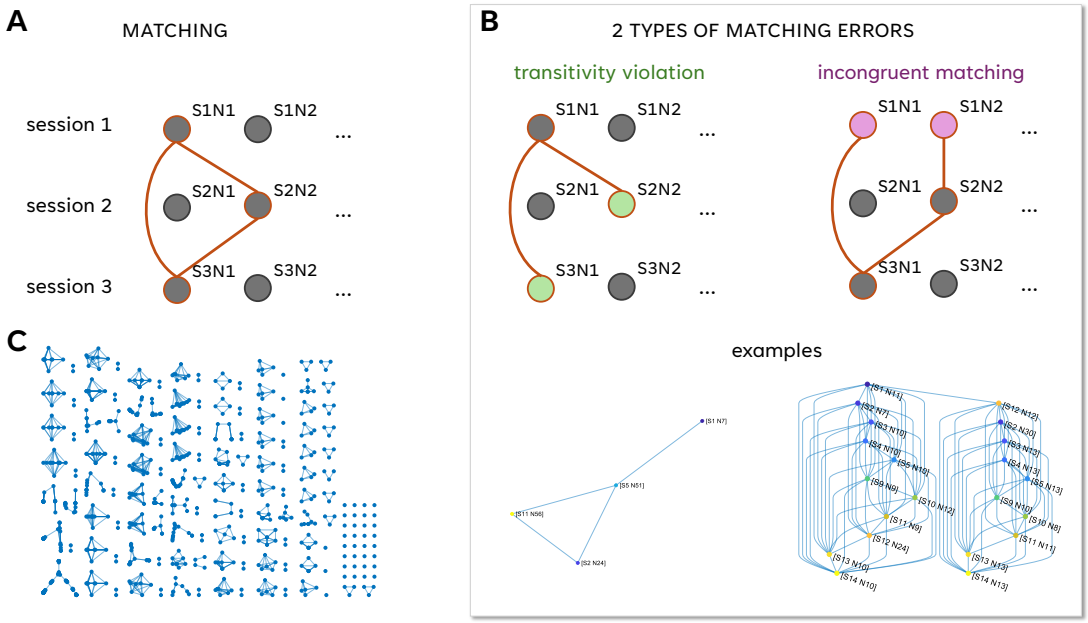


Figure 4.9. Across sessions neuron matching. A. A proper matching. Each node has a unique identifier (S: ID of the recording session, N: ID of the ROI) **B**. Left side: green ROIs violate transitivity. Right side: pink ROIs are incongruent. **C**. An example graph's factorization to its connected subcomponents.

After constructing the graph, the algorithm identifies all matchings (??C) by detecting the connected components (using the built in matlab function: `conncomp`). It then checks whether these matchings meet the two conditions.

Specifically, transitivity is tested by checking whether the number of edges (E) and the number of nodes (N) in a matching is consistent with that of a complete graph, which amounts to testing the following relationship:

$$E = \frac{N(N-1)}{2} \tag{4.18}$$

To test for congruence, each node is assigned an attribute corresponding to the session's ID. Congruence is verified by ensuring that the session IDs are unique within each set of matched nodes.

Chapter 5

General discussion

I conclude my Thesis by recapitulating its central hypothesis and the objectives I set to support it, as well as providing an overview of what has been accomplished and what remains to be done. Lastly, I take a broader perspective and outline potential directions for future research, aiming to make investigations both more natural and more effective.

5.1 Primary objectives and achievements

The central hypothesis of the thesis is grounded in two well-established ideas. First, it is increasingly believed that humans and animals rely on an internal model well adapted to the environment and perform approximate probabilistic computations on it to handle the abundant uncertainty inherent in the world (Fiser et al., 2010). Second, since the problems they encounter daily are complex, interrelated, and modular in nature, recent proposals suggest that the structure of internal models should mirror this modularity, which would enable efficient generalisation across the wide variety of everyday tasks (Tenenbaum et al., 2011; Ho et al., 2019; Lake et al., 2017). Combining these two ideas leads to the concept of task-independent probabilistic representations that can be flexibly factorized in different ways upon the changing task demands.

In Chapter 1, I argued that fully Bayesian models, which exhibit the highest degree of modularity among probabilistic recognition models, offer outstanding data- and memory-efficiency,

which makes them promising candidates for being the brain's recognition model. However, I also highlighted the lack of direct evidence supporting that the brain actually employs such models. Thus, my Thesis outlined and undertook a plan to gather the first pieces of evidence regarding the fully Bayesian brain hypothesis, which comprised three major steps.

First, in Chapter 2, I tested one of the prerequisites for fully Bayesian representations: the ability to represent the posterior distribution of more than one variable simultaneously. I verified that humans can maintain at least the marginal posterior distributions of several perceptual variables concurrently, similar to what has been demonstrated in working memory before. Second, in Chapter 3, I confirmed that the internal model adapts to changes in line with fully Bayesian models when confronted with situations that evoke complex internal representations. Finally, in Chapter 4, I introduced, implemented, and demonstrated a novel data-driven approach to identify the neural traces of rich probabilistic representations, including posteriors related to variables beyond the decision variable. I provided preliminary evidence showing that mouse V1 encodes perceptual rather than decision posterior distributions using a temporal rather than spatial code, though these findings requires further validation. The results of these studies were consistent with the fully Bayesian brain hypothesis, offering the first experimental evidence in its support.

5.2 An outlook on the role of proxies

The main objective of my thesis was to clarify where the brain's recognition model is positioned within the spectrum of possible probabilistic recognition models. To address this, in Chapter 1 Fig. 1.1, I tabulated possible probabilistic recognition models in the range between fully probabilistic to hybrid to task-dependent and finally non-probabilistic models. In my thesis, I clarified more precisely how different models in this tabulation can be realized and for this process, I introduced the concept of proxies.

The definition of proxies is useful because fully Bayesian recognition models excel in data- and memory efficiency but they come with prohibitive computational costs due to the large number of variables needed to adequately describe the environment. As a direct consequence,

even a "globally" fully Bayesian brain, which is in principle capable of computing the posterior of each of its latent variables, is likely to resort to hybrid strategies "locally", in the context of a particular task, to save on computational cost. I posit that the use of proxies is one such hybrid strategy. Proxies serve as shortcuts to the uncertainty of particular internal variables by leveraging point estimates of other variables that directly influence those uncertainties. This strategy substitutes (at least partially) the proper Bayesian evaluation of the full joint posterior distribution with simpler computations, albeit at the cost of introducing greedy approximations.

Proxies can be handy in two scenarios: (1) when the benefits of having a properly evaluated posterior distribution of an internal variable are outweighed by the computational costs of the inference process or (2) when the approximation of the posterior distribution is simply too unreliable, making proxies a valuable source of additional information. For illustration, consider the example of inferring a vehicle's motion in poor visibility conditions that reduce contrast. In such a case, a fully Bayesian brain would infer the joint posterior over both the motion direction and contrast. In contrast, the first strategy would bypass this probabilistic inference by computing a point estimate of the contrast and using it as a substitute for the uncertainty of the motion estimate. While this approach is manageable with a single factor (e.g. contrast), it quickly becomes infeasible when multiple interacting factors (e.g speed and vehicle size) are involved, which is arguably more reminiscent to the typical circumstances of perception. The second strategy offers a more refined approach by still computing an approximate joint distribution of the two internal variables, while using proxies solely to enhance the quality of the inference. This could be particularly useful, if the approximation algorithm is time-consuming – especially in hierarchical systems where a consistent probabilistic representation across interacting variables must be reached under tight time constraints (Lee and Mumford, 2003; Haefner et al., 2016). In my example, the point estimate of the contrast could provide a valuable and fast immediate prior for estimating motion uncertainty.

Previous studies exploring the possibility of proxy-based strategies while testing the probabilistic nature of perceptual decision making overlooked the implications of the approximate nature of probabilistic inference and thus, were mainly concerned with the first scenario

(Barthelmé and Mamassian, 2010; Meyniel et al., 2015a; Adler and Ma, 2018). As a consequence, they compared two extreme cases: where proxies are entirely excluded from the computation of decision posterior (they called it Bayesian or probabilistic model), or where the decision posterior computation is entirely replaced by the utilization of proxies (they called it non-Bayesian "heuristics" model). This oversimplification led to the premature conclusion that using proxies automatically implies a non-Bayesian internal model. In contrast, my theoretical work points out that using proxies is the right strategy even for the most sophisticated fully Bayesian models, given the approximate nature of inference and the resource constraints of the brain, while my experimental work confirms that humans likely rely on such strategies.

The potential utility of proxies calls for further investigations as it triggers many new questions such as whether the brain uses proxies for all variables or just a selected few. Assuming the answer that it is the latter, a subsequent challenge is identifying the factors that determine which variables are employed as proxies in specific tasks. Interestingly, proxy-based strategies can be seen as a form of loss-calibrated inference (see Chapter 1), if the demands of the specific task determine which variables' proper Bayesian inference is substituted by the use of proxies. Further questions concern the behavioral implications of using only a subset of variables as proxies and the interpretations of such strategies in terms of the probabilistic nature of the model. These open questions underscore the importance of a new look and future research on the usage of proxies in perception and cognition to fully understand the mechanisms behind it.

5.3 On the need for more complex and realistic experiments

In Chapter 1, I proposed a set of normative criteria that can serve as a basis for testing the probabilistic nature of the brain's recognition model. Although the studies I've conducted so far utilized these criteria to some extent, the full potential of the proposed normative framework has yet to be realized in earnest. In Chapter 2, I assessed task flexibility, but due to the

paradigm's limitations, I was only able to test the marginals of a joint posterior. In Chapter 3, I showed that learning exhibits a fully Bayesian characteristic; however, rather than testing for the efficiency advantages proposed in Chapter 1, I demonstrated the consistency of (group-level) behavior with a fully Bayesian model across different task conditions. Finally, in Chapter 4, I did not require the mice to display clear signs of utilizing the advantages of fully Bayesian models. Instead, I began with the assumption that mice employ fully Bayesian representations to estimate posteriors from their behavior, and searched for traces of these representations in their neural activity. Therefore, despite the progress I made, many of the experimental proposals outlined in Chapter 1 have has not yet been implemented.

However, most of the desired tests would only be feasible with a substantial increase in the complexity of the experimental paradigms. Take, for instance, the fully Bayesian model's advantage in temporal information fusion: it remains memory-efficient even when the fluctuation of a nuisance variable introduces correlation between separate observations. However, the model's efficiency in these situations can already improve a lot if the nuisance variable is represented at all, even if only as a point estimate. Therefore, to test the fully Bayesian brain hypothesis, tasks are needed in which the inference of the nuisance variable's uncertainty has a significant impact on performance. Such complex strategies are rarely needed in typical psychophysical experiments, but they may be needed all the more often in everyday situations that the perceptual system is accustomed to due to the complex interactions of numerous environmental variables regulating these situations.

One of the main barriers to use high complexity tasks in cognitive and neuroscience experiments has been the difficulty of analysing them, which requires vast computational power and sophisticated data-analysis tools. However, significant progress has been made in these areas recently. The rapid increasing of raw computational power of scientific computing has been the catalyst for the development of more powerful data analysis techniques. For example, it is now possible to analyse normative Bayesian cognitive models using Bayesian data analysis tools as well by a doubly-Bayesian technique known as cognitive tomography (CT), (Houlsby et al., 2013). CT demands significant computational resources, so applying it to complex Bayesian models – such as the ones in my experiments – will require the use of

efficient computational tools, including probabilistic programming languages (PPL), (Paszke et al., 2017; Bingham et al., 2019; Phan et al., 2019). PPLs efficiently automate Bayesian parameter estimation, allowing data analysts to focus solely on defining generative models, while the PPL handles the underlying inference. The use of PPLs is spreading rapidly due to their broad application in machine learning but in some areas of research the full exploitation of their power is lacking. For example, to use them successfully for CT, PPLs should be employed twice in the task. First, they would handle the estimation of the Bayesian model's parameters, and second, they would be used at each iteration step of this estimation process to evaluate the likelihood of the fitted Bayesian model parameters. Unfortunately, this combined application of the PPLs remains a challenge that has not yet been fully resolved.

More recently, a technique called continuous psychophysics has been developed to measure behaviour through the continuous interaction of the participants with the task, rather than through discrete trials (Straub and Rothkopf, 2022). This method offers more efficient data collection and enables more intricate and ecological tasks designs than the standard psychophysical experiments. It not only allows researchers to tackle complex questions – including those in my thesis – but also has the potential to enhance the participant's engagement with the task. The importance of the latter aspect is gaining increasing attention (Allen et al., 2024), as natural behavioural processes are arguably better reflected in more ecological tasks, building on the internal motivation of the participants.

Finally, there have also been significant advances in studying animal cognition. Modern machine learning techniques now allow for tracking real time natural animal behavior (Mathis et al., 2018), and for breaking down complex motion patterns to elementary behavioural "syllables" (Lin et al., 2024). These techniques take experimentation to more ecological domains. At the same time, other efficient computational tools are being developed to interpret neural activity in terms of the behavior it brings about (Schneider et al., 2023). However, the use of these tools has so far been somewhat limited, primarily focusing on identifying low-dimensional latent embedding spaces of high-dimensional neural activity that correspond to low-dimensional behavior, and the application of these advanced tools beyond descriptive analysis has yet to be developed.

In sum, while steady advances were made in multiple domains of brain sciences, further improvements are desired and the lack of developing new computational tools and techniques for analyzing the available quantity and complexity of data appears to be a major obstacle in this process. Despite the recent highest recognition of the work and results amassed in this domain, there is a dire need for new synergistic and sophisticated experimental designs and adequately suited analysing methods for faster advances. I hope my work can provide a suitable stepping stone in this endeavour.

Appendices

Appendix A

Supplementary Materials to Chapter 2

A.1 Circular statistics: the basics

Here, I introduce those fundamental circular statistical quantities and identities that I utilize for deriving the equations of the IEA models. All definitions and identities are based on (Jammalamadaka and Sengupta, 2001).

Basic Quantities

Let $\mathcal{P}(x)$ be a continuous probability density function over the circular variable x, and consider the following expected value:

$$z = \int e^{i\,x}\,\mathcal{P}(x)\,\mathrm{d}x \tag{A.1}$$

We define the circular mean of $\mathcal{P}(x)$ as:

$$\mu = \arg(z) \tag{A.2}$$

and the circular precision of $\mathcal{P}(x)$ as:

$$\rho = |z| \tag{A.3}$$

After drawing N independent samples from $\mathcal{P}(x)$:

$$\tilde{x}_n \sim \mathcal{P}(x) \tag{A.4}$$

we call the sum

$$\mathbf{R} = \sum_{n=1}^{N} e^{i\,x_n} \tag{A.5}$$

the resultant vector. The population average is the resultant vector's argument:

$$\bar{x} = \arg(\mathbf{R}) \tag{A.6}$$

and its length

$$R = |\mathbf{R}| \tag{A.7}$$

is called the resultant length.

Circular normal (or von Mises) distribution

The von Mises probability density function is given by:

$$\mathrm{vM}(x;\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)}\, e^{\kappa\cos(x-\mu)} \tag{A.8}$$

where I_0 is the modified Bessel function of the first kind of order 0, and κ is the concentration parameter that is tied to the distribution's precision in the following way:

$$\rho = \frac{I_1(\kappa)}{I_0(\kappa)} \tag{A.9}$$

In some derivations I use the following identities:

$$\mathbb{E}_x[\cos x] = \int \mathrm{d}x\, \mathrm{vM}(x;\mu,\kappa)\cos x = \frac{I_1(\kappa)}{I_0(\kappa)}\cos\mu = \rho\cos\mu \tag{A.10}$$

$$\mathbb{E}_x[\sin x] = \int \mathrm{d}x\, \mathrm{vM}(x;\mu,\kappa)\sin x = \frac{I_1(\kappa)}{I_0(\kappa)}\sin\mu = \rho\sin\mu \tag{A.11}$$

based on which the circular precision can be expressed in the following way

$$\rho(\kappa) = |\mathbb{E}_x[\cos x] + i\mathbb{E}_x[\sin x]| \tag{A.12}$$

The distribution of N independent von Mises samples:

$$\mathrm{vM}(x_{1:t}; \mu, \kappa) = \prod_{n=1}^{N} \mathrm{vM}(\tilde{x}_n; \mu, \kappa) = \frac{2\pi I_0(R\,\kappa)}{(2\pi I_0(\kappa))^N} \mathrm{vM}(\bar{x}; \mu, R\,\kappa) \tag{A.13}$$

and the convolution of two von Mises distributions:

$$\mathrm{vM}(\theta; \mu_1, \kappa_1) * \mathrm{vM}(\theta; \mu_2, \kappa_2) = \frac{1}{2\pi I_0(\kappa_1) I_0(\kappa_2)} I_0\left(\sqrt{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2\cos(\theta - (\mu_1 + \mu_2))}\right) \tag{A.14}$$

A.2 Ideal evidence accumulators' posterior

The following derivations are based on the work of David Zoltowski (Zoltowski, 2016).

Using Bayes rule and exploiting that orientation has a uniform prior, we can write the evidence accumulator's posterior over the orientation (x) in the following way:

$$\mathcal{P}(x \mid x_{1:t}, \tilde{y}) \propto \mathcal{P}(x_{1:t} \mid x, \tilde{y}) = \int \mathcal{P}(x_{1:t} \mid x, \rho_S)\ \mathcal{P}(\rho_S \mid \tilde{y})\,\mathrm{d}\rho_S \tag{A.15}$$

The second term in the integral is the posterior of the sampling distribution's precision given the observed stimulus strength, and it is given by the following integral:

$$\mathcal{P}(\rho_S \mid \tilde{y}) = \int \mathcal{P}(\rho_S \mid y)\ \mathcal{P}(y \mid \tilde{y})\,\mathrm{d}y \tag{A.16}$$

where y's posterior can be computed from the Bayes rule again.

The first "likelihood" term in the integral on the right hand side of Eq. A.15 is the probability of the samples given x and ρ_S. The exact formula depends on which model variant is being

used. For the noise model it is:

$$\mathcal{P}_{\mathrm{N}}(x_{1:t} \mid x, \rho_S) = \prod_n \mathrm{vM}(\tilde{x}_n; \kappa_S) = \frac{2\pi I_0(\kappa_S R)}{(2\pi I_0(\kappa_S))^N} \mathrm{vM}(\bar{x}; x, \kappa_S R) \tag{A.17}$$

where I used Eq. A.13 and the notation $\kappa_S := \kappa(\rho_S)$.

For the signal model, the bias of the sampling distribution's mean (measured as its distance from the ground truth) has to also be taken into account, therefore:

$$\mathcal{P}_{\mathrm{S}}(x_{1:t} \mid x, \rho_S) = \int \prod_n \mathrm{vM}(\tilde{x}_n; \mu_S, \kappa_S)\ \mathrm{vM}(\mu_S; x, \kappa_S)\, \mathrm{d}\mu_S \tag{A.18}$$

$$= \frac{2\pi I_0(\kappa_S R)}{(2\pi I_0(\kappa_S))^N} \int \mathrm{vM}(\bar{x} \mid \mu_S, \kappa_S R)\ \mathrm{vM}(\mu_S; x, \kappa_S)\, \mathrm{d}\mu_S \tag{A.19}$$

A.2.1 Ideal evidence accumulators' orientation estimate and certainty

The ideal observer's (noiseless) behavioral reports are the angle (orientation estimate) and magnitude (certainty) of the first trigonometric moments of its posterior. The first trigonometric moment:

$$\eta = \mathbb{E}_{x|x_{1:t},\tilde{y}}[\cos x] + i\, \mathbb{E}_{x|x_{1:t},\tilde{y}}[\sin x] \tag{A.20}$$

Noise model

$$\eta = \mathbb{E}_{x|x_{1:t},\tilde{y}}[\cos x] + i\, \mathbb{E}_{x|x_{1:t},\tilde{y}}[\sin x] = \tag{A.21}$$

$$= \int \frac{1}{2\pi \mathcal{P}(x_{1:t} \mid \tilde{y})} \int \frac{2\pi I_0(\kappa_S R)}{(2\pi I_0(\kappa_S))^N} \mathrm{vM}(\bar{x}; x, \kappa_S R)\, P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S\ (\cos x + i \sin x)\, \mathrm{d}x \tag{A.22}$$

utilizing Eq. A.10 and Eq. A.11 after exchanging the integrals

$$= \int \frac{1}{\mathcal{P}(x_{1:t} \mid \tilde{y})} \frac{I_0(\kappa_S R)}{(2\pi I_0(\kappa_S))^N} \frac{I_1(\kappa_S R)}{I_0(\kappa_S R)} P(\rho_S \mid \tilde{y}) \left(\cos \bar{x} + i \sin \bar{x}\right) \mathrm{d}\rho_S = \tag{A.23}$$

$$= \frac{1}{(2\pi)^N \mathcal{P}(x_{1:t} \mid \tilde{y})} \int \frac{I_1(\kappa_S R)}{I_0(\kappa_S)^N} P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S \ \left(\cos \bar{x} + i \sin \bar{x}\right) \tag{A.24}$$

where

$$\mathcal{P}(x_{1:t} \mid \tilde{y}) = \int \mathcal{P}(x_{1:t} \mid x, \tilde{y}) \ \mathcal{P}(x) \, \mathrm{d}x = \frac{1}{2\pi} \iint \mathcal{P}(x_{1:t} \mid x, \rho_S) \ \mathcal{P}(\rho_S \mid \tilde{y}) \, \mathrm{d}\rho_S \mathrm{d}x \tag{A.25}$$

$$= \frac{1}{2\pi} \int \frac{2\pi I_0(\kappa_S R)}{(2\pi I_0(\kappa_S))^N} \int \mathrm{vM}(\bar{x}; x, \kappa_S R) \, \mathrm{d}x \ P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S \tag{A.26}$$

$$= \frac{1}{(2\pi)^N} \int \frac{I_0(\kappa_S R)}{I_0(\kappa_S)^N} \mathcal{P}(\rho_S \mid \tilde{y}) \, \mathrm{d}\rho_S \tag{A.27}$$

Taken together

$$\eta = \frac{\int \frac{I_1(\kappa_S R)}{I_0(\kappa_S)^N} P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S}{\int \frac{I_0(\kappa_S R)}{I_0(\kappa_S)^N} P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S} \left(\cos \bar{x} + i \sin \bar{x}\right) \tag{A.28}$$

So the orientation estimate is

$$\mu = \arg(\eta) = \bar{x} \tag{A.29}$$

and certainty is

$$\rho = |\eta| = \frac{\int \frac{I_1(\kappa_S R)}{I_0(\kappa_S)^N} P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S}{\int \frac{I_0(\kappa_S R)}{I_0(\kappa_S)^N} P(\rho_S \mid \tilde{y}) \mathrm{d}\rho_S} \tag{A.30}$$

Signal model

For the signal model in Eq. A.22 instead of the integral

$$\int \mathrm{vM}(\bar{x}; x, \kappa_S R) \left(\cos x + i \sin x\right) \mathrm{d}x = \frac{I_1(\kappa_S R)}{I_0(\kappa_S R)} \left(\cos \bar{x} + i \sin \bar{x}\right) \tag{A.31}$$

the following integral has to be performed:

$$\iint \mathrm{vM}(\bar{x} \mid \mu_S, \kappa_S R)\ \mathrm{vM}(\mu_S; x, \kappa_S)\,\mathrm{d}\mu_S\,(\cos x + i \sin x)\,\mathrm{d}x \tag{A.32}$$

$$= \int \mathrm{vM}(\bar{x} \mid \mu_S, \kappa_S R)\,\frac{I_1(\kappa_S)}{I_0(\kappa_S)}\,(\cos \mu_S + i \sin \mu_S)\,\mathrm{d}\mu_S = \tag{A.33}$$

$$= \frac{I_1(\kappa_S)}{I_0(\kappa_S)}\frac{I_1(\kappa_S R)}{I_0(\kappa_S R)}\,(\cos \bar{x} + i \sin \bar{x}) \tag{A.34}$$

otherwise the derivation is the same.

The only difference is in the certainty reports

$$\rho = \frac{\int \frac{I_1(\kappa_S R)}{I_0(\kappa_S)^N}\frac{I_1(\kappa_S)}{I_0(\kappa_S)} P(\rho_S \mid \tilde{y})\mathrm{d}\rho_S}{\int \frac{I_0(\kappa_S R)}{I_0(\kappa_S)^N} P(\rho_S \mid \tilde{y})\mathrm{d}\rho_S} \tag{A.35}$$

A.3 Certainty scaling

.

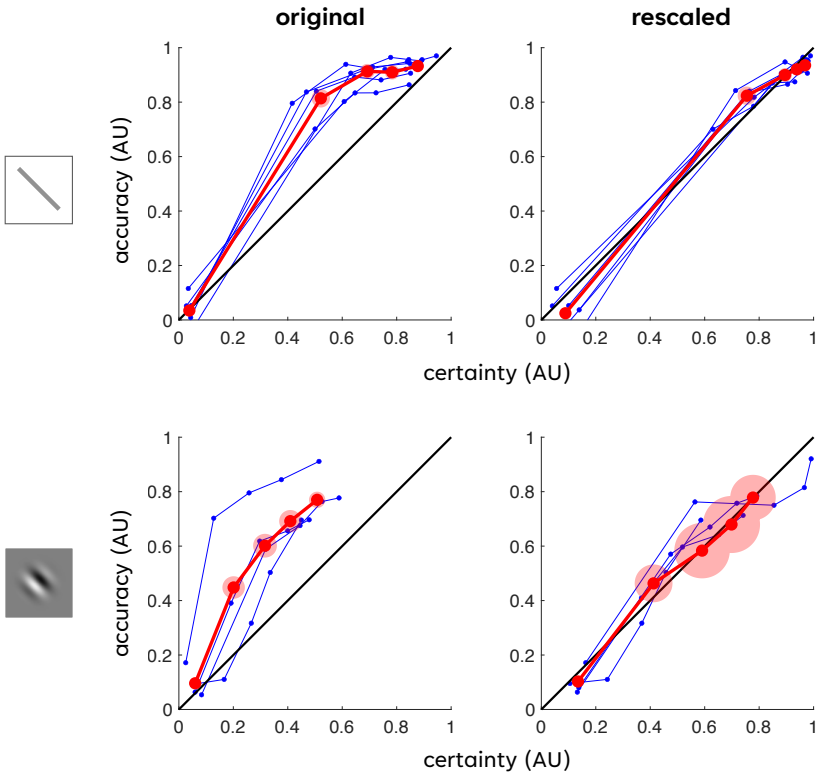


Figure A.1. Accuracy-certainty plots before and after the certainty was rescaled. Blue connected dots represent individual participants, while red connected dots indicate the averages across all participants. Red error ellipses show the standard error of the mean.

A.4 Orientation-dependent response bias

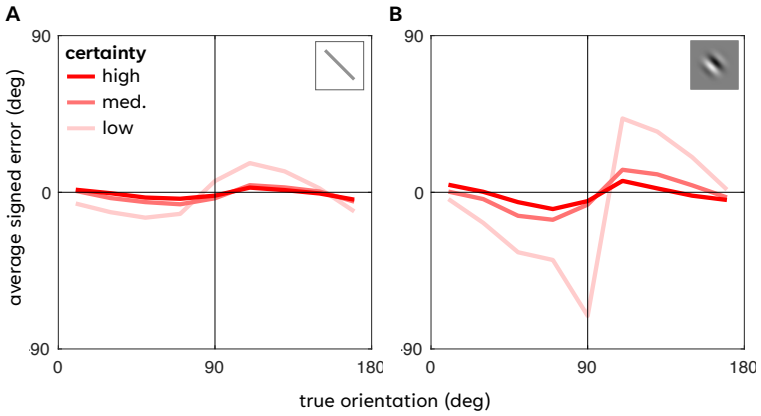


Figure A.2. Orientation-dependent response bias. Orientation-dependent response biases at different levels of certainty.

Appendix B

Supplementary Materials to Chapter 3

B.1 Short term serial effects

The $STSE_r(n)$ metric quantifies how much the probability of a "frequent" response increases when the response n trials earlier was also frequent, compared to when it was rare, while counterbalancing for what was the actual stimulus n trials back:

$$STSE_r(n) = \frac{1}{2}\sum_{l=-1}^{1} F(r_t = 1 | r_{t-n} = 1, z_{t-n} = l) - \frac{1}{2}\sum_{l=-1}^{1} F(r_t = 1 | r_{t-n} = -1, z_{t-n} = l) \tag{B.1}$$

where $F(r_t = i|r_{t-n} = j, z_{t-n} = k)$ is the empirical ratio of response $i \in \{\pm 1\}$ given that the response and stimulus n trial before were $j \in \{\pm 1\}$ and $k \in \{\pm 1\}$, respectively.

When measuring the effect of past stimuli, instead of past response, with $STSE_z(n)$, r_{t-n} and z_{t-n} are interchanged.

(B.2)

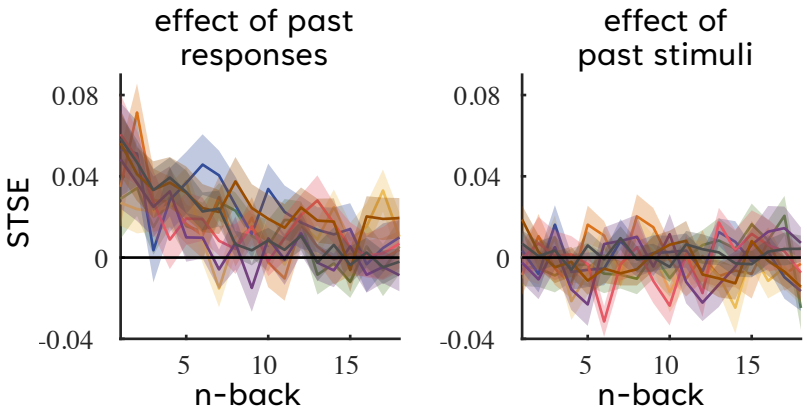


Figure B.1. Short term serial effects for all experiments.

B.2 Accounting for the potential difference in the relative visibility of objects

Different abstract shapes that were used in the experiments might not be equally well detectable under the same amount of noise. If that is the case, then the pair's true relative visibility is skewed, which biases behavior. To mitigate potential effects stemming from differences in shape detectability, for each participant, we randomly selected which two shapes formed a pair and what the two shapes role was (i.e. which shape was more frequent). Nevertheless, to eliminate the possibility that any of the results we found is due to improper counterbalancing, we explicitly took into account the effect of differing shape detectability in our analysis.

To assess the relative differences in shape detectability, we resorted to post-hoc analysis by comparing the biases induced by the same shape in different pair configurations while its role was reversed across different participants. Given that each shape appeared infrequently within each experimental condition and in various pairings, we pooled the data across experimental conditions, while being careful to separate the experimental condition's impact on the bias from the influence of the shape pairs as much as possible.

We employed two different methods for the psychometric function fitting and the static Bayesian model fitting in accordance with what method suited best the given task. For the psychometric analysis, the biases were corrected after curve fitting. For the Bayesian analysis, we first estimated what might be the AP that the participants actually observed and just then we fitted the model with this 'effective' AP, because it was important to know where the maximum likelihood ridge fell for each participants for fitting the complex static model.

B.3 Converting noise (γ) to stimulus strength (y)

The noise parameter's distribution is uniform, which is the standard deviation of the Gaussian pixel noise:

$$\gamma \sim \mathrm{Uniform}(a, b) \tag{B.3}$$

Stimulus strength is the inverse noise:

$$y = \frac{1}{\gamma} \tag{B.4}$$

The probability density distribution of stimulus strength:

$$P_Y(y) = P_\Gamma\Big(\gamma^{-1}(y)\Big) \cdot \left|\frac{\mathrm{d}\gamma^{-1}(y)}{\mathrm{d}y}\right| \propto \frac{1}{y^2} \tag{B.5}$$

B.4 Comparing different versions of the bounded evidence accumulation model.

To verify that the complex model was indeed necessary to explain the response time data, we compared its performance with that of the simple models. To do this, we first fitted the simple static Bayesian models to the stimulus-response pair data (now, without constraining the single prior parameter to the maximum likelihood point). We then fitted the bounded evidence accumulator (BEA) model using the priors obtained, exactly as we did for the complex model, but now with either the w_{bias} or v_{bias} parameter set to their unbiased values (0 or 0.5, respectively), depending on whether AP or RV (respectively) was the relevant parameter in the simple Bayesian model.

Additionally, we also tested the performance of a BEA model variant that accounted only for the STSEs. In this version, both w_{bias} and v_{bias} parameters were set to their unbiased values.

To quantify the model's performance, we computed their Bayesian information criterion (BIC) score (Fig. B.2).

B.4.1 Constructing the conditional distribution of observations.

Eq. 3.11 has no unique solution, introducing an arbitrariness in the definition of the biased ($RV \neq 0.5$) likelihood function, $P(x|z;RV)$. We choose a definition that introduces the most distortion to the likelihood function (relative to the $RV = 0.5$ unbiased case) at the observations, x, where a completely unbiased observer ($AP = 0.5 \,\&\, RV = 0.5$) would be most uncertain in its decisions.

In subsequent derivations, we omit the explicit indication of the parameters in the notation of the marginal distribution of x, such that:

$$P(x) := P(x; AP = 0.5, RV = q) = P(x; AP = q, RV = 0.5) \tag{B.6}$$

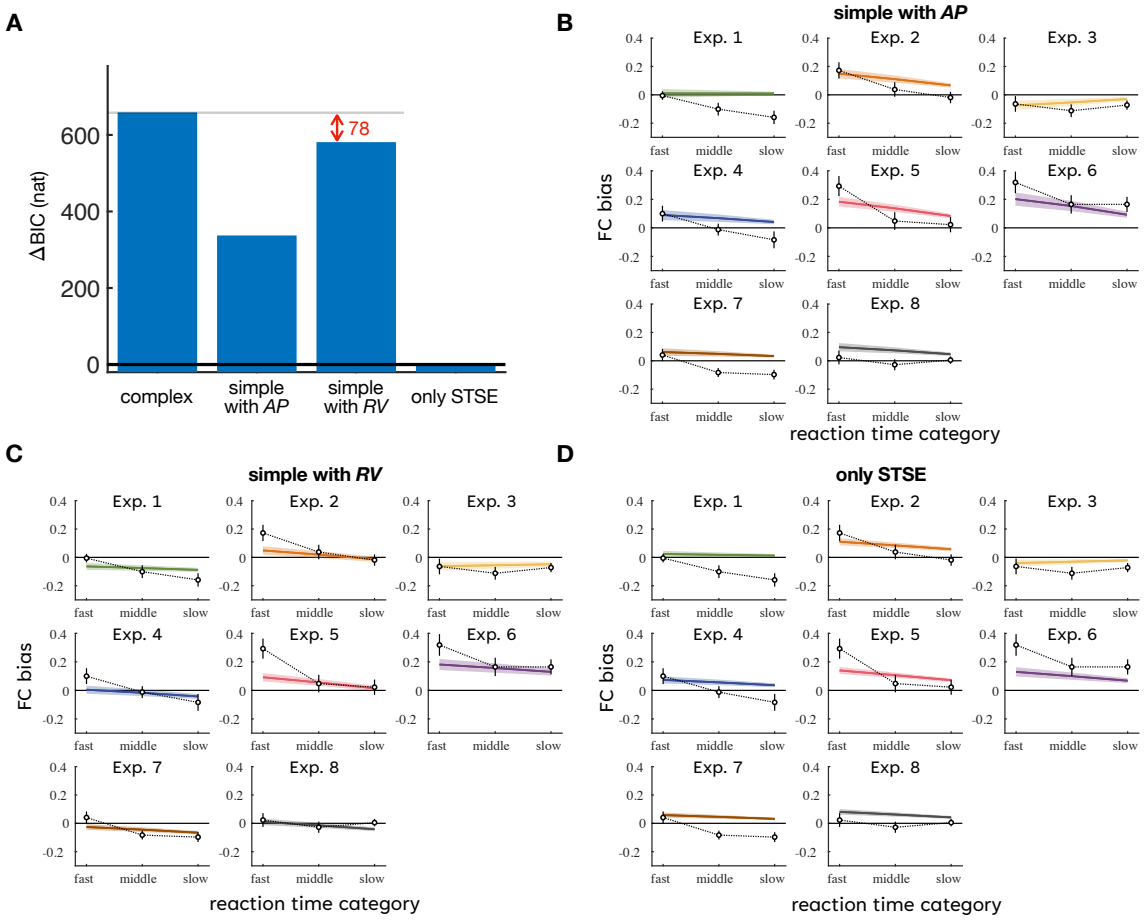


Figure B.2. The comparison of different versions of the bounded evidence accumulation model. A. Bayesian information criterion (BIC) scores for the complex model, the simple models, and a model accounting only for the STSEs. Higher values indicate better model fit,and the values are shown relative to the worst model's (only STSE) BIC score. **B-D**. Model fits corresponding to the models in panel A.

We define the biased likelihood via the posterior $P(z = 1|x; AP = 0.5, RV = q)$ utilizing the following equality:

$$P(x|z = 1; RV = q) = 2 \cdot P(z = 1|x; AP = 0.5, RV = q) \cdot P(x) \tag{B.7}$$

We construct the posterior distribution of z under the parameters $\{AP = 0.5, RV = q\}$ as the sum of the posterior under the unbiased parameters $\{AP = 0.5, RV = 0.5\}$ and a "correction" term $-a_q \cdot h(x)$:

$$P(z = 1|x; AP = 0.5, RV = q) = P(z = 1|x; AP = 0.5, RV = 0.5) - a_q \cdot h(x) \tag{B.8}$$

For conciseness, we will use the following notation henceforth:

$$P_{UB}(z=1|x) := P(z=1|x; AP=0.5, RV=0.5)\left(= \frac{P(x|z=1; RV=q)}{P(x|z=1; RV=q) + P(x|z=0; RV=q)} \right) \tag{B.9}$$

where UB stands for the term "unbiased".

The function $h(x)$ is defined in such a way that the distance between the two posteriors is the largest for those xs at which the unbiased observer is the most uncertain (i.e. it's value is nearest to 0.5 for both options):

$$h(x) := \frac{1}{2}^2 - \left(P_{UB}(z=1|x) - \frac{1}{2}\right)^2 \tag{B.10}$$

Following Eq. B.7, the norm of the conditional probability distribution of x, $P(x|z=1; RV=q)$, imposes the following constraint on a:

$$a_q = \frac{\int \mathrm{d}x P_{UB}(z=1|x) \cdot P(x) - \frac{1}{2}}{\int \mathrm{d}x h(x) \cdot P(x)} \tag{B.11}$$

The constraint that $P(z=1|x; AP=0.5, RV=q)$ has to be between 0 and 1 for every x is satisfied if and only if $a_q \in [-1, 1]$. Therefore we replace a_q with $\tilde{a}_q$:

$$\tilde{a}_q = \max(-1, \min(1, a_q)) \tag{B.12}$$

This means that when a is negative or greater than one, a bias in the likelihood cannot completely replace the bias in the prior. When fitting the model or running the simulations, this was never the case.

B.5 Scaling the noise distribution

When fitting the model, we assumed that participants knew the exact distribution of the noise ($P^\star(y)$). However, this distribution changes continuously during training due to the

adaptive staircase method applied, and then undergoes a sudden change at the beginning of the test, so it is possible that participants do not know exactly its value during the test phase.

A uniform noise distribution was used in the experiment:

$$P^{\star}(y) = \mathrm{Unif}(y_{\min}, y_{\max}) \tag{B.13}$$

This results in the following distribution over stimulus strength ($s = 1/y$):

$$P^{\star}(s) \propto \frac{1}{s^2} \tag{B.14}$$

To model participants' potential misestimation of this distribution, we allowed the power to deviate from 2, leading to the modified distribution:

$$P^{\star}(s) \propto \frac{1}{s^{\eta}} \tag{B.15}$$

Changing the η parameter did not significantly alter the fitted models' likelihood, meaning we could not reliably estimate its value. Therefore, we tested how this model indeterminacy effected the inference of the two key parameters, AP and RV.

For the complex model, when the AP-RV parameter combination was sought along the maximum likelihood ridge that passed through the unbiased point ($AP = 0.5$ and $RV = 0.5$), the η had no effect on the parameter estimates (Fig. B.3A, E3 and E6). For the other maximum likelihood ridge, it linearly shifted the estimated AP-RV combinations along that ridge by the same constant value, Δ, across all experiments (Fig. B.3A, E1, E2, E4, E5, E7, E8):

$$AP' = AP + \Delta \tag{B.16}$$

$$RV' = RV - \Delta \tag{B.17}$$

For the simple models, in which only one of the key parameters was fitted, we did not restrict the parameter search to the trivial maximum likelihood point. There, the effect of η was

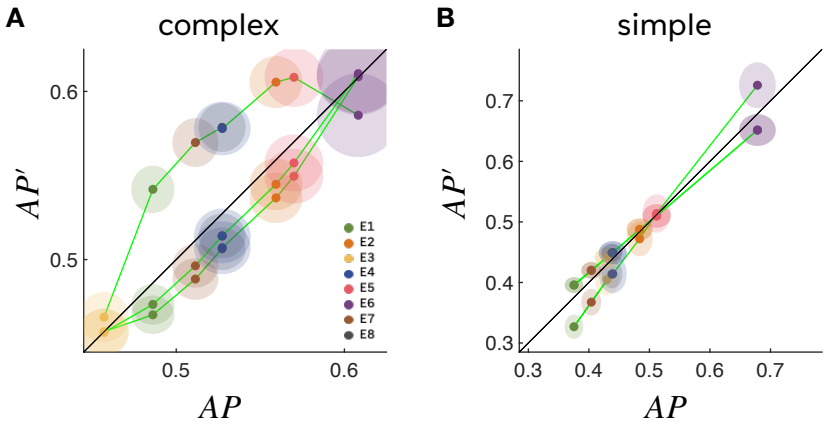


Figure B.3. Influence of noise distribution on parameters. A. The estimate of the complex model's AP parameter (AP') when η is either 1 or 3. (RV is trivial given the maximum likelihood constraint.) **B**. Simple model with AP parameter. The estimate of the complex model's AP parameter (AP') when η is either 1 or 3. Fitting the simple model with RV parameter would result in an almost identical pattern (not shown).

multiplicative (Fig. B.3B):

$$AP' = AP + (AP - 0.5) * \Delta \tag{B.18}$$

or

$$RV' = RV + (RV - 0.5) * \Delta \tag{B.19}$$

with Δ again shared across all experiments.

When fitting the BEA model, we allowed these transformations of the AP and RV parameters. Crucially, this reduced the constraints imposed by the static Bayesian model by only one degree of freedom.

Appendix C

Supplementary Materials to Chapter 4

C.1 Bayesian inference

The ideal observer computes both the perceptual posterior:

$$P(x|\tilde{x},\tilde{y}) \propto \sum_{z=0}^{1} \int P(\tilde{x}|x,y)P(\tilde{y}|y)P(y)P(x|z)P(z)\mathrm{d}y \tag{C.1}$$

and the decision posterior:

$$P(z|\tilde{x},\tilde{y}) \propto \iint P(\tilde{x}|x,y)P(\tilde{y}|y)P(y)P(x|z)P(z)\mathrm{d}x\mathrm{d}y \tag{C.2}$$

First, I derive Eq.C.1. Starting with the Bayes rule:

$$P(x|\tilde{x},\tilde{y}) \propto P(\tilde{x},\tilde{y}|x)P(x) \tag{C.3}$$

using $P(A,B|D) = \int P(A,B|C,D)P(C|D)\mathrm{d}C$:

$$P(\tilde{x},\tilde{y}|x)P(x) = P(x) \int P(\tilde{x},\tilde{y}|x,y)P(y|x)\mathrm{d}y \tag{C.4}$$

using that x and y are independent and $\tilde{x}$ and $\tilde{y}$ are conditionally independent given y

$$P(x)\int P(\tilde{x},\tilde{y}|x,y)P(y|x)\mathrm{d}y = P(x)\int P(\tilde{x}|x,y)P(\tilde{y}|y)P(y)\mathrm{d}y \tag{C.5}$$

Finally, using $P(A)=\sum_B P(A|B)P(B)$:

$$\int P(\tilde{x}|x,y)P(\tilde{y}|y)P(x)P(y)\mathrm{d}y = \sum_{z=0}^{1} P(x|z)P(z)\int P(\tilde{x}|x,y)P(\tilde{y}|y)P(y)\mathrm{d}y \tag{C.6}$$

C.2 can be derived in a similar way.

C.2 Fitting the model

We use the BADS optimizer (Acerbi and Ma, 2017) to find the parameter set ($\theta = \{x^\star, \rho_\kappa, \kappa_{\mathrm{amp}}, \kappa_{\mathrm{min}}, \lambda, \epsilon_{\mathrm{cont}}, \epsilon_{\mathrm{apert}}, \alpha, \beta, \gamma, \delta\}$) that maximizes the probability of the observed right lick ratios (RLR) during stimulus presentation while treating the total number of licks observed:

$$P(\mathbf{RLR}|\mathbf{x},\mathbf{y},\mathbf{N},\theta) = \prod_{t=1}^{T} P(RLR_t|x_t,y_t,N_t,\theta) \tag{C.7}$$

where the bold fonts denote vectors containing data for every trial t. For a single trial:

$$P(RLR_t|x_t,y_t,N_t,\theta) = \int P(RLR_t|\tilde{x}_t,\tilde{y}_t,N_t,\theta)P(\tilde{x}_t,\tilde{y}_t|x,y,\theta)\mathrm{d}\tilde{x}_t\mathrm{d}\tilde{y}_t \tag{C.8}$$

and

$$P(RLR_t|\tilde{x}_t,\tilde{y}_t,N_t,\theta) = \binom{N_t}{RLR_t} f(\tilde{x}_t,\tilde{y}_t)^{N_t,RLR_t}\left(1-f(\tilde{x}_t,\tilde{y}_t)\right)^{N_t-N_t,RLR_t} \tag{C.9}$$

where f is the probability that any given lick on that trial is to the right:

$$f(\tilde{x}_t,\tilde{y}_t) = (1-\alpha)\,S\left(\tilde{x}_t,\tilde{y}_t;\beta,\gamma\right) + \alpha\,\delta \tag{C.10}$$

C.3 Estimation of the behavioural posteriors

If we knew the animal's observations, we could easily compute an estimate of its perceptual and decision posterior distributions based on the inferred internal model parameters. However, the animal's actual observations are never accessible to us, the experimenters. Thus, the best we can do is to infer what these observations might be on a given trial i based on the stimulus ($S_i := \text{direction}, , \text{contrast}, , \text{aperture}$) and the animal's behavior ($RLR_i :=$ right lick ratio during stimulus presentation):

$$\mathcal{P}(\tilde{x}_i, \tilde{y}_i \mid S_i, RLR_i) \tag{C.11}$$

and use this distribution, which is the experimenter's posterior over the observations, for estimating the perceptual and decision posteriors of the animal. However, it's not immediately clear how to use the experimenters' inferred observations to estimate the animal's beliefs. The strategy that we chose was to average the inferred model's perceptual and decision posterior under our posterior distribution of observations (omitting the i index for clarity):

$$\text{post}_{\text{animal}}(l) \approx \pi(l) = \int \mathcal{P}(l \mid \tilde{x}, \tilde{y}) \; \mathcal{P}(\tilde{x}, \tilde{y} \mid S, RLR) \, \mathrm{d}\tilde{x} \, \mathrm{d}\tilde{y} \tag{C.12}$$

where l can stand for x and z. The drawback of this choice is that the average posterior's uncertainty (measured as its variance) is greater then the average uncertainties of the individual posteriors:

$$\mathbb{V}_L \left[l \mid S, RLR \right] \geq \mathbb{E}_{\tilde{X},\tilde{Y}} \left[\mathbb{V}_L [l | \tilde{x}, \tilde{y}] \mid S, RLR \right] \tag{C.13}$$

leading to a general overestimation of the animal's uncertainty. Another possibility would be to use only the MAP estimate of the observations, but the risk is that the ground truth posteriors and the estimated posteriors may have little or no overlap on a significant proportion of the trials.

The Section C.3 inequality is a direct consequence of the law of total variance:

$$\mathbb{V}_A \left[a \right] = \mathbb{V}_B \left[\mathbb{E}_A [a \mid b] \right] + \mathbb{E}_B \left[\mathbb{V}_A [a \mid b] \right] \tag{C.14}$$

where all terms are positive.

C.4 Normalized mismatch statistics.

perceptual: spatial - temporal

mouse	mean	sem	p	t	df
1	0.042	0.017	0.016	2.430	138
2	0.038	0.057	0.506	0.666	143
3	0.051	0.071	0.475	0.716	119

temporal: decision - preceptual

mouse	mean	sem	p	t	df
1	0.008	0.019	0.686	0.405	138
2	0.150	0.059	0.012	2.550	143
3	0.116	0.068	0.092	1.697	119

perceptual: momentary temporal - temporal

mouse	mean	sem	p	t	df
1	0.058	0.006	0.000	9.215	138
2	0.150	0.012	0.000	12.878	143
3	0.174	0.016	0.000	10.813	119

perceptual: shuffled - temporal

mouse	mean	sem	p	t	df
1	0.461	0.053	0.000	8.728	138
2	0.218	0.089	0.015	2.46	143
3	0.044	0.107	0.679	0.414	119

Figure C.1. Comparison of different coding scheme - latent variable combinations. Each comparison is a paired, two-tailed t-test.

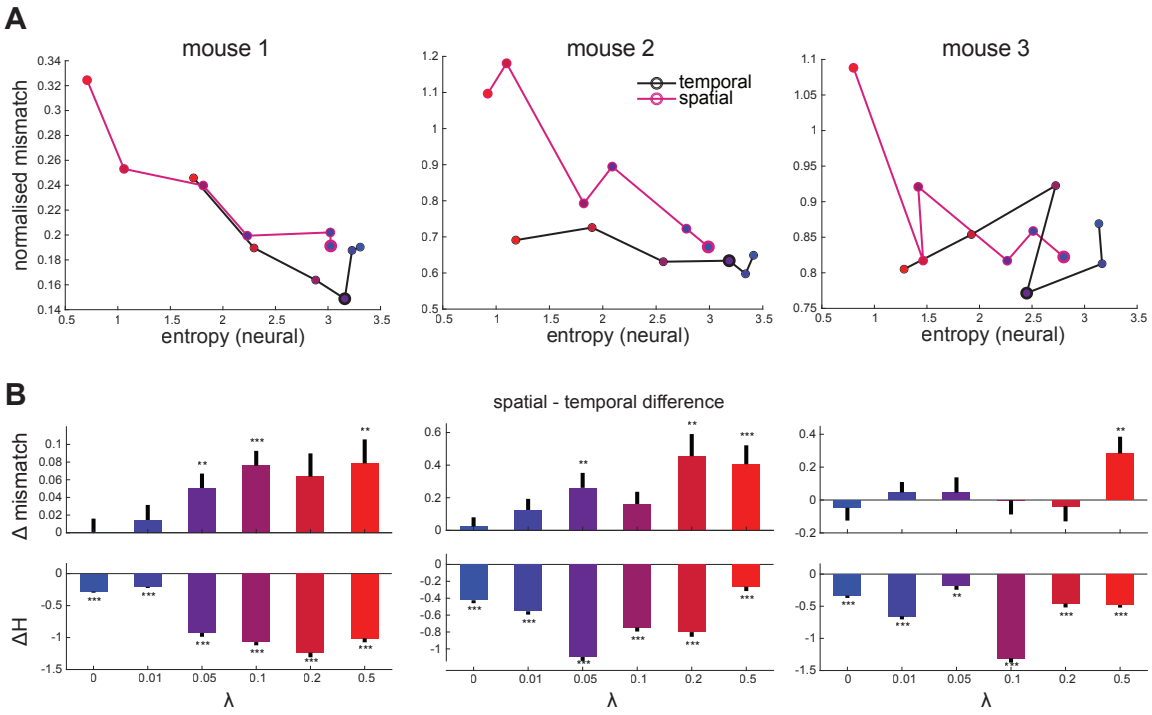


Figure C.2. Entropy-mismatch trade-off. A. Normalized mismatch between the temporal (black) and spatial (magenta) neural posteriors and the behavioral perceptual posterior, plotted as a function of total neural entropy at varying values of the momentary entropy penalty's λ multiplier (dots). **B.** Difference in raw mismatch (upper) and total neural entropy (lower) between the spatial and temporal codes across different values of the momentary entropy penalty's λ multiplier.

Bibliography

Abbott, L. F. and Dayan, P. (2001). Theoretical neuroscience. Comput Math Model Neural, 60:489–95.

Acerbi, L. and Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. Advances in Neural Information Processing Systems, 30:1834–1844.

Adams, W. J., Graf, E. W., and Ernst, M. O. (2004). Experience can change the'light-from-above'prior. Nature neuroscience, 7(10):1057–1058.

Adler, W. T. and Ma, W. J. (2018). Comparing bayesian and non-bayesian accounts of human confidence reports. PLoS computational biology, 14(11):e1006572.

Aitchison, L. and Latham, P. E. (2014). Bayesian synaptic plasticity makes predictions about plasticity experiments in vivo. arXiv preprint arXiv:1410.1029.

Aitken, F., Turner, G., and Kok, P. (2020). Prior expectations of motion direction modulate early sensory processing. Journal of Neuroscience, 40(33):6389–6397.

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Current biology, 14(3):257–262.

Aldous, D. J. (1985). Exchangeability and related topics. In École d'Été de Probabilités de Saint-Flour XIII—1983, pages 1–198. Springer.

Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., Griffiths, T. L., Hartshorne, J. K., Hauser, T. U., Ho, M. K., et al. (2024). Using games to understand the mind. Nature Human Behaviour, pages 1–9.

Amvrosiadis, T. (2023). Representation of perceptual uncertainty in mouse primary visual cortex. PhD thesis, The University of Edinburgh.

Arató, J. (2018). Active learning as a link between environmental statistics and the development of internal representations. PhD thesis, Central European University.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., and Griffiths, T. L. (2015). Structure and flexibility in bayesian models of cognition. Oxford handbook of computational and mathematical psychology, pages 187–208.

Bach, D. R. and Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. Nature reviews neuroscience, 13(8):572–586.

Barthelmé, S. and Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. Proceedings of the National Academy of Sciences, 107(48):20834–20839.

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. Scientific reports, 6(1):19203.

Bays, P. M., Schneegans, S., Ma, W. J., and Brady, T. F. (2024). Representation and computation in visual working memory. Nature Human Behaviour, pages 1–19.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic population codes for bayesian decision making. Neuron, 60(6):1142–1152.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. Nature neuroscience, 10(9):1214–1221.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science, 331(6013):83–87.

Berniker, M. and Kording, K. (2008). Estimating the sources of motor errors for adaptation and generalization. Nature neuroscience, 11(12):1454–1461.

Berniker, M. and Kording, K. P. (2011). Estimating the relevance of world disturbances to explain savings, interference and long-term motor adaptation effects. PLoS Comput Biol, 7(10):e1002210.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. Journal of machine learning research, 20(28):1–6.

Bosch, E., Fritsche, M., Ehinger, B. V., and de Lange, F. P. (2020). Opposite effects of choice history and evidence history resolve a paradox of sequential choice bias. Journal of vision, 20(12):9–9.

Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. Science, 340(6128):95–98.

Chen, Z. et al. (2003). Bayesian filtering: From kalman filters to particle filters, and beyond. Statistics, 182(1):1–69.

Chopin, A. and Mamassian, P. (2012). Predictive properties of visual adaptation. Current biology, 22(7):622–626.

Cicchini, G. M., Mikellidou, K., and Burr, D. C. (2024). Serial dependence in perception. Annual Review of Psychology, 75(1):129–154.

Clark, J. J. and Yuille, A. L. (2013). Data fusion for sensory information processing systems, volume 105. Springer Science & Business Media.

Courville, A. C. (2006). A latent cause theory of classical conditioning. Carnegie Mellon University.

Dayan, P. and Hinton, G. E. (1996). Varieties of helmholtz machine. Neural Networks, 9(8):1385–1403.

Dayan, P. and Kakade, S. (2000). Explaining away in weight space. Advances in neural information processing systems, 13.

de Gardelle, V., Le Corre, F., and Mamassian, P. (2016). Confidence as a common currency between vision and audition. Plos one, 11(1):e0147901.

De Gardelle, V. and Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? Psychological science, 25(6):1286–1288.

Deneve, S. (2005). Bayesian inference in spiking neurons. In Advances in neural information processing systems, pages 353–360.

Denison, R. N., Adler, W. T., Carrasco, M., and Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. Proceedings of the National Academy of Sciences, 115(43):11090–11095.

Devkar, D., Wright, A. A., and Ma, W. J. (2017). Monkeys and humans take local uncertainty into account when localizing a change. Journal of Vision, 17(11):4–4.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. Journal of Neuroscience, 32(11):3612–3628.

Drugowitsch, J., Moreno-Bote, R., and Pouget, A. (2014). Relation between belief and performance in perceptual decision making. PloS one, 9(5):e96511.

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. Neuron, 92(6):1398–1411.

Echeveste, R., Aitchison, L., Hennequin, G., and Lengyel, M. (2020). Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. Nat Neurosci.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870):429–433.

Fischer, J. and Whitney, D. (2014). Serial dependence in visual perception. Nature neuroscience, 17(5):738–743.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends in cognitive sciences, 14(3):119–130.

Forstmann, B. U., Ratcliff, R., and Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. Annual review of psychology, 67.

Fritsche, M., Mostert, P., and de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. Current Biology, 27(4):590–595.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., and Latham, P. E. (2014). The perception of probability. Psychological Review, 121(1):96.

Gershman, S. and Goodman, N. (2014). Amortized inference in probabilistic reasoning. In Proceedings of the annual meeting of the cognitive science society, volume 36.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2012). Multistability and perceptual inference. Neural computation, 24(1):1–24.

Gilbert, C. D. and Li, W. (2013). Top-down influences on visual processing. Nature Reviews Neuroscience, 14(5):350–363.

Glaze, C. M., Kable, J. W., and Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. Elife, 4:e08825.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. Annual review of neuroscience, 30.

Goldstone, R. L. (1998). Perceptual learning. Annual review of psychology, 49(1):585–612.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Grabska-Barwińska, A., Barthelmé, S., Beck, J., Mainen, Z. F., Pouget, A., and Latham, P. E. (2017). A probabilistic approach to demixing odors. Nature neuroscience, 20(1):98–106.

Grabska-Barwinska, A., Beck, J., Pouget, A., and Latham, P. (2013). Demixing odors-fast inference in olfaction. In Advances in Neural Information Processing Systems, pages 1968–1976.

Griffiths, T., Kemp, C., and B Tenenbaum, J. (2008). Bayesian models of cognition.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 29.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. Trends in cognitive sciences, 14(8):357–364.

Haefner, R. M., Berkes, P., and Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. Neuron, 90(3):649–660.

Harris, K. D. and Mrsic-Flogel, T. D. (2013). Cortical connectivity and sensory coding. Nature, 503(7474):51–58.

Heald, J. B., Lengyel, M., and Wolpert, D. M. (2021). Contextual inference underlies the learning of sensorimotor repertoires. Nature, 600(7889):489–493.

Heilbron, M. and Meyniel, F. (2019). Confidence resets reveal hierarchical adaptive learning in humans. PLoS computational biology, 15(4):e1006972.

Hemmer, P. and Steyvers, M. (2009). A bayesian account of reconstructive memory. Topics in Cognitive Science, 1(1):189–202.

Hénaff, O. J., Boundy-Singer, Z. M., Meding, K., Ziemba, C. M., and Goris, R. L. (2020). Representation of visual uncertainty through neural gain variability. Nature communications, 11(1):2513.

Henschke, J. U., Dylda, E., Katsanevaki, D., Dupuy, N., Currie, S. P., Amvrosiadis, T., Pakan, J. M., and Rochefort, N. L. (2020). Reward association enhances stimulus-specific representations in primary visual cortex. Current Biology, 30(10):1866–1880.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The" wake-sleep" algorithm for unsupervised neural networks. Science, 268(5214):1158–1161.

Ho, M. K., Abel, D., Griffiths, T. L., and Littman, M. L. (2019). The value of abstraction. Current opinion in behavioral sciences, 29:111–116.

Honig, M., Ma, W. J., and Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. Proceedings of the National Academy of Sciences, 117(15):8391–8397.

Houlsby, N. M., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., and Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. Current Biology, 23(21):2169–2175.

Hoyer, P. O. and Hyvärinen, A. (2003). Interpreting neural response variability as monte carlo sampling of the posterior. In Advances in neural information processing systems, pages 293–300.

Huszár, F. (2018). Note on the quadratic penalties in elastic weight consolidation. Proceedings of the National Academy of Sciences, page 201717042.

Jammalamadaka, S. R. and Sengupta, A. (2001). Topics in circular statistics, volume 5. world scientific.

Jardri, R., Duverne, S., Litvinova, A. S., and Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. Nature communications, 8(1):1–13.

Jaynes, E. T. (1996). Probability theory: the logic of science. Washington University St. Louis, MO.

Kahneman, D. (2013). A perspective on judgment and choice: Mapping bounded rationality. Progress in Psychological Science around the World. Volume 1 Neural, Cognitive and Developmental Issues., pages 1–47.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. Proceedings of the National Academy of Sciences, 105(31):10687–10692.

Khalvati, K., Kiani, R., and Rao, R. P. (2021). Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. Nature communications, 12(1):5704.

Kiani, R., Corthell, L., and Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. Neuron, 84(6):1329–1342.

Kiani, R., Hanks, T. D., and Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. Journal of Neuroscience, 28(12):3017–3029.

Kiani, R. and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. science, 324(5928):759–764.

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

Koblinger, Á., Fiser, J., and Lengyel, M. (2021). Representations of uncertainty: where art thou? Current Opinion in Behavioral Sciences, 38:150–162.

Kok, P., Brouwer, G. J., van Gerven, M. A., and de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. Journal of Neuroscience, 33(41):16275–16284.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. Nature, 427(6971):244–247.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. Trends in cognitive sciences, 10(7):319–326.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. Neuron, 93(3):480–490.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual review of vision science, 1:417–446.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105.

Lacoste-Julien, S., Huszár, F., and Ghahramani, Z. (2011). Approximate inference for the loss-calibrated bayesian. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 416–424.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., and Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. Neuron, 84(1):190–201.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and brain sciences, 40:e253.

Lange, R. D., Shivkumar, S., Chattoraj, A., and Haefner, R. M. (2023). Bayesian encoding and decoding as distinct perspectives on neural coding. Nature neuroscience, 26(12):2063–2072.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. nature, 521(7553):436–444.

Lee, S., Gold, J. I., and Kable, J. W. (2020). The human as delta-rule learner. Decision, 7(1):55.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. JOSA A, 20(7):1434–1448.

Lengyel, G. and Fiser, J. (2019). The relationship between initial threshold, learning, and generalization in perceptual learning. Journal of Vision, 19(4):28–28.

Lengyel, G., Shivkumar, S., and Haefner, R. M. (2024). A general method for testing bayesian models using neural data. In UniReps: the First Workshop on Unifying Representations in Neural Models.

Lengyel, M., Koblinger, Á., Popović, M., and Fiser, J. (2015). On the role of time in perceptual decision making. arXiv preprint arXiv:1502.03135.

Leptourgos, P., Notredame, C.-E., Eck, M., Jardri, R., and Denève, S. (2020). Circular inference in bistable perception. Journal of vision, 20(4):12–12.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behavioral and brain sciences, 43:e1.

Lieder, F., Griffiths, T. L., and Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. Psychological review, 125(1):1.

Lin, S., Gillis, W. F., Weinreb, C., Zeine, A., Jones, S. C., Robinson, E. M., Markowitz, J., and Datta, S. R. (2024). Characterizing the structure of mouse behavior using motion sequencing. Nature Protocols, pages 1–50.

Ma, W. J. (2012). Organizing probabilistic models of perception. Trends in cognitive sciences, 16(10):511–518.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nature neuroscience, 9(11):1432–1438.

Ma, W. J., Husain, M., and Bays, P. M. (2014). Changing concepts of working memory. Nature neuroscience, 17(3):347.

Ma, W. J. and Jazayeri, M. (2014). Neural coding of uncertainty and probability. Annual review of neuroscience, 37:205–220.

Maloney, L. T. and Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing bayesian transfer. Visual neuroscience, 26(1):147–155.

Maloney, L. T., Trommershäuser, J., and Landy, M. S. (2007). Questions without words: A comparison between decision making under risk and movement planning under risk.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience, 21(9):1281.

Mellers, B. A., Schwartz, A., and Cooke, A. D. (1998). Judgment and decision making. Annual review of psychology, 49(1):447–477.

Meyniel, F., Schlunegger, D., and Dehaene, S. (2015a). The sense of confidence during probabilistic learning: A normative account. PLoS computational biology, 11(6):e1004305.

Meyniel, F., Sigman, M., and Mainen, Z. F. (2015b). Confidence as bayesian probability: From neural origins to behavior. Neuron, 88(1):78–92.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. Proceedings of the National Academy of Sciences, 108(30):12491–12496.

Mostert, P., Kok, P., and de Lange, F. P. (2015). Dissociating sensory from decision processes in human perceptual decision making. Scientific reports, 5(1):18253.

Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. Journal of Neuroscience, 30(37):12366–12378.

Navarro, D. J. and Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. Journal of mathematical psychology, 53(4):222–230.

Newell, B. R., Lagnado, D. A., and Shanks, D. R. (2022). Straight choices: The psychology of decision making. Psychology Press.

Newell, B. R. and Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. Behavioral and brain sciences, 37(1):1–19.

Niell, C. M. and Stryker, M. P. (2008). Highly selective receptive fields in mouse visual cortex. Journal of Neuroscience, 28(30):7520–7536.

Orbán, G., Berkes, P., Fiser, J., and Lengyel, M. (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. Neuron, 92(2):530–543.

Orbanz, P. and Teh, Y. W. (2010). Bayesian nonparametric models. Encyclopedia of machine learning, pages 81–89.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In NIPS-W.

Paulun, V. C., Schütz, A. C., Michel, M. M., Geisler, W. S., and Gegenfurtner, K. R. (2015). Visual search under scotopic lighting conditions. Vision research, 113:155–168.

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in numpyro. arXiv preprint arXiv:1912.11554.

Piray, P. and Daw, N. D. (2020). A simple model for learning in volatile environments. PLoS computational biology, 16(7):e1007963.

Pitkow, X. (2016). Probability by time. Neuron, 92(2):275–277.

Prat-Carrabin, A., Meyniel, F., and da Silveira, R. A. (2024). Resource-rational account of sequential effects in human prediction. Elife, 13:e81256.

Qamar, A. T., Cotton, R. J., George, R. G., Beck, J. M., Prezhdo, E., Laudano, A., Tolias, A. S., and Ma, W. J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. Proceedings of the National Academy of Sciences, 110(50):20332–20337.

Raju, R. V. and Pitkow, Z. (2016). Inference by reparameterization in neural population codes. In Advances in Neural Information Processing Systems, pages 2029–2037.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. Neural computation, 20(4):873–922.

Resulaj, A., Kiani, R., Wolpert, D. M., and Shadlen, M. N. (2009). Changes of mind in decision-making. Nature, 461(7261):263–266.

Sahani, M. and Dayan, P. (2003). Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. Neural computation, 15(10):2255–2279.

Schneegans, S., Taylor, R., and Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. Proceedings of the National Academy of Sciences, 117(34):20959–20968.

Schneider, S., Lee, J. H., and Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. Nature, 617(7960):360–368.

Shadlen, M. N. and Kiani, R. (2013). Decision making as a window on cognition. Neuron, 80(3):791–806.

Shadlen, M. N. and 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.

Shadlen, M. N. and Shohamy, D. (2016). Decision making and sequential sampling from memory. Neuron, 90(5):927–939.

Shushruth, S., Zylberberg, A., and Shadlen, M. N. (2022). Sequential sampling from memory underlies action selection during abstract decision-making. Current Biology, 32(9):1949–1960.

Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. arXiv preprint arXiv:1803.08533.

Sotiropoulos, G., Seitz, A. R., and Seriès, P. (2011). Changing expectations about speed alters perceived motion direction. Current Biology, 21(21):R883–R884.

Stocker, A. A. and Simoncelli, E. P. (2008). A bayesian model of conditioned perception. In Advances in neural information processing systems, pages 1409–1416.

Straub, D. and Rothkopf, C. A. (2022). Putting perception into action with inverse optimal control for continuous psychophysics. Elife, 11:e76635.

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

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. science, 331(6022):1279–1285.

Tomić, I. and Bays, P. M. (2024). A dynamic neural resource model bridges sensory and working memory. Elife, 12:RP91034.

Van den Berg, R., Yoo, A. H., and Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. Psychological review, 124(2):197.

Vértes, E. and Sahani, M. (2018). Flexible and accurate inference and learning for deep generative models. In Advances in Neural Information Processing Systems, pages 4166–4175.

Vértes, E. and Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. In Advances in Neural Information Processing Systems, pages 13714–13724.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., and Kording, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. Current Biology, 22(18):1641–1648.

Walker, E. Y., Cotton, R. J., Ma, W. J., and Tolias, A. S. (2020). A neural basis of probabilistic computation in visual cortex. Nature Neuroscience, 23(1):122–129.

Walker, E. Y., Pohl, S., Denison, R. N., Barack, D. L., Lee, J., Block, N., Ma, W. J., and Meyniel, F. (2023). Studying the neural representations of uncertainty. Nature neuroscience, 26(11):1857–1867.

Wei, X.-X. and Stocker, A. A. (2015). A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. Nature neuroscience, 18(10):1509–1517.

White, C. N. and Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(2):385.

Whiteley, L. and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. Journal of vision, 8(3):2–2.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2020). The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. Cell, 183(5):1249–1263.

Wyart, V., De Gardelle, V., Scholl, J., and Summerfield, C. (2012). Rhythmic fluctuations in evidence accumulation during decision making in the human brain. Neuron, 76(4):847–858.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3):356–365.

Yang, S. C.-H., Lengyel, M., and Wolpert, D. M. (2016a). Active sensing in the categorization of visual patterns. Elife, 5:e12215.

Yang, S. C.-H., Wolpert, D. M., and Lengyel, M. (2016b). Theoretical perspectives on active sensing. Current opinion in behavioral sciences, 11:100–108.

Yoo, A. H., Acerbi, L., and Ma, W. J. (2021). Uncertainty is maintained and used in working memory. Journal of vision, 21(8):13–13.

Yuille, A. and Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In Perception as Bayesian inference, pages 123–161. Cambridge University Press.

Zemel, R. S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. Neural computation, 10(2):403–430.

Zoltowski, D. M. (2016). On the role of time in perception. Master's thesis.

Zylberberg, A., Wolpert, D. M., and Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. Neuron, 99(5):1083–1097.