The best-laid models of mice and men:
Towards a holistic characterisation of animal behaviour

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Sebastian A. Bruijns
aus Bensheim, Deutschland

2026

Tag der mündlichen Prüfung:	17.9.2025
Dekan der Math.-Nat. Fakultät:	Prof. Dr. Thilo Stehle
Dekan der Medizinischen Fakultät:	Prof. Dr. Bernd Pichler
1. Berichterstatter:	Prof. Dr. Peter Dayan
2. Berichterstatter:	Prof. Dr. Jonathan Pillow
Prüfungskommission:	Prof. Dr. Peter Dayan
	Prof. Dr. Jonathan Pillow
	Prof. Dr. Jakob Macke
	Prof. Dr. Eric Schulz

Erklärung / Declaration: Ich erkläre, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

"**The best-laid models of mice and men:**
Towards a holistic characterisation of animal behaviour"

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

I hereby declare that I have produced the work entitled "***The best-laid models of mice and men: Towards a holistic characterisation of animal behaviour***", submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, den ....

Datum / Date Unterschrift / Signature

ACKNOWLEDGMENT

A simple thank you, no matter how verbose, would never do. This work owes its existence, and I the academic part of mine, to Peter. Your unwavering support, neverending kindness, great humor in both give and take, and your bottomless well of knowledge, insights, and ideas, make you the point at infinity for my lines of scientific and personal growth.

The third main chapter owes a lot to the additional supervision of Maria, thank you for your friendly guidance. Thank you to Eric and Jakob for their support as part of my advisory committee. Thank you to Jonathan for your advice and support as my IBL supervisor.

I owe an immense debt of gratitude to my emotional support network of dear friends, without whom this would have been much harder on me. Thank you Friedi, for everything you do and are. Thank you Fabi, for being an even better friend than neuroscientist (paradoxical, I know). Thank you Julia and Dominik, thank you Louis and Rebekka, thank you Markus and Miri. Gratitude is not my strong suit, but I doubt anyone could help but feel inadequate in the face of the debt of gratitude I owe to my parents, for having enabled me to get where I am in every way. Thank you for that.

It was (and is) an exceptional privilege and pleasure to have been part of the lab from its very beginning in Tübingen. I am shaped for the better by my extraordinary colleagues, who have always made life in the lab fun and rewarding, and a time I am already nostalgic for. I am particularly grateful to my closest friends in the lab: to Misi for his mentoring and dry humor, to Franzi for being my great first office buddy, to Tianyuan for finding me, an unproductive loser, a job, to Charline for livening up the lab and IBL trips, to Sofiya for being a great student, to Susan for keeping the lab running, to Andrew for keeping me grounded, to Noémi, and to Peter for, among other things, putting punctuation marks in my sentences, when given the chance.

I thank the many IBL members for their input from afar and fun times when we came together. Especially Jonathan for extra supervision and being part of my paper board, thanks for that also goes to Anne. I thank Mayo and Julia for reading my code and publishing my data, and a huge thank you to Charles for a code review on the MCMC details.

I thank the MPI IT team, Walter, Joachim, Martin, and Haydar, for magnanimously putting up with my last minute poster printing and other assorted technical problems. Thank you to Dagmar and Bianca for bureaucratic support.

And, though it will never do, thank you to Peter, for everything you have done for me.

STATEMENT OF CONTRIBUTIONS

The work presented in the chapters of this thesis is my own and was written by me, though it enjoyed the support of various much appreciated colleagues. The first main chapter presents a new modelling framework for capturing the learning trajectory of animals on a new task. It was conceived by me as a combination of previous modelling efforts, under the supervision of Peter Dayan. The code to implement it was written by me and represents an extension of the code of Matthew J. Johnson and Alan S. Willsky, who were not involved in this work. The validation analyses for this model were performed by me, under the supervision of Peter Dayan. This project was developed within the framework of the International Brain Laboratory (IBL) and the theoretical side of this work benefited from the feedback of IBL members, particularly those of the Theory group, Jonathan Pillow and Anne Churchland who served as our advisory board. Charles Findling conducted a code review.

The second chapter presents the application of this model to the extensive data set of the IBL. I was not involved in the collection of this data, but am instead grateful to my colleagues who did so: Kcénia Bougrova, Inês C. Laranjeira, Petrina Y. P. Lau, Guido T. Meijer, Nathaniel J. Miska, Jean-Paul Noel, Alejandro Pan-Vazquez, Noam Roth, Karolina Z. Socha, and Anne E. Urai. The results of the model fits were analysed by me, under the supervision of Peter Dayan. I am again grateful for feedback on the project from many IBL members, particularly again the members of my advisory board, Jonathan Pillow and Anne Churchland. I also received help from the IT staff of the IBL in interfacing with the data and fixing issues with the behavioural data. This and the previous chapter combined were also written up and published as a preprint on bioRxiv. I wrote the first draft for this and was heavily involved in the editing of the manuscript, together with Peter Dayan. Again, we received valuable feedback from IBL members and members of Peter's lab.

The third chapter contains my work on applying the hybrid neural network approach of Maria Eckstein to the behaviour of expert IBL mice. I designed and implemented the code for this, under the supervision of Maria Eckstein and Peter Dayan, working off of code provided by Maria Eckstein for a different task. The analyses of the results were performed by me, under the supervision of Maria Eckstein and Peter Dayan. I acknowledge helpful feedback from the IBL and local lab members.

CONTENTS

Acknowledgements vii
Statement of contributions ix

Summary xiii

I Introduction
1 Introduction 3
1.1 Modelling aspect: Learning 5
1.2 Modelling aspect: Motivation 7
1.3 Modelling aspect: Inter-individual differences 8
1.4 Modelling aspect: Perseveration 9
1.5 Our model of learning: The dynamic infinite hidden Markov model 10
1.6 Modelling aspect: Model limitations 12
1.7 Our model of expert behaviour: Hybrid networks 13
1.8 Outline 13

II Behavioural Task
2 Behavioural Task 17

III Modelling the learning trajectory
3 Learning trajectory model setup 23
3.1 Infinite hidden semi-Markov model 25
3.2 Dynamic logistic regression prior and sampling 27
3.3 Aggregation and interpretation of chains 30
3.4 Cross-validation and ablations 33
3.5 Posterior predictive checks 35
3.6 Model recovery 37

IV Behavioural fits
4 Learning trajectory fit 43
4.1 Single animal fit 43
4.1.1 Posterior uncertainty 45
4.2 Fits across the population 46
4.2.1 Inter-individual differences and variability 50
4.3 Discussion 51

V Modelling expert behaviour
5 Hybrid neural network modelling of expert behaviour 57
5.1 Hybrid network ladder 58
5.2 Model results – Climbing the ladder 63
5.2.1 Network 100 64
5.2.2 Network 111 65
5.2.3 Network 200 66
5.2.4 Network 202-20 69
5.2.5 200-iv 70
5.3 Model analysis 71
5.3.1 Individual differences versus trial differences 73
5.4 Discussion 75

VI Discussion
6 Discussion 81

VII Supplemental material for diHMM behavioural fits
A Supplemental material for diHMM behavioural fits 89
A.1 Psychometric type classification 89
A.2 Bias training analysis 90
A.3 Supplemental figures 90

VIII Supplemental material for hybrid neural networks
B Supplemental material for hybrid neural networks 97
B.1 Other network architectures 97
B.2 Network training and evaluation 97
B.3 Additional figures 98

Bibliography 103

SUMMARY

The mathematical modelling of behaviour enables the formalisation of theories of cognition. Removed from the details of neural implementation, we can abstractly reason about properties of the algorithms employed by the brain which transform the presented inputs into the observed actions. However, there are a number of complications, both in behaviour itself and the process of modelling it, which impede a comprehensive characterisation of behaviour. In this thesis, we present two separate modelling approaches which deal with some of these issues. We showcase these frameworks on the International Brain Laboratory (IBL) data of over 100 mice performing a perceptual decision-making task. Mice learn the basic contingencies of this task over a number of sessions and many thousands of trials. Afterwards, the task gains a biased block structure, requiring the animals to track this hidden state to improve their task performance.

We first build a highly flexible model which deals with the issues of non-stationary behaviour due to learning and motivation, along with individual differences. This is achieved using an infinite hidden Markov model (iHMM) which provides a state based description of behaviour, with a non-parametric Bayesian structure. The latter allows for the introduction of new states in response to drastic changes in behaviour (such as learning through a sudden insight or motivational fluctuations). We fit this model to individuals independently, exploiting automated complexity control. Dynamics in the characterisation of the behavioural states additionally imbue the model with the capacity to capture gradual learning. This allows us to identify distinct stages of learning which are present throughout the population of IBL mice. We also find substantial inter-individual differences in our model-based characterisations, and quantify the limited predictability of the course of learning.

The second model we present uses neural networks to overcome the inherent rigidities of models such as the iHMM, by progressively removing restrictions from the class of modellable functions. This amounts to hybridising the neural networks with a classical model of expert mouse behaviour on our task, to maintain interpretability. We use this to find a simple extension of the classical model which outperforms it, and thus provides a powerful but interpretable full model of task behaviour. Amongst other insights, it shows how motivational fluctuations represent a substantial source of behavioural variability for which any complete model will have to account.

We thus provide tools which bring the field closer to a holistic modelling approach of animal behaviour.

Part I

INTRODUCTION

1 INTRODUCTION

Intelligence is what lies between the sensory inputs an organism or agent receives and the actions it takes. It is the goal of behavioural modelling to understand the mediating processes, purely by considering directly observable inputs and outputs (Daw, 2011; Wilson et al., 2019). David Marr famously delineated three levels of understanding for such processes (Marr, 2010): (i) The *computational level* is about the question of what a computation achieves and which purpose it thereby serves. (ii) The *algorithmic level* considers how a computation turns input into outputs, what the intermediate representations are, and how they are manipulated. (iii) Lastly, the *implementational level* details how the abstract computations of the algorithmic level are actually realised, in the case of animal cognition through neuronal dynamics. Marr's idea was that true understanding of an information processing system requires complementary answers at all three levels. Other systems of levels are related – for instance, with Tinbergen's notion of 'function' being closely tied to Marr's computational level, and of 'mechanism' to the algorithmic and implementational levels (MacDougall-Shackleton, 2011).

Behavioural modelling is best suited to questions at the algorithmic level. For this, it is usually applied to data collected in the well-controlled setting of a laboratory experiment (Skinner, 1953): this generally allows researchers to determine the sensory inputs and to register an animal's outputs through the well-defined channels of the experiment, pinning down the start and end point of a computation. We can then formulate and test models of how the animal performs the required transformation, via the touchstone of empirical data. To this end, behaviour can be elicited in response to an enormous range of experimental paradigms, from the basal perceptual processing probed in psychophysics (Fechner, 1860) to the abstract considerations of choices under probabilistic uncertainty (Tversky et al., 1974). Thus, a great many cognitive processes are amenable to this scientific method. Behavioural modelling is also associated with Marr's other two levels, as the question of what a computation aims to achieve in the first place is unavoidable in experimental design, model specification, and beyond, and insights at the algorithmic level may help constrain neural details at the implementation level. Behavioural modelling is thus a crucial component in a holistic understanding of cognition (Krakauer et al., 2017).

Orthogonal to Marr's levels of analysis is a different axis of understanding (Craver, 2006): our models can be a description or summary of the behavioural data (a descriptive or phenomenological model), or they may provide a mechanism for how the observed data come about (a mechanistic or cognitive model). Descriptive behavioural models aim to arrive at a useful and succinct formal description of behaviour, making it amenable to other analyses. While not directly furthering the understanding of cognition, these models can integrate out the noise of behaviour, enabling understanding through their simpler or more direct representations. Thus, with a descriptive model we can extract quantities of interest with more clarity than may be possible from raw behaviour (such as the time spent at a certain level of understanding, as we will discuss later).

By contrast, cognitive behavioural models aim to establish a form of isomorphism to the actual processes of cognition, emulating the mapping from sensory cues to behavioural output as closely as possible within a restricted formal language. Descriptive and mechanistic models are two extremes on a spectrum, an actual model will generally aim to provide mechanistic explanations for some aspects of the behaviour, but appealing to descriptions of underlying components.

The classical approach to either sort of behavioural modelling consists of defining a manageable set of equations which tie together abstract representations of the observations, latent variables representing concepts of interest for or about the agent being modelled, and probabilities over actions. These equations are steered by a small set of parameters, which are fitted to the administered or observed input sequence and the output sequences of the agent. A model thus provides a rigid structure in the form of equations, which can be somewhat adjusted to a specific context, individual, or group via the setting of parameters. Importantly, these equations, latent variables, and parameters are usually designed in such a way as to yield easily interpretable models of the underlying process. We can illustrate this with the popular mechanistic example of a Q-learning agent (Watkins et al., 1992), which acquires the capacity to choose actions to maximize rewards or minimize punishments. Here, the latent concepts are the future cumulative rewards an agent expects from taking actions (the Q-values), which are estimated via a simple update equation or learning rule, with the magnitude of each update being directly modulated by a learning rate parameter.

Parameters can be fit in a variety of ways. In general, the goal is to have the model assign high probabilities to the actually performed actions. An important distinction in parameter estimation is whether they are optimised via maximum likelihood, or in a Bayesian manner: Maximum likelihood simply sets the parameters such that the probability for chosen actions is maximal, yielding point estimates of the free parameters. A Bayesian fit finds the posterior over the parameters of interest, treating them as random variables. This requires the specification of a prior over parameters, introducing, depending on one's point of view, a source of subjectivity into the modelling process, or a useful way to embed existing knowledge. More problematically, the computation of the posterior requires the computation of generally intractable integrals (namely the evidence, i.e. the marginal probability of the data themselves). This necessitates the use of approximate inference, such as through variational inference or Markov-Chain Monte-Carlo methods. The benefit of using Bayesian methods is that we do not just arrive at a point-estimate for each parameter, but rather infer a full joint distribution over them. This makes explicit the uncertainty inherent in estimates, and can reveal issues such as parameters trading off with one another.

The resulting fitted parameters within the framework of a model represent a succinct summary of the observed behaviour. Within a descriptive framework, these parameters simply serve the purpose of describing the taken actions (e.g. providing a distribution over choices at each point), without necessarily making causal claims about underlying processes. A cognitive model goes further, surfacing parameters and latent constructs (such as Q-values) that may be detectable in the brain using the methods of neuroscience. However, in a vacuum, we may only know that the fitted parameters have the most appropriate values within the given modelling framework, but not how well the model itself, or the model plus the fitted parameters, fits the behaviour on an absolute scale. If the model is itself misspecified, i.e., is not an accurate reflection of the kind

of processes going on during decision-making, the inferred parameters may not be very useful. A partial answer is for behavioural modelling to engage in fitting and comparing against each other multiple candidate models, to find the generally most faithful framework for the modelled behaviour. However, this still leaves open the question of the quality of untested models, an issue we will revisit later.

With a best-performing model selected and parameters fitted, we have arrived at our desired model of behaviour. We can then use it to extract latent variables which we assumed to be present in the brain and/or mind of a subject (like the Q-values discussed earlier) and correlate them with other measures of interest (e.g. neural activity, to go beyond pure behaviour). The same can be done with the fitted parameters of the model (e.g. a learning rate).

It is a critical step in modelling to decide on the nature and level of detail at which the data should be described and fitted, and adopting an appropriate degree of simplification. Making simplifying assumptions is an unavoidable part of the modelling process. An agent may not be so simplistic as to be captured by a stationary algorithm which is completely uniformly implemented across the population. We will highlight four prominent issues in behavioural modelling, stemming from complexities within the modelled animals: learning, motivational fluctuations, inter-individual differences, and perseverative biases. We visit these issues in turn, discussing their effects and how researchers have dealt with them in the past. Afterwards, we will outline the ideas which we will apply in this work to address these difficulties, in our effort to model the learning trajectories in the mouse perceptual decision-making task that we consider throughout (the so-called international brain lab or IBL task, The International Brain Laboratory et al., 2021). Following this, we will confront another problem, this time stemming from the inherent limits of the model design process itself. Capturing processes through a manageable and interpretable set of equations and variables imposes structure upon the process being modelled and it can be difficult to ascertain whether the made assumptions are justified or not. We will introduce our approach to overcome the issues of model assumptions, while also accounting for the previously discussed issues, that we will apply to the expert behaviour of IBL mice.

1.1 MODELLING ASPECT: LEARNING

Engaging with a new environment or experiment raises a multitude of questions: Which sensory signals are pertinent to the task at hand, and which are just noise? What actions are relevant to performance? How should observations inform actions? As an agent gains experience with an environment, its behaviour will naturally change, as it starts to form answers to these questions and moves from uninformed to proficient behaviour. The details of learning are intimately tied to the question of intelligence. It has been argued that the most remarkable aspect of animal intelligence is its accuracy and robustness in the face of few samples (Lake et al., 2017), in stark contrast to the extensive training required by deep neural networks trained via backpropagation (Lake et al., 2015). A better understanding of the processes which mediate learning thus represents a tantalizing goal.

Learning constitutes a fundamental non-stationarity in behaviour, as the process of decision-making adapts to the current best understanding of the animal. During the initial task acquisition period, an animal has to learn the relevant structure of the world (Gershman et al., 2010b) and how to act in it through the use of feedback (Sutton et al., 2018). These improvements

through experience entail great changes in behaviour, and it is immediately obvious that behaviour is not stationary in this period. Learning is generally seen as a normative process, in the sense that an increase in information *should* lead to changes in behaviour. However, learning comes in a number of subtly different forms, some more desirable to the modeller than others.

Some tasks will purposefully induce learning even after basic acquisition has (more or less) completed, in order to model learning processes. Examples include protocols such as reversal learning, in which an established conditioned association is suddenly altered (Izquierdo et al., 2017), or the ongoing tracking of non-stationary probabilities in a multi-armed bandit task (Daw et al., 2006). While these are interesting in their own right, they represent changes in modest facets of a task, not requiring completely de novo learning. It is typical in such cases that modelling only begins after most of the problem has been solved, with subjects who failed to acquire overall competence having been discarded.

However, even in stationary tasks, it is of course never a reasonable assumption to make that the world is 100% unchanging and subjects will likely engage in at least some amount of inference about the current world model. While this may still be normatively correct, it is nevertheless undesirable for researchers who are interested in the behaviour itself, rather than continual learning or adaptation around it. This ongoing learning may be sufficiently negligible to be ignored, and we may assume that each response, during the period under study, is produced via the same process.[1]

One of the main reasons for not tackling the problem of learning in its full breadth, is that each animal provides only one sample of a learning curve, whereas for fully acquired behaviour, every trial can typically be viewed as another sample from the learned behaviour. This means that learning curve data are generally sparse, further aggravating issues through large inter-individual variability (which we highlight below). We make use of the large data set collected by the IBL (The International Brain Laboratory et al. (2021)), to build a rigorous descriptive model of the multi-session learning curves of more than 100 mice coming to solve a perceptual decision-making task. This puts us in an exceptional position to tackle the sparsity of data.

Previous work on quantifying acquisition has sought to achieve somewhat more modest goals, like finding the point in time at which an animal can be said to have "learned" a task, often defined as reliably above chance performance (Gallistel et al., 2004; Smith et al., 2004). Methods for solving this kind of change-point detection task (e.g., Papachristos et al. (2006) and Jang et al. (2015)) typically make a binary distinction between uninformed and learned behaviour, rather than describing the strategies that are used in detail, or finding possible intermediate stages. However, there is previous work addressing strategy inference more specifically, which does consider learning (e.g., Maggi et al., 2022). In this paper, the authors perform inference on a trial-by-trial basis over a set of simple, pre-selected strategies, decaying evidence exponentially over time to track the arrival and departure of various strategies. Their framework does however require formalising possible strategies ahead of time and cannot incorporate probabilistic strategies. There has also been work on forms of progressive meta-learning in humans (Jain et al., 2023), which similarly works with a set of predefined strategies, their connection to performed actions, and a prior over strategy sequences, enabling inference with a focus on individual learning trajectories.

1 Note that, in a way, this is trivially true, as the animal itself instantiates this one single process. However, in practice, there will be fluctuations in this singular process which change response statistics in a sufficiently modest way that they are omitted from models.

One of the interesting questions about learning is whether it progresses through sudden jumps or gradual accumulations (Moore et al., 2022). Simply considering the population-averaged increase in performance over time, one might be inclined, from the smoothness of such a curve, to conclude that it is a gradual process. For instance, one of the most influential models of learning, the Rescorla-Wagner rule, implements learning as the gradual, step-wise updating of weights which associate, for example, stimuli and rewards (Rescorla, 1972). There are indeed notable examples of people learning extremely gradually over a great number of trials, for instance, when the underlying pattern is rather subtly obscured (Éltető et al., 2022). However, it has also been recognised that apparent slow learning can be an artefact of averaging over many animals: Individuals themselves often exhibit quite drastic increases in performance over a short time – which one might call insights (Maier, 1931; Köhler, 1948; Epstein et al., 1984) – but which, however, wash out into a smoothly increasing population average (Gallistel et al., 2004). Similarly, whereas the population may appear to progress as a Bayesian observer, individuals do not. This has given rise to the hypothesis that the population works as a particle filter approximation to Bayesian inference, with each individual holding on to just a single posterior particle, which undergoes occasional update jumps (Daw et al., 2007). Of course, the speed of progress may also differ between the algorithmic substrate and behavioural expression, as analysed in Löwe et al. (2024). In that work, the dissociation is a consequence of a thresholding process in decision-making.

Connected to these questions on the nature of change is the study of intermediate learning stages: Does an animal exhibit distinct intermediate behaviours, representing an incomplete advancement over initial uninformed behaviour, in which it remains for sufficiently long and sufficiently consistently to be identified as a distinct phase of understanding? And how do these different strategies reflect in the underlying representations which the animal uses? The existence of such stepping stones probably strongly depends on task details; however, in the right setting, they have been observed. Intermediate stages have been reported in motor skill learning, which has also been related to fast and slow learning (Luft et al. (2005), though the terms are used differently from the way we will use them). Liebana Garcia et al. (2023) observed distinct intermediate stages, and even entire trajectories connecting them, in a task very similar to the IBL task that we analyse. Intermediate strategies were also observed in humans by Dekker et al. (2022), in a task that required them to generalise observations across two dimensions. While most participants were best described by a model progressing directly from random behaviour to correct generalisation across both dimensions, a sizeable minority exhibited a distinct intermediate stage of generalising over one dimension only. Strategies at multiple levels of sophistication might even co-exist, which would then require a model with a repertoire of behaviours (e.g. Ashwood et al. (2022))

The dynamics of learning thus represent both a highly interesting target for research questions and possibly problematic stumbling block to our modelling efforts. We will present a model with the express purpose of dealing with these pervasive difficulties.

1.2 MODELLING ASPECT: MOTIVATION

Another major aspect of behavioural non-stationarity can be coarsely summarised as a fluctuation in the degree of motivation during task performance. This summarises many different phenomena, such as the waxing and waning

of attention a person may commit to a lengthy task, momentary bouts of inattenion on individual trials, the effects of an initial warm-up period when starting a well-known task after a short break, and the decrease in motivation of an animal which has acquired sufficient reward during a session so as to be sated at the end (after having initially been kept engaged as a consequence of restricted access to food or water). Whereas the drivers behind learning are part of a task, this is less so for motivation, making it harder to capture these processes mechanistically (but see, for instance, Meyniel et al., 2013; Brielmann et al., 2022).

The need to take motivational effects during decision-making tasks into account has been recognised for some time now. Wichmann et al. (2001) model momentary lapses in attention on single trials of perceptual decision-making tasks as an ever-present chance of a random response. They particularly highlight how this small chance of a mistake at any stimulus intensity may considerably affect the estimation of asymptotic performance on extremely salient stimuli, requiring explicit handling to accurately fit model parameters. In further work, Fründ et al. (2011) analyse non-stationarities more generally (also considering temporally extended motivational effects and learning), discussing how to detect them via simulations and how to correct for them by adjusting the credible intervals of estimated parameters of a psychometric function. Notably, lapses have also been contextualised as a form of exploration (Pisupati et al., 2021), as they have shown themselves to be manipulable in a one-sided manner by one-sided reward manipulations, which a pure noise process should not be. This would place them in the previous category of non-stationarity through learning. Evidently, lapses may have many different drivers.

Ashwood et al. (2022) extend the original lapse rate framework via a model which explicitly tracks multiple behavioural states, modelling behaviour as a mixture of policies. This allows for a more precise description of the disengaged strategies (rather than a uniform distribution over responses) and a temporal autocorrelation of the different modes of behaviour (rather than modelling lapses as independent events across trials).

As all these approaches highlight, considering all trials equally when fitting a model is likely to confound multiple modes of operation, leading to flawed parameter estimates, no matter how well trained a subject is, and therefore this needs to be accounted for in a comprehensive model of behaviour.

1.3 MODELLING ASPECT: INTER-INDIVIDUAL DIFFERENCES

When fitting a model to a population of subjects, we may do so at different levels of granularity. We can pool all individuals together and only fit an overall set of parameters, leading to a model which describes the population as a whole. Alternatively, model fits can be individualised by fitting parameters to animals independently, returning separate parameter estimates for each individual. This will better reflect differences in the approaches of the subjects, but will also tend to overestimate the true variability in the population, as noise in behaviour or other incidental factors can affect the fitting to a greater extent in the comparatively smaller data set of a single individual. An hierarchical model fitting procedure can help with this, as it partially pools information across individuals in a principled manner (Gelman, 2013). One step further still, different animals may be best described by different models, i.e. their decision-making processes do not just rely on different parameters, but differ-

ent equations entirely. Taking this stance necessitates an individualised model comparison.

Determining what exactly should be attributed to individual differences (through nature or nurture), versus natural noise, or day-to-day variability, is a subtle question (Boogert et al., 2018). However, there is evidence for real inter-individual differences within a population, not attributable to behavioural plasticity alone, which do play a sizeable role in affecting behaviour, called "animal personality" (Dingemanse et al., 2010) or "behavioural syndromes" (Sih et al., 2012). Dougherty et al. (2018) performed a large meta-analysis, studying the connection between animal personality and different learning characteristics (such as performance after training or learning speed). Interestingly, they found significant, if modest, effects, which were however so variable across studies, that they averaged out to be insignificantly different from zero.

Inter-individual differences also interact with the earlier highlighted issue of learning, as individuals progress at different speeds and over distinct intermediate stages (Piaget, 1952). Inter-individual differences during learning are a known phenomenon (Papachristos et al., 2006; The International Brain Laboratory et al., 2021), though they are rather less commonly studied (though, for instance, see Kastner et al. (2022) and Akiti et al. (2022)). Even if the ultimate behaviour is indistinguishable across individuals, the initial variability can make comparisons across groups during learning challenging (The International Brain Laboratory et al., 2021). More generally, the particularities of the learning path may never be fully overcome, so that their trace is detectable in performance even after learning has finished and behaviour has stabilised (Dayan et al., 2020).

Thus, with our later goals of accurately describing the highly individualised learning trajectories, and capturing as much variance as possible in expert behaviour, it will also be important to account for individual differences.

1.4 MODELLING ASPECT: PERSEVERATION

Animals show a tendency to repeat previous actions, as famously formalised in Thorndike's law of exercise (Thorndike, 1911). In behavioural experiments, this is often not appropriate, as researchers want to treat their data as independent and identically distributed (as many statistical methods require this assumption), and so they present trials in an independent manner. Systematically persevering in spite of this harms the reward rate of the animal and breaks the independence assumption which a researcher may make. It has been argued that this pervasive perseverative tendency is the result of a bias towards simpler strategies, which may explain why it persists in spite of causing a lower reward rate (Gershman, 2020a).

Even though it breaks one of the fundamental assumptions of many models, perseveration is a somewhat less concerning issue as compared to the others. It is reasonably well understood and straightforward to model. There are statistical methods for detecting such tendencies (Fründ et al. (2014), which quantifies intertrial dependencies more broadly), and they can be modelled as an increased probability to repeat the previous action (Roy et al., 2021; Ashwood et al., 2022) or as a bias towards a weighted average over previous actions (Akam et al., 2021).

Complicating this for us, however, is that in the later stages of the IBL task, perseveration is a partially desirable trait. This is because the task induces specific autocorrelations in optimal behaviour, as it involves blocks of trials in which one of two possible stimulus conditions is probabilistically more

prevalent. This induces autocorrelations into expert choices, and a measured degree of perseverative bias is thus a reasonable, though, as we will see, not optimal prior for the animal. Mice have shown themselves capable of appropriately adapting their perseverative tendencies in a very similar task (Fritsche et al., 2024), so we may expect them to settle on an efficient, though improveable, strategy which makes use of the biased block structure.

1.5 OUR MODEL OF LEARNING: THE DYNAMIC INFINITE HIDDEN MARKOV MODEL

We identified four broad categories of behavioural complexities as our focus for this work: learning, inter-individual differences, motivational fluctuations, and perseverative biases. As already highlighted for individualised learning, these complexities can interact: motivational fluctuations are individualised and cannot be described in a standardised way across the population, and motivation of course also influences the when and how of learning.

While we cannot hope to solve these difficulties in their entirety, the goal of this thesis is to provide methods which explicitly address them, taking into account the dynamics of behaviour through various processes, and generally increasing the capacity of existing modelling frameworks.

Our first major goal is the descriptive modelling of the initial learning trajectories of the population of IBL mice. To handle the difficulties present during learning that we outlined above, we draw upon multiple previous approaches: We use a non-parametric Bayesian model for its flexibility in encompassing degrees of complexity. This has an arbitrary number of states, each of which parameterises a (potentially partly perseverative) strategy. Sudden switches between states can capture immediate changes in performance, as animals grasp something new about the task or change their levels of attention. Slow changes in the parameters of behavioural models within the states allow for gradual improvement in the quality of choices. The rest of this section discusses these foundations in turn, so we can bring them together in chapter 3.

Non-parametric Bayesian models represent an important extension of the classical Bayesian framework. Classical parametric models, like a Gaussian mixture model, fix their complexity at the outset, enjoying a fixed number of parameters. The fitting procedure then adapts these parameters to the given data. Classical non-parametric models on the other hand, such as Gaussian kernel density estimation, do not specify the number of parameters ahead of time, but simply smooth around the existing data points (in a way, each data point becomes a parameter). A Bayesian non-parametric model offers an intermediate solution: The fitted model is capable of growing with the data, but has a preference for simpler solutions (less complexity, fewer parameters). This removes the need for trying out different levels of complexity within a parametric framework (varying the number of used Gaussians) and comparing their performance, as this becomes part of the inference process itself. The inference does become more involved though, as we now have to consider the complexity prior as well, which requires special procedures to cope with the varying dimensionality of the model.

Of particular relevance for us are non-parametric Bayesian extensions of the hidden Markov model (HMM, Baum et al. (1966)). The basic HMM aims to find a probabilistic characterisation of a sequence of observations, through the use of latent states. Each state is equipped with an observation distributions, which specifies how an observation might be generated on each time step on which the state is active, and there is a transition matrix which governs

the probabilistic transitions between states. Non-parametric versions of this model use a hierarchical Dirichlet process (Teh et al., 2006), to implement an infinite transition matrix, enabling complexity inference over the number of states (Beal et al., 2001). We focus particularly on a specific inference scheme which also imbues the latent states with duration distributions (turning the hidden Markov model into a hidden semi-Markov model), as presented in Johnson et al. (2013) and Johnson (2014). This framework has proven its great flexibility and capacity by informatively describing the movement patterns of animals and how they lead into one another, making full use of the infinite HMM framework (Wiltschko et al., 2015).

The non-parametric Bayesian approach has also been applied as a cognitive model, as complexity inference is a natural component of many real world problems: How do we support a possibly unbounded number of objects or problem instances, while effectively sharing information across instances (Austerweil et al., 2015; Gershman et al., 2012b). Putting this problem into a Bayesian framework allows us to tackle this difficult question in a normative way. One of the most notable applications of this framework is the modelling of fear acquisition and extinction through context inference (Gershman et al. (2010a), along with a more general approach which appropriately deals with time by modelling context persistence by Lloyd et al. (2013b)). The experimental paradigm is initially simple conditioning: an animal is repeatedly exposed to a tone, followed by a shock (acquisition) and it will begin to exhibit a fear reaction in response to the tone alone. This is followed by an extinction phase – many trials during which the tone is delivered without a subsequent shock (extinction). If the experimenter then lets a number of days pass (spontaneous recovery) or delivers a shock just by itself, no previous tone (reinstatement), and then plays the tone again, the exhibited fear response has surprisingly returned, being considerably stronger than it was at the end of extinction.

Latent cause models explain this by assigning observations (e.g., tone plus shock or no tone plus shock) to inferred latent causes, via a combination of the observation likelihood and a Chinese-restaurant process prior (a non-parametric structure closely related to the Dirichlet process, Gershman et al. (2015a)). Since the observations between the acquisition and extinction phase differ considerably, it may occur that the animal assigns them to two separate states. In this case, the fear memory is not overwritten during extinction, and may resurface at a later time. This paradigm and model have been used to model individually different learning trajectories (Gershman et al., 2015b), though the Bayesian non-parametric component was used mechanistically in that work, whereas we use it descriptively.

One noteworthy addition to the nonparametric Bayesian cognitive model literature is the contextual inference (COIN) model (Heald et al., 2021). This learns motor behaviour by inferring the current context via a non-parametric Bayesian structure, and tracks changes within these contexts through inference with a Gaussian random walk prior (which will turn out to be closely related to our descriptive learning model).

For the second ingredient of our model, we turn to state-based modelling. Here, we take inspiration from Ashwood et al. (2022) (see also Calhoun et al. (2019)), who described decision-making performance after learning with an HMM. Each latent state of the model captures a single component of behaviour as a map from task-relevant variables to a distribution over choices, via logistic regression. In the case of perceptual decision-making, this generalises a psychometric function to include other factors (e.g. perseveration). The overall description of behaviour is in terms of a mixture of different policies

that can switch rapidly, allowing, for instance, for the effects of motivational fluctuations in engagement.

The final component idea of our model is Roy et al. (2021), which effectively considers just a single state, but allows the weights of the logistic regression to be dynamic, tracking changes in behaviour through appropriate changes in the weights.

Taken together, these components address a sizeable proportion of the behavioural complexities during learning and allow for a comprehensive description of behaviour during it, as we show in the later chapters of this thesis.

1.6 MODELLING ASPECT: MODEL LIMITATIONS

The final critical aspect of modelling does not have to do with issues of animal behaviour specifically, but rather is a pair of inherent issues for the standard modelling approach itself: We design models for interpretability, and can only evaluate them comparatively. The models we have considered so far are specifically focussed on making a process understandable, which they achieve through a small set of parameters, each of which fills a distinct, easily understandable, role in the equations governing the model. This makes the modelled process interpretable, but naturally restricts model expressiveness, as the starting point is a set of simple equations, expanded upon by researcher ingenuity and creativity. Even the model comparison process is not without issues, as a held-out data set needs to be maintained properly to prevent data leakage and correlations within our data (as naturally occur in time-series data) may allow some models to gain an untoward advantage over others. However, even assuming models can be properly evaluated, this still does not answer the question of whether better, untested, models exist (Nassar et al., 2016) (although evidence of fitting failures can be provided by generative testing of models (Palminteri et al., 2017)).

Increasing the expressiveness of models such that they capture as much variance of the data as possible is actually fairly straightforward and can be achieved via neural networks, which are universal function approximators (assuming an appropriately chosen training procedure and model architecture) (Hornik et al., 1989). They can even be powerful enough to adapt their predictions to the behaviour of individuals as they progress through the trials of single sessions. That is, they can implement a kind of hierarchical fit (or possibly even a model comparison procedure), as they are tuned to the overall statistics of the population, and can then fine-tune as they observe behaviour from a single individual, in case of successful generalisation even one which was not in the training set.

However, the way they achieve this flexibility typically clashes directly with the goal of model interpretability. As in the case of individual differences, neural networks can capture and intermingle all sources of variance, further complicating the interpretation and possibly capturing processes we are not directly interested in. For the most part, models can be put on a spectrum from interpretable but limited in their predictive power, as in Q-learning, to predictively powerful but opaque, as in neural networks. However, the opacity comes with some advantages, most notably the confidence that all relevant aspects of the modelled process are captured, and that our explanations are not limited by the creativity and exhaustive experimentation of the model designer.

It is therefore a tantalizing goal to try to combine both of these aspects, which we have so far described as being mutually exclusive, into a single

treatment which offers both qualities. One noteworthy approach to achieving this involves methods for making sense of neural networks after they were trained, such as through distillation into smaller networks (Schaeffer et al., 2020), or by studying simulated behaviour from the fitted networks, using them as an easily manipulable simulacrum of the actual behaviour (Dezfouli et al., 2019b). A second approach is to design networks with interpretability in mind from the outset. This can take the form of training networks with very few hidden units (Ji-An et al., 2023), adding an information disentanglement objective to the loss function (Miller et al., 2024; Dezfouli et al., 2019a), or designing networks with a restricted information flow (Eckstein et al., 2024). The work of Dezfouli et al. (2019a) is particularly relevant for us, as they also focussed on individual differences. They achieved this by explicitly encoding subjects' behaviour into a low-dimensional and interpretable space, where they could find individuals differently sensitive to reward and varying in their inclination to persevere or switch.

All the mentioned approaches reduce the complexity of the fitted network by using a few computational elements or granting access to only limited parts of the input information. Through this, one limits the expressiveness of the models again, but we can always compare performance against a completely unrestricted network to estimate how close the model is to the noise ceiling of the given data set (assuming an appropriate architecture and training procedure).

1.7 OUR MODEL OF EXPERT BEHAVIOUR: HYBRID NETWORKS

We adapt the approach of Eckstein et al. (2024), as it transparently decomposes a decision making process into individual functions (which are realised through neural networks), into hybrid neural networks. These separate functions remain interpretable by virtue of only representing a relatively manageable step in the entire process. As the IBL decision-making process naturally decomposes into separate streams (involving the current stimulus and past stimuli/choices/rewards), the hybrid network approach lends itself naturally to our task.

A particular focus of our model will be on the details of how the history of trials affects the current choice for expert mice. As mentioned, the latent block structure of the IBL task introduces autocorrelations into the contrast sides, making action perseveration a useful tendency (Findling et al., 2023). However, adhering to one's own previous actions is not the ideal way to implement this strategy and animals would do better if they combined the deterministic feedback information they receive to infer the actual stimulus location. Additionally, they should apply Bayesian change point detection to these inferred stimuli to make inferences about the current block (Behrens et al., 2007). Of course, this would require complicated inference machinery and a precise knowledge of the generating process, which might not be worth the relatively modest increase in reward that they would afford. In fact, there is still a vast space of strategies between the exponential filtering of previous actions and full Bayesian inference of the current contrast prior, and we use the hybrid network approach to explore it.

1.8 OUTLINE

In the next, brief, Chapter 2, we present the task protocol employed by the IBL, the different phases of which underlie the subsequent chapters. We then develop in detail our flexible and extensible modelling framework to capture

learning behaviour as comprehensively as possible in Chapter 3. This model is applied to a large population of 134 IBL mice in Chapter 4, in which we showcase the capabilities of the model and use it to gain insights into the trends across the population, while also highlighting the substantial inter-individual differences. These two chapters are based on the work published in Bruijns et al. (2023). We then move on to the study of expert behaviour in Chapter 5, where we apply neural networks to arrange for tight control over the effects of restrictions we impose upon our model and to find an interpretable, but predictively powerful, model of behaviour. We conclude with a discussion in Chapter 6.

Part II

BEHAVIOURAL TASK

2 BEHAVIOURAL TASK

We applied our modelling frameworks to different phases of the International Brain Laboratory (IBL) task protocol. Due to its international collaborative nature, the IBL provides a large dataset of > 100 mice, performing the same perceptual decision-making task, trained under (mostly) the same rigorous protocol (The International Brain Laboratory et al., 2021). In this task, head-fixed mice were shown a sinusoidal grating of a controlled contrast on either the right or left side of a screen. This was accompanied by a tone which signified the beginning of a trial. They then had to center the grating (within 60 s) by turning a steering wheel in the appropriate direction see **Fig. 2.1(a)**. Note, that an animal has to move the wheel leftwards (or rather, counterclockwise), when the stimulus is on the rightwards side. This serves to establish an intuitive connection between the wheel movement and the stimulus, since the latter moves when the wheel does. To keep things simple, we will call the correct choice on a trial with a rightwards stimulus a rightwards choice, ignoring the direction of wheel movement. Successful trials led to water reward; unsuccessful trials to a noise burst and a 1 s timeout. Trials were self-paced, with mice signalling their readiness by keeping the wheel still for a period.

The IBL divides performance of the task into a number of stages, so as to encourage successful learning of the task and similar endpoints of behaviour. Our first modelling target was the initial learning period, from the first day on which the mouse interacts with the task until the animal reaches full proficiency with the basic task contingencies. Our second model focussed on proficient behaviour after learning. Mice learned the task according to a rigorous shaping protocol, which introduced more difficult stimuli gradually and actively removed action biases. Accordingly, shaping started with gratings of the highest contrasts: 100% and 50%. At this initial stage, there was no perceptual difficulty, but the animals had to learn the basic contingencies and requirements of the task. Once they had reached sufficient performance on these contrasts ($\geq$ 80% correct for each contrast type on the last 50 trials), 25% strength contrasts were introduced. After performance was good on this extended set as well (same criterion for each contrast), the remaining contrasts were introduced in a staggered manner: 12.5%, 6.125%, and 0%, while the 50% contrast was dropped from the task. For the 0% contrast, one side was rewarded randomly (with probabilities 50% each). A debiasing protocol increased the probability of repeating the stimulus that was just shown when the mouse made a mistake on an easy (100% or 50%) contrast. This served to motivate the animals away from perseverative or biased strategies, and could lead to reward rates lower than 50% (as would otherwise be expected from pure chance).

An animal moved on to the next stage of training (which we did not consider with our learning model), after having been introduced to the full set of contrasts, and fulfilling the following criteria: Accuracy on easy trials (100% or 50% contrast) on the last three sessions each needed to be better than 90%, each sessions needed to include at least 400 trials (there are automatic stopping criteria for a session, checking whether accuracy or reaction time drops too much), and a computed psychometric function needed to have a certain level of steepness, low bias and low lapse rate across the last three

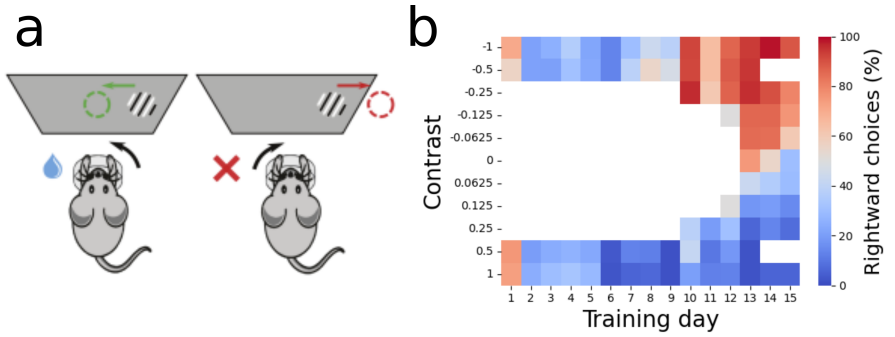

Figure 2.1: Basics of the IBL task. **(a)** left: the mouse performs a rightwards action (moving the wheel counterclockwise), thereby centering the stimulus on the rightwards side of the screen, earning a water reward. Right: A clockwise motion moves the stimulus outside of the screen, resulting in a noise burst and timeout before the onset of the next trial. **(b)** Representative behaviour of mouse KS014 (also used later). This shows the improvements in behaviour and concomitant extensions of the contrast set across the days on which the animal learns the contingencies of the task.

sessions. In this new stage, the biased training, the basic task stayed the same, but instead of contrasts appearing equiprobably left or right, there were now unsignalled alternations between blocks, lasting 20-100 trials following a truncated geometric distribution, during which a contrast was 80% likely to appear on one side versus 20% on the other, see **Fig. 2.2**. The geometric distribution was chosen to produce a flat hazard rate across the block, not facilitating the anticipation of a block switch This is obviously invalidated by the truncation, but it seems implausible that the mice are actually able (not to mention willing) to learn these restrictions with sufficient certainty. Blocks only started after the first 90 trials, these first trials still had an even distribution of left and right stimuli. The block structure was of course particularly helpful for 0% contrasts, on which an animal could now reach a much higher reward rate than chance, given a suitable block inference mechanism (a detailed analysis of their actual algorithm was performed in Findling et al., 2023). Mice completed this part of training when their behaviour was appropriately modulated during the two types of blocks, as indicated in **Fig. 2.1(c)**.

After animals passed this stage of training as well, their behaviour was deemed sufficiently stationary and they moved on to a new experimental rig, in which it was possible to perform neural recordings during task performance. The IBL used Neuropixels multi-electrode probes (Jun et al., 2017) to perform recordings covering large parts of the brain, which so far included 279 brain areas in the left fore- and midbrain, and the right hindbrain (The International Brain Laboratory et al., 2023). As this represents the peak of stationarity and performance in mouse behaviour on this task, we used these recorded sessions as the target for our second model, which aimed at a comprehensive description of proficient behaviour in this perceptual task combined with a hidden state inference component. Mice performed varying numbers of sessions while being recorded, ranging from just one, up to sixteen sessions, with the mode being at five.

Just like the criteria for when a mouse moves to a new stage of the task, there is also a set of rigorous criteria for when a session ends (a mouse has at most one session per day). A session can end for one of three reasons: (i) the mouse has been training for more than 90 minutes, (ii) the mouse fails to do more than 400 trials in the first 45 minutes, or (iii) the mouse completes

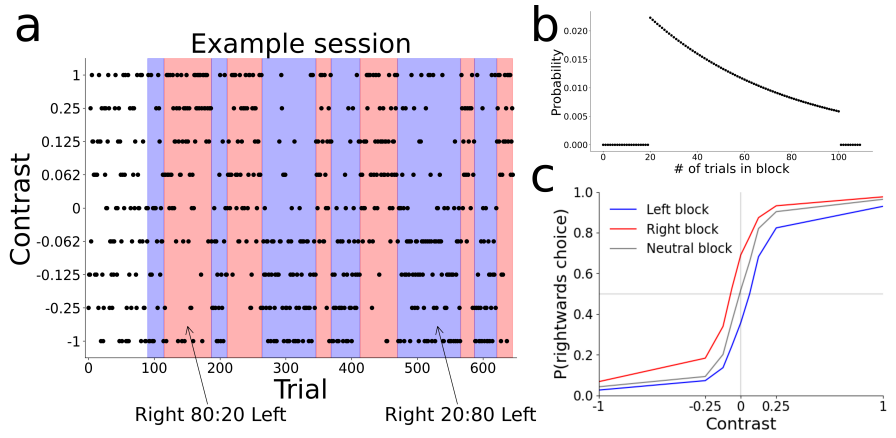

Figure 2.2: Basics of the hidden blocks in the IBL task. **(a)** An example session, showing which contrast were shown on which trial, and which block is currently active, through the background colour (white: neutral first 90 trials, blue: left biased, red: right biased). Inferring the current block correctly is especially helpful for 0% contrast trials, since they get rewarded probabilistically according to the current block (i.e. in a rightwards block, 80% of rightwards choices on a 0% contrast will be rewarded. **(b)** The distribution over block lengths. A block lasts at least 20 trials and at most 100, the distribution is a truncated geometric distribution. **(c)** The psychometric functions of expert animals, conditioned on the blocks (we call the first 90 trials the neutral block). Clearly, their behaviour is modulated by the block identity. Here, we plotted the different contrast strengths on a linear scale. Throughout this work, however, we will plot them equidistantly, for visual clarity.

over 400 trials in the session and the median response time (defined as the time from stimulus onset to the registration of the response) over the past 20 trials is over five times longer than the median response time over the entire session. This has two important consequences: If a session has less than 400 trials, behaviour will usually have been rather poor, therefore we exclude such sessions in our second model which focusses on proficient behaviour. The other consequence is that behaviour will usually degrade towards the end of a session, since the end of a session tends to come about due to slow responses, meaning the animals tended to have disengaged, possibly due to satiation. We will see a reflection of this decreased performance towards the end in our second model and discuss disengagement in general with our first model.

For the purpose of reproducibility, the IBL employed the exact same sessions for all mice while they were recorded from (the training sessions were created randomly on the spot). With "exactly the same", we mean that the contrasts which were presented and the choices which were rewarded followed a fixed sequence, so we call these prototypical sessions "session types". There were twelve different session types, which were always presented to the mice in the same order (though some mice, presumably accidentally, were presented the same session types twice). As a consequences, the early session types were considerably more frequent in our data set than later ones, the count of session types in order is: 107, 98, 96, 90, 32, 30, 30, 24, 13, 8, 6, 5. It is the prevalence of the first four session types which introduced the visible trends into some of our performance plots. For example, since there were only so many block switches to observe, if the contrast on the third trial after a switch was, by chance, usually a strong one in the most frequent session types, the predictions

on this trial will be relatively better than surrounding trials, as seen later in **Fig. 5.9**.

Part III

MODELLING THE LEARNING TRAJECTORY

3

LEARNING TRAJECTORY MODEL SETUP

Our first model focusses on the initial learning period of the animal, from its first day interacting with the task, to the session at which it reaches sufficient proficiency to move on to bias training. We will begin with a description of the high-level intuition of our model, before developing it in detail in the following sections. For our description of the behaviour, we follow previous work (Ashwood et al., 2022; Roy et al., 2021) and formalise the response probabilities for the binary choices of mice through logistic regression (omitting the rare trials in which the animal timed out by not responding within 60 s). Trial t of session n is described by features $\boldsymbol{f}_{n,t}$ comprising: (i) the stimulus, i.e. the contrast on the left and right of the screen; separated, to allow for different sensitivities to leftwards and rightwards stimuli, as mice were frequently differently sensitive to the screen sides in this task (ii) task history, in the shape of an exponentially decaying average over the last actions; interestingly, mice only employed a perseverative bias, but did not use reward information to implement a win-stay lose-shift strategy (as was also observed in Miller et al. (2021), Beron et al. (2022), and, in the same task at a later stage, by Findling et al. (2023)) (iii) a bias term to allow for side preferences regardless of other features. Labelling the state that is active on this trial as $x_{n,t}$, the response $y_{n,t} \in \{\mathrm{L}, \mathrm{R}\}$ (for Left and Right) is modelled by the distribution

$$P(y_{n,t} = \mathrm{R}) = \mathrm{sig}\left(\boldsymbol{f}_{n,t} \cdot \boldsymbol{w}_{x_{n,t},n}\right), \tag{3.1}$$

where the weights of the states $\boldsymbol{w}_{x,n}, \forall x$ are also indexed by session n, as they can drift across sessions. Here $\mathrm{sig}(\cdot)$ is the standard logistic sigmoid function.

In order to characterise the course of learning across every trial, we developed a flexible model which segments the behaviour of an animal into discrete states that last for variable numbers of trials within a session and can recur in multiple sessions. As this is a descriptive model, we equate a behaviour with its corresponding state, and generally will not distinguish between the two in the text. We first describe how a single state generates choice probabilities on a trial for which it was responsible (**Fig. 3.1** within circles); and then how we treat multiple states (**Fig. 3.1** arrows).

The model generalises a standard hidden Markov model (HMM) in three ways that make it especially suited to describe the phases of learning, see also **Fig. 3.1**: (i) it is non-parametric about the number of states, i.e. the number of states describing the behaviour of each individual is separately determined, accommodating inter-individual differences. This characteristic also allows the model to capture sudden changes in behaviour, as it is able to introduce a new state when behaviour changes notably (we call this the 'fast process', see section 3.1). (ii) States are dynamic over sessions, allowing the behaviour implied by a state to change gradually across session boundaries (Roy et al., 2021) (the 'slow process', see section 3.2). (iii) Whereas for HMMs, the numbers of trials for which a single state remains active always follows a geometric distribution, we adopt a semi-Markovian approach, allowing for more general distributions. The prior over these duration distributions encourages temporally extended states, in order to extract persistent behavioural modes, rather than single trial deviations which are more likely noise. Taking all these additions together we end up with a dynamic infinite input-output hidden semi-Markov model (abbreviated as diHMM).

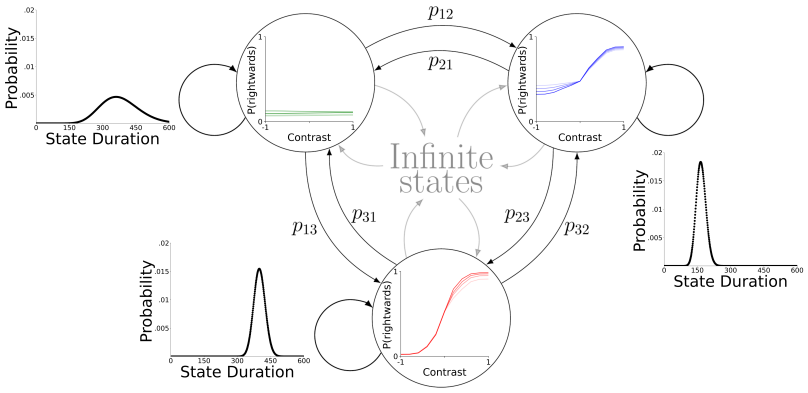

Figure 3.1: Visual representation of the main components of the model. Each state, represented by a circle, has an associated observation distribution, shown inside its circle. This is implemented via logistic regression, to consider the contrast of the current trial and a weighted history of previous choices (the latter is not shown here). The weights underlying these regressions can change from session to session, resulting in shifts of the psychometric functions (PMFs) they represent; we depict this evolution here with varying shades of colour. States are connected to other existing states via transition probabilities p. In addition to that, states also have the option to transition into a new state, to describe a type of behaviour which is not well captured by any of the existing states. Lastly, staying in the same state for more than one trial is not modelled via a self-transition probability, but instead each state has its own duration distribution.

The transition matrix over a flexible number of states and the evolution of the psychometric weights are defined by priors, and the Bernoulli observation model provides a likelihood for each trial, allowing for approximate Bayesian inference (details in Methods). We performed this via a Markov chain Monte-Carlo algorithm, namely Gibbs sampling. For a single animal, the entire response and feature data across all training sessions were fitted together. Individuals were fitted separately, meaning a large number of subjects is not necessary for the application of our model. Integrating across a number of Gibbs samples from multiple Markov-chains led to a set of behavioural states defined by their session-varying weights $\boldsymbol{w}_{x,n}$ and duration distributions, as well as a hard assignment of every trial onto one of these states. Due to the flexibility of the nature and number of states on any given sample, which usually expressed themselves in a multi-modal posterior, we had to develop techniques which were able to appropriately summarise a collection of samples. As a result of this necessary post-processing, it became more complex to form e.g. a posterior over which state is active on which trial, which we will elaborate on below. While all other relevant random variables are specified hierarchically or ruled by vague priors, the variance for slow changes within states is set, as inference over this variable proved problematic; we revisit this parameter in the discussion.

We structure these methods as follows: We begin with a detailed description of the infinite hidden semi-Markov model, i.e. the specification of the priors and how they relate to the other random variables in the hierarchy. In section 3.2 we describe inference for the logistic regression observation distributions, with a focus on the details of resampling. These two components together give us the full dynamic infinite input-output hidden semi-Markov model (diHMM). We cover how to integrate the collection of generated samples in section 3.3, dealing with the hurdles of label switching and multi-modality, ultimately

obtaining well-defined states from the samples. This concludes the details of the model itself. Section A.1 elaborates how we assign states and their PMFs to the three types. Extending slightly beyond the scope of this work, section A.2 relates properties of the initial learning considered here with the next training step, the biased block training. We conclude with important validation analyses: we elaborate on our cross-validation procedure for parameter and prior selection and present ablations of our full model in section 3.4, discuss posterior predictive checks in section 3.5, and finally by showing successful recoveries of various generative models in section 3.6.

3.1 INFINITE HIDDEN SEMI-MARKOV MODEL

We start by describing the diHMM, focussing on Bayesian inference over its random variables. Following Johnson et al. (2013), we use Gibbs sampling, a Markov chain Monte-Carlo algorithm (MCMC), to realise an iterative resampling scheme over the model components, including the PMFs of the hidden states and the assignments of the individual trials onto those states. For this purpose, all distributions are paired up with conjugate priors in this section, to enable simple resampling steps. The posterior distribution is ultimately represented by a collection of samples, with every component being assigned an explicit value in each sample.

Overview over quantities and counters		
Quantity	Meaning	Counter
N	number of sessions	n
T_n	number of trials within a session n	t, m
J	number of MCMC samples	j
L	number of states in the model	i
S	number of states employed within a session	s

We first describe all the relevant random variables (using the iterator notation from the table above).

The technical backbone of an infinite HMM is a hierarchical Dirichlet process. At the top of the hierarchy of this process is the prototypical transition vector

$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma), \tag{3.2}$$

where GEM (named after Griffiths, Engen, and McCloskey) is a Dirichlet process without a base distribution, a pure stick-breaking process which samples a probability vector over infinitely many elements (which will be states in our case). The concentration parameter γ probabilistically determines the size of the individual sticks, and therefore how many states are practically relevant, with a higher γ encouraging more states. We put a vague Gamma prior on γ, making it, and thereby the propensity to introduce new states, part of the inference as well, with $\gamma \sim \mathrm{Gamma}(0.01, 0.01)$.

At the next level we sample the transition vectors, a classical HMM component, π_i of the individual states i. These are tied together via $\boldsymbol{\beta}$, which is used as the base distribution for a second Dirichlet process

$$\boldsymbol{\pi_i} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \qquad i = 1, 2, ..., L, \tag{3.3}$$

$$\boldsymbol{\pi_0} \sim \mathrm{GEM}(3) \tag{3.4}$$

α is another concentration parameter and determines how closely the π_i are related to $\boldsymbol{\beta}$. Sampling the individual state transition vectors from this

common source formalises an overall kind of state popularity. The higher α, the more like $\boldsymbol{\beta}$ is $\boldsymbol{\pi_i}, \forall i$, and so the more the bias in the frequency of state i' in the particular sample $\boldsymbol{\beta}$ will be reflected in the transitions from i to i', and so the more popular i' will be overall. We put another vague Gamma prior on it, $\alpha \sim \text{Gamma}(0.01, 0.01)$. The initial state distribution $\boldsymbol{\pi_0}$ is drawn entirely separately, with a concentration parameter of 3 as a trade-off between allowing new states but not encouraging the invention of new states at the start of sessions.

For our inference scheme, we make use of the weak-limit approximation which puts an upper limit $L = 15$ on the number of states, rather than employing the full infinite process. This simplifies the resampling scheme, while still behaving similarly to an infinite HMM if L is sufficiently large. Across the entire population, there were only three mice with 12 states, after applying our hierarchical state clustering procedure (section 6.3); all other mice used fewer states. Furthermore, the minimum fraction of trials captured in states (as described further below) is 99.38% (mean is 99.97%), justifying the choice of $L = 15$ (though a higher limit would possibly allows us to capture motivational fluctuations better). In particular, we still perform inference over the realized state complexity. In the weak-limit framework, Eq. 3.2, 3.3, and 3.4 turn into L-dimensional Dirichlet distributions

$$\boldsymbol{\beta} \sim \text{Dir}(\gamma/L, ...\gamma/L), \tag{3.5}$$

$$\boldsymbol{\pi_i} \sim \text{Dir}(\alpha\boldsymbol{\beta}_1, ..., \alpha\boldsymbol{\beta}_L), \qquad i = 1, 2, ..., L, \tag{3.6}$$

$$\boldsymbol{\pi_0} \sim \text{Dir}(3/L, ..., 3/L) \tag{3.7}$$

The transition structure within a session is given by

$$z_{n,1} \sim \boldsymbol{\pi}_0, \tag{3.8}$$

$$z_{n,s} \sim \boldsymbol{\pi}_{z_{n,s-1}}, \tag{3.9}$$

where $z_{n,s} \in \{1 \dots L\}$ is an indicator for the sth state within a session n (which does not align with the trial number), and π_0 is the initial state distribution.

Given the transition vectors, the workings of the hidden semi-Markov model are fairly standard, except that the duration distributions are specified explicitly rather than being drawn from a geometric distribution (as in a regular HMM). We therefore prohibit self-transitions, which makes a data augmentation scheme for resampling necessary, as described in Johnson et al. (2013). Nevertheless, as in a standard HMM, durations are statistically independent of the target state of transitions. Durations are drawn from a negative-binomial distribution, with state-specific random variables, coming from their own priors

$$r_i \sim U(5, 6, 7, ..., 704), \qquad i = 1, 2, ..., L, \tag{3.10}$$

$$p_i \sim \text{Beta}(1, 1), \tag{3.11}$$

$$d_{n,s} \sim \text{NB}(r_{z_{n,s}}, p_{z_{n,s}}). \tag{3.12}$$

(Note the difference between state names i, which hold for the entire model, and the session specific state counters s, which can be used to find the current state name via the indicator $z_{n,s}$). We chose a uniform prior over a large range of numbers for the possible values of r, to enable long durations, but excluded small values for r (in particular $r = 1$ would give the geometric distribution). Small values of r encourage transitions after a very small number of trials, which would capture the statistics of the presentation of left and right stimuli by the experimenter rather than the longer-lasting states that we sought. Using cross-validation we ensured that enabling larger values of r did not benefit the fits.

States stay active and generate observations for as long as the drawn duration indicates

$$t_n(s) = \sum_{k=1}^{k<s} d_{n,k}, \tag{3.13}$$

$$x_{n,t_n(s)+1:t_n(s)+d_s} = z_{n,s} \tag{3.14}$$

$$P(y_{n,t} = \mathrm{R}) = \mathrm{sig}\left(\boldsymbol{f}_{n,t} \cdot \boldsymbol{w}_{x_{n,t},n}\right), \tag{3.15}$$

where we defined $t_n(s)$ to return the trial on which the sth state of a session n starts, which allows for the definition of $x_{n,t}$, the state on any given trial t. We denote the logistic sigmoid function as sig. This takes the dot product between the state weights $\boldsymbol{w}_{s,n}$ (which we discuss in the next section) and the input features of the current trial $\boldsymbol{f}_{n,t}$ and produces the probability over the observation $y_{n,t}$. The binary response variable y has 0 representing a leftwards, and 1 a rightwards, choice. See also **Fig. 3.2** for a visual summary of these variables.

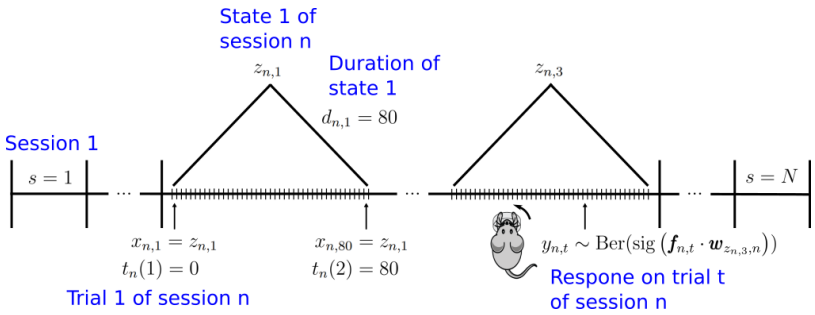

Figure 3.2: Visualisation of the different variables across training, with some explanatory text in blue.

We summarise this collection of variables as

$$\Theta = \{\gamma, \alpha, \beta, \boldsymbol{\pi}_0, \{\boldsymbol{\pi}_i, r_i, p_i, \{\boldsymbol{w}_{i,n}\}_{n=1}^{N}\}_{i=1}^{L}, \{\{x_{n,t}\}_{n=1}^{N_t}\}_{t=1}^{T}\},$$

where N is the total number of sessions. The connections between these variables are visualized in the form of a graphical model in **Fig. 3.3**. The result of inference is a set of samples $\{\Theta_j\}_{j=1}^{J}$. Each sample is a full instantiation of the listed random variables, which we can treat as a representation of the posterior. Gibbs sampling works by iterating through all variables, and re-sampling them from their distribution, given all other variables of the model. After updating all variables, the result is one new sample within the MCMC-chain. The details of how to resample the individual components can be found in (Johnson et al., 2013).

3.2 DYNAMIC LOGISTIC REGRESSION PRIOR AND SAMPLING

Gibbs sampling resamples each random variable conditioned on all others. Thus, inference over the observation distributions of the states is separate from almost all the rest of the model, only using the information as to which trial is currently assigned to which state. We drop the explicit state dependence i in $w_{i,t}$ for this section, but it is important to keep in mind that this sampling scheme is applied to every state individually, with each state s being influenced only by trials for which $x_{n,t} = s$ in the current sample. We implement slow

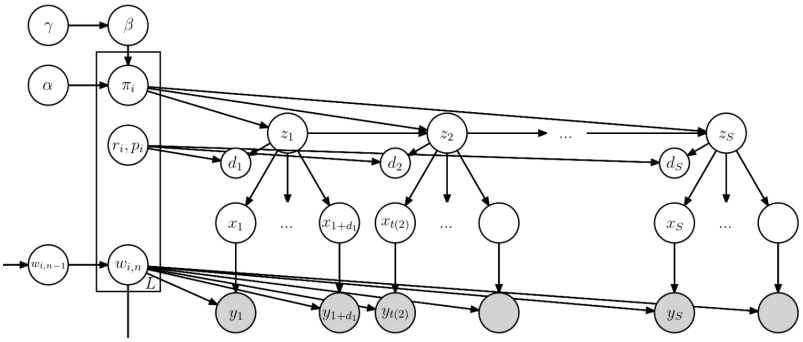

Figure 3.3: Visualisation of the random variables as a graphical model, showing the variables $z, d, x,$ and y of a specific session n. Variables in shaded circles are observed, unshaded variables are inferred. We suppressed the dependence on session n for all relevant variables other than the state weights w, for which we highlighted that they can indeed change across sessions.

changes in the characteristics of the states by putting a Gaussian random walk prior on the weights $\boldsymbol{w}_n$, allowing for modest change across session boundaries, parameterised by the variance σ. We choose a diffuse initial distribution for the weights, and use cross-validation to select the inter-session variance $\sigma = 0.04$ (we performed cross-validation on a range of small values, in order to limit the state adaptation process to small changes):

$$\boldsymbol{w}_1 \sim \mathcal{N}(0, 8 \cdot I), \tag{3.16}$$

$$\boldsymbol{w}_{n+1} \sim \mathcal{N}(\boldsymbol{w}_n, \sigma \cdot I), \tag{3.17}$$

where I denotes the identity matrix. If a state has no trial assigned to it in a particular session, its weights are held fixed during the next transition, preventing states from morphing radically during a prolonged absence.

Inference for the logistic regression weights is performed using Pólya-Gamma data augmentation, which allows for efficient inference in settings with binomial likelihoods (Polson et al., 2013; Linderman et al., 2015), since it is not possible to choose a conjugate prior. We review the relevant computations here, for a full treatment we refer to (Windle et al., 2013). In the first step of the resampling scheme we sample pseudo-observations. This uses a Pólya-Gamma distribution PG, by first sampling $\omega_n \sim \mathrm{PG}(b_n, \boldsymbol{\psi}_n)$, where $\boldsymbol{\psi}_n = \boldsymbol{f}_n \cdot \boldsymbol{w}_n$ is the dot product of features and weights, and b_n is the total number of times this exact instantiation of features was observed in session n. However, the same state is associated with more than just one specific instantiation of features (i.e., including contrasts of different strengths and side, and different response histories). To handle this, we treat a single session as multiple different time points, but prevent weight changes between time points that belong to the same session. In this way, the observations from different features within the same session are effectively aggregated. To complete the pseudo-observation generation, we need $\kappa_n = a_n - b_n/2$, where a_n is the number of rightwards answers observed for the current $\boldsymbol{\psi}_n$ under consideration. Now $z_n = \kappa_n/\omega_n$ can be treated as if they were drawn from $\mathcal{N}(\boldsymbol{\psi}_n, 1/\omega_n)$.

This data-augmentation serves the purpose of having the $\boldsymbol{w}_n$ emit observations with Gaussian noise (after combination with the features $\boldsymbol{x}_n$ into $\boldsymbol{\psi}_n$). Since the prior on $\boldsymbol{w}$ is a Gaussian random walk, this places inference in the well studied realm of Kalman filtering. To resample the $\boldsymbol{w}_n$ we use the forward filter backward sample algorithm (FFBS, Carter et al. (1994) and Frühwirth-

Schnatter (1994)), which filters forwards through all the observations using a Kalman filter, then samples the sequence of $\boldsymbol{w}_n$ backwards through time. A single resampling step therefore consists of first drawing the Pólya-Gamma variables to create pseudo-observations, then using them to sample the $\boldsymbol{w}_n$ using the FFBS algorithm.

We consider four features for the logistic regression: the contrast on the left side, the contrast on the right side, an exponentially weighted history over all previous choices, and a bias. Separating the features for left and right contrast allows the sensitivities to the two sides to be different. Since the notional contrast values do not match the psychophysical difficulty of the contrasts (100% and 50% are both virtually equally easy to perceive, not a factor of 2 apart), we apply a transformation to have a better alignment. For this, we follow Roy et al. (2021) and use a tanh-transformation, mapping the actual contrast c onto the input $\tilde{c}$ for our logistic regression through $\tilde{c} = \tanh(pc)/\tanh(p)$, where we follow their recommendation and set $p = 5$, which scales the steepness of the transformation. This maps the contrasts (1, 0.5, 0.25, 0.125, 0.0625, 0) onto (1, 0.987, 0.848, 0.555, 0.302, 0).

The regressor for previous answers, enabling perseveration as a strategy, proved to be beneficial for cross-validated performance. It is associated with the famous law of exercise (Thorndike, 1911; Gershman, 2020b), and has also been found to be exhibited by the mice in the asypmtotic regime that arises after the sessions that we are presently analysing (Findling et al., 2023). The same analyses showed no general statistical support for a regressor sensitive to the interaction between past choice and past reward, as would be reflected, for instance, in win-stay, lose-shift behaviour. We implement the perseveration regressor as an exponentially weighted sum over all past trials. We found that weighting previous trials with an exponentially decaying filter with smoothing factor 0.25 worked best (though slightly different parameter settings have almost equal cross validation performance). Thus, we compute this feature on session n and trial m as such:

$$\frac{1}{Z}\sum_{k=1}^{m-1}\exp\left(-0.25\cdot k\right)\cdot\left(2*y_{n,m-k}-1\right), \tag{3.18}$$

where $Z = \sum_{k=1}^{m-1}\exp\left(-0.25\cdot k\right)$ is a normalisation constant, such that the entire exponential filter adds to 1. The transformation $2*y-1$ serves to encode responses as -1 and 1, for the purpose of having the perseverative feature sway the current response appropriately. Therefore, this feature reaches its maximal value of 1 if all previous responses were rightwards and -1 if they were all leftward, putting it on the same scale as the other features. Time-out trials, where the animal did not respond before 60 s had passed, while skipped for the logistic regression of responses, are taken into account for the previous answer regressor, encoded as 0.

We experimented with a parameter that determined for how many sessions a state is allowed to change its weights through the slow process after it was last observed. The idea was that a behaviour may be inactive during any given session, but is still malleable, or even being improved through e.g. replay mechanisms. However, it turned out that allowing states to change in their absence allowed the model too much freedom to simply re-use old states with considerably different PMFs in the later parts of the learning trajectory, leading to undesirable behaviour in the usage of states, even for low settings of the parameter. We therefore decided to only allow the weights of a state to change after a session during which it actually occurred.

3.3 AGGREGATION AND INTERPRETATION OF CHAINS

We generally generated 48,000 samples from each of 16 chains (with different starting points), discarding the first 4000 as burn-in. We assessed convergence of the chains using the classical measure $\hat{R}$ (Gelman et al., 1992), and generated more samples by continuing each chain if necessary (though not all animals ever reached a sufficiently low $\hat{R}$ score). $\hat{R}$ compares intra- and inter-chain variability of bespoke, state-independent features of the chains. To detect differences in the variances of the chains, and other problems which $\hat{R}$ is known to miss, we also used folded-$\hat{R}$ and rank-normalised-$\hat{R}$ (Vehtari et al., 2021). We reduced the memory cost by thinning the chain, using only every 25th sample (we did this purely for memory reasons, not because it is necessary for MCMC algorithms (Link et al., 2012)). For a first pass, we sought to discard chains which differed substantially from other chains in the explored region in parameter space, either because they never reached the relevant parts of it, or because they spent disproportionate amounts of time in some modes over others. This is a known problem for MCMC algorithms in multi-modal environments, and can be mitigated by taking non-mixed chains and combining them via stacking (Yao et al., 2022). However, since our goal here is not prediction, we still want to focus on finding and visualising the most important modes of the posterior, which we did by combining the (possibly not perfectly mixed) chains, and considering the regions of probability space in which they collectively spent the most time. Given the slow transitions between different modes, we also did not split our individual MCMC chains when computing $\hat{R}$, as the two halves of the chains were often too different.

As scalars underlying $\hat{R}$, we used the concentration parameters: α and γ, as they are independent of states, and, as general properties of the fit: the number of trials assigned to the state with the most trials, and the second-most trials, as well as the overall numbers of states with more than 20%, and more than 10% of trials assigned to them (we chose multiple cutoffs to gain information about the fit at different levels of resolution). By greedily discarding the chains which increase $\hat{R}$ the most, we reduced the number of chains under consideration from 16 to at least 8. For this we considered all features and all variants of $\hat{R}$ (normal, folded, rank-normalised) at once, so we were minimising the maximum over all these $\hat{R}$s. We only further processed the chains when $\hat{R} < 1.05$, which is more conservative than some recommendations, but, in light of the strong multi-modality, more lenient than the newest ones (Vehtari et al., 2021).

However, it is still not trivial to extract information from the remaining chains given the multi-modality. There are two main sources of multi-modality: (i) genuine uncertainty in the usage of states or the exact setup of the random variables of the states, and (ii) mode equivalence with permuted labels (e.g., state $i = 1$ in the first chain might explain roughly the same set of trials as state $i = 2$ in the second). Although the second source makes evaluating the results more complicated, it is in fact just the sampling scheme working correctly, as there is nothing special about the particular state labels – solutions with permuted state labels are functionally equivalent. For the same reason, even within a single chain, a relatively consistent set of trials might be explained by one label for some part of the chain, but by a different label in another. Indeed, we frequently observed this kind of label switching, where one state completely took over the trials of another within a few sampling steps. In the limit of infinitely many samples, we can expect any trial to have a uniform distribution over the state label assigned to it; the only important question is

which other trials were usually accounted for by the same state as the given trial within suitably similar samples.

To formalise the necessary abstraction from direct state assignments, we computed co-occupancy matrices C^j for each sample j. C^j is a matrix of size $T \times T$, with T being the total number of trials across all sessions of a mouse, whose t, m^{th} entry reports whether trials t and m (for convenience, dropping the additional session label) used the same state in sample j

$$C^j_{t,m} = \mathbb{1}(x_t = x_m). \tag{3.19}$$

We used these co-occupancy matrices as a basis for two different processing steps: (i) at a coarser resolution across trials, we applied dimensionality reduction to find posterior modes; (ii) at full resolution, we averaged C^j across similar samples j to derive a matrix that describes the mutual affiliation of trials, allowing us to overcome the labelling issues. Both steps are reminiscent of representational similarity analysis (Kriegeskorte et al., 2008), in that, instead of comparing two samples directly, we compare state co-occurrence within the samples.

In principle, to explore the posterior, we could have flattened each C^j into an T^2 vector and applied principal components analysis (PCA). However, there were too many trials per mouse (of the order of 15000) to do this at full resolution, so we binned the trials into 200 bins, ignoring session boundaries, and then used the Wasserstein distance to measure state co-occurence between the bins. That is, we define modified matrices C'^j as

$$C'^j_{t,m} = \sum_{i=1}^{L} 1 - |p^j_{t,i} - p^j_{m,i}| \tag{3.20}$$

where $p^j_{t,i}$ is the proportion of trials in bin t which is assigned to state i in sample j. C'^j reduces to C^j for bins comprising a single trial. We then plotted individual samples in the first three dimensions of the PCA-space arising from flattened versions of C'^j, as shown in **Fig. 3.4**.

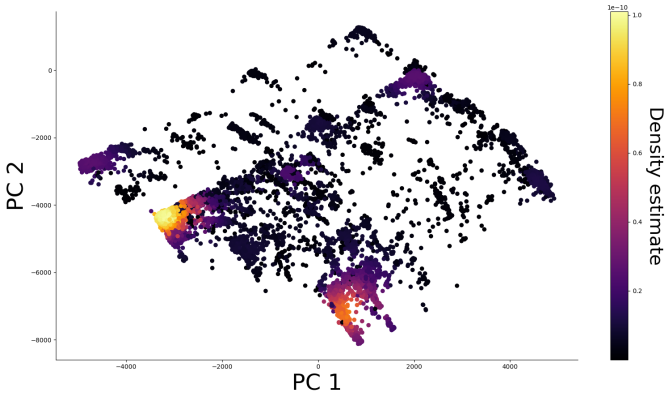

Figure 3.4: Individual MCMC-samples can be scattered in 2D principal component (PC) space, to find regions of high probability. To make those regions more salient, we colour individual samples according to a Gaussian density estimation (the density estimation occurs in 3D PC space). There are multiple modes of varying importance, with one mode being particularly dominant.

In doing this, we found that the posterior for a number of animals wanders itinerantly between different modes, reflecting true uncertainty. These modes

are distinct solutions and should not be blended. To isolate them, we performed Gaussian density estimation in the 3D PCA space to identify the ones that were most prevalent, as the regions of highest estimated density. We used this clustering to select samples $j \in \mathcal{J}^\eta$ that were sufficiently similar as to comprise an individual mode $\mathcal{J}^\eta$. For now, we did this by hand; however, the process could be made more formal by fitting a mixture of Gaussians to the posterior, and then selecting samples around the means of the Gaussians with sufficiently large mixture weights. We selected at least 400 samples from a mode to form a representative collection.

Next, we sought to understand how trials within that mode were co-assigned to states. To do this, we averaged the co-occurrences $C^\eta = \frac{1}{|\mathcal{J}^\eta|}\sum_{j\in\mathcal{J}^\eta} C^j$, and treated $\tilde{C}^\eta = 1 - C^\eta$ as a distance matrix, where trials were close if they share a state in most samples in the mode. We then performed hierarchical clustering on $\tilde{C}^\eta$, using as a cluster distance $d(v,\nu) = \max\left(\tilde{C}^\eta_{v[k],\nu[l]}\right), k \in v; l \in \nu$, which took as distance between clusters the maximum distance between any two trials in the clusters v and ν. The result of the hierarchical clustering was a tree on the individual trials; cutting this tree at a certain level leads to a specific clustering. Thus, cutting at, say, 0.6 means that we only have clusters in which every trial was explained by the same state in at least 40% $(1 - 0.6)$ of the samples. For our plots we cut at 0.95, which empirically returned good results. Even though this meant that trials needed to use the same state in only 5% of samples to be in one cluster, most trials were assigned to the same state much more frequently; see **Fig. 3.5**. This also shows a number of alternative clusterings from different thresholds, demonstrating that there is little change across a wide range of thresholds: The 95% threshold leads to 8 states with 100% trial coverage, an 80% cutoff leads to 9 states and 99.92% coverage, a 50% cutoff gives 12 states with 98.77% coverage, and lastly a threshold at 20% gives 15 states and 95.27% coverage. We can thus see that low criteria led to trials becoming unassigned and some states splitting apart, which is why we chose a rather high cutoff. A further verfication that the procedure and its threshold gave a faithful representation of the collection of samples comes from comparing the overall solution against individual solutions from single samples. Empirically, these did indeed align. Our later recovery analyses also used this approach.

The states we show are therefore defined at heart by sets of trials. To compute the PMFs of such a set, we first considered a single MCMC-sample, and noted which states it assigns to the trials within this set on a session-by-session basis (though each individual trial only had one state assigned in a single sample, for the whole set of trials it usually will not just have been a single state, due to random fluctuations, but mostly a single state). We turned the psychometric weights of these states into PMFs, over which we then averaged (in a weighted manner, considering how often any state occurred in the set of trials). For a single sample, this resulted in an average PMF of that state for each session. This then got averaged across samples within a cluster (evenly over all selected samples of a mode), to obtain the ultimate PMFs of this state.

To determine how closely a single trial is connected to its assigned state, we averaged the proportions of samples in which it was in the same state as all the other trials assigned to this state. That is, for a given trial t, we took a row of the consistency matrix C^η_t, and considered only the entries corresponding to other trials within the state under consideration. We then averaged over those entries, yielding the average proportion of co-assignment. We think of this as a proxy of the posterior over which state a trial is assigned to, and show it in **Fig. 4.2**.

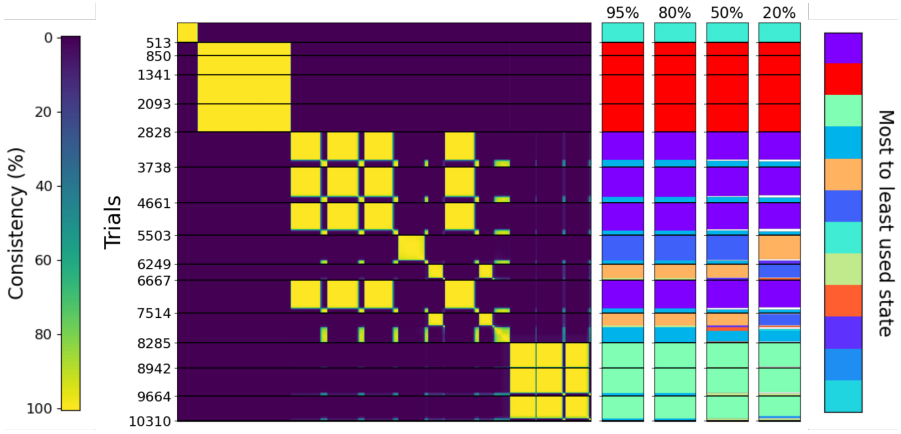

Figure 3.5: Consistency matrix C^{η} of the animal seen in **Fig. 4.1**, with different state assignments, based on cutting the hierarchical clustering tree at different levels, as noted above the colouring assignment. The ticks on the left mark session boundaries, with the numbers indicating the total number of trials so far. State colour on the bars to the right was determined based on the ranking of how many trials are assigned to the state, therefore a colour change need not be a major change in state assignments. As can be seen, state assignments are robust in a large range of cutoff values, with large states staying particularly consistent. Most of the change comes from the splitting of smaller states, and some trials losing a state assignment altogether (a state needed at least 40 trials to be coloured; if the state of a trial had fewer than that, we colour it white).

3.4 CROSS-VALIDATION AND ABLATIONS

Our model contains a number of free parameters which we set using a cross-validation procedure. We used this most notably for the variance σ of the normal distribution over how much the logistic weights of a state can change from session to session and the decay constant of the exponential filter over previous actions, which are fixed parameters that are not inferred during the inference procedure. This inference procedure is itself guided by priors, which we set to be vague, exerting minimal influence upon the ultimate posterior. However, their precise setting can nevertheless also be evaluated via cross-validation. This applies to the two gamma distributions over the concentration parameters α and γ and the priors over the parameters of the states' duration distributions. Cross-validation also allowed us to verify that our usage of the weak-limit $L = 15$ did not hurt our model fits, and that including a win-stay lose-switch feature, indicating which side was or would have been rewarded on the previous trial, was not beneficial in capturing animal choices during learning.

We used a 10-fold cross-validation scheme, randomly masking 10% of trials on each session. Since we were not interested in the details of the fits, we only ran one chain of 10.000 samples for each parameter combination and cross-validation fold we wanted to test and evaluated the quality of the fit through the summed negative log-likelihood on the last 4.000 samples on the held-out trials, which was sufficient for a stable estimation of the held-out log-likelihood. Despite this time saving strategy, there were too many combinations of parameters to check exhaustively, so we employed a manual heuristic search over promising combinations, finding an optimal setting and verifying that any relevant deviations from it only lowered the negative log-likelihood, see **Fig.**

3.6 left. As another measure to save computation we only evaluated 2 folds of each animal for each parameter setting, but since we evaluated our model on all 134 mice, we still evaluated on a substantial number of folds in total.

We tested the perseveration decay constant over the set of values (0.15, 0.2, 0.25, 0.3, 0.35, 0.4), the variance σ over the set (0.01, 0.02, 0.03, 0.04, 0.06, 0.12, 0.24), representing the small range which we found desirable for a consistent state identity, as well as some larger values to ensure that they did not outperform smaller variances. The search also included a larger support for the r parameter of the duration distribution (running from $r = 2$ to $r = 905$), and different settings of the α and γ concentration priors, which were independently varied over the set ((0.1, 0.1), (0.01, 0.01), (0.001, 0.001)).

Many of the parameter settings performed at a similarly high level. Note that the parameter setup which is closest to the selected model is one which simply allows for higher r values in the duration distribution, representing a strict extension of the model, which however does not improve fit. When studying the correlations of the same animal and same cross-validation fold across two different parameter settings, we found extremely strong correlations, with only slight offsets from the identity line and a small handful of outliers making the difference. This provided evidence that the fits were fundamentally the same, and different mice did not significantly benefit from different settings, allowing us to simply take the best among many good settings and proceed with it for the population wide fit. These settings were the ones specified throughout the paper: perseveration $= 0.25$, $\sigma = 0.04$, $r_i \sim U(5, 6, 7, ..., 704)$, and both α and $\gamma \sim \text{Gamma}(0.01, 0.01)$

In addition to finding the best parameters for our fit, we also used this approach to ablate the most important model components, verifying that all aspects of the model were necessary to provide as good a fit as possible within our framework, see **Fig. 3.6 right**. In particular, we tested the best parameter setting we found, but did not allow for change in weights between sessions (effectively removing the slow process of the model), both with 3 states (thus emulating the work of Ashwood et al. (2022), though with duration distributions) and with the usual upper bound of 15 states. Allowing for 15 states but no slow process led to only somewhat worse performance than the full model, but did so at the cost of significantly increasing the usage of short-lived states. We tested this by considering how many states explained more than x% of trials of an individual animal (which can be read directly from the cross-validation samples, not requiring the sample aggregation procedure described previously). The full model makes more use of highly prevalent states which explain more than 20% of trials: 1.7 ± 0.53 (mean $\pm$ s.d.) per animal, versus 1.07 ± 0.67 of such states for a model without the slow process (two-sided Mann-Whitney-U-Test, $U = 21858.5$, $p < 1e{-}30$, effect size$= 1.18$ (standardized mean difference with standard deviation over full model state number)), but fewer overall states, such as any that explain more than 2% of trials: 5.16 ± 1.62 versus 9.13 ± 2.14 (two-sided Mann-Whitney-U-Test, $U = 87773.5$, $p < 1e{-}73$, effect size$= 2.45$). Thus, while the removal of the slow process can mostly be made up for by an increased reliance on new states (for which our model has plenty of capacity), the slow process benefits the fits by tying together highly similar trials across short time scales, rather than arbitrarily separating them when behaviour gradually changes too much to be accommodated by a single state.

We also allowed for only one state (but including the slow process), removing the notion of multiple states from the fit. This model performed perhaps surprisingly well, but since a session is usually dominated by a single state,

a single adaptable state may perform somewhat well. We tested whether a Win-stay lose-switch feature, indicating which choice was or would have been rewarded on the last trial, was beneficial, which it was not, and whether the perseveration feature could be removed, which it could not. Lastly, we also tested the improvement due to the duration distributions (which replaced the implicit geometric duration distribution of an HMM). This test proved somewhat problematic within our framework, as restricting the model to implement durations through the transition matrix led many of the posteriors to settle on an unsatisfying solution. In this solution, states were extremely strongly biased leftwards or rightwards and rapidly alternated, depending on the choice of the animal. Such a model has of course almost no predictive power on held-out trials. This is seemingly a consequence of the hierarchical nature of the transition matrix: if we often transition into a state (and without duration distributions we have a state transition after every single trial, with most of them being self-transitions), it becomes generally attractive in the iHMM framework, encouraging transition distributions which are much closer to uniform than one would expect for a reasonable notion of temporally extended states. We thus implemented geometric distributions which prefer longer states by fixing $r = 1$, but biasing the prior over p. We performed another small cross-validation sweep, and present here the best model found in this way.

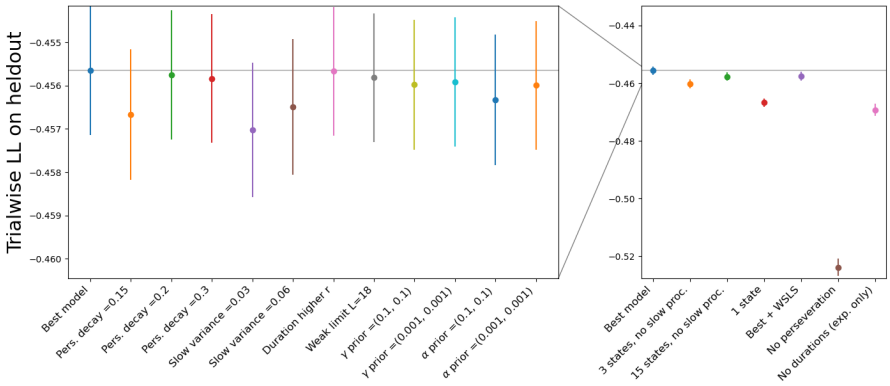

Figure 3.6: Left, Cross-validation results over the best model, and all variants which represent a small change either way in one of the relevant model parameters (not all tested parameter combinations are depicted). Error bars represent 1 standard error of the mean. The best model uses these parameters: perseveration $= 0.25$, $\sigma = 0.04$, $r_i \sim U(5, 6, 7, ..., 704)$, and both α and $\gamma \sim \text{Gamma}(0.01, 0.01)$. **Right,** Cross-validation results for ablated versions of the model. Note that the y-axis is considerably zoomed out from the plot on the left.

3.5 POSTERIOR PREDICTIVE CHECKS

To identify any mismatches between our modelling assumptions and actual behaviour, we performed posterior predictive checks using multiple test statistics. The goal of this analysis was to study if responses, when generated purely from posterior samples, reproduced the behavioural trends present in the actual data. We simulated behaviour for each session of an animal by taking each sample from our selected posterior mode, initialising with the state which was the actual state on the first trial for that sample, and then generated responses. We needed to initialise with the true state, since the model uses a static initial

state distribution π_0, so a random initialisation would lead to an unstructured mix of proficient and inexperienced behaviour. However, after intialising the first state, the model ran completely independently: we drew a duration from the duration distribution of that state, using posterior parameters, randomly sampled a next state from the transition matrix once a state ended, and sampled responses from the observation distribution of the current state, given the current features. These features included the contrast that was presented on that trial and a recomputed perseveration feature based on the choices of the current run of the simulation (so notably not the perseveration feature based on the choices of the animal). This unguided generation of behaviour thus represents a very stringent test of the posterior fit.

We visualised the results by plotting actual behaviour in relation to the distribution created by simulating behaviour three separate times with each sample (since we use at least 400 samples from a mode, this equates to > 1200 simulations). As metrics of interest we chose the percentage of correct choices in a session and the percentage of rightwards choices for each contrast. We plot the accuracy of a single individual (the mouse of Fig. 4.1) in **Fig. 3.7a** and the PMF on the last session of that animal in **Fig. 3.7b**. As we can see here, behaviour simulated from the posterior generally provides both a tight as well as accurate estimate around the true behaviour.

To summarise the relationship between true behaviour and the simulated distribution across the population, we calculated the percentiles of the empirical values within the simulated distribution, visualised in **Fig. 3.7c & d**. In an ideal case, the histograms over these percentiles would be uniform, indicating that the posterior provides an unbiased and calibrated estimate for the true behaviour. This is not quite true here: We can see that accuracy has a modest tendency to be overestimated (i.e. the true accuracy tends to fall onto lower percentiles of the simulated distribution). As mentioned in the discussion, behaviour often degrades towards the end of a session (almost by necessity, as it is one of the session termination criteria), but this was not always acknowledged with a separate state by the model, perhaps because behaviour degrades in a gradual and inconsistent manner across sessions. We suggest this as an interesting direction for a possible extension of our framework, by combining the states with a mechanism for change on a shorter time scale, similar to the work of Roy et al. (2021). However, implementing this in a way which keeps states distinct and has them retain their identity over long time periods seems challenging, in the face of motivational changes which occur gradually but can change behaviour quite notably on the order of tens of trials. Note that the overestimation of accuracy also occurs on sessions on which the model does ultimately include a state that reflects a substantial reduction in performance. This happens since the model sometimes fails to make a transition to this worse state appropriately (given that it averages over all sessions). Thus, accuracy in free simulations can be too great.

While the percentage of rightwards choices across contrasts forms a seemingly uniform distribution, splitting the histogram over the different contrasts reveals that there is a modelling assumption which biases the estimates for the different contrasts somewhat, as shown in **Fig. 3.7e**. Most notably, for the 100% contrasts, the model underestimates how accurate the animals are (by overestimating the % rightwards choices on leftwards contrasts and vice versa). Note though, that the insets for these contrasts show that the actual deviation is very small. Somewhat more subtly, the opposite occurs for the respective 50% contrasts. These deviations arise from the psychophysical transform we borrowed from Roy et al. (2021) and Ashwood et al. (2022), namely the tanh-

transformation on the raw contrast values. The 100% and 50% contrasts are mapped onto very similar values (1 and 0.987 respectively), strongly coupling the percentage of rightwards choices for the two contrasts, requiring them to take on almost the same value. This is intuitively desirable: allowing a smoothing over the different contrast strengths and reducing the number of parameters in our logistic regression (using a general "leftwards sensitivity", rather than having a separate parameter for each contrast). While 100% and 50% are very different in terms of absolute value, they are both highly visible, meaning their difference from a psychophysical perspective is rather minor (Fechner, 1860). Nevertheless, as it turns out, some mice can occasionally exhibit rather different behaviour on the two contrasts (see **Fig. 3.7e insets**), leading to an underestimation for the stronger contrast, and an overestimation on the weaker one.

The 0% contrast plot on the other hand exemplifies a posterior predictive check without such reservations: there is no noticeable bias and the posteriors appear correctly calibrated. The predictive checks thus serve as an important tool to study the limitations of our modelling approach, highlighting that degrading behaviour is not fully captured by the model, and that the smoothing over contrasts imposes some structure onto the PMFs that biases the performance estimates. To study further the effect these biases in the model have upon the fits, we analysed the magnitude of the bias imposed, see **Fig. A.3**. As we can see, most of the differences fall within a close range around the posterior mean.

3.6 MODEL RECOVERY

We tested the model and our inference procedures by fitting to data for which the ground truth was available. For this we instantiated all the random variables of the model to specific values and generated responses from it. This was performed for multiple different variable settings, to assess the accuracy of the fitting procedure in all relevant regimes, and using input data (i.e. contrast sequences) from actual training trajectories. The data generated this way were processed in exactly the same way as those of the IBL mice.

We paid particular attention to assessing the strength of the inductive biases of the inference procedure - particularly in terms of the number of states it inferred (given that this could be potentially unbounded, within our weak-limit approximation) and the degree of change between sessions (since slow and fast state changes could interact). We tested multiple settings in which all the data were actually generated from a single state, to test whether the model would incorrectly split behaviour into multiple states. In one setting, the psychometric weights of the state stayed constant throughout all sessions, in another, the weights gradually evolved from poor performance to proficiency (at constant steps of a magnitude that corresponds to a variance of 0.0311, the variance of the fitting procedure was fixed to 0.03). Both fits recovered their ground truth successfully, explaining virtually all trials with a single state, as can be seen for the example of the changing state in **Fig. 3.8**. We also tried a variation of the latter situation, in which the psychometric weights changed in (proportionally smaller) steps on every single trial, rather than all at once at a session boundary (as the model assumes). This, too, was recovered by the model with only one state (which we consider the best possible solution, given that the generative process was outside the model class).

We also successfully recovered settings from 2 to 9 states, with and without session-to-session variation on the weights, with strongly varying trial pro-

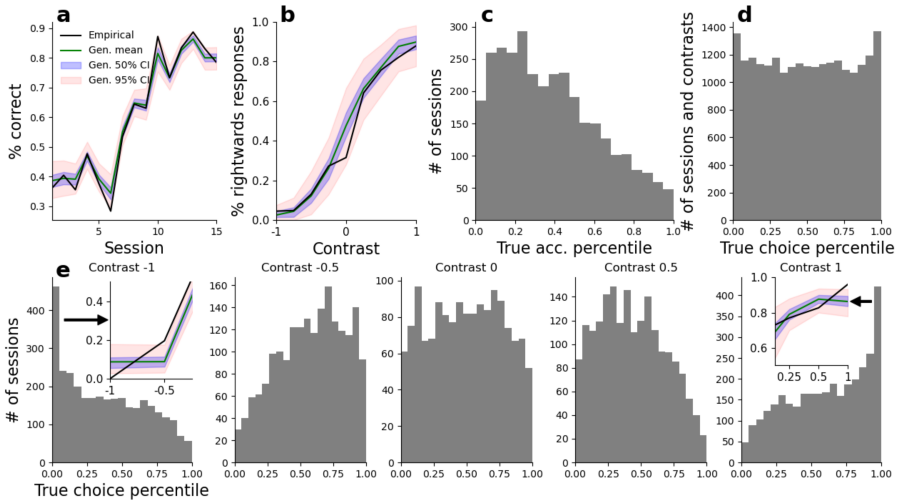

Figure 3.7: Posterior predictive checks. **(a)** True session-wise accuracy of the animal shown in Fig. 4.1, together with the mean posterior accuracy, as well as 50% and 95% credible intervals, created through simulation. **(b)** The true PMF of the same animal on its last session, together with the posterior distribution over the PMF. Both accuracy and PMF are well captured by the posterior. **(c)** By computing the percentile onto which the true accuracy falls within the posterior for each session across all animals, we can visualise the posterior fit on a large scale. This reveals a tendency to overestimate the accuracy of the animal – see discussion in text. **(d)** We can do the same thing as in (c) for the PMF, now over sessions and contrasts. Here the posterior seems close to uniform, as is desired. **(e)** Splitting (d) by the different contrasts, however, reveals a bias that arises from the psychophysical transform we apply, which maps strong contrasts onto very similar values. This can lead to an underestimation of performance on the strongest contrasts, and an overestimation on weaker ones (see the insets, which show an example PMF as in panel (b), zoomed in on the relevant contrasts). When performance on these contrasts is quite different, as shown in the insets, the desired smoothing over behaviour can introduce biases – see further discussion in text.

portions between the different states (see **Fig. 3.10**), and of varying overall training lengths (particularly to test whether long training trajectories lead the model to impose fewer states, making more use of the slow process), as seen in **Fig. 3.9**. The model was also tested on a setting with completely implausible PMFs, but with the added difficulty of having a larger number of states active within each session, **Fig. 3.11**. This, too, was captured accurately. These successful recoveries suggest that the model is able to uncover states that truly correlate with distinctly different modes of behaviour in animals.

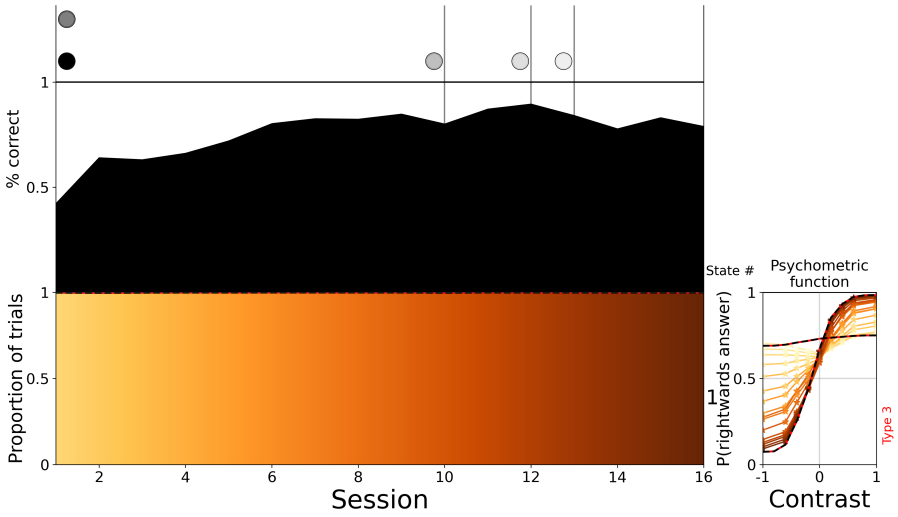

Figure 3.8: Model recovery for a case in which behaviour was generated according to the model, using only one state with a changing PMF. The red and black outlines indicate ground truth, which was basically perfectly recovered. For the changing PMF, we only show the first and last ground truth (the initial PMFs are not incorrectly recovered for low contrasts, but are not uniquely identifiable due to the limited contrast set during this period)

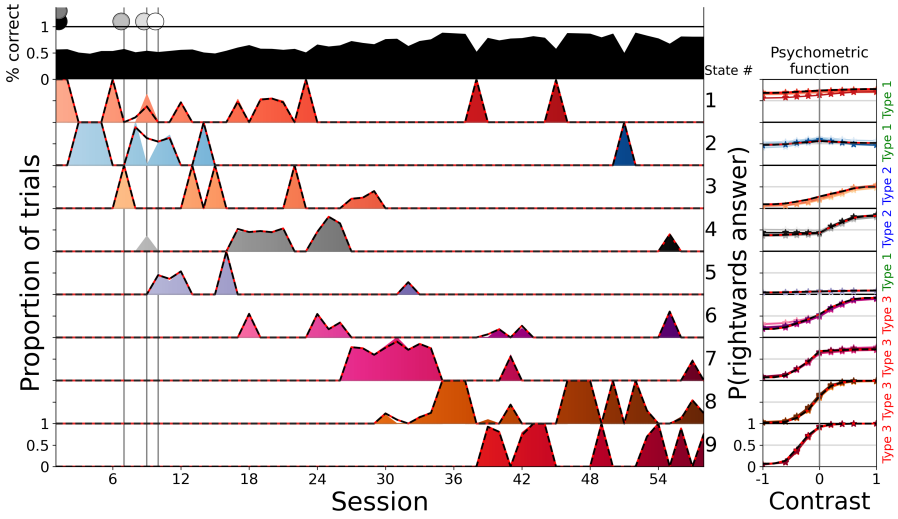

Figure 3.9: Model recovery where behaviour was generated using 9 different states, on a large number of sessions. The red and black outlines indicate ground truth, which is recovered close to perfectly, in particular correctly recovering the number of states. On session 9, state 2 is incorrectly split between state 1 and 4, the only major flaw.

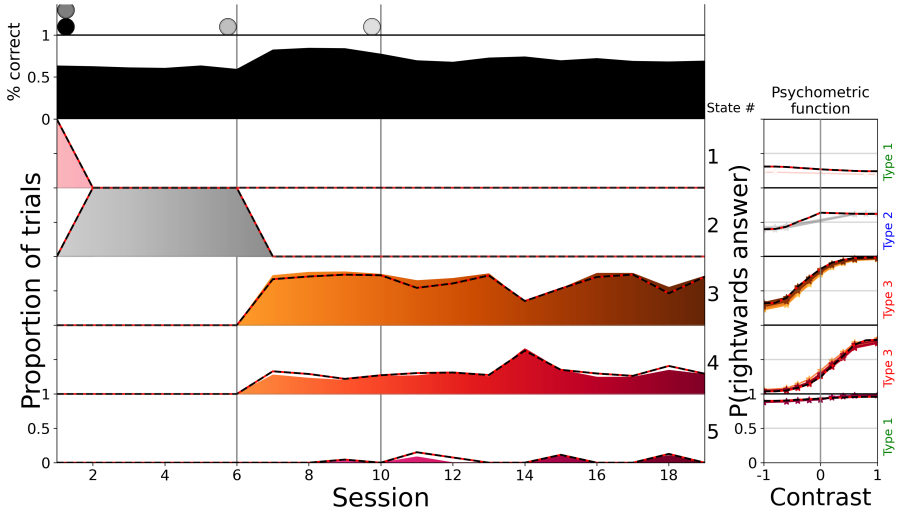

Figure 3.10: Model recovery of 5 states in a typical progression, the red and black outlines indicate ground truth. This example shows the model's ability to distinguish between similar states (3 and 4) within a session. A small number of trials were incorrectly assigned between states 3, 4, and 5, in particular the rare state 5 missed some trials.

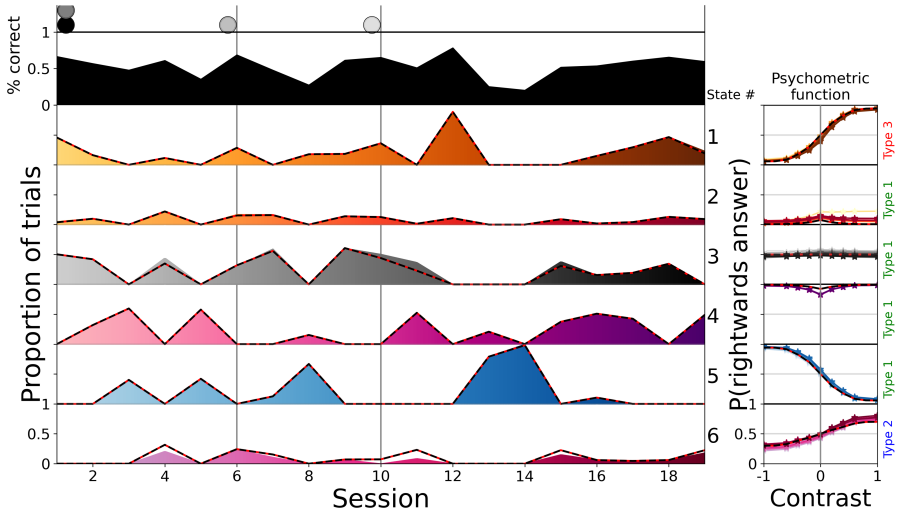

Figure 3.11: Model recovery of 6 states with some unusual PMFs, which made the states more distinguishable, but the recovery is made more difficult by the fact that many states co-occurred in single sessions. The red and black outlines indicate ground truth. The model found the correct number of states for this recovery and almost flawlessly discovered the boundaries between the states in the sessions. Only trials of state 3 and 6, which were the most similar (noting that state 3 had the highest variance in its responses), were sometimes noticeably misplaced.

Part IV

BEHAVIOURAL FITS

4 LEARNING TRAJECTORY FIT

We analysed the choices of 134 mice learning a perceptual decision-making task, each of them going through on average 24.4 (total: >3200) sessions and on average ∼14800 trials (total: >1.9 million) (The International Brain Laboratory et al., 2021). In this task, head-fixed mice were shown a sinusoidal grating of a controlled contrast, which had equal probability of being on either the right or left side of a screen, see **Fig 3.1a**. They then had to center it (within 60 s) by turning a steering wheel in the appropriate direction. Successful trials led to water reward; unsuccessful trials to a noise burst and a 1 s timeout. Trials were self-paced, with mice signalling their readiness by keeping the wheel still for a period.

4.1 SINGLE ANIMAL FIT

We visually summarise the model fit for mouse KS014 at the resolution of entire sessions in **Fig. 4.1**. This animal exemplified many of the interesting properties that can be found across the population. The inferred model contains eight states, but these states were generally active for only a small number of sessions, before being replaced by others. We number them in order of appearance. In a typical session, the majority of trials was explained by a single one; at most a few were active. Later states generally represented more adept behaviour, though not exclusively. The mouse started out with state 1 that exhibited a flat psychometric function (PMF; far right of the plot), indicating that the animal did not take into account the side of the sensory input. This state was promptly replaced by state 2 in the next session, which also had a flat PMF, though shifted. This change in bias was strong enough to warrant a new state (rather than using the slow process to change the existing state), but there was no evidence that the animal advanced in its understanding of the underlying task.

State 2 lasted for four sessions, meaning behaviour stayed relatively consistent during this time. It was then predominantly replaced by state 3 which started with a mostly flat, and strongly biased, PMF (leading to a lower reward rate, due to the bias correction), but improved considerably over the next few sessions, as can be seen in the evolving PMF (with darker colours showing later sessions). It seemingly only took sensory information from the left side into account when making its choice, performing increasingly randomly if that side was uninformative. The random behaviour was doubly beneficial, as the animal will sometimes have been correct, but also got foiled less by the bias correction protocol. State 3 was accompanied by state 4, describing the behaviour at the ends of the next few sessions (and later also the ends of session 14 and 15). Puzzlingly, this state had a good PMF on both sides and a higher reward rate than state 3, but even though this better state was available, the animal seemed incapable or unwilling to use it for the majority of a session.

The last major step in learning appeared abruptly as state 6, with good performance on both sides (albeit differently from state 4). Along with state 6 we saw the introduction of state 7, which captured a strong but transient decline in the quality of behaviour. Lastly, state 8 represented another notable

change in behaviour, as the performance on 100% contrasts increased sufficiently abruptly to warrant a new state, which allowed the mouse to conclude this part of training.

Various aspects of our model cannot be reproduced by existing treatments. The Psytrack model of Roy et al. (2021) can fit incremental changes in behavioural characteristics, but, lacking a concept of state, does not natively support the identification of recurring behavioural patterns. We find that there are many states which occur, then disappear, before reoccurring on a later session, such as state 3 and state 6 of the animal shown in **Fig. 4.1**. This re-emergence of previous strategies is an important feature of learning. Similarly, the static GLM-HMM of Ashwood et al. (2022), which is aimed at asymptotic behaviour, does not determine the number of states automatically. This implies that model selection is required for each individual animal, which the relatively small number of datapoints can make challenging. Furthermore, states cannot adapt their PMFs, which is a second important feature of learning. Without this, the GLM-HMM would tend to split states when behaviour changes gradually but sufficiently as to elude a single set of weights.

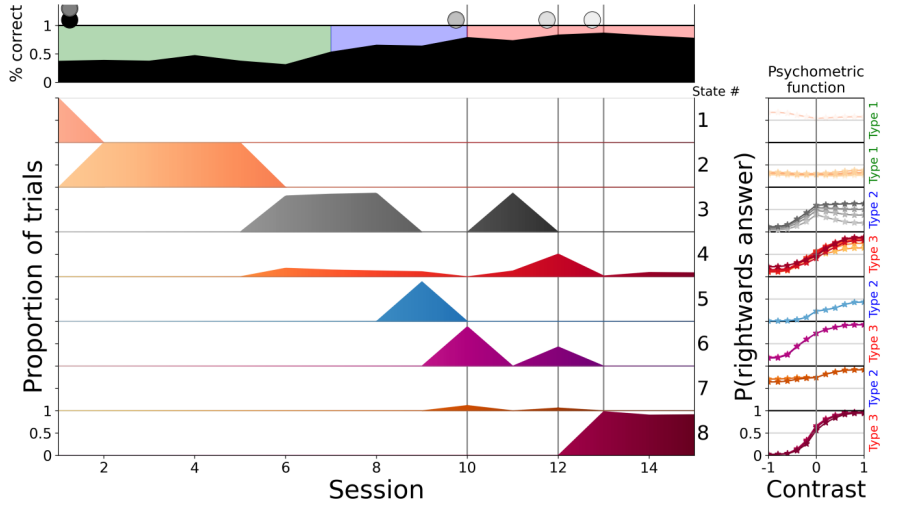

Figure 4.1: Dynamic infinite hidden Markov model (diHMM) fit to mouse KS014. The topmost row shows the overall performance during the session, as percent correct, and the current stage of learning as the background colour (we elaborate on learning stages later in the text). Vertical lines with shaded circles at the top indicate the sessions during which new contrasts were introduced. The remaining rows show the prevalent behavioural states (label to the right) ordered by appearance, indicating which percentage of trials they explained during each session. To the far right of every state we show its psychometric functions (PMFs) across time, ignoring the contribution from the history of previous choices. The saturation of the colours of the states indicates successive appearance, and match the PMF plots.

Our model also affords a fine-grained look at the use of behavioural states within a session. Although the diHMM provides a full posterior over the states for each trial, this is not directly useful, due to technicalities of the sampling procedure. We therefore processed the chains of samples to extract a measure of how much a trial belonged to a state (details in Chapter 3). We show an excerpt of this, for session 12 of mouse KS014 in **Fig. 4.2**. This shows two clear transitions between states. The reasons for the animal to have made such a transition are probably multi-faceted, and may have been both internal (e.g. insights or motivational fluctuations (Berditchevskaia et al., 2016)), as

well as external (e.g. a number of low contrast, perchance unrewarded trials demotivating the animal). We do not model these reasons, and instead only describe observed changes.

The within-session fit showcases that the model is able to find temporary, but strong deviations in behaviour. State 7 only explained a couple of dozen trials in two sessions, but represented extremely biased behaviour (comparable, but flipped relative to the earlier state 2, albeit lasting for many fewer trials). We speculate that state 7 arose from a form of inattention, since the animal had previously shown itself capable of performing appropriately. This change in behaviour is directly evident in the response patterns of the animal.

We can also capture subtler differences in behaviour. The model used different states to explain behaviour before and after state 7, even though performance looked about equally good. However, the model identified different error rates on easy contrasts for the two states, and this can be found in the choices: state 6 was associated with more incorrect responses to contrasts on the left side, whereas in state 4, performance on leftwards contrasts was good, but there were frequent lapses on rightwards contrasts (in an ANOVA for responses during state 4 or 6 in session 12, with factors signed contrast and state, the factor state is significant with $p < 0.0004$ ($F=12.980669$, $df=1$), $\eta^2 = 0.009$), see also the responses marked by arrows in **Fig. 4.2)**.

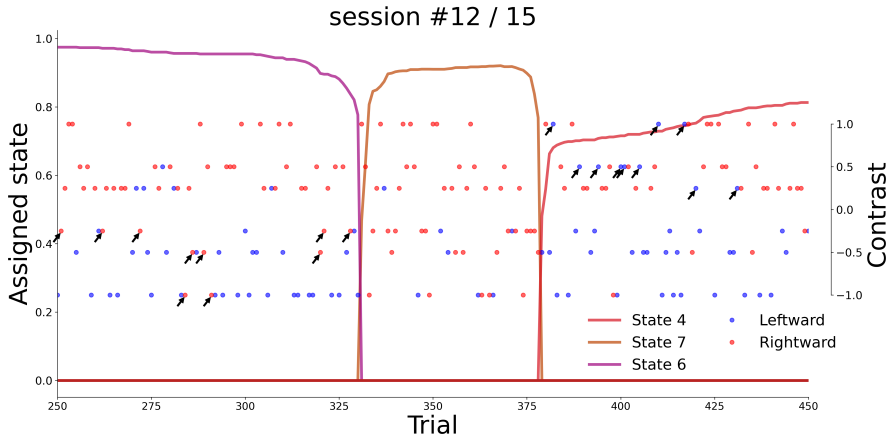

Figure 4.2: Excerpt of state assignments in session #12 from mouse KS014, also shown in **Fig. 4.1**. The left y-axis serves as a scale for how connected a trial is to the other trials of that state (see Methods for details). The right y-axis shows the contrast. The dot colour indicates the animal's response. One can see how the drastic and sudden change in the response patterns, rightwards (red) answers for leftwards (negative) contrasts, from trial ∼330 to ∼380 was detected by the model with a transition to state 7. The PMFs of state 4 and 6 looked similar, but did in fact represent significantly higher rates of errors on the right and left side respectively. These mistakes are highlighted with arrows.

4.1.1 *Posterior uncertainty*

Our MCMC-sampling approach gives us access to a full posterior of the model. For the majority of animals one mode of the posterior clearly dominates, and we used samples from this mode to represent it within the following population analysis. However, sometimes there are secondary and even tertiary modes which capture a substantial number of samples. The animal we have focussed

on so far is a good example of this, **Fig. 4.3** shows the visualisation of the entire fit for the three modes of this animal, highlighting points of uncertainty within the posterior.

As we can see the fits are extremely similar, revealing only some uncertainty about the extent of state 4. It appears to be the case the the behaviour in the first session of state 4 (speaking from the perspective of the first mode) is somewhat distinct, and could be considered its own state. On the other hand, behaviour on session 9 is sufficiently similar to state 4 behaviour that it could be merged with it. It employs a somewhat degraded version of the PMF which state 4 arrives at towards the end of its lifetime, but similar to the PMFs state 4 has exhibited along its trajectory. Assigning this behaviour therefore seems a subtle distinction. Since the fits are so similar and most animals are clearly described by a single mode, we will not consider other modes in this work.

4.2 FITS ACROSS THE POPULATION

The three-fold progression we observed throughout learning in **Fig. 4.1**, from flat PMFs, to "one-sided" behaviour, to generally good performance, is typical for the population of mice we fitted. To define this more objectively, we clustered the states into these three types, based on their reward rate on easy trials (see section A.1 for details). The boundary between type 1 and 2 is at 60% reward rate, and between type 2 and 3 at 78% reward rate (details in A.1). We show an overview and examples of the different types in **Fig. 4.4**.

In addition to the state types, we define the *stage* at which an animal is on any given session, as the highest type it has so far used for the majority of trials of any session up to this point. For instance, if up to session $n-1$, an animal only used type 1 states, or type 2 states for fewer than 50% of trials, then it would be in stage 1 for those sessions. If on session n, it then used type 2 states for more than 50% of trials, it would switch to stage 2 on that session. Since the state types delineate different aspects of task understanding, the stages allow us to determine for how many sessions the animals stayed at a certain level of understanding. Whereas the progression through state types was not monotonic, (e.g. session 11 of **Fig. 4.1**), the stage classification is, by definition, monotonically increasing.

Stage 1 consisted of states with flat PMFs of various biases, generally ignoring the contrast location. Stage 2 almost always involved asymmetric states, responding well to one side of the screen, but close to uniform guessing for the other (PMFs assigned to **Fig. 4.4**b and c account for 86% of those in stage 2, see section A.1 for details). Only rarely were intermediate PMFs nearly equally good on both sides (the 30 in **Fig. 4.4**d). Lastly, in stage 3, the animals started apparently paying attention to both sides. Generally, it took some further refinement of initial type 3 states, through the reduction of errors on easy trials on either side, to master this stage of training and progress to the next phase of shaping.

It is not directly obvious that a symmetric type 2 (**Fig. 4.4**d) belongs with the other type 2 PMFs, or is rather a very early stage of a type 3 PMF, and simply misclassified by our reward rate criterion. To make sure our classification is appropriate for these special cases, we analysed when during training the different types tended to be active, see **Fig. 4.5**. The fact that asymmetric type 2's have an appearance profile which is extremely similar to that of symmetric type 2's suggests that the grouping we impose on them handles these two cases appropriately.

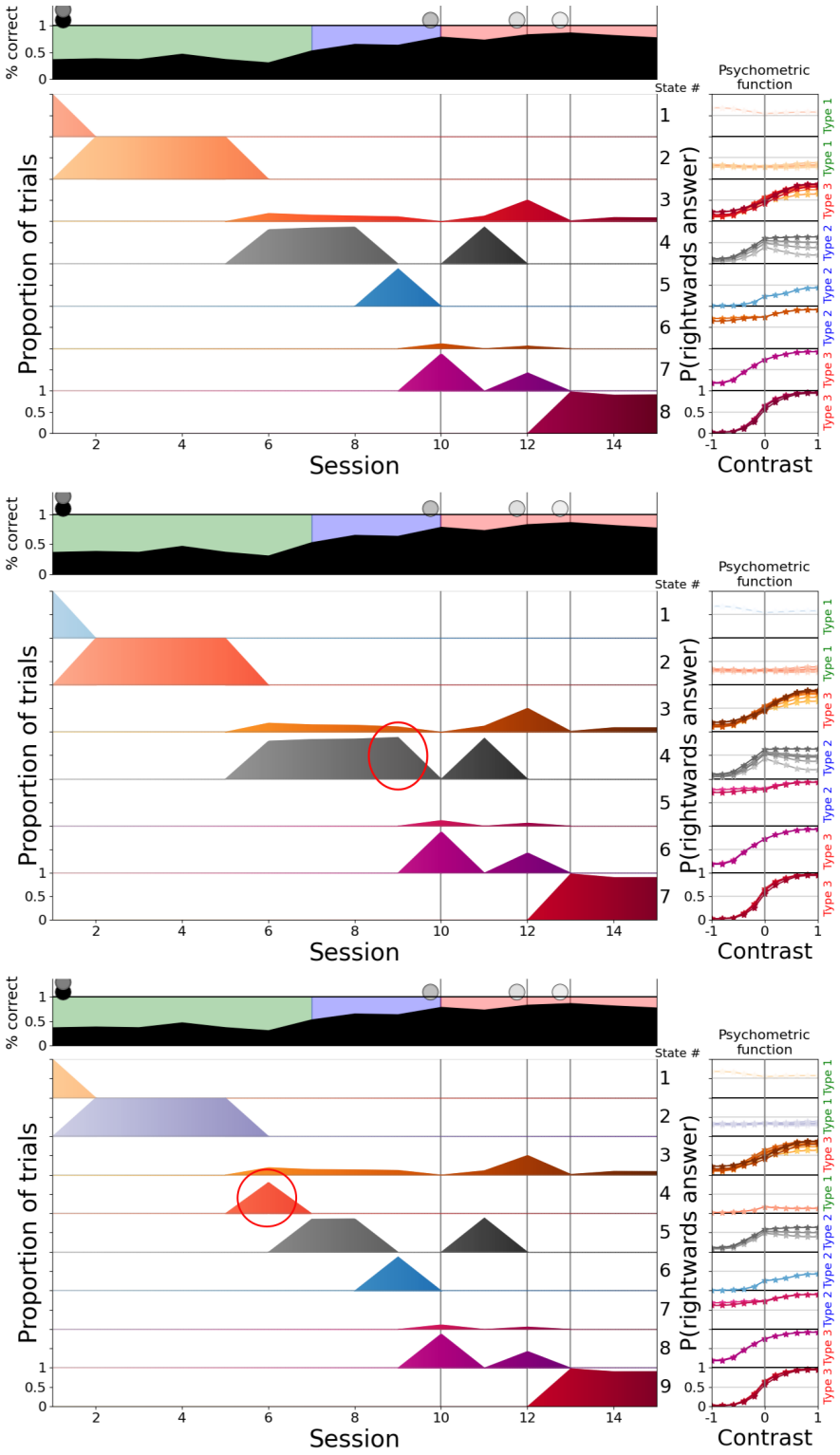

Figure 4.3: Comparison of the posterior modes of mouse KS014, first, second, and third from top to bottom. The fits are extremely similar, the red ellipses mark the points of disagreement to the first mode. As we can see, there is some uncertainty about state 4, it can possibly absorb state 5, or its first session can be split into a different state.

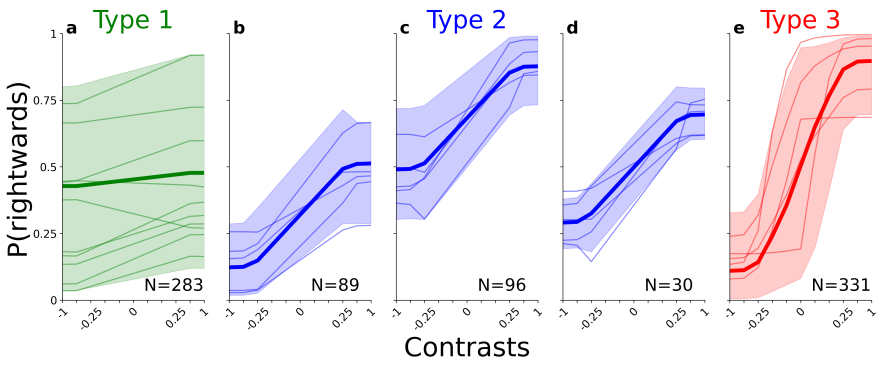

Figure 4.4: Summary of the PMFs associated with the different types, for this we use the first PMF of every state in every animal (thereby representing response characteristics after a notable discontinuity in behaviour). Each subplot shows a specific type: **(a)** type 1 in green; **(b-d)** type 2 in blue, further split by whether the PMF is left-biased (b), right-biased (c), or symmetric (d); and **(e)** type 3 in red. The thick lines indicate the overall mean over PMFs of the type, which shows representative behaviour of that type , the shaded regions show the range in which 95% of the PMFs fell (computed separately for each contrast level), and the thin lines show samples of individual PMFs of these types. Text in the bottom right indicates how many PMFs of each type were present across the entire population.

The three stages segment the learning process. We can analyse the proportion of training time the animals spent in the different stages, by showing these proportions on a simplex (**Fig. 4.6**; see also the linear representation in **Fig. A.2**). The large majority of animals spent some time in each of the stages (i.e. only few mice are assigned to the edges of the simplex). Most animals spent the longest time in stage 3 – going from moderately competent performance to passing the stringent training criteria. No fundamental change in understanding was necessary for this, unlike the changes from stage 1 to 2 or 2 to 3 (where the animal had presumably to learn to pay attention to the Gabor patch on one or both sides of the screen). However, reaching the required accuracy seemed difficult, even once the principles of the task were understood (possibly due to the small increase in reward rate afforded by the extra accuracy). Some of the longest trajectories (the largest circles) were associated with especially many sessions in stage 3, but overall the average fractional occupation was remarkably consistent across training lengths (the mean relative occupancy for stage 1, 2, and 3 respectively were (0.24, 0.17, 0.59) and (0.21, 0.14, 0.65) for the shorter and longer halves of a median split on the total number of training sessions). Stage 2 consistently lasted for the fewest sessions, implying that the mice managed to pay attention to both sides not too long after starting to pay attention to one side.

Connected to this is the question of how slow and fast changes characterize behaviour. We analyzed gradual changes within a state by comparing its PMF weights on its first and last appearances. We analyzed new state introductions by comparing their PMF weights to those of the closest previous state, as determined by the Wasserstein metric on their resulting PMFs (ignoring the perseverative weight). To highlight the changes, we focussed on states which brought the animal into a new stage. These weight evolutions, split by the different types, can be seen in **Fig. 4.7**. As the main driver of performance, contrast sensitivities reliably increased both over the lifetime of a state and when new states were introduced. Surprisingly however, both the bias and

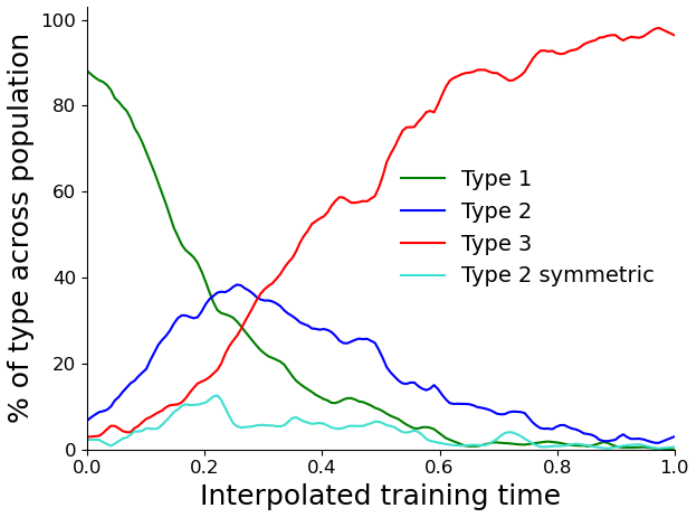

Figure 4.5: Prevalence of the different types across the entire population. We standardised each mouse's trajectory onto the range from 0 to 1, to average across varying training lengths. As can be seen, the asymmetric type 2 states behave just like type 2 states in general in terms of their appearance, lending credence to our grouping.

perseverative weights were stable within a state. This was markedly different for the fast process: the changes through this were significantly larger (**Fig. A.5** and **Fig. A.6**; one-sided Mann-Whitney-U-test on absolute weight changes, 2 biases, 2 fast change points, 3 slow change processes; the fast process had significantly larger changes for all twelve comparisons at a 0.05 significance level, after applying the Benjamini–Hochberg procedure to control the false discovery rate, with effect sizes ranging from 0.43 to 0.94, quantified as standardized mean difference, using the fast change standard deviation). We also see that the perseveration weight played a small, but consistent role throughout learning (though its relative influence waned as the sensitivities grew).

The introduction of new states signifies notable changes in behaviour, so by studying the patterns of their occurrences, we gain insight into when behaviour was volatile or when substantial progress was made. The histograms in **Fig. 4.8** show when new states first appeared across normalised training and session times. In later sessions, gradually fewer states were introduced, indicating that behaviour saw fewer drastic changes as training progressed. We noted earlier that animals spent most of their time in stage 3, i.e. perfecting their behaviour, and we can now conclude that gradual improvements played an important role in this, more so than sudden marked changes. The pattern of introductions within sessions is even more striking: the majority of states were introduced at the very start of a session. This resonates with previous findings about changepoints in behaviour occurring at session boundaries (Papachristos et al., 2006). Apart from this strong trend, there seems to be a slight tendency for new states to get introduced towards the end of sessions, which might have partly been demotivated states which the animals often fell into at the end of a session, and which the model sometimes picked up on when they were consistent and long enough (the end of a session is triggered when performance degrades).

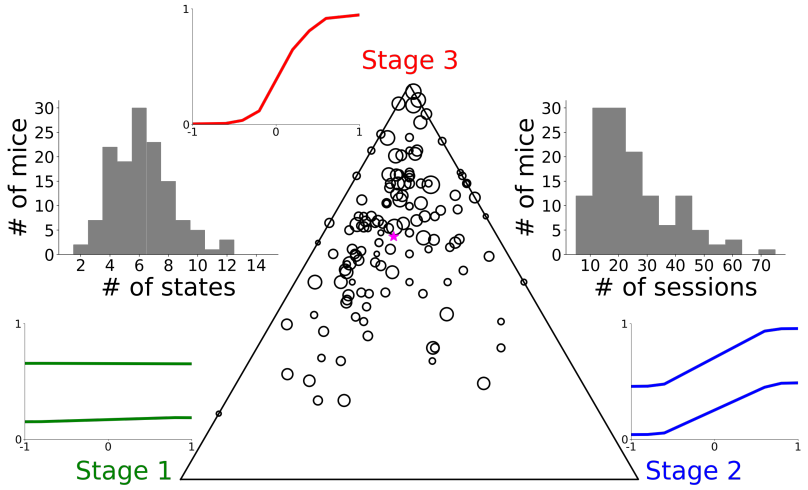

Figure 4.6: Proportions of sessions it took each mouse to reach the next major step in training, as defined by the 3 stages, scattered as circles on a simplex (the larger the proportion of sessions within a specific stage, the closer the dot for that animal is to that corner of the simplex). Simplex corners are identified by example PMFs from the type of that stage. Marker area indicates the total number of sessions (min=5, max=75). The magenta star marks the average proportion, the size of the star indicates the mean number of sessions (which was 24.4). See Fig. A.2 for a linear version of this plot. The histogram on the left shows the distribution over the number of states used by the model per mouse. The histogram on the right shows the distribution over the number of training sessions.

4.2.1 *Inter-individual differences and variability*

So far, we have highlighted general patterns during learning, but perhaps even more salient than these similarities was the wide-ranging variability across animals. Such differences are already visible in many of the plots above. Biases in type 1 states spanned the entire range of possible response patterns. Similarly, type 2 PMFs appeared to have been randomly biased to one side or another, or, rarely, symmetric. We were particularly surprised that we could find no regularity between type 1 and type 2 biases. Of the 56 mice in which type 2 onset occurred suddenly, 31 had expressed the same direction of bias (average choice from the PMF being more than 5% away from 50%) as the new type 2 state in any previous type 1 state, whereas 25 had not (two-sided Binomial test for whether the proportion of previously expressed biases differs from 0.5 gives p=0.504). Thus, we were unable to predict future biases of the animal from its stage one biases.

The number of sessions mice required to learn varied greatly, spanning an order of magnitude. The number of sessions spent in the different stages was similarly highly variable. To gain insight into the factors underlying the learning steps between the stages, we analysed the correlations between the numbers of sessions spent in them. The simplex plot does not strongly indicate any patterns, we quantify this as such: duration of stage 1 to stage 2: Pearson's r=0.21, p=0.015; stage 1 to stage 3: Pearson's r=0.04, p=0.685; stage 2 to stage 3: Pearson's r=0.14, p=0.095. Notably, the main chunks of training time, stage 1 and 3, show no correlation whatsoever. A speedy understanding of the basic contingency of the task therefore did not tend to go along with the ability (or will) to perfect this behaviour quickly, suggesting that they required

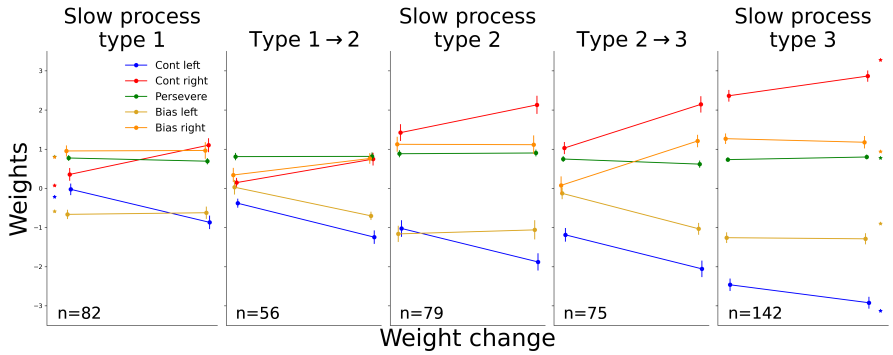

Figure 4.7: Evolution of the weights of states on average, through slow and sudden changes; errorbars indicate ± 1 SEM (lines are slightly offset along the x-axis for visibility). Subplots titled by a type represent the weight changes from the first appearance of a state of this type to its last, so only showing state-internal slow changes (and only including states which were present on between 5 and 15 sessions, as extremes would skew these averages). Subplots with a title indicating a transition from one type to another show how much each weight of the new state differed from the weights of the closest previously existing state, and are based exclusively on the states which first brought the mouse into a new stage (i.e. for "Type 1 $\rightarrow$ 2", we only took into account the first type 2 state exhibited by the mouse, and only when that state was type 2 from its inception, for mouse KS014 this was state 4, which started as type 2 before using the slow process to become type 3). Coloured stars on the leftmost and rightmost plot indicate the average value of the weights of the very first state of each mouse and of the dominant state on the last session, respectively. To prevent biases from cancelling out across the population, we split the bias weights into two groups: starting out below 0 (bias left) or starting out above 0 (bias right). Whereas contrast sensitivities increased both through fast and slow changes, it is noticeable that biases stayed almost constant throughout the lifetime of a state on average, but changed more noticeably through sudden transitions.

different competences. The strongest correlation exists between stage 1 and 2, which makes sense in so far as they were both concerned with discovering how to make use of the stimulus information.

This suggests that learning about the biased blocks tapped into yet another type of skill, unaddressed by the requirements of the pre-bias protocol.

4.3 DISCUSSION

We presented a highly flexible model which describes the stages of learning from the very first day an animal interacts with a task until it becomes an expert. Using it on the shaping sessions of the IBL decision-making task, we distinguished fast, abrupt transitions in behaviour, and slower, gradual ones. Learning on this task decomposed into three distinct stages, through which almost all animals went: Initial, undifferentiated, and often biased behaviour, partial, one-sided understanding of task contingencies, and lastly full understanding of the task. While these broad-stroke characteristics were consistent across mice, and indeed resonate with recent results from other tasks (Dekker et al., 2022; Liebana Garcia et al., 2023), the details of behaviour in these stages differed considerably across the population. Similarly, the way they progressed through these stages differed widely in duration and composition of sudden and gradual steps.

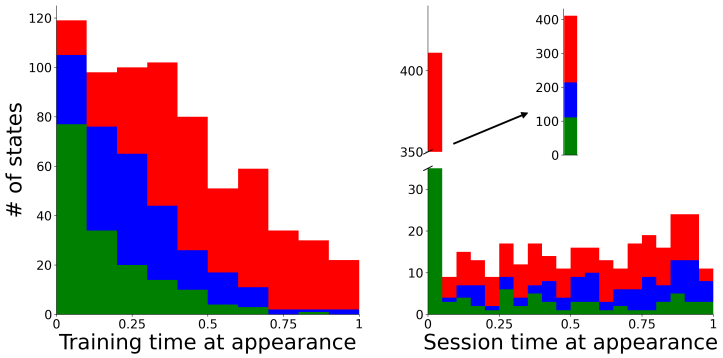

Figure 4.8: Histograms of all state introductions, excluding the first state of every animal (which necessarily occurred on the first trial of the first session), across all of training (left) and within sessions (right). We colour by state type (green, blue, red for types 1, 2, 3 respectively) and normalise the entire length of training of an animal, and all individual sessions, onto the range between 0 and 1 for comparison purposes. The inset on the right plot shows the bar of the first time bin uninterrupted.

We found only a weak correlation between the time it took individual mice to progress through some of the behavioural stages, suggesting that they had to draw upon largely different skills to learn the requirements of the task. Similarly, animals expressed varying, largely uncorrelated, biases across the stages of learning. They might therefore have different sources: in stage 1, when the mice paid no attention to the stimulus, biases might be motoric; in stage 2 they could have been an expression of the side that individual mice happened to notice first as being informative, and in stage 3, they might have stemmed from differences in sensory acuity. In the IBL training scheme, after the sessions we analysed, the mice underwent a further phase ('biased block training', in which left or right stimuli dominated in blocks of 20-100 trials). Consistent with our other results, the length of this phase also turned out not to positively correlate with the total pre-bias training duration, nor to any of the stage durations. At most, there was a negative correlation between the overall bias training time and the stage 3 duration (see section A.2 for details). This again shows that learning was influenced by a large number of factors in our setting.

We originally expected that mice who took very many sessions to train would be characterized by very many states. However, even though recovery analyses show that the model can cope effectively with long trajectories (see **Fig. 3.9**), this was not always the case. Instead, we often saw that few states were taken a long way via the slow process, from uninformed to proficient (see **Fig. A.4**). It will be important to assess the underlying nature of these states and their progression, by tracking neural data through the course of learning.

It is important to note that our model does not require as large a data set as we used. Individuals were fitted by themselves, the model proved flexible enough to accommodate considerably different numbers of training sessions, and our cross-validation indicates that the fits are not critically sensitive to hyper-parameter selection, the only part which made use of all subjects combined. Nevertheless, our modelling approach does have a number of limitations. First, the setting of the slow change variance parameter which determined how much the behaviour of a state could change from one session to the next plays a critical role in steering the trade-off between introducing a new state

versus adapting an existing one. We optimised this parameter in terms of cross-validation performance for the entire population (see Methods). However, the magnitude of slow changes may depend on the individual or vary across training time, and thus a more differentiated treatment might be appropriate. Furthermore, slow changes may also occur within a session (Roy et al., 2021), which could be incorporated into the model by adding additional time points at which weights can change. Another desirable extension would be to allow the duration distributions to change over sessions. As training progresses, an animal might, for instance, be able to use a high performant state for longer. Similarly, a dynamic transition matrix and dynamic initial state distribution might better capture the evolution of state usage across training.

The model may be extended by adding more observations for the states to explain, as the binary choice behaviour may limit the power to distinguish different behavioural modes. One obvious possibility are the reaction times of the animal's choices; in principle, this would only require adding a suitable distribution to produce times for each state (e.g., from a drift diffusion decision-making process; Ratcliff et al., 2008; Gold et al., 2002). It would likely be necessary to make the distributions dynamic, as the reaction times improve with training. Other possibilities include pupil dilation or even body posture (Wiltschko et al., 2015).

Previous work using an HMM-based approach discovered demotivated states in behaviour during the first 90 unbiased trials per session in the subsequent phase of IBL behaviour (Ashwood et al., 2022). The prevalence of sizable blocks of trials during which the animal performs at a decreased level will, if left unaccounted for, lead to confounded estimates of model parameters and a flawed understanding of the animal's current skill development stage, making it an integral component of a good behavioural model of this task. We also find such states, characterized by reduced sensitivity to the contrast feature of at least one side, and a strong bias in extreme cases, leading to higher than normal lapse rates on strong contrasts. However, these were not as pervasive as might have been expected from Ashwood et al. (2022). For us, a majority of sessions were dominated by a single state. The model sometimes acknowledged the dip in performance of the animals at the ends of sessions for tens of trials with a separate state (seen in **Fig. 4.1** on multiple sessions). We analyze aspects of these trials in section 3.5. However, frequently we just see a decrease in the prevalence of all sufficiently represented states. The main source of behavioural variability in our data came from learning and other large jumps in psychometric space, therefore the model used its capacity to capture these.

Besides task acquisition, our approach to capturing behavioural evolution, which has conceptual relations to those used in the animal conditioning (Gershman et al., 2012a; Lloyd et al., 2013a) and motor learning (Heald et al., 2023; Luft et al., 2005) literatures, should be well suited to model other progressive changes, such as those occurring during ageing (Nyberg et al., 2012). Furthermore, our framework can be flexibly adapted to other cases of long-run learning. For instance, it is possible to tune the model to capture minute changes within sessions as opposed to broad stroke states across sessions, as here, by adjusting the propensity to infer new states for small changes in behaviour. Equally, the modular resampling procedure of the model allows it to be adapted to different kinds of observations, e.g. multinomial or Gaussian, by simply swapping out the inference mechanism of this component (although only some distributions are convenient for the gradual dynamics). We therefore

hope that the tool we developed here will enable a wide range of researchers to study the development of behaviour in a systematic and revealing manner.

Part V

MODELLING EXPERT BEHAVIOUR

5

HYBRID NEURAL NETWORK MODELLING OF EXPERT BEHAVIOUR

In the final stage of the IBL protocol, expertly trained mice perform the full task. By this time, they have had ample exposure to the biased block contingencies (see chapter 2) and their behaviour is considered sufficiently stationary and consistent across the population to allow for pooling of the behavioural data across labs (The International Brain Laboratory et al., 2021). In the full task, optimal performance would involve combining two components: (i) current contrast processing, which takes into account the side of the currently presented stimulus; (ii) the posterior probability of the identity of the current block. This posterior should ideally arise via a Bayesian changepoint analysis of past stimuli sides, whose identity the animals can infer from the trial-by-trial feedback (which is deterministic),

The starting point for our investigation is the observation (Findling et al., 2023) that the animals seem not to calculate this Bayesian prior. Findling et al. (2023) suggested that they make two critical approximations: one is to use their own previous choices rather than the actual stimuli sides; the other is to enter these past choices into an exponential low-pass filter. This leads to a form of perseveration – something that we already observed during shaping. The difference is that perseveration was deleterious during shaping, since stimulus sides were chosen independently at random. The biased blocks introduce autocorrelations between the contrast sides, and perseveration, at least with the appropriate learning rate for the low pass filter (which Findling et al. (2023) showed the animals generally adopted), is only slightly worse than optimal.

Since this exponential filtering is not normative, but Findling et al. (2023) showed that the animals were also not Bayes optimal, we sought to examine whether they had adopted some form of intermediate strategy. We did this whilst staying close to the previous framework, learning a mapping of input features (involving contrast strengths and a summary of past trials) onto logits, which we then add and pass through a softmax function to obtain probabilities for the choice on the current trial. To open up the space of expressible functions, we move to a by design unrestricted class of models, namely neural networks (Hornik et al., 1989). Within this framework, the only limits to the predictive performance of our models are the amount of training data, the capacity of the chosen architecture, the suitability of the training procedure, and the inherent noisiness of the data. We will use feedforward and recurrent neural networks (an LSTM (Hochreiter et al., 1997) for the latter) which, given their potentially unrestricted natures, should not impose a limit upon performance through their architecture. Given the large amount of data embodied by the IBL dataset and a comprehensive exploration of architectures and training procedures, we can hope to achieve predictive performance close to the noise ceiling of the behaviour which we study.

The downside of this move to unrestricted models is, of course, the inability to readily interpret a fitted model. To alleviate this issue, along with an unrestricted recurrent neural network, we also employ the hybrid neural network approach of Eckstein et al. (2024): rather than fitting one monolithic network to the data, we experiment with splitting the network into distinct

sub-networks, limiting the ways in which each network accesses information or influences the ultimate logits, thereby assigning them distinct roles in explaining behaviour. The goal is to keep the sub-networks interpretable not by limiting their expressiveness, but by keeping the role they play in the larger whole sufficiently small, so as to be understandable by considering only the mapping from inputs to outputs which they implement. In this scheme, we cover an entire range from the simplest model (similar to the one presented in Findling et al. (2023)) to the completely unrestricted network. Not only do the relative performances at each level provide insights, but also, by ensuring that our ultimate model performs at a level similar to the unrestricted neural network, we can protect against imposing unwarranted extra structure. This mitigates against the risk of inadvertently limiting the space of functions which we can model.

Making use of the flexibility which neural networks afford, we also extend the binary choice framework which we used previously into three dimensions, explicitly modelling the rare timeout responses (in the process generalising the logistic function to a softmax).

We fit each variant of our network architecture to all sessions performed by fully trained mice (at least the ones in our training set), as we are searching for the overall strategy of mice through the use of a huge dataset, rather than specifically tuning to individuals with much less data (though our final network may do this to some degree, as discussed later). It is important to keep in mind that we train our networks to predict mouse choices, not to perform the task itself.

The rest of this chapter will first discuss our hybrid network approach, laying out the levels of sophistication and where they come in. Afterwards, we will retrace our progression through the most important insights obtained through our model fitting, by briefly discussing a number of specific architectures in more detail. This section ends with a relatively simple, but interpretable model, which nevertheless achieves performance on par with an unrestricted recurrent neural network. We will then analyse the workings of this final hybrid network and compare it against the classical exponential filter model to study its sources of improvement, before discussing our results. An appendix to this chapter contains more details on the other architectures which we did not present, the details of our fitting procedure, and some supplemental figures.

5.1 HYBRID NETWORK LADDER

We can think of our hybrid networks as progressively higher rungs on a ladder of flexibility. At the very bottom is the classical, simple, and interpretable exponential filter in an extended logistic regression framework. The neural network architectures are designed as explicit extensions of this model, with three distinct points for adding complexity through networks. We begin by detailing the simplest model, and then introduce the different levels of neural network extensions.

THE BASE MODEL: For the processing of the current stimulus, we produce contrast logits by applying the tanh transform (as discussed previously) elementwise to the contrasts $\mathbf{c}_t$ on trial t. $\mathbf{c}_t$ is a two-dimensional vector encoding the strength of the contrast on the left and the right side of the screen (at most one of which will be non-zero). Since we consider timeout choices as well, we have to slightly extend our logistic regression framework of the previous chapters. We now have to produce three logits to turn into probabilities over three

possible choices. The tanh is thus followed by a learnable linear transformation comprising a matrix $\mathcal{M}_c \in \mathbb{R}^{3,2}$ and a bias $\mathbf{b}_c \in \mathbb{R}^3$:

$$\hat{\mathbf{c}}_t = \mathcal{M}_c \frac{\tanh(5\mathbf{c}_t)}{\tanh(5)} + \mathbf{b}_c. \tag{5.1}$$

We use $\hat{}$ to indicate that raw information has been processed into a logit representation associated with the action propensities for the different choices. Note that the linear transform corresponds to the logistic regression weights of the previous chapters and the added bias implements the bias feature.

The contrast logits are supplemented by history logits, which summarise the recent history of choices, biasing the current one. This can reflect largely appropriate considerations about the current biased block, but also other perseverative tendencies of the animals. An action is represented as a three-dimensional one-hot vector. The first entry encodes whether a leftwards choice occurred, the second a timeout "choice", and the last dimension indicates a rightwards choice. After a trial, the current memory over previous choices $\hat{\mathbf{h}}_{t-1}$ is updated by multiplying it with a decay constant $0 < d < 1$, and adding the most recent action $\mathbf{a}_{t-1}$ into memory:

$$\hat{\mathbf{h}}_t = \hat{\mathbf{h}}_{t-1} \times d + \mathbf{a}_{t-1}. \tag{5.2}$$

The repeated multiplication implements an exponential decay of the influence of past choices. This is, of course, equivalent to the perseveration feature of the previous chapters. The ultimate choice probabilities $\mathbf{p}_t$ are generated by adding the two components (weighting the history by a learneable scalar w) and passing them through a softmax function:

$$\mathbf{p}_t = \sigma(w \cdot \hat{\mathbf{h}}_t + \hat{\mathbf{c}}_t). \tag{5.3}$$

With the softmax defined as

$$\sigma(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}. \tag{5.4}$$

This approach is summarised visually in **Fig. 5.1**, under the header "Level 0 processing", as the most basic approach to modelling our behaviour of interest. During model fitting, we minimise the negative log-likelihood of the probability assigned to the action that was actually chosen on each trial, across all sessions in the training set.

The three extensions to this classical model are:

FLEXIBLE CONTRAST PROCESSING: Rather than picking a specific functional form for the mapping from contrasts to logits, we input the current contrast into a network (contrast network), which then produces a three-dimensional output vector. This contrast network N^c (using N to denote the operation of a feedforward multilayer perceptron, or MLP) thus performs the simple transformation:

$$\hat{\mathbf{c}}_t = N^c(\mathbf{c}_t). \tag{5.5}$$

Our MLPs always involve a single hidden layer, following:

$$N(\mathbf{x}) = \mathcal{M}_2(\tanh(\mathcal{M}_1\mathbf{x} + \mathbf{b}_1)) + \mathbf{b}_2. \tag{5.6}$$

with an elementwise tanh non-linearity. We tried various network sizes, as discussed in B.2. The matrices $\mathcal{M}$ and biases b are of the appropriate dimensions to map the input first to the hidden units and from there onto the three output logits.

FLEXIBLE PREVIOUS TRIAL PROCESSING: In the classical model, the encoding of the previous trial that enters the exponential filter over past trials considers only the choice (left, right, or timeout), ignoring the contrast strength and whether the choice was rewarded or not. While previous work established that mice do not use the feedback to infer and memorise the true stimulus side in the context of an exponential filter (Findling et al., 2023), it did not consider more complex or subtle kinds of interactions. Possibly, rewarded trials have a stronger effect upon future choices than unrewarded ones, or stronger contrasts might incorrectly be interpreted as stronger evidence towards block identity by the animal. We open up the space of effects of trial history upon memory by replacing this one-hot vector of choice with a neural network, the encoding network N^e. The input to this network can be anything that happened on the previous trial (choice $\mathbf{a}_{t-1}$, stimulus strength and side $\mathbf{c}_{t-1}$, reward r_{t-1}), and the output is a three-dimensional vector, just like it was for the classical model. The most comprehensive version of this network produces this trial encoding $\hat{\mathbf{e}}_t$:

$$\hat{\mathbf{e}}_{t-1} = N^e(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, r_{t-1}). \tag{5.7}$$

FLEXIBLE HISTORY PROCESSING: The encoding of the previous trial needs to be combined with the previous memory. In the classical exponential filter, the one-hot vectors decay by a constant proportion, meaning that on each trial, memory is multiplicatively weakened through the decay constant d. Here, again, we instead employ a neural network, the decay network N^d: The three-dimensional vector of memory over past trials is passed through this network to change over time. Here, we describe the simpler, "infinite" memory version, but we also experimented with a more interpretable finite memory, elaborated at the end of this section. This network thus simply maps three-dimensional vectors onto three-dimensional vectors, but instead of the constant rate of decay, it can vary the magnitude of decay and implement more complex patterns than the straightforward decay from one-hot towards zero. Thus, the full memory update using a neural network N^d takes this form:

$$\hat{\mathbf{h}}_t = N^d(\hat{\mathbf{h}}_{t-1}) + \hat{\mathbf{e}}_{t-1}. \tag{5.8}$$

The final choice probabilities are produced just as in eq. 5.3. The replacement of classical functions with neural networks is visualised in **Fig. 5.1** under the heading "Level 1". While the networks themselves can implement arbitrary functions (within the limits of the number of hidden units), the individual networks remain interpretable by always producing three-dimensional representations, which directly relate to the ultimate output probabilities. As we will see, however, interactions between these networks can create rather opaque dynamics, making it important for us to maintain overall interpretability.

On top of these extensions of previously existing components, we also found it necessary to add another separate network to achieve the highest possible predictive performance. This network runs in parallel to the previously discussed components, and provides crucial information to them, by succinctly summarising past information (i.e. excluding the current contrast) into single scalars. These then modulate the other networks by serving as an augmentation of their input. This network is a recurrent one (we use LSTMs (Hochreiter et al., 1997)), passing its own hidden state to itself on the next trial. Therefore, this network is much more powerful than the others and could, in the worst case, absorb the explanatory dynamics within its difficult-to-interpret hidden states. We limit the power of this network by allowing it to pass only a single

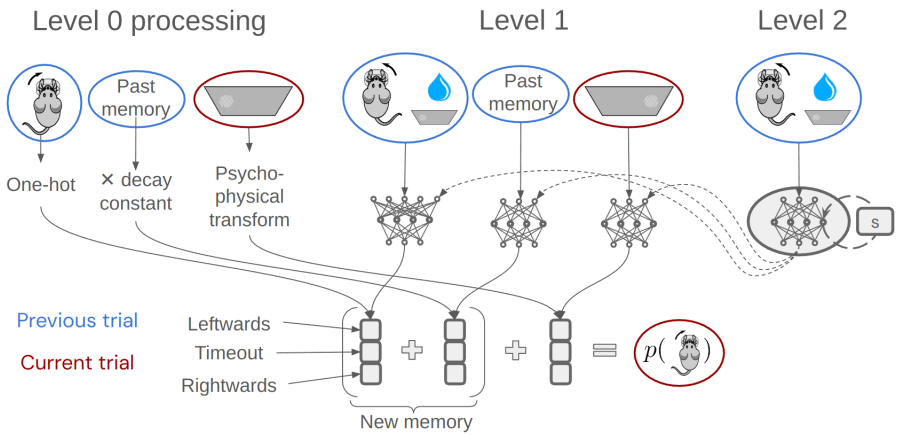

Figure 5.1: Visualisation of the different levels of replacing classical model components with neural networks. All versions ultimately produce an encoding of the most recent trial, an updated memory, and logits for the currently presented contrast (bottom). These can either be produced in the classical way (level 0), handled by neural networks (level 1), or handled by neural networks which also receive summarising input from a parallel LSTM (level 2).

scalar to the other networks, reducing its expressiveness, and we will later see that it does not take over the tasks of the MLPs.

The introduction of this modulatory network is necessary in the first place because we find that temporary fluctuations in the behaviour of the animal (presumably through factors like motivation) and individual differences across animals are substantial components of the explainable variance within behaviour. If we do not provide our architecture with an explicit mechanism to handle them, they get loaded into the other networks as much possible, requiring them to be needlessly complex and complicating the dynamics of their representations, as we will discuss further below. Thus, while this additional network provides an additional component to interpret and an additional input within other networks to make sense off, handling it in this way at least makes that component accessible, and as we will see leads to a highly interpretable modulation of other networks.

If we want to pass the summarising contrast scalar s_t^c and the summarising history scalar s_t^h to all networks, the equations become:

$$\hat{\mathbf{c}}_t = N^c(\mathbf{c}_t, s_t^c), \tag{5.9}$$

$$\hat{\mathbf{e}}_{t-1} = N^e(\mathbf{a}_{t-1}, \mathbf{c}_{t-1}, r_{t-1}, s_t^h), \tag{5.10}$$

$$\hat{\mathbf{h}}_t = N^d(\hat{\mathbf{h}}_{t-1}, s_t^h) + \hat{\mathbf{e}}_{t-1}. \tag{5.11}$$

The relevant scalars are simply appended to the other inputs, with the history scalar going to both the encode and decay network. The scalars are produced by separate learned linear readouts from the hidden state of an LSTM, which receives as its input the whole list of $\mathbf{a}_{t-1}, \mathbf{c}_{t-1}$, and r_{t-1}:

$$s_t^c = \mathcal{M}_c \mathbf{h}_t^{LSTM} + \mathbf{b}_c, \tag{5.12}$$

$$s_t^h = \mathcal{M}_h \mathbf{h}_t^{LSTM} + \mathbf{b}_h, \tag{5.13}$$

where $\mathbf{h}_t^{LSTM}$ is the hidden state of the LSTM. We work with a standard LSTM architecture. This additional network is visualised in **Fig. 5.1**, under the header "Level 2".

We presented the networks in their most general form, but there are many variations, depending on which parts of the classical model are replaced with networks and which networks get which additional input. We therefore introduce the following notation: each architecture is represented by a three-digit number, indicating which level of network usage is employed for (i) contrast processing, (ii) trial encoding, and (iii) memory decay. Thus, for example, 210 would denote a network which uses a neural network for processing the current contrast which also incorporates a summarising scalar from the LSTM (the '2'), while encoding the current trial through a network without such summarising input (the '1'), and simply decays the memory by multiplying it with a constant (the '0'). Architecture 000, which is the classical model, might be considered a pathological case of a neural network, as it only employs a linear transformation and an exponential filter, no non-linearities or hidden layers, but we will still refer to it as a network, as part of the hybrid network approach.

One subtle remaining distinction is memory length. The exponential filter of the classical model has an implicit infinite memory, as it is simple to just decay the sum of all previous choices via a multiplication with the decay constant and add the one-hot encoded newest action. In this way, the influence of past actions never fully vanishes, though of course it exponentially decreases to become effectively meaningless. This works since decaying the sum of memories is equivalent to decaying all components individually and then taking the sum. However, for more powerful decay functions, this will generally not be the case. We can sum up the memories regardless, as described in eq. 5.8, which allows for complex interactions between the memory traces of past trials, as the content of all memories is available while updating them. Even though the expressiveness of these memories is formally limited by the fact that they are used to produce choice probabilities, the three-dimensional logit representation is over-parameterised (we only need two numbers to specify three probabilities), giving the memory dynamics a degree of freedom which it, as we will see later, seems to exploit.

There are two ways to mitigate this. First, we can keep and decay the memories of past trials entirely separately, allowing for an easier interpretation, but requiring us to pick a bound upon the length of memory. We will at one point discuss a network which tracks only the last 20 trials, and denote this special restriction as XXX-20 as an addition to the normal network name. The memory update equation 5.8 thus becomes:

$$\hat{\mathbf{h}}_t = (\sum_{i=1}^{20} (N^d)^i(\hat{\mathbf{e}}_{t-1-i})) + \hat{\mathbf{e}}_{t-1}, \tag{5.14}$$

where the function N^d is applied i times to the original encoding of the trial i steps in the past, to produce the current state of that trials' memory. Only after this individualised decay are the trials summed up.

The other way of simplifying memory decay is by cutting out out the mediating N^d network, and instead use the LSTM output directly as a decay constant d (such as in eq. 5.2). This component is very easy to interpret, as the simple multiplication of past memories with the LSTM output is a complete description of what happens to the stored memories, and we can then track the extent of this decay on a trial-by-trial basis. A version of this will be our most predictive and interpretable architecture. Since multiplication distributes over addition, we can use an implicit infinite memory for such networks again.

We denote this variant as XXX-is, as they infer the decay scalar. The memory update equation 5.8 turns into:

$$\hat{\mathbf{h}}_t = s_t^h \cdot \hat{\mathbf{h}}_{t-1} + \hat{\mathbf{e}}_{t-1}. \tag{5.15}$$

While this setup can in theory be combined with an encoding network N^e, in practice we did not do this and simply used the one-hot encoding directly.

As a slight extension to this simplicity, we will also use a decay vector $\mathbf{s}_t^h$, which is multiplied elementwise with the history, to grant some more flexibility, while still maintaining interpretability. We call this XXX-iv, since it infers a decay vector (presented here in its most general form, though we will mostly use a simplified version which uses the direct action encoding, $\hat{\mathbf{e}}_{t-1} = \hat{\mathbf{a}}_{t-1}$):

$$\hat{\mathbf{h}}_t = \mathbf{s}_t^h \cdot \hat{\mathbf{h}}_{t-1} + \hat{\mathbf{e}}_{t-1}, \tag{5.16}$$

where $\cdot$ denotes a Hadamard product. The LSTM decay constant s_t^h or constants $\mathbf{s}_t^h$ come from a learned linear read-out of the LSTM hidden state again, but passed through a sigmoid function (possibly elementwise), to restrict the decay constant(s) between 0 and 1.

The effectiveness of this decay vector model also made us implement an architecture 100-sv, as a simple extension of the classical model. This combines simple contrast processing through a static neural network, but rather than using a single static decay constant, we use a static decay vector. Importantly, this vector is not always applied in the same way to the history memory, but in an action-dependent manner – it implements separate decay constants for memories relating to the just previously chosen action versus the unchosen action (as this turned out to be the more relevant distinction, as opposed to left versus right directly). Timeouts are handled separately, we thus update eq. 5.2 and use three separate decay vectors with the following symmetry:

$$\hat{\mathbf{h}}_t = \hat{\mathbf{h}}_{t-1} \times \mathbf{d}_t + \mathbf{a}_{t-1} \tag{5.17}$$

$$\text{with} \tag{5.18}$$

$$\mathbf{d}_t = (a, c, b) \qquad \text{if } a_{t-1} = (1,0,0), \tag{5.19}$$

$$\mathbf{d}_t = (d, e, f) \qquad \text{if } a_{t-1} = (0,1,0), \tag{5.20}$$

$$\mathbf{d}_t = (b, c, a) \qquad \text{if } a_{t-1} = (0,0,1). \tag{5.21}$$

Thus, a and b are the chosen and unchosen decay constant, respectively.

The highest rung on the hybrid network ladder is a completely unrestricted recurrent neural network (RNN), which consumes all input data at once and indiscriminately on every trial ($\mathbf{c}_t, \mathbf{a}_{t-1}$, and r_{t-1}) and directly produces output probabilities over the three choices. In particular, this network is able to implement arbitrary interactions between the input streams. We use an LSTM for this, but to differentiate it from the LSTM of the level 2 architectures, we will refer to it as the unrestricted RNN throughout.

With all formalities laid out, we can explore the space of models in terms of their performance and interpretations.

5.2 MODEL RESULTS – CLIMBING THE LADDER

We show the performance of the relevant networks in **Fig. 5.2** (the four dots are for four different random number seeds, where this matters). Two important landmarks are the completely classical model on the very left (000), establishing a baseline for performance, and the completely unrestricted RNN on the very right, marking the upper bound of explainable variance given our

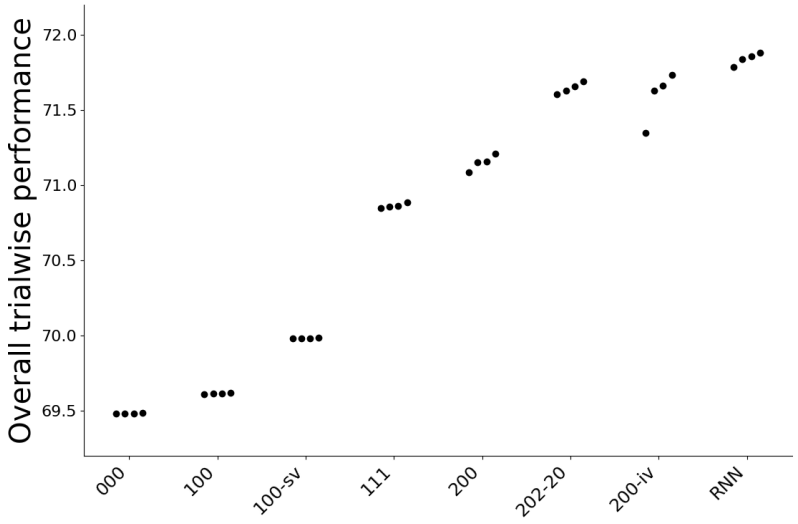

Figure 5.2: Model comparison of various relevant architectures, showing the trial-wise exponentiated log-likelihood on the validation set as percentages of each of four seeds with the same hyperparameters. Models are sorted by increasing flexibility and predictive capability from left to right. We see a notable jump in performance at architecture 111, which is when the models become capable of accounting for individual differences and motivational fluctuations (as made explicit in network 200 with similar performance), as discussed in the text. Network 200-iv achieves high performance while being interpretable.

data and training protocol. The goal of our other networks is therefore to obtain comparable performance through an interpretable architecture.

However, before we get there, we will briefly discuss some of the steps on the way to our best hybrid network, which will elucidate why the different components provide an improved fit.

5.2.1 *Network 100*

Network 100 simply replaces the psychophysical tanh-transform of contrast value with a neural network, enabling a bit of extra flexibility. The model comparison indicates that this, reassuringly for the standard model, provides only a tiny increase in performance. We compare the PMFs (i.e. the function implemented by the contrast network, in the case of 100) of the two models in logit space in **Fig. 5.3**. For this we compute the difference between the rightwards and leftwards logit for each contrast, as any constant offset applied to all logits will be removed by the softmax, meaning only the relative difference between logits is relevant (this hearkens back to the additional degree of freedom in our representation discussed earlier, an issue which will recur).

The contrast processing of both architectures is mostly in agreement, but we can see a difference in their encoding of 100% contrasts in particular. The more flexible neural network prefers logits which lead to slightly higher probabilities for the correct answer. The neural network can of course set the logits for each contrast effectively independently, while the psychophysical transform is tied to the typical contrast representation which employs the tanh-transformation (eq. 5.1). It seems this representation forces a trade-off upon network 000, somewhat more extreme values would be desirable for the the 100% contrasts, but this would come at the cost of having values which are too extreme for the

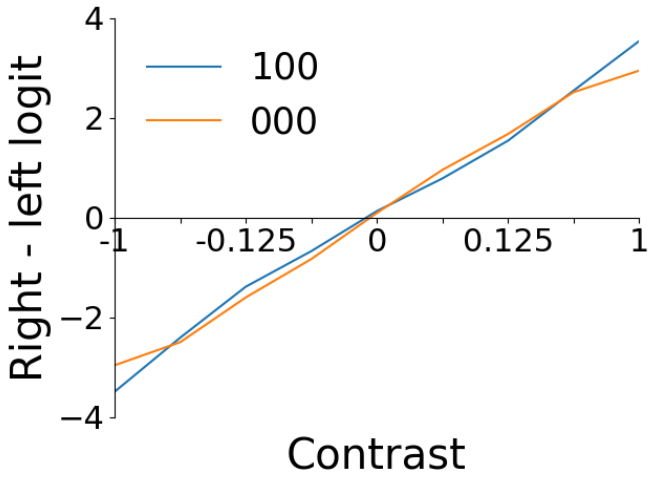

Figure 5.3: PMF, as the difference between the rightwards and leftwards logit for each contrast, of network 000 and network 100. The additional flexibility afforded by the neural network does not show substantial deviations for most contrasts. For the very strongest contrasts however, the neural network prefers slightly stronger responses. The psychophysical transform seems forced into a trade-off, the weaker contrasts take on every so slightly more extreme logits than the neural network assigns.

other non-zero contrasts. We can see that the logits for these contrasts are ever so slightly stronger than those of network 100. This might be related to the bias we noticed in the posterior predictive checks in our learning trajectory model in section 3.5. It is however important to keep in mind that the logits are exponentiated when turning into probabilities, meaning that small differences at larger values play a minor role, which is why the advantage is minor.

5.2.2 *Network 111*

The combination of all three types of static neural networks, architecture 111, achieves a notably better performance than the simplest networks. Unfortunately, even though it combines relatively simple networks, it turns out to be rather tricky to interpret.

To see this, consider **Fig. 5.4**, which depicts the function implemented by the encoding network N^e of 111. The encoding is clearly highly structured, with choice identity being the major distinction. Reassuringly, a leftwards choice results in an encoding which favours future leftwards choices (the leftwards logits are larger than the rightwards one), and vice versa for rightwards choices. The second important distinction is whether a choice was rewarded or not. Within these four groups the trials appear ordered: if the animal answers correctly for a weak contrast, that pushes the history component strongly towards the direction of that correct choice. Stronger contrasts have a progressively weaker effect. However, this order is exactly reversed for incorrect choices: an incorrect response on a strong contrast pushes the history component in the direction of that wrong choice even more strongly.

This hints at the problematic nature of these encodings: They seem to be more about inference of the motivational state of the animal, rather than reflecting true patterns of trial memorisation. A wrong response on a strong contrast is probably not a sign for the animal to repeat such choices, but rather

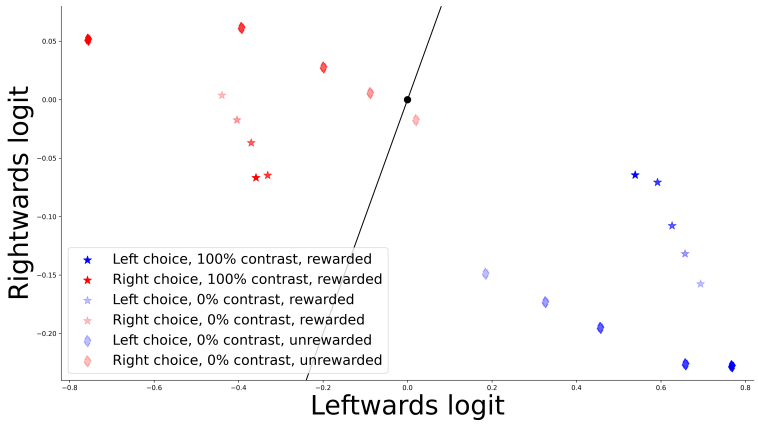

Figure 5.4: Representations of all trial types after processing with the encoding network of 111. We scatter the encodings according to their leftwards and rightwards choice logit (the first and third dimension of $\hat{\mathbf{e}}_t$), ignoring timeouts due to their small effect and to keep the plot 2D. The different trial types exhibit relevant patterns (ignoring the bias away from the identity line, as this was cancelled out by the contrast network). There are four major groups of trials, created by two separations: most saliently, the trials are split by choice, red versus blue. Secondly, trials are split by whether the animal was rewarded or not, stars versus diamonds. The ordering within these groups however calls into question whether these encodings represent the causal effect of the current trial upon future choices, or rather inferences about the motivational state of the animal.

it reveals that the animal is currently acting in a strongly biased manner and is likely to repeat such mistakes in the near future. It is hard to disentangle causality and correlation in this network and it did not provide a clear description of behaviour either, which is why we generally avoided the usage of an encoding network in our efforts for interpretability.

However, network 111 did reveal that there is a need for the networks to be able to account for motivational effects. Thus, rather than having them emerge in an uncontrolled manner, we made room for this process explicitly, allowing us to monitor it and its effects directly.

In addition to these difficulties, the additional degree of freedom afforded by the logit representation complicates things. Notice that the position of the encodings and the subsequent decay process (not shown) along the identity line are completely irrelevant for the ultimate probability output, as they are removed by the softmax, it is only the distance to the identity line which matters. However, this does leave the network room to encode something else along this dimension. One possibility is that the model uses this axis to store longer term information about the individual animal and its motivational state. However, we found this hard to prove, highlighting again the benefit of making information explicit, as in the level 2 networks.

5.2.3 *Network 200*

The introduction of a contrast scalar s_t^c, supplied to the contrast network by an LSTM which has processed all relevant bits of information from previous trials, imbues network 200 with an extra degree of adaptability, hinted at by the result of network 111. Up to this point, networks consisted of effectively static components which had to fit the entire population. This resulted in a

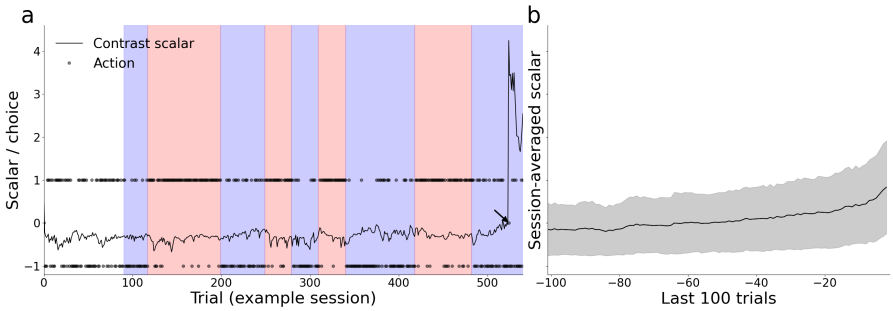

Figure 5.5: (a) Trajectory of the contrast scalar of network 200 during an example session. We also plot the true block in the background with red (rightwards block) and blue (leftwards), as well as the choices of the animal with circles (1 corresponds to a rightwards action, -1 leftwards, and 0 timeout). The contrast scalar evolves relatively smoothly, up to the point at which the animal lets a trial time out (see arrow). After observing this, the scalar drastically changes into a different regime. **(b)** Average contrast scalar on the last 100 trials of all sessions. The shaded region indicates one standard deviation around the mean. The scalar s^c exhibits a tendency to increase towards the end of a session.

PMF which averages over all individuals and trials, and a similarly averaged decay process (or emergent model capabilities, as in 111). The summarising LSTM opens the door to capture motivational effects as they become noticeable during the course of a session, and also to adapt to the idiosyncrasies of a given individual, if those are prevalent, from early in a session. For now, we restrict the deployment of these dynamics to the contrast network, where they are easiest to interpret.

It is necessary to be careful when introducing the influence of an LSTM into our architecture. A network with the power of a generic RNN can capture as much explainable variance as possible, which we see confirmed in **Fig. 5.2**. A danger, therefore, is that the LSTM might effectively short circuit our network. That is, the LSTM might handle inference in its entirety and agree upon a complex code with the contrast network that bypasses the need for the encoding and history networks altogether. In this case, the contrast network would unpack the information from the LSTM and combine it with the current contrast, which the LSTM does not see. This is a valid concern, but there are a couple of important considerations addressing it: (i) network 200 does not reach the performance of the unrestricted RNN, it is thus at least somewhat limited. (ii) As we will see, the workings of the contrast network when processing augmented contrast input appear to be interpretable, rather than implementing some obscure decoder, allowing us insights into the changing contrast sensitivity of the animals throughout the session. (iii) The contrast scalar maintains a relatively smooth trajectory (barring one specific type of trial). It thus seems to extract a slowly changing, relevant axis of behaviour, rather than jumping around wildly, fulfilling trial by trial demands of a hard to understand LSTM. In particular, the scalar (generally) seems hardly affected by the current block, which is doubtlessly one of the major causes of behavioural changes across a session (this is especially so for the subsequent networks with better history processing). In network 200, the local history is still handled by the one-hot encoding and exponential filter of this network however.

We visualise the trajectory of the contrast scalar during an example session in **Fig. 5.5a**. We can see that the contrast scalar fluctuates mostly smoothly, albeit jumping dramatically right after the first timeout response, towards the

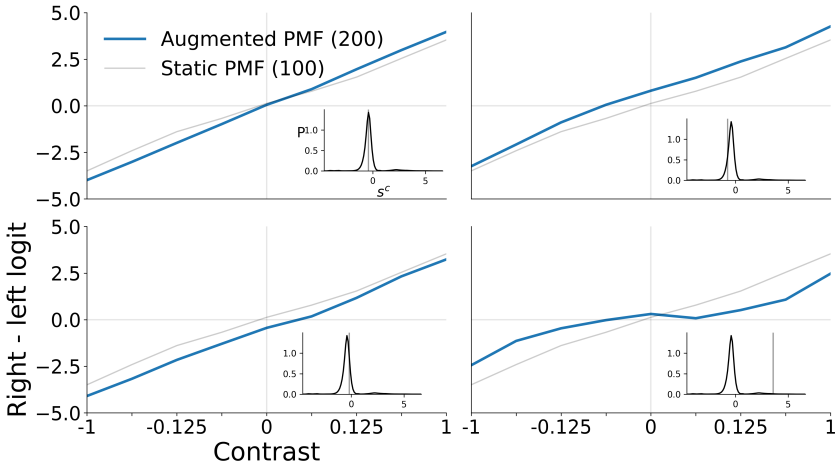

Figure 5.6: Output of the contrast network of 200, as a function of the contrast, for four different values of s^c. All plots show the PMF of network 200 as well as the static PMF of network 100 for comparison. The inset depicts the distribution over the contrast scalar (obtained via a Gaussian kernel density estimation), and the scalar value that was used for the given plot (vertical line). At the mode of the contrast scalar (top left), the network exhibits a slightly more steeply tuned PMF than the static network. Scalar values that are slightly less (top right) and more (bottom left) than the mode cause the PMF to take on a bias. In the long right tail of the contrast scalar distribution we find a PMF which is considerably less sensitive to the current contrast (bottom right).

end of the session. In fact, this is a general trend, as can be seen in the average over the last trials of all sessions which we show in **Fig. 5.5b**. The hidden block, on the other hand, does not seem to affect the scalar in a visually detectable manner, but we will analyse this in more detail in the final architecture.

The effects of the contrast scalar s^c in the contrast network are shown in **Fig. 5.6**, together with the overall distribution over the contrast scalar. To obtain this plot we chose representative points from the distribution of contrast scalar values and presented each contrast, with the chosen value concatenated, to the contrast network. This then led to the difference in logits that is shown. At the mode of the contrast scalar distribution, the PMF is unbiased and somewhat more sensitive to the current contrast than the static PMF of network 100. This static PMF of course has to accommodate the entirety of behaviour, and thus had to settle on a somewhat average PMF. The dynamic network reveals that during most trials, mice are actually more sensitive to the current contrast than the classical model indicates.

Network 200 does use s^c to impute biases into the PMF, as can be seen in the top right and lower left panel of **Fig. 5.6**. These depict values of s^c_t slightly to the left and right of the mode, and show a well tuned PMF with a right and left bias respectively. The presence of these biases indicates that some of the hidden block adaptation is off-loaded into the LSTM component (though an adaptive sensitivity to current task statistics might also be an explanation (Summerfield et al., 2015)). We do not find this in the later networks with more powerful history inference.

Finally, the strong jump in the contrast scalar following the timeout response is shown by the PMF which receives a contrast scalar from the long right tail of the distribution. The PMF becomes considerably flatter, i.e. less sensitive to the contrast, indicating that the animal has mostly disengaged from the task. Letting a timeout occur is sensibly a strong sign of this. Performing explicit

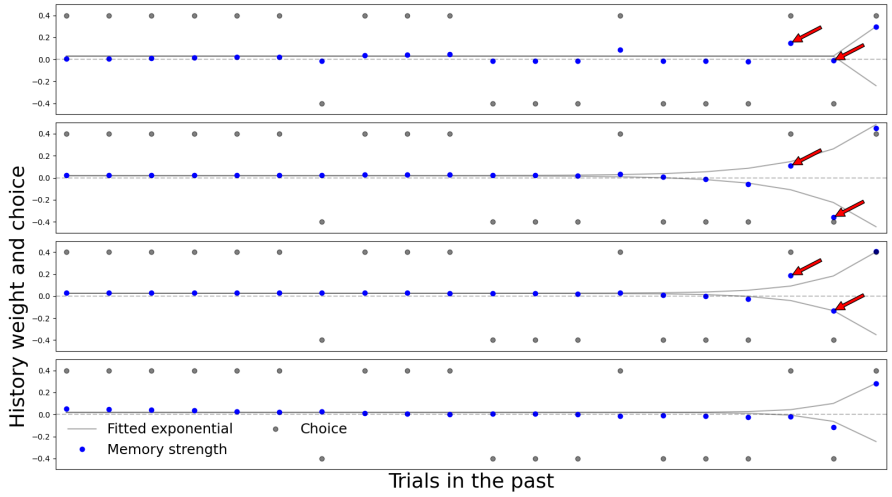

Figure 5.7: Visualisation of the effect of past trials on a specific choice, the four rows are the best 202-20 networks (one per seed, validation set performance from top to bottom: 71.69%, 71.66%, 71.63%, 71.61%). We show the choices on the previous 20 trials (the length of the history filter), with positive values encoding rightwards choices and negative values encoding leftwards choices. We also show the magnitude of the effect their memories have on the current choice, i.e. the difference between their rightwards and leftwards logit after all the steps of decay that they have gone through. Despite their differences in weighting, they make similar predictions on this trial, assigning these probabilities for a rightwards choice from top to bottom: 65.2%, 62%, 70%, 59%. The faint grey lines show the best fitting exponential to this memory strength (and a mirrored version for visualisation), with a fitted offset from 0. These networks implement a long-term bias through their offset from 0, but the information from individual trials decays quite rapidly. Interestingly, the most recent past trials exhibit strong deviations from the fitted exponential, and notably do so in a choice-dependent manner, as highlighted by the red arrows in the first three rows.

inference over the current state of the PMF thus allowed us to see what pure, engaged behaviour looks like, without it being influenced by later, disengaged trials.

5.2.4 *Network 202-20*

Network 202-20 represents another step up in performance. It combines flexible contrast sensitivity with a similarly flexible memory decay process. We avoid the usage of a sophisticated encoding network, due to the interpretative complications it introduces. Instead of focussing on the precise workings of this model, we will study how it implements the memory decay process specifically, extracting insights into its mechanism. This analysis leads us directly to our final model.

To this end, we show in **Fig. 5.7** how the memories of past trials affect the current choice, for the four seeds of the best 202-20 architecture. We picked an example trial which shows the important effect saliently. The weights of the memories reach their asymptote fairly quickly, anything more than four trials ago seems scarcely to matter. However, this is somewhat deceptive, as these history traces actually implement a notable bias, by not converging to 0, but rather a slight offset. Importantly for our next architecture, we see that the single fitted exponential does not capture well the effect of the

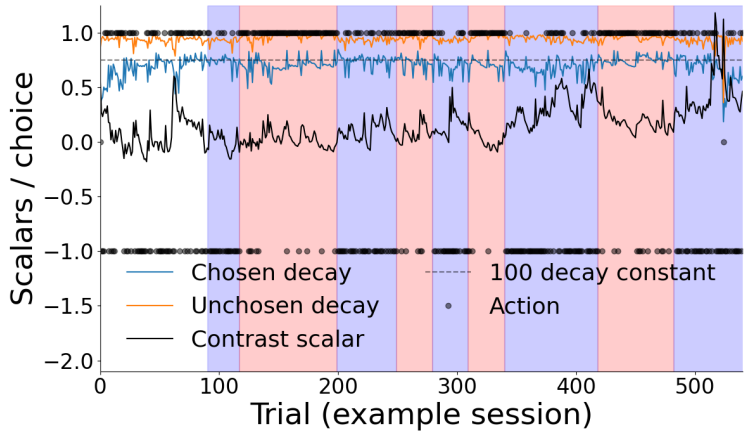

Figure 5.8: Trajectory of the contrast scalar and decay factors of the architecture 200-iv during the same example session as used in Fig. 5.5. We plot the true block in the background, as well as the choices of the animal. The contrast scalar evolves slowly over time, except for a big change in response to the first timeout trial. The decay constants for the chosen versus unchosen action are rather consistent across time, albeit sitting at quite different values for the two types of choices and with occasional single-trial deviations. The grey dotted line indicates the constant decay factor fitted in network 100. It is not clearly visible here, but the timeout response also causes a notable decrease in both decay constants, meaning that the strength of memory is reduced drastically on this trial.

most recent trials: Depending on the choice which was made on a past trial, their associated memory decays at different speeds, deviating from the overall exponential (marked by red arrows). The four different seeds do not implement a common solution to this though, as the deviations they exhibit from the fitted exponential differ.

5.2.5 *200-iv*

With the insights gained from the 202-20 network, we were able to craft a simpler model which achieves the same level of performance, if not even slightly better. This new model does not use an entire decay network, but again works with multiplicative decay, similar to the original classical model. However, it employs a three dimensional decay vector $\mathbf{s}_t^h$ that is the dynamic output of an LSTM. As discussed earlier, we implemented this as separate decay factors for the chosen and the unchosen action type (with timeouts being handled separately). We also applied this logic of action-dependent decay to the classical model directly, and fitted the architecture 100-sv, which uses a static action-dependent decay vector and performs notably better than the classical model.

The behaviour of the contrast scalar and the trial-by-trial decay constants are shown, for the same example session studied earlier, in **Fig. 5.8**. We can clearly see the distinct values the two decay constants take on, establishing that memories of trials in which the currently unchosen action was taken are barely decayed, whereas memories of the same action as the one that was chosen decayed considerably more. Of course, the history component of the current action is strengthened by the fact that this side was just chosen, counteracting the stronger decay of the previous same actions. We visualise this interplay for

an artificial action sequence using the static decay vector model in **Fig. B.1**. When comparing to the single decay constant used in network 100, we see that it employs a rate which is somewhat on the high side of the typical chosen action decay rate, but within its normal range.

We have thus arrived at an architecture which provides a highly accurate description of mouse behaviour on the given task, while maintaining (or actively establishing) a high degree of interpretability. Along the trajectory of building this model, we saw that mouse behaviour changes considerably during the course of a session, and the explicit handling of this allowed us to describe their sensitivity to the sensory aspects of the task during periods of engagement and disengagement much more accurately. Additionally, this latest architecture highlighted that the classical exponential decay process actually elides two separate decay processes, which differentially govern the decay of memories of the action which was just chosen and the one which was not. This opens up further avenues to pursue, but we will for now settle on this model and consider its overall performance in more detail.

5.3 MODEL ANALYSIS

With our winning hybrid network in hand, we can study its performance in comparison to the classical model. Of particular interest is the performance of the networks around the block switches. That the performance of the mice drops around the switch (**Fig. 5.9a**) shows that they are indeed tracking and using some information about the prevailing block. This characteristic of their behaviour is captured by the exponential filter component of the classical model. However, if the mice actually perform block tracking via a different computation (like the dual decay rate framework of 200-iv), this should be notable during these periods, i.e. where the block tracking component undergoes the fastest change.

Fig. 5.9b shows that the LSTM and network 200-iv predict behaviour better throughout a session, as can be seen by the difference in performance before the block switch, but they have an even bigger relative advantage right after the block switch, when the gap in performance grows considerably larger.

Even more insight into the improvements afforded by our architecture can be gleaned by studying which types of trials enjoy better predictions under 200-iv. To this end we consider the performance differences between the networks in relation to different task aspects. **Fig. 5.10** compares the performance of model 200-iv and model 100 on each contrast type individually, split by whether the animal chose correctly (upper plots) or not (lower plots). We note that, irrespective of contrast side, unrewarded trials with a weaker contrast generally contributed more to the difference in log-likelihood between the models (i.e. the % of log-likelihood difference between models is much larger than the % of trials of those contrasts). This does not however peak at 0% contrast, but instead is the most notable on the weak contrasts, 12.5% and 6.25%. This is due to the fact that 100 cannot assign high probabilities to an incorrect action.

One can think of 100 as a model which is tied to its PMF, which gets modified, to a degree, by the history component. However, this history component is somewhat limited in its magnitude, and cannot, no matter how overwhelming the evidence, predict a rightwards choice with any certainty for a weak contrast on the left. We therefore see a large gap of missing high probability predictions away from the current contrast. This even affects the correct choice on 0% contrasts, as the history component itself is unable to bias the predicted

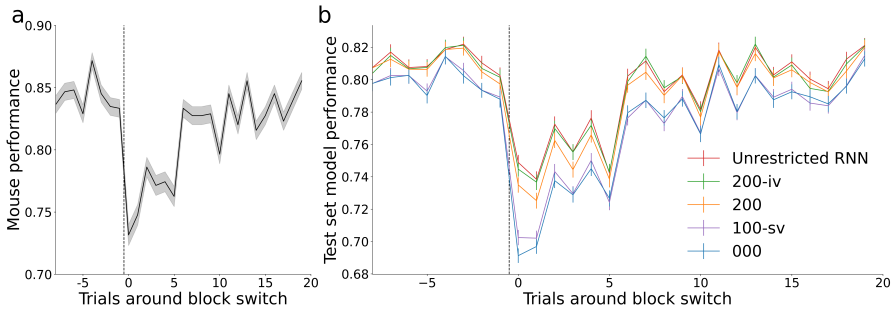

Figure 5.9: Mouse and model performance around the validation set block switches (see **Fig. B.2** for a version of this plot on 0% contrasts only). Error bars and shaded region indicate one standard error the mean. **(a)**: Mouse performance averaged around all block switches (left to right or right to left). We can see how the animals perform at a relatively stable level of around 85% accuracy towards the end of blocks. This drops right after the block switch to just over 70% accuracy, as the history information now points in the wrong direction, before recovering over the course of around 10 trials. **(b)**: We can see a similar time course of performance drop and recovery in the prediction quality of the models. Note that as the mice become noisier (the conflict between current contrast information and recent history leads to overall closer to 50:50 chance responses), the models also necessarily become worse, as the underlying process becomes harder to predict. However, the stronger models see a much less sharp decline in performance. The somewhat jagged evolution of performances is a result of the fact that the IBL only uses twelve different, fixed, schedules of contrast presentations as sessions, leading to insufficient variation.

probability strongly enough towards one choice or the other on such trials. The considerations of these extremes explains the gaps which are evident in 5.10 for network 100, but does not preclude subtler advantages for network 200-iv on other trials as well. By contrast, network 200-iv has the advantage of being able to tune its PMF, it can be both highly sensitive and highly insensitive, given the circumstances. Similarly, the history component reaches higher values if there is evidence for strongly persistent behaviour, allowing the model much more overall flexibility.

Adding to this, **Fig. B.3** depicts how performance differences spread out over the first, second, and last third of the sessions. We know that behaviour tends to degrade particularly towards the end of a session, thus this plot indicates the degree to which motivational effects might explain the performance differences. We can see that the last third of all sessions contributes the most to the difference between models, but there are sizable effects in the first and second third as well (performance can of course also occasionally drop at any point in a session, but does so most notably in the last third).

Next, we analyse in more detail how the LSTM inferred values, contrast scalar and decay vector, serve to implement a more faithful description of behaviour, starting with the contrast scalar. Similarly to what we observed for network 200, s^c serves to tune the sensitivity of the PMF. Unsurprisingly, the contrast scalar thus exhibits a clear relationship to the overall reward rate of a session, as shown in **Fig. 5.11a**. Sessions with higher reward rates have a distribution over s^c which is notably shifted to the right.

The combination of the LSTM chosen contrast scalar and the augmented PMF is well calibrated to actual behaviour, as visualised in **Fig. 5.11b**. Here, we split trials based on the overall distribution octile into which their associated s^c falls, and then compute the empirical PMF on those trials as well as the 200-iv

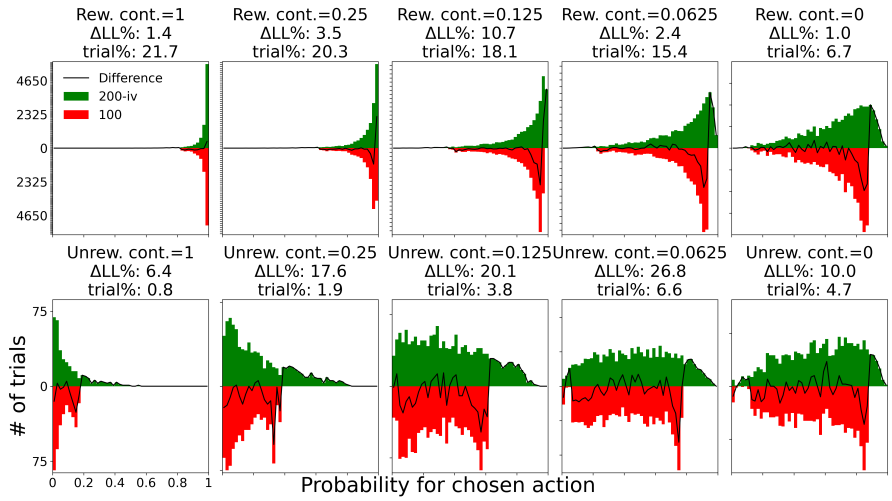

Figure 5.10: Histograms comparing the prediction quality on the validation data set of model 200-iv against 100. We split the trials based on absolute contrast strength and whether the choice was correct or not, and depict histograms over the probability that was assigned to the actually chosen action on that trial (thus, numbers closer to one are better), upwards in green for model 200-iv and downwards in red for model 100. We also plot a grey line which represents the difference between the histogram bars at each bin. The titles denote the trial subgroup, which percentage of log-likelihood difference this group accounts for, and the percentage of trials in this group. The comparison between the latter two numbers allows us to discern whether a type of trials contributes disproportionally to the log-likelihood difference. As we can see, wrong choices on weak contrasts contribute more to the log-likelihood difference overall.

predicted PMF (by averaging over the probabilities assigned to a rightwards action). For comparison, we also show the predicted PMF on these trials for network 100. While 200-iv is very well aligned with the empirically observed PMFs across the board, from relatively flatter to steeper PMFs, network 100 lacks the needed flexibility (the action perseveration mechanism proves insufficient here) and only performs appropriately for the central octiles. Thus, the contrast scalar reflects very real effects in the perceptual sensitivity of the animals, and relates directly to the overall quality of behaviour.

The decay factors exhibit a less clear connection to the reward rate, see **Fig. B.4** (though the distribution over unchosen action decays becomes more diffuse for worse sessions). To better understand their workings, we plotted their dynamics around the block switch in particular, see **Fig. B.5**. Intuitively, we might expect effects such as an increase in the decay rates right after a switch, as the mice notice that a change has occurred in the environment and so increase their rate of updating to accommodate it. However, this is not clearly the case. At most, the unchosen action decay on a left to right block switch shows a slight increase after a switch and the contrast scalar seems to react somewhat, but the effects are overall quite subtle. If the LSTM is implementing noteworthy dynamics around the block switch, it is not clearly identifiable at this level of detail.

5.3.1 *Individual differences versus trial differences*

Given the strong flexibility afforded to 200-iv by the inferred scalars, this raises the question to which extent they are used to infer and adapt to individual

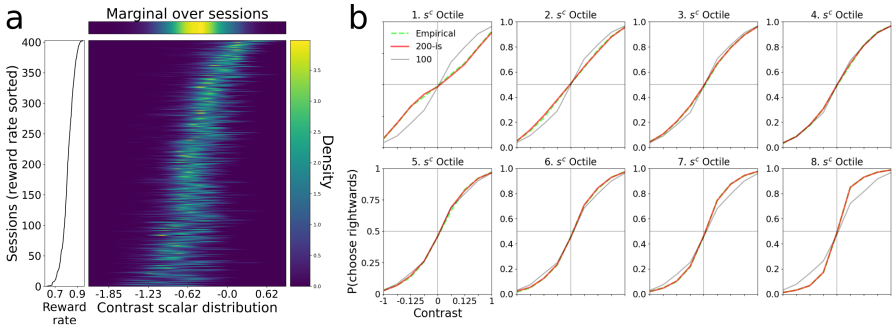

Figure 5.11: (a) Relationship between s^c of 200-iv and the reward rate across sessions. The heatmap shows the distribution over all 403 training sessions, sorted by their reward rate, decreasing from top to bottom (plot on the left depicts the reward rate). A generally higher contrast scalar clearly goes along with a higher reward rate. **(b)** Calibration of the predicted PMF of 200-iv and 100 versus the empirical PMF, across an octile split of the trials based on s^c. The effect of the scalar upon 200-iv is obviously very well calibrated, as reflected by the close agreement between the predicted and empirical PMF. Naturally, 100 is much less flexible (though it does have some adaptability due to the perseveration component), and thus does not accurately reflect behaviour for the more extreme octiles. Also note that there is a very slight leftwards bias on 0% contrasts for some of the octiles (1., 5., 6., and 7.).

idiosyncrasies in task strategy across sessions versus the better tracking of fluctuations of attention within sessions. This is a tricky question, as these two sources of model advantages are entangled, the difference between a motivational fluctuation and a different strategy is not entirely clear. We nevertheless try to address this question by considering model performance on the most prevalent session type (thereby at least holding the task component of behaviour fixed) in **Fig. 5.12**. In this figure, we also separately average across trials and sessions, to get a clearer picture of the trial-wise versus session-wise model improvements. We can see that both 200-iv and 100 make generally better predictions on sessions with better animal performance, however 200-iv benefits more from this performance increase. In the lower half of sessions in terms of performance, we can see a number of sessions which are equally well described by both models. At least for the first 400 trials thus, some animals are captured very well by the classical exponential filter.

Trials on which the population performs better are better predicted by both models too, though there is no dissociation between the models here, 200-iv gains its advantage over the entire range of trials via many incremental improvements. This does not settle the question of within and across session improvements, but this decomposition of model advantages at least gives us a clearer picture of where to search for further insights on this difficult question.

For a cursory impression of the consistency of individuals in terms of their inferred LSTM scalars, we consider a small set of animals with a large number of sessions in the training set (at least 7), visualising where their mean inferred scalars fall within the population distribution, see **Fig. 5.13**. This depicts only a small sample size, but within these animals we can see that the different scalars generally span a considerable range of the overall distribution (though this is of course confounded with the different session types). Nevertheless, some animals exhibit tendencies. Most notably, the animal marked by blue stars tends to have higher contrast scalars and lower chosen decay factors (almost creating the lower tail of that distribution entirely by itself). From this

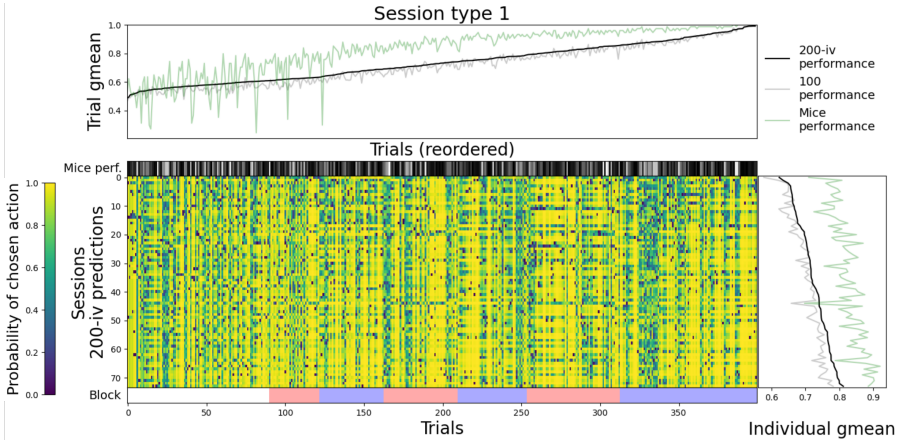

Figure 5.12: Performance advantage of 200-iv split over trials and individuals, for the first 400 trials of the most prevalent session type (74 sessions in the training set). The central heatmap visualises the probability which 200-iv assigned to the actions chosen by the animals. Sessions are sorted by the mean predictive quality of the model over those 400 trials (using the geometric mean), with the best predicted session at the bottom. The additional color line at the bottom shows the block on each trial and the color line at the top indicates the trial-wise averaged performance of the mice (with darker colours being closer to 1). The lineplots at the sides average over the trials (top, trials are resorted by their predictive quality) and over individuals (right, same sorting as heatmap). The lineplots also show the predictive quality of network 100 and the reward rate of the mice. Sessions with better mouse performance are generally better predicted by both models, but 200-iv gains a particular advantage on the best sessions. More rewarding trials are also predicted better, though there are some notable outliers on which the mice perform poorly, but the models predict this.

plot we would conclude that animals exhibit only a loose consistency in terms of their inferred LSTM scalars across sessions.

5.4 DISCUSSION

Using the hybrid network approach, we spanned the range from classical and directly interpretable models to maximally predictive, but opaque, neural networks. By purposefully designing the intermediate steps (Eckstein et al., 2024), we arrived at an architecture which achieves close to maximal performance, while being interpretable in most aspects. We were additionally able to distill a particular insight, the presence of separate decay factors for chosen versus unchosen actions, into a simple extension of the single-constant exponential filter model. This straightforward extension provided better fits and further insight into the dynamics of perseveration computations, a general feature of animal behaviour (Fründ et al., 2014; Gershman, 2020a), which mice have shown to be able to tune adaptively (Fritsche et al., 2024).

We also found pervasive effects of the level of perceptual sensitivity. Our inferred scalar s^c was crucial for modulating the PMF used by our final architecture, highlighting the stark differences between high and low acuity, which simpler models have to average over. This connects to the literature on motivational fluctuations, as discussed in Wichmann et al. (2001) and Ashwood et al. (2022).

The differences in predictive performance between baseline 000 and the unrestricted RNN may seem somewhat modest, we see a 2.5% percentage

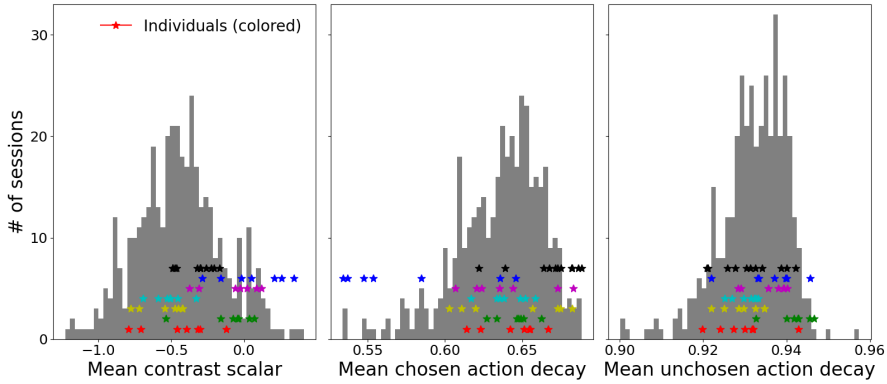

Figure 5.13: Histograms over the means of the different LSTM inferred scalars (s^c, chosen decay, and unchosen decay) over the first 400 trials of each session. Overlaid with different coloured markers are the mean values of seven individuals with a large number of sessions. The values of an individual span a considerable portion of the population distribution, though there are also some broad trends in some individuals.

point increase, from $\sim$ 69.5% to $\sim$ 72% (the average trial-wise increase per validation session between 000 and 200-iv was 2.34 percentage points, with a minimum of -1.0 and a maximum of 9.89, only two sessions were predicted worse by 200-iv). However, this is not an unsubstantial improvement. For comparison, an extremely simple model which only uses the current contrast (i.e. makes its predictions based on the empirical distribution over leftwards, rightwards, and timeout choices in the training set, split for each contrast) already reaches an evaluation performance of 65%. Thus, the task-essential exponential filter also only affords a 4.5 percentage point increase, and further improvements are ever harder to come by. We thus think that relevant and important mechanisms may well be hiding in the 2.5 percentage point increase we see, such as the more sophisticated memory decay mechanism we uncovered, and therefore view a thorough investigation of these improvement as a fruitful endeavour.

There is still more to explore within our modelling framework. Ideally we would like a better understanding of how exactly the LSTM arrives at its contrast scalar and decay vector. As it is set up currently, the LSTM in way serves as a descriptive model within a mechanistic model: It provides the relevant contrast and memory updating context, without however granting access to the mechanisms which make them come about. One option may be to attempt to decompose this network again, applying the hybrid network approach to this smaller network. One prevalent component within this LSTM is presumably the tracking of the overall quality of behaviour, which could be relatively easily replaced with a local estimate of the recent reward rate. In future work, we will further simplify the workings of this summarising LSTM.

Better understanding this component might also help us assess how much of the advantage of our best model comes from the capture of motivational effects, inter-individual differences, or a more faithful implementation of the population-wide cognitive processes. Full RNNs are certainly capable of inferring individual idiosyncrasies while predicting, however what exactly they extract is not accessible. Work by Dezfouli et al. (2019a) elegantly reveals these components, by training an encoder to represent individuals within a low-dimensional and disentangled space. A similar approach might be useful for us to separate out individual differences versus session-specific motivational

fluctuations. While we have begun to tackle these questions on within- and across-session differences, there are still many aspects left open for further exploration. When looking at the trajectory of the individual decay components in **Fig. 5.8**, it is evident that there are notable patterns across trials. Insights into what causes these deviations might allow us to further tune our simple models and improve our understanding of this latent state adaptation process.

In terms of more general insights for behavioural modelling, we found the usage of an unrestricted RNN to gauge the predictive ceiling of our data most helpful in assessing our modelling progress and consider it an essential tool for the field. While the hybrid network approach was most useful in generating insights and improving models, it was far from an automatic process in our hands: Designing the intermediate models required deliberation and we ended up with a network which is handcrafted to our particular task. We also tried out other architectures, which did not end up relevant to achieving our goals. Thus, the problem of model discovery remains a difficult one, though the design of hybrid networks presents a workable approach of broad applicability to tackling it.

Part VI

DISCUSSION

6 DISCUSSION

SUMMARY: In this thesis, we provided two different approaches towards a more holistic modelling of animal behaviour. With our first model, we extended a state-based descriptive framework to enable it to capture the vicissitudes of the acquisition of task competence. In particular, we introduced the capacity to describe fast and slow changes in behaviour, to capture both prevalent modes of progression (as well as possible motivational effects). We also allowed for individualised modelling thanks to inference about inherent complexity that is performed in the fitting procedure. We then applied the model to the session data of individual animals. This allowed us to capture rather comprehensively details of the learning trajectories of a large number of animals on a perceptual decision-making task, extracting high-level commonalities, such as intermediate learning stages, as well quantifying the considerable individual differences. The automatic discovery of intermediate strategies, without having to pre-specify possible expressions, is a particularly promising aspect of our work and we are excited to see results on this direction in the future.

The second approach employed neural networks to lift in a systematic manner the restrictions imposed by the simple and interpretable equation approach of classical modelling (Eckstein et al., 2024). This is in line with the ideas presented in Agrawal et al. (2020) for working with large data, in that we aim to reach a level of predictive performance with simple models which we know to be achievable through the performance of powerful but opaque methods. We used this to search for unaccounted sources of variance in the behaviour of expert mice on a later stage of the same perceptual decision-making task. We extended the class of expressible functions, analysing their workings with an eye towards interpretable components, and then distilled our insights into a simpler architecture. Through this, we were able to find a meaningful extension of a classical model with significantly better performance on our data. Additionally, in a more elaborate architecture, we were able to capture the effects of high and low motivation on the processing of stimulus information. With both of these presented approaches, we were able to extend the range of addressable components of behaviour.

MECHANISTIC AND DESCRIPTIVE MODELS: Both of our models combine mechanistic and descriptive aspects and it is a promising direction to try to extend what is captured in a mechanistic manner. The diHMM formalises the effect of the contrast upon the current choice in what qualifies as a mechanistic way, and it also does so for the effect of previous choices (through its weight for the perseveration feature). On the other hand, state changes in the diHMM, be they motivational- or learning-based, are entirely without mechanism, and since learning was the focal point of this model, it is largely a descriptive one. To make this component mechanistic, we would have to formalise the process of state change: when do they occur and what causes them, e.g., why and when does the animal disengage from the task or have insights into it. Specifically for motivational effects, there is recent work by Mohammadi et al. (2024) on formalising the mechanism behind state changes. For the slow process of weight changes within states, we would have to implement a gradual process for changing weights, a temporal-difference learning rule might well

be appropriate here (Schultz et al., 1997), though the extent of weight updates might not be uniform across time, requiring a more sophisticated treatment (for example, something similar to Piray et al. (2020)).

Similarly, the hybrid network 200-iv provides a formal mechanism for the effect of the current contrast. It also grants insights into the mechanism of the influence of previous actions (through the multiplicative decay of memory), but partly remains at a descriptive level, in that the decay factors are produced by an LSTM through an opaque process. It is an explicit goal of our approach to make potentially mechanistic some of the insights we gain from this rather descriptive LSTM, and we have already had success with that in the form of model 100-sv. However, some descriptive aspects might be irreducible: it is possible that individual differences are present without an apparent reason, at least given the limited behavioural data we have studied here. Data on the developmental history of an animal or its current social situation might, in theory, afford a more thorough mechanisation (Dingemanse et al., 2010), or we could use the description for an animal on this task to make predictions for the same animal on a different task, to test its reliability. Similarly, while motivation may have some normative or otherwise predictable components, there may also be factors at play in its expression which are beyond our ability to capture. Again, more data, like e.g. information on how much water an animal consumed in the period before the task, could be helpful in finding reasons for the when of motivational changes.

TASK ACQUISITION: There are many interesting phenomena during learning which we were unable to explore more deeply in this work. One particularly intriguing aspect is the presence of behavioural regressions, when an animal uses a state which performs worse than the best state it has previously exhibited, for a considerable fraction of a session. Such regressions were a frequent occurrence. In the animal we showcased in **Fig. 4.1**, we saliently see the animal persist for a long time in state 3, which is worse than the parallel state 4, and even appears quite late on session 11. This may relate to the work of Kuchibhotla et al. (2019), who showed that animals can possess a better latent task understanding than they express during standard, reinforced trials (as revealed by unrewarded 'probe trials'). They speculated that this worse performance may be the result of mal-adaptive impulsivity on rewarded trials (they showed faster reaction times) or a reduced need for exploration on anyway unrewarded trials. We did not analyse these behavioural regressions in detail, as it was difficult to quantify them in a satisfactory manner, but the study of what causes them (such as, for example, an unusually long time between training sessions) and their consequences for learning seem a most appealing target for future research.

The presence of intermediate stages of proficiency is another highly interesting aspect of learning. In our specific task, this took the shape of uninformed behaviour at first, to a one-sided consideration of the stimulus presenting screen next, to finally full stimulus responsiveness. The stage-wise progression through a lengthy process is well documented in developmental psychology (Piaget, 1952) and has also been observed in the much shorter acquisitions of behavioural experiments (Dekker et al., 2022). However, what qualifies as an intermediate stage is a subtle question. It should be a distinct pattern of behaviour which is in some way persistent for an extended period of time. But would we preclude any amount of gradual improvement during this period? Does the transition between stages need to occur in a sudden manner (Epstein et al., 1984), or can it include continuous changes? And are we only consider-

ing behavioural expression for these distinctions, or also the representational substrate, as these can show distinct dynamics (Löwe et al., 2024)?

While this ultimately depends on the research question, further insights into the process of learning would elucidate the relevant questions to ask. The type of learning we consider here will in large part be driven by reinforcement (Sutton et al., 2018), to learn both the values of actions, but also the relevant sensory aspects and structure of the world (Gershman et al., 2010b). This has been analysed behaviourally in combination with neural data by Liebana Garcia et al. (2023), who study learning on a very similar task. They too find intermediate one-sided stages, and relate dopamine activity in the dorsolateral striatum to these biases. They further formalise learning in a deep linear neural network model, which forms predictions and computes multiple types of prediction errors (inspired by their observations on dopamine activity) and is able to capture the different types of individual learning trajectories. Through this elegant analysis, they find that there are a number of meta-stable points in weight-space, which allow behaviour to remain stable for a while, but also suggest the dynamics and sequence of the transitions between them. This provides a precise answer for the nature of intermediate stages in such a task from a mechanistic perspective, as distinct locations in the weight space of the policy and the representations of task elements, validating our model for finding them descriptively. Notably however, we found that earlier biases are generally not predictive of later ones. Future research will have to investigate whether this is a consequence of task or modelling details.

We already mentioned the somewhat puzzling behavioural regressions, the learning perspective may offer another explanation: These regressions towards a worse, possibly previously exhibited, behaviour, may be an attempt at using a simpler strategy again. It is sensible for the animal to trade-off the gained reward against the costs a policy imposes, thus it makes sense to try whether a simpler strategy will do (Piray et al., 2021). A particularly interesting consideration in this light is the presence of perseverative tendencies in animal behaviour (Gershman, 2020a). In **Fig. 4.7**, we saw the perseverative weight of our behavioural states remains remarkably consistent across learning, though its relative impact waned. Of course, as discussed, perseveration is a useful strategy component once the task includes the biased blocks, which are introduced in the phase of training right after the learning we analysed. This raises the interesting question of how the perseverative weight (or its mechanism more generally) changes with the onset of biased training, as it opens up a valid alley for a regression towards more perseverative behaviour.

Given the insights from our hybrid network modelling, it would be worthwhile to test whether the chosen-unchosen decay vector mechanism is also appropriate during learning, or whether it is part of an adaptation to the block structure specifically. Similarly, fitting the hybrid neural networks to entire learning trajectories of animals may be insightful as well. While the diHMM appropriately handled individual differences, its flexibility also made comparisons across the population more difficult. Casting general behavioural trends during learning into the common language of an LSTM scalar, as we have done for expert behaviour, may be helpful in uncovering such trends after all.

SHAPING: Apart from general insight into learning dynamics, our characterisation could help with the fine-tuning of shaping protocols to facilitate task acquisition. For instance, we observed that the animals initially needed to learn to connect the stimulus side information with their behaviour. There-

fore, if quickly reaching expert behaviour is the only goal, it would seem ideal to make this association as salient as possible, for instance by making it dynamic ('wiggling') in early trials. Conversely, having the more difficult contrasts present from the start might make the task much more difficult to learn, as the connection between actions and feedback would seem more probabilistic to the animals. Further, increasing the salience between a successful trial, the animal's movement, and the relevant stimulus aspect, might lead to an accelerated progression from type 1 to 2 and 2 to 3. We also saw that the last stage of training takes the longest time, in which the animal brings its performances to a consistent enough level to pass this stage of training. We found that the biggest contributor to this, i.e. the aspect of behaviour which held them back the longest (for the IBL-specific requirements), was insufficient consistency on easy contrasts. To accelerate this process it might be necessary to manipulate the reward rate, as the animals get a reasonably high reward rate without being too consistent. It might therefore be worth trying to only reward the animal after two correct responses in a row by the time it gets to stage 3. An additional possibility afforded by our model is a personalised shaping procedure: since the model describes behavioural states on session-by-session and trial-by-trial basis, we always know what the animal has acquired so far (at least in so far as this is expressed in behaviour) and what it currently struggles with, allowing for a dynamic adaptation of task details. To realise this possibility it would however be necessary to study how well the model works on incomplete learning trajectories, which we have not yet done.

NEURAL IMPLICATIONS: An important step, which we left for future research, will be to assess the relevance of the uncovered states for neural activity. While neural activity is not recorded during learning in most mice, one project within the IBL did perform recordings using fiber photometry during learning (Pan-Vazquez et al., 2024). This work provided insights into how dopaminergic activity in the striatum predicted the future learning trajectory, as responses to specific visual stimuli preceded learning. Other members of the IBL are currently using a wide-field imaging approach to collect the activity of a large number of neurons, also during the learning period. It will thus be instructive to relate the signals from these different modalities with our behavioural states, and see in how far they can be used to segment neural activity. We would certainly expect a strong influence of the different state types upon neural activity, as we think of them as directly relating to what the animal attends to: In a type 1 state, we expect a much more limited response to the onset of visual stimuli as compared to a type 3 state. Similarly, we expect the response of a type 2 state to be one-sided and in agreement with its expressed bias. Relatedly, the expectations over future rewards in response to a visual stimulus (or the trial onset tone) should be quite different across state types (as was seen for comparable intermediate stages in Liebana Garcia et al. (2023)).

The separability of different states beyond these broad distinctions will be most interesting, and critical, to assess the relevance of the states for neural analyses. One difficulty in this may be that it is not directly evident how the states relate to neural data: in the simplest case, neural activity could be different across states in some way throughout the states' presence. But it might also be the case that neural data mostly reflects the switches between states, or even mostly the introduction of a new state (which might be difficult to disentangle from other factors at session onset, which is when most states are introduced). The search for a relationship between states and neural activity

may thus not be a simple one. If successful however, the state estimates would provide a valuable regressor to inform neural data analyses.

LIMITATIONS: While we addressed a number of modelling limitations with our diHMM framework, there is notable space for further improvement. We already briefly discussed that dynamic duration distributions, a dynamic transition matrix and particularly a dynamic initial state distribution could be helpful. The latter would particularly fix the issue of states not being localised in time across sessions which we encountered in our posterior predictive checks. Progress in this direction has been made by Cuturela et al. (2024).

The hybrid network approach has, by design, no issues with limitations upon its expressiveness. However, it does have to contend with limitations of interpretability. In future work, we will further analyse the workings of the LSTM of architecture 200-iv, which provides the crucial scalars as context for the other networks. The aim is to understand the different sources which contribute to these scalars, such as inter-individual differences, motivational effects, and deviations from the mechanistic model implemented by the history decay vector component. We will attempt to achieve this by simplifying the LSTM as much as possible and thereby formalising any insights. As mentioned previously, capturing individual differences in an explicit encoding, like in the work of Dezfouli et al. (2019a), may also help us disentangle these different sources of variability.

As it stands currently, the parallel LSTM of 200-iv may well tailor its outputs to individual particularities, as they become evident during a session. This makes the description of individuals more complex (as compared to, for example, the from the start individualised description of the diHMM), as the LSTM becomes increasingly informed about the individual throughout a session. Such capacity is a general feature of sufficiently powerful recurrent or auto-regressive models (Binz et al., 2024). Individual differences may also express themselves only transiently: an animal may be more prone to become disengaged, or fall into an idiosyncratic pattern while doing so, but behave typically otherwise. A thorough disambiguation of individual differences and motivational effects thus represents an involved question for the future, though the automatic complexity inference provided by our diHMM approach represents one approach to tackle it.

Another issue for the hybrid network framework is data set size: We have not formally analysed how our networks cope with less data, but this method generally relies on an expansive amount of data. This made the IBL data set attractive for this sort of modelling in the first place. Especially the summarising LSTM probably relies on diverse behaviour to gain its capabilities. This is in contrast to our diHMM, which can be fit to individuals independently and deals well with a large range of training lengths. For that model, we only used the large number of animals to draw more meaningful conclusions about the population, not because the model required it.

MOTIVATIONAL EFFECTS: Motivational effects, particularly in the form of a reduced level of accuracy, were a prominent feature during both periods of the IBL task that we analysed. However, we captured them in two separate ways: as discrete states during learning (similar to Ashwood et al. (2022)) and with a continuous scalar during expert behaviour (close in spirit to Roy et al. (2021)). This raises the question of whether motivational effects fall on a continuous spectrum or can be discretised. In our modelling efforts on learning, we discussed that behaviour tends to degrade over time, which can

cause issues for our model (and lead to accuracy overestimation in posterior predictive checks). At the same time, on a specific session, see **Fig. 4.2**, we found disengagement with relatively sharp temporal boundaries, which was captured appropriately by a single state. Expert behaviour also degraded towards the ends of sessions, though as we saw in **Fig. 5.5**, at least sometimes the model captures this as a rather drastic, single-trial change in its workings (which subsequently continued changing more gradually). We conclude that motivation can evolve in both ways: it exhibits dramatic changes from one trial to another, has periods of relative stability, but there are also other times when it changes consistently over an extended period, such that discrete states are not ideally suited to capture it.

Taking a slightly broader perspective, there is a clear place for discrete states in the description of behaviour (Dayan, 2012). This can take the shape of motivational states with differences in the rewards they value, resulting in stark differences in behaviour across such states (Loewenstein et al., 2004). Even during a single behaviour such as foraging, there is evidence for alternation between distinct phases of exploration and exploitation, with different behavioural patterns and neural activity (Marques et al., 2020). The dorsal raphe nucleus seems to play a unique role for these kinds of states, differing across states (Marques et al., 2020) and signalling switches (Priestley et al., 2025), making it a prime candidate to search for the neural component of our uncovered states.

CONCLUSION: We hope that the suite of modelling approaches we have provided here, either in their entirety, their components, or just in their aspirations, can help researchers achieve a more comprehensive modelling of their studied behaviour. In so doing, we will be able to get a clearer picture of the underlying cognitive processes and further our understanding of natural intelligence.

Part VII

SUPPLEMENTAL MATERIAL FOR DIHMM BEHAVIOURAL FITS

A

SUPPLEMENTAL MATERIAL FOR DIHMM BEHAVIOURAL FITS

A.1 PSYCHOMETRIC TYPE CLASSIFICATION

We observed by eye that the psychometric functions (PMFs) that the model found for the behavioural states had a tendency to fall into one of three characteristic classes: flat (type 1), half-tuned (type 2), and fully-tuned (type 3). However, the boundaries between the classes were blurry, so we sought an objective distinction, recognizing its inevitable arbitrariness.

The measure we used in the main paper is the mean reward rate implied by the PMF on easy trials (100% and 50%), ignoring the effects of perseveration (and the debiasing protocol). We chose the reward rate, since this tends to grow as the animals proceed from ignorance to competence. We chose to assess only the easy trials since early PMFs were not defined on the lower contrasts (since these stimuli were not presented) and including more difficult contrasts can lead to lower reward rates for more broadly defined PMFs even though they are better on easy contrasts. **Fig. A.1** shows the distribution of such reward rates across all states. It is apparent that there is a rather clear grouping of PMFs with reward rates below 0.6, defining type 1. The boundary between types 2 and 3 is somewhat less evident, implying that edge cases will be hard to assign. The threshold reward rate of 0.78 served reasonably, as evidenced in **Fig. 4.4**. We further split type 2 PMFs on whether they were left-biased, right-biased, or symmetric. We considered a PMF symmetric if its error rate on 100% leftwards contrasts was within 10 percentage points of the error rate on 100% rightwards contrasts.

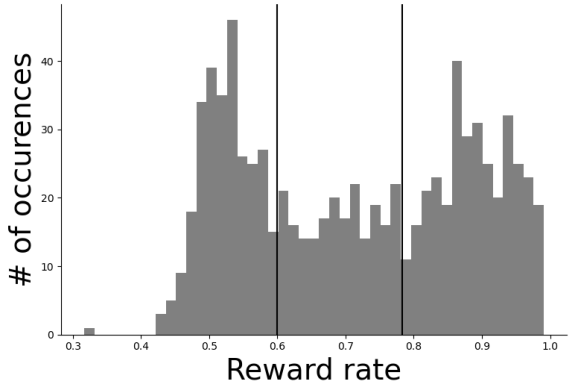

Figure A.1: Histogram of the mean reward rates on easy trials of the PMFs of all states, at the moment they first appeared. Vertical lines indicate the boundaries we used to classify states into the three types. The boundaries we drew do align with points of low density in the histogram.

A.2 BIAS TRAINING ANALYSIS

As elaborated, the learning stages exhibited by the animals seem rather independent, without strong patterns across the stages. This becomes even clearer when considering the next phase of learning, involving biased training: The basic task stayed the same, but instead of contrasts appearing equiprobably left or right, there were now unsignalled alternations between blocks, lasting 20-100 trials following a truncated exponential distribution, during which a contrast was 80% likely to appear on one side versus 20% on the other. This was of course particularly helpful for 0% contrasts, on which an animal could now reach a much higher reward rate than chance, given a suitable block inference mechanism (a detailed analysis of their actual algorithm was performed in Findling et al. (2023)). To finish this part of training, mice had to exhibit behaviour that was sufficiently modulated by the current block. The number of sessions it took them to achieve this was not correlated with the time spent in stage 1 or 2, although there was a slight negative correlation with the number of sessions in stage 3 (Correlation to stage 1 duration: Pearson's r=-0.06, p=0.47; to stage 2 duration: Pearson's r=-0.005, p=0.96; to stage 3 duration: Pearson's r=-0.2, p=0.02). This suggests that learning about the biased blocks tapped into yet another type of skill, but extensive pre-training with the full contrast set gave some mice a slight edge here.

A.3 SUPPLEMENTAL FIGURES

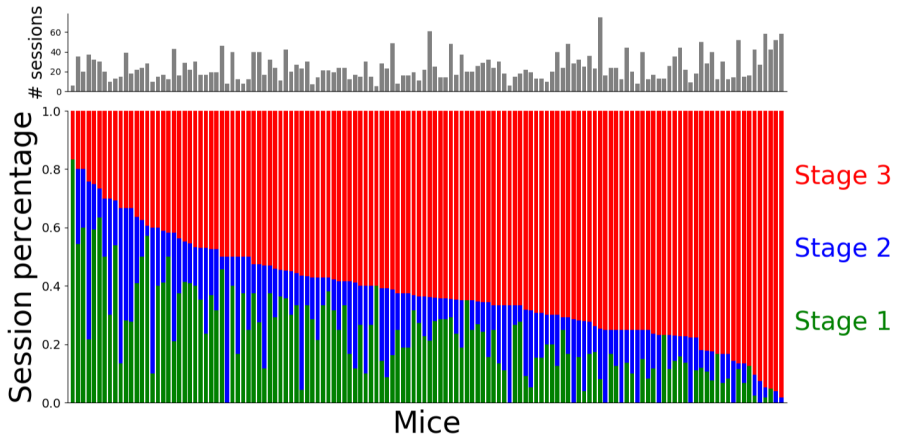

Figure A.2: The proportion of sessions per behavioural stage sorted by the proportion in Stage 3 **(bottom)**, with a histogram of the total number of sessions **(top)**. No strong trend between any of the stages or the total number of session emerges.

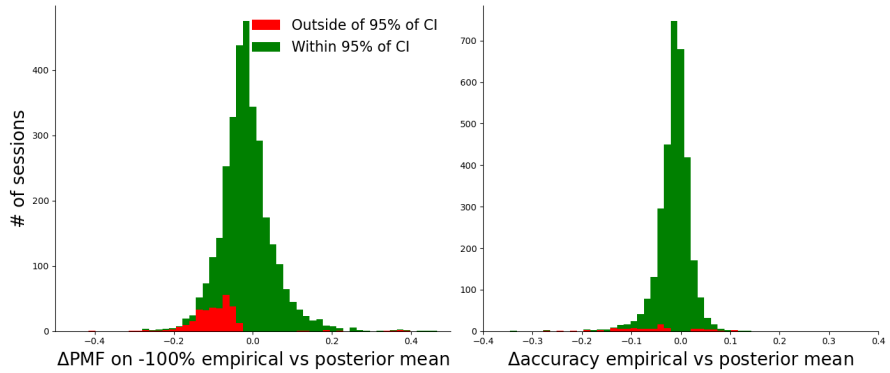

Figure A.3: To examine the biases evident in the posterior predictive checks, we plot the distance between the posterior mean and the empirically observed PMF on -100% contrasts **(left)**, and overall accuracy **(right)**. The green part of the histogram shows sessions which fall within 95% of the credible interval of the posterior, and red those that fall outside of it. Note that we expect 5% of sessions to fall outside of the 95% CI, but especially for the PMF on extreme contrasts this rate is elevated. Also note that large differences are not necessarily problematic, as behaviour can simply be noisy.

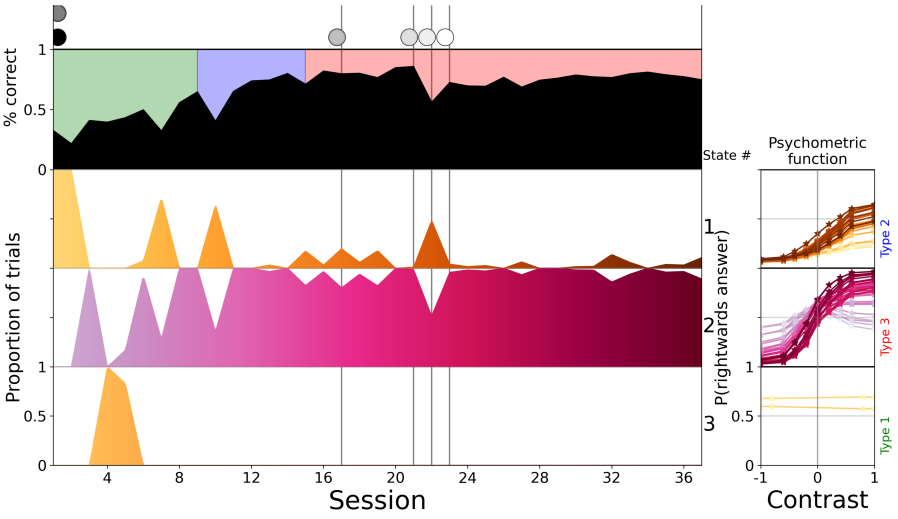

Figure A.4: State overview for a mouse with a larger number of sessions. As showcased here, especially for these long training trajectories the model sometimes took some states all the way from uninformed to proficient behaviour, through the slow change process alone.

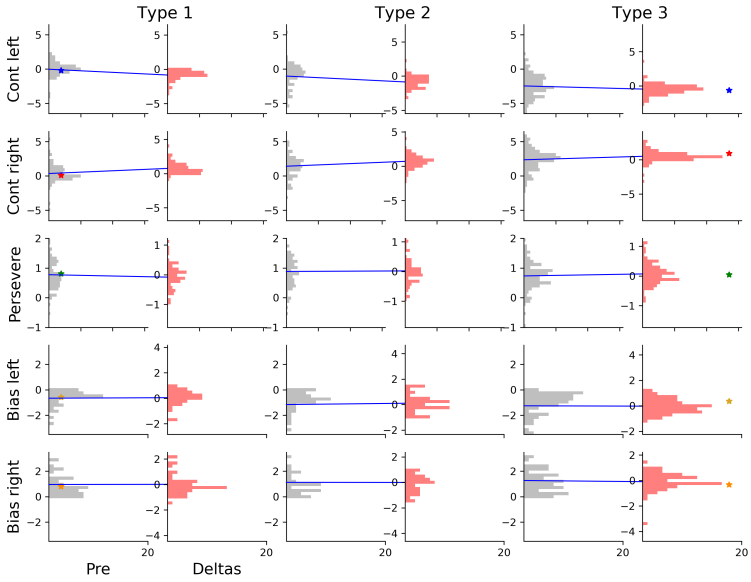

Figure A.5: Distributions of initial weights of states and how they changed across their lifespan, split by types. The red distributions to the right show the weight change, not the final distribution, to highlight the change through the slow change process. The blue lines connect the means of two related distributions (we use the mean of the initial distribution as the 0 point of the delta distributions). The bias is split as in **Fig. 4.7**, and the x-axis is shortened due to this, though the x-ticks are at the same distances across all plots. Red stars mark the average weight of the first and last state of every animal, as in **Fig. 4.7**

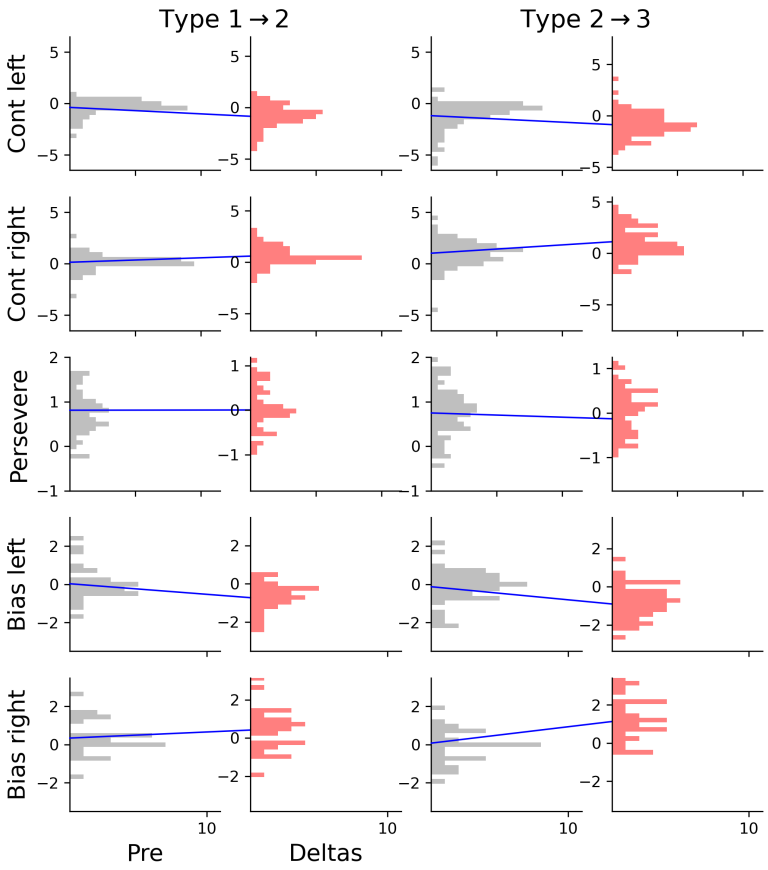

Figure A.6: Distributions of weights of the closest previous states and how they differed in the first introduced state of a new type. The red distributions to the right show the weight difference, not the final distribution, to highlight the change through the fast process. The blue lines connect the means of two related distributions (we use the mean of the previous distribution as the 0 point of the delta distributions). The bias is split as in **Fig. 4.7**, and the x-axis is shortened due to this, though the x-ticks are at the same distances across all plots.

Part VIII

SUPPLEMENTAL MATERIAL FOR HYBRID NEURAL NETWORKS

B

SUPPLEMENTAL MATERIAL FOR HYBRID NEURAL NETWORKS

B.1 OTHER NETWORK ARCHITECTURES

While we focussed on model 200-iv, as a particularly performant and interpretable model version, part of the hybrid network approach consists of using the diversity of possible architectures to study which restrictions harm performance, and which do not. We have of course also fitted all other versions of networks within our naming scheme (from 000 to 222), and have tried additional variations, which we briefly mention here for completeness sake.

Of particular note is a variation on the usage of the summarising LSTM (which provides the additional scalar to all networks which employ a type-2 network): Before settling on a version which produces a separate contrast scalar and history scalar (or multiplicative decay vector directly), we initially only returned a single scalar which went to both types of network. While this is a simpler architecture in some ways, it did have the side-effect of making the scalar harder to interpret, as it now entangled history processing changes and contrast processing changes.

Noticing the marked biases the networks sometimes like to introduce in their components which tend to cancel out in their sum, we created a network which prohibited such asymmetries. For this we targeted the history network, leaving the contrast network to absorb any biases which were actually present, since biases are particularly easy to read off from this network. However, these networks did not perform quite as well as the other interpretable architectures. Furthermore, we noticed that such networks rapidly modulated their history scalar s^h based on the previous choice, effectively circumventing the symmetry restriction we tried to impose.

B.2 NETWORK TRAINING AND EVALUATION

For our model training we made use of a total of 539 sessions, documenting the behaviour of 139 mice. All of these were performed by mice which were deemed fully proficient, and, at the same time, were being recorded from with Neuropixel probes. This data ultimately entered into the analysis of The International Brain Laboratory et al. (2023). Our sessions were selected from the total of 693 sessions in the full database based on two exclusion criteria, loosely following The International Brain Laboratory et al. (2023): we removed sessions with fewer than 400 trials (which is more stringent than The International Brain Laboratory et al. (2023)), as well as sessions which had poor performance on easy contrasts, namely fewer than 90% correct responses on 100% contrasts. We did not remove sessions or trials based on reaction times or neural data, unlike The International Brain Laboratory et al. (2023).

We trained our networks on a training data set, consisting of 403 sessions (75% of all sessions, in total encompassing 262,569 trials). We evaluated the fits on a validation set, comprising 68 sessions (representing 12.5% of all sessions, 43,278 trials). We also have a final test set of the same size, which was not used yet (68 sessions, 43,968 trials).

We stratified our data split according to two important criteria: session length and session type. As discussed in Chapter 2, a session usually ends when reaction times reach a relatively high level, thus the end of a session is almost always marked by degraded behaviour. The length of a session is thus correlated with the extent of good quality behaviour, and we made sure that each split has even proportions of all different types of session lengths. To this end, we divided sessions into quartiles based on the number of trials within them, and distributed sessions from each quartile into our train-validation-test split (following the 75%-12.5%-12.5% proportions). Behaviour had to be very poor to lead to a session less than 400 trials, which is why we excluded such sessions. However, sessions which were just more than 400 trials may still have had somewhat poor behaviour. Therefore we had a separate bin for all sessions with 405 trials or fewer (of which there were 28), which we also spread out approximately uniformly across the splits. The other criterion was the session type (as elaborated in Chapter 2). To preclude our models from getting specialised to just a subset of the presented session types, we ensured that these were distributed relatively evenly across the data sets, introducing another level of stratification into our splitting scheme.

We performed an independent hyperparameter sweep for each network architecture, to determine the ideal model capacity and training procedure. We varied the batch size of the number of sessions on which to evaluate the gradient (since we were also training recurrent networks, sessions had to be evaluated in their entirety), trying out batch sizes of 8, 16, 32, and 64 sessions. We used the Adam optimizer (Kingma et al., 2017), with different learning rate initialisations, considering $10^{-3}, 10^{-4}, 10^{-5}$, and regularised using different strengths of weight decay $10^{-3}, 10^{-4}, 10^{-5}$. We also varied the number of units in the hidden layers of the network, trying out 4, 8, 16, and 32. This applied separately to any of the networks present: the dimension of the hidden layer of the contrast processing network, the decay network, the encoding network, and the dimensionality with which the LSTM operates. The overall number of values to set was too vast for an exhaustive search in a reasonable amount of time. Luckily, a few parameter settings established themselves as generally favorable across a range of network types: A learning rate of 10^{-3} clearly outperformed other settings, and the number of hidden units was generally best set at the intermediate values of 8 or 16, so we sometimes limited all searches to these values.

To gauge the reliability of our training procedure, we paired each hyperparameter combination with four different seeds, allowing us to compare performance across four random initialisations. As seen in **Fig. 5.2**, the performance of those seeds for the stronger models could differ. To ensure that models have trained to completeness (though this can be hard to determine, as steep improvements in performance may occur after many training steps without obvious improvement (Power et al., 2022)), we trained for 400.000 epochs, by which point it appeared that the networks had thoroughly converged in terms of both training and validation loss. We evaluated the loss of each network on the validation data set at least every 20 training steps, and used the best-performing model on this metric as the representative for its architecture (ignoring possible differences to worse performing seeds).

B.3 ADDITIONAL FIGURES

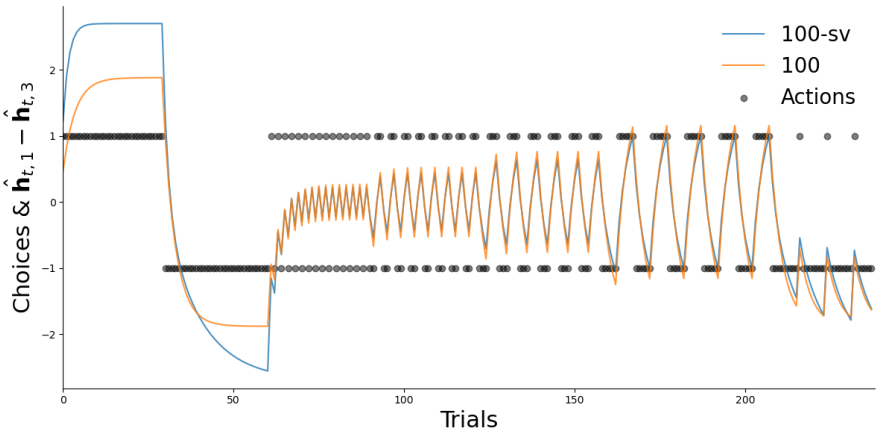

Figure B.1: Visualisation of the history adaptation implemented by the decay vector of 100-sv (calculated as the difference between the rightwards and leftwards logit of the history component $\hat{\mathbf{h}}_t$), on a sequence of artificial actions. We compare this to the classical, single decay constant of network 100. The static decay process of 100-sv settles on a chosen decay factor of 0.55, an unchosen decay of 0.91, and a history weight w of 1.23. Network 100 uses a decay factor of 0.75 and a smaller weight w of 0.47. We can see that the decay vector model can afford a considerably larger history term in response to extremely biased behaviour. More subtly, the history term of 100-sv is somewhat slower to adapt to swings in the choice behaviour. This latter case will be more relevant to actual behaviour.

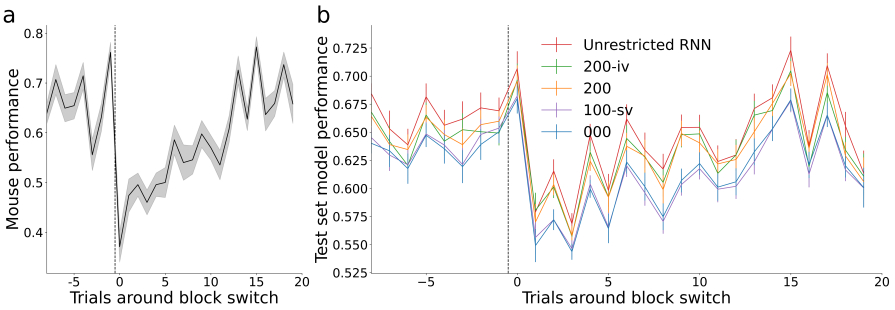

Figure B.2: Mouse and model performance around the validation set block switches on 0% contrasts only, compare with Fig. 5.9. Error bars and shaded region indicate one standard error of the mean. **a**: Mouse performance is generally worse on 0% contrasts (see before the block switch), as there is no contrast information available. If the first contrast after a block switch is a 0% contrast, the mice naturally drop notably below chance level on such a trial. **b**: The models are also generally worse on 0% contrast trials, as the mice are noisier on such trials. Note that the models still perform well on the first trial after a block switch, if it is a 0% contrast one: the mice are still predictable here, as they have yet to infer the block switch.

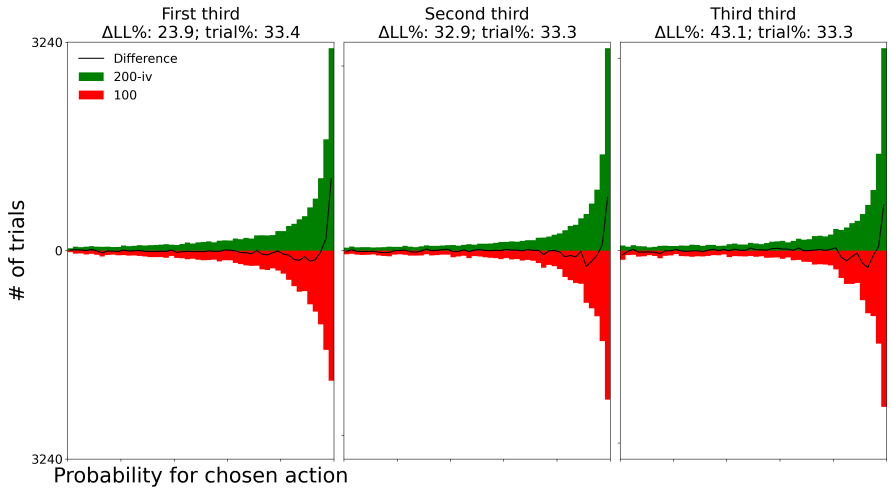

Figure B.3: Histograms comparing the prediction quality on the validation data set of model 200-iv against 100. We split the trials based on when in a session a trial occurs (first, second, or third third) and apply the same plotting logic as in Fig. 5.10. 200-iv performs better than 100 throughout a session, with somewhat of an increase towards the end, indicating that it particularly deals better with the deteriorating behaviour towards session end.

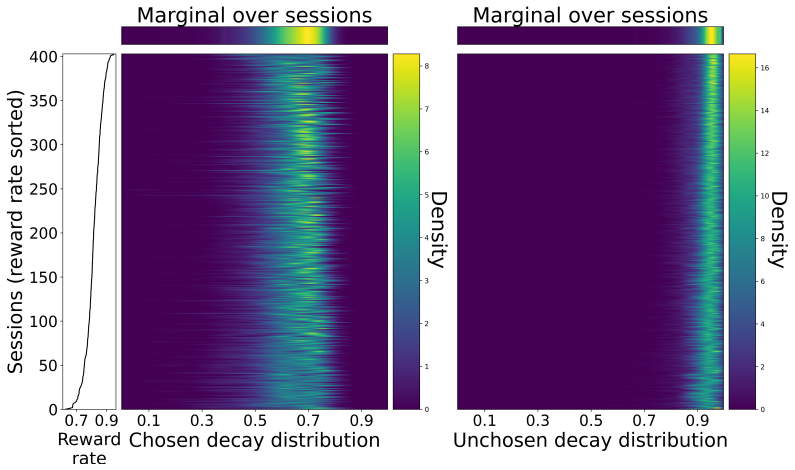

Figure B.4: Relationship between chosen action and unchosen action decay of 200-iv and the reward rate across sessions, same logic as in Fig. 5.11. These values have a less clear relationship to the session reward rate as compared to s^c. Particularly for the unchosen decay factor however, the distribution becomes more diffuse on sessions in which animals performed worse.

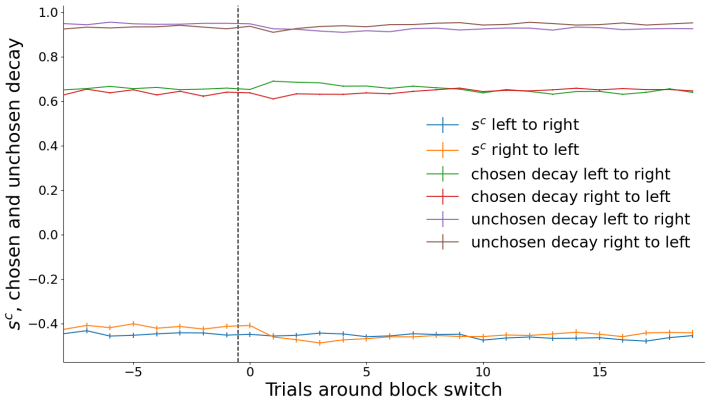

Figure B.5: Behaviour of the LSTM inferred scalars of 200-iv around block switches. We show the contrast scalar s^c, chosen, and unchosen decay, split by whether the block flipped from left to right or vice versa. Error bars represent $\pm$ 1 standard error of the mean. Block switches have a subtly different effect depending on the direction of the flip.

Agrawal, Mayank, Joshua C. Peterson, and Thomas L. Griffiths (2020). "Scaling up psychology via Scientific Regret Minimization." *Proceedings of the National Academy of Sciences* 117.16, pp. 8825–8835. DOI: 10.1073/pnas.1915841117 (cit. on p. 81).

Akam, Thomas et al. (2021). "The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection." *Neuron* 109.1, 149–163.e7. DOI: https://doi.org/10.1016/j.neuron.2020.10.013 (cit. on p. 9).

Akiti, Korleki et al. (2022). "Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction." *Neuron* 110.22, pp. 3789–3804 (cit. on p. 9).

Ji-An, Li, Marcus K. Benna, and Marcelo G. Mattar (2023). "Automatic Discovery of Cognitive Strategies with Tiny Recurrent Neural Networks." *bioRxiv*. DOI: 10.1101/2023.04.12.536629 (cit. on p. 13).

Ashwood, Zoe C. et al. (Feb. 2022). "Mice alternate between discrete strategies during perceptual decision-making." en. *Nature Neuroscience* 25.2, pp. 201–212. DOI: 10.1038/s41593-021-01007-z (cit. on pp. 7–9, 11, 23, 34, 36, 44, 53, 75, 85).

Austerweil, Joseph L., Samuel J. Gershman, and Thomas L. Griffiths (Apr. 2015). "187Structure and Flexibility in Bayesian Models of Cognition." *The Oxford Handbook of Computational and Mathematical Psychology*. Oxford University Press. DOI: 10.1093/oxfordhb/9780199957996.013.9 (cit. on p. 11).

Baum, Leonard E and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains." *The annals of mathematical statistics* 37.6, pp. 1554–1563 (cit. on p. 10).

Beal, Matthew, Zoubin Ghahramani, and Carl Rasmussen (2001). "The Infinite Hidden Markov Model." *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker, and Z. Ghahramani. Vol. 14. MIT Press (cit. on p. 11).

Behrens, Timothy E. J. et al. (Sept. 2007). "Learning the value of information in an uncertain world." *Nature Neuroscience* 10.9, pp. 1214–1221. DOI: 10.1038/nn1954 (cit. on p. 13).

Berditchevskaia, A., R. D. Cazé, and S. R. Schultz (June 2016). "Performance in a GO/NOGO perceptual task reflects a balance between impulsive and instrumental components of behaviour." *Scientific Reports* 6.1, p. 27389. DOI: 10.1038/srep27389 (cit. on p. 44).

Beron, Celia C. et al. (2022). "Mice exhibit stochastic and efficient action switching during probabilistic decision making." *Proceedings of the National Academy of Sciences* 119.15, e2113961119. DOI: 10.1073/pnas.2113961119 (cit. on p. 23).

Binz, Marcel et al. (2024). *Centaur: a foundation model of human cognition* (cit. on p. 85).

Boogert, Neeltje J. et al. (2018). "Measuring and understanding individual differences in cognition." *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1756, p. 20170280. DOI: 10.1098/rstb.2017.0280 (cit. on p. 9).

Brielmann, Aenne A and Peter Dayan (July 2022). "A computational model of aesthetic value." en. *Psychol Rev* 129.6, pp. 1319–1337 (cit. on p. 8).

Bruijns, Sebastian A. et al. (2023). "Dissecting the Complexities of Learning With Infinite Hidden Markov Models." *bioRxiv*. DOI: 10.1101/2023.12.22.573001 (cit. on p. 14).

Calhoun, Adam J, Jonathan W Pillow, and Mala Murthy (Nov. 2019). "Unsupervised identification of the internal states that shape natural behavior." en. *Nat Neurosci* 22.12, pp. 2040–2049 (cit. on p. 11).

Carter, C. K. and R. Kohn (1994). "On Gibbs Sampling for State Space Models." *Biometrika* 81.3, pp. 541–553 (cit. on p. 28).

Craver, Carl F (2006). "When mechanistic models explain." *Synthese* 153.3, pp. 355–376 (cit. on p. 3).

Cuturela, Lenca I., The International Brain Laboratory, and Jonathan W. Pillow (2024). "Internal states emerge early during learning of a perceptual decision-making task." *bioRxiv*. DOI: 10.1101/2024.11.30.626182 (cit. on p. 85).

Daw, Nathaniel D. (May 2011). "Trial-by-trial data analysis using computational models." English (US). *Decision Making, Affect, and Learning*. Publisher Copyright: © The International Association for the study of Attention and Performance, 2011. All rights reserved. Oxford University Press. DOI: 10.1093/acprof:oso/9780199600434.003.0001 (cit. on p. 3).

Daw, Nathaniel D. and Aaron C. Courville (2007). "The pigeon as particle filter." *Proceedings of the 21st International Conference on Neural Information Processing Systems*. NIPS'07. Vancouver, British Columbia, Canada: Curran Associates Inc., pp. 369–376 (cit. on p. 7).

Daw, Nathaniel D. et al. (June 2006). "Cortical substrates for exploratory decisions in humans." *Nature* 441.7095, pp. 876–879. DOI: 10.1038/nature04766 (cit. on p. 6).

Dayan, Peter (2012). "How to set the switches on this thing." *Current Opinion in Neurobiology* 22.6. Decision making, pp. 1068–1074. DOI: https://doi.org/10.1016/j.conb.2012.05.011 (cit. on p. 86).

Dayan, Peter, Jonathan P. Roiser, and Essi Viding (Oct. 2020). "The first steps on long marches: The costs of active observation." en. *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine*. Ed. by Julian Savulescu et al. Oxford University Press, pp. 213–228. DOI: 10.1093/med/9780198789697.003.0014 (cit. on p. 9).

Dekker, Ronald B., Fabian Otto, and Christopher Summerfield (2022). "Curriculum learning for human compositional generalization." *Proceedings of the National Academy of Sciences* 119.41, e2205582119. DOI: 10.1073/pnas.2205582119 (cit. on pp. 7, 51, 82).

Dezfouli, Amir et al. (2019a). "Disentangled behavioural representations." *Advances in neural information processing systems* 32 (cit. on pp. 13, 76, 85).

Dezfouli, Amir et al. (2019b). "Models that learn how humans learn: The case of decision-making and its disorders." *PLoS computational biology* 15.6, e1006903 (cit. on p. 13).

Dingemanse, Niels J. et al. (2010). "Behavioural reaction norms: animal personality meets individual plasticity." *Trends in Ecology & Evolution* 25.2, pp. 81–89. DOI: https://doi.org/10.1016/j.tree.2009.07.013 (cit. on pp. 9, 82).

Dougherty, Liam R and Lauren M Guillette (Sept. 2018). "Linking personality and cognition: a meta-analysis." en. *Philos Trans R Soc Lond B Biol Sci* 373.1756 (cit. on p. 9).

Eckstein, Maria et al. (July 2024). *Hybrid Neural-Cognitive Models Reveal How Memory Shapes Human Reward Learning*. DOI: 10.31234/osf.io/u9ks4 (cit. on pp. 13, 57, 75, 81).

Éltető, Noémi et al. (Nov. 2022). "Tracking human skill learning with a hierarchical Bayesian sequence model." *PLOS Computational Biology* 18.11, pp. 1–28. DOI: 10.1371/journal.pcbi.1009866 (cit. on p. 7).

Epstein, R. et al. (Mar. 1984). "'Insight' in the pigeon: antecedents and determinants of an intelligent performance." en. *Nature* 308.5954, pp. 61–62. DOI: 10.1038/308061a0 (cit. on pp. 7, 82).

Fechner, Gustav Theodor (1860). *Elemente der psychophysik*. Vol. 2. Breitkopf u. Härtel (cit. on pp. 3, 37).

Findling, Charles et al. (2023). "Brain-wide representations of prior information in mouse decision-making." *bioRxiv*. DOI: 10.1101/2023.07.04.547684 (cit. on pp. 13, 18, 23, 29, 57, 58, 60, 90).

Fritsche, Matthias et al. (Aug. 2024). "Temporal regularities shape perceptual decisions and striatal dopamine signals." *Nature Communications* 15.1, p. 7093. DOI: 10.1038/s41467-024-51393-8 (cit. on pp. 10, 75).

Frühwirth-Schnatter, Sylvia (1994). "DATA AUGMENTATION AND DYNAMIC LINEAR MODELS." *Journal of Time Series Analysis* 15.2, pp. 183–202. DOI: https://doi.org/10.1111/j.1467-9892.1994.tb00184.x (cit. on p. 28).

Fründ, Ingo, N. Valentin Haenel, and Felix A. Wichmann (May 2011). "Inference for psychometric functions in the presence of nonstationary behavior." *Journal of Vision* 11.6, pp. 16–16. DOI: 10.1167/11.6.16 (cit. on p. 8).

Fründ, Ingo, Felix A. Wichmann, and Jakob H. Macke (June 2014). "Quantifying the effect of intertrial dependence on perceptual decisions." *Journal of Vision* 14.7, pp. 9–9. DOI: 10.1167/14.7.9 (cit. on pp. 9, 75).

Gallistel, Charles R., Stephen Fairhurst, and Peter Balsam (Sept. 2004). "The learning curve: Implications of a quantitative analysis." en. *Proceedings of the National Academy of Sciences* 101.36, pp. 13124–13131. DOI: 10.1073/pnas.0404965101 (cit. on pp. 6, 7).

Gelman, Andrew (2013). *Bayesian data analysis* (cit. on p. 8).

Gelman, Andrew and Donald B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7.4, pp. 457–472. DOI: 10.1214/ss/1177011136 (cit. on p. 30).

Gershman, Samuel J, David M Blei, and Yael Niv (Jan. 2010a). "Context, learning, and extinction." en. *Psychol Rev* 117.1, pp. 197–209 (cit. on p. 11).

Gershman, Samuel J and Yael Niv (2012a). "Exploring a latent cause theory of classical conditioning." *Learning & behavior* 40, pp. 255–268 (cit. on p. 53).

— (2010b). "Learning latent structure: carving nature at its joints." *Current Opinion in Neurobiology* 20.2. Cognitive neuroscience, pp. 251–256. DOI: https://doi.org/10.1016/j.conb.2010.02.008 (cit. on pp. 5, 83).

Gershman, Samuel J, Kenneth A Norman, and Yael Niv (2015a). "Discovering latent causes in reinforcement learning." *Current Opinion in Behavioral Sciences* 5. Neuroeconomics, pp. 43–50. DOI: https://doi.org/10.1016/j.cobeha.2015.07.007 (cit. on p. 11).

Gershman, Samuel J. (2020a). "Origin of perseveration in the trade-off between reward and complexity." *Cognition* 204, p. 104394. DOI: https://doi.org/10.1016/j.cognition.2020.104394 (cit. on pp. 9, 75, 83).

Gershman, Samuel J. (Nov. 2020b). "Origin of perseveration in the trade-off between reward and complexity." en. *Cognition* 204, p. 104394. DOI: 10.1016/j.cognition.2020.104394 (cit. on p. 29).

Gershman, Samuel J. and David M. Blei (Feb. 2012b). "A tutorial on Bayesian nonparametric models." en. *Journal of Mathematical Psychology* 56.1, pp. 1–12. DOI: 10.1016/j.jmp.2011.08.004 (cit. on p. 11).

Gershman, Samuel J. and Catherine A. Hartley (Sept. 2015b). "Individual differences in learning predict the return of fear." *Learning & Behavior* 43.3, pp. 243–250. DOI: 10.3758/s13420-015-0176-z (cit. on p. 11).

Gold, Joshua I. and Michael N. Shadlen (2002). "Banburismus and the Brain: Decoding the Relationship between Sensory Stimuli, Decisions, and Reward." *Neuron* 36.2, pp. 299–308. DOI: https://doi.org/10.1016/S0896-6273(02)00971-6 (cit. on p. 53).

Heald, James B, Daniel M Wolpert, and Máté Lengyel (2023). "The computational and neural bases of context-dependent learning." *Annual Review of Neuroscience* 46.1, pp. 233–258 (cit. on p. 53).

Heald, James B., Máté Lengyel, and Daniel M. Wolpert (Dec. 2021). "Contextual inference underlies the learning of sensorimotor repertoires." *Nature* 600.7889, pp. 489–493. DOI: 10.1038/s41586-021-04129-3 (cit. on p. 11).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory." *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (cit. on pp. 57, 60).

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators." *Neural Networks* 2.5, pp. 359–366. DOI: https://doi.org/10.1016/0893-6080(89)90020-8 (cit. on pp. 12, 57).

Izquierdo, A. et al. (2017). "The neural basis of reversal learning: An updated perspective." *Neuroscience* 345. Cognitive Flexibility: Development, Disease, and Treatment, pp. 12–26. DOI: https://doi.org/10.1016/j.neuroscience.2016.03.021 (cit. on p. 6).

Jain, Yash Raj et al. (2023). "A computational process-tracing method for measuring people's planning strategies and how they change over time." *Behavior Research Methods* 55.4, pp. 2037–2079 (cit. on p. 6).

Jang, Anthony I et al. (Aug. 2015). "The Role of Frontal Cortical and Medial-Temporal Lobe Brain Areas in Learning a Bayesian Prior Belief on Reversals." en. *J Neurosci* 35.33, pp. 11751–11760 (cit. on p. 6).

Johnson, Matthew J. (2014). "Bayesian time series models and scalable inference." PhD thesis. MIT (cit. on p. 11).

Johnson, Matthew J. and Alan S. Willsky (Feb. 2013). "Bayesian Nonparametric Hidden Semi-Markov Models." *J. Mach. Learn. Res.* 14.1, pp. 673–701 (cit. on pp. 11, 25–27).

Jun, James J et al. (2017). "Fully integrated silicon probes for high-density recording of neural activity." *Nature* 551.7679, pp. 232–236 (cit. on p. 18).

Kastner, David B. et al. (2022). "Spatial preferences account for inter-animal variability during the continual learning of a dynamic cognitive task." *Cell Reports* 39.3, p. 110708. DOI: https://doi.org/10.1016/j.celrep.2022.110708 (cit. on p. 9).

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization* (cit. on p. 98).

Köhler, Wolfgang (1948). "The mentality of apes, 1917." Century psychology series. East Norwalk, CT, US: Appleton-Century-Crofts, pp. 497–505. DOI: 10.1037/11304-054 (cit. on p. 7).

Krakauer, John W et al. (Feb. 2017). "Neuroscience Needs Behavior: Correcting a Reductionist Bias." en. *Neuron* 93.3, pp. 480–490 (cit. on p. 3).

Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini (2008). "Representational similarity analysis - connecting the branches of systems neuroscience." *Frontiers in Systems Neuroscience* 2. DOI: 10.3389/neuro.06.004.2008 (cit. on p. 31).

Kuchibhotla, Kishore V. et al. (May 2019). "Dissociating task acquisition from expression during learning reveals latent knowledge." *Nature Communications* 10.1, p. 2151. DOI: 10.1038/s41467-019-10089-0 (cit. on p. 82).

Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum (2015). "Human-level concept learning through probabilistic program induction." *Science* 350.6266, pp. 1332–1338. DOI: 10.1126/science.aab3050 (cit. on p. 5).

Lake, Brenden M. et al. (2017). "Building machines that learn and think like people." *Behavioral and Brain Sciences* 40, e253. DOI: 10.1017/S0140525X16001837 (cit. on p. 5).

Liebana Garcia, Samuel et al. (2023). "Striatal dopamine reflects individual long-term learning trajectories." *bioRxiv*. DOI: 10.1101/2023.12.14.571653 (cit. on pp. 7, 51, 83, 84).

Linderman, Scott W., Matthew J. Johnson, and Ryan P. Adams (2015). "Dependent Multinomial Models Made Easy: Stick Breaking with the Pólya-Gamma Augmentation." *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Montreal, Canada: MIT Press, pp. 3456–3464 (cit. on p. 28).

Link, William A. and Mitchell J. Eaton (Feb. 2012). "On thinning of chains in MCMC: *Thinning of MCMC chains*." en. *Methods in Ecology and Evolution* 3.1, pp. 112–115. DOI: 10.1111/j.2041-210X.2011.00131.x (cit. on p. 30).

Lloyd, Kevin and David S Leslie (2013a). "Context-dependent decision-making: a simple Bayesian model." *Journal of The Royal Society Interface* 10.82, p. 20130069 (cit. on p. 53).

— (2013b). "Context-dependent decision-making: a simple Bayesian model." *Journal of The Royal Society Interface* 10.82, p. 20130069. DOI: 10.1098/rsif.2013.0069 (cit. on p. 11).

Loewenstein, George and Ted O'Donoghue (2004). "Animal spirits: Affective and deliberative processes in economic behavior." *Available at SSRN 539843* (cit. on p. 86).

Löwe, Anika T. et al. (Oct. 2024). "Abrupt and spontaneous strategy switches emerge in simple regularised neural networks." *PLOS Computational Biology* 20.10, pp. 1–29. DOI: 10.1371/journal.pcbi.1012505 (cit. on pp. 7, 83).

Luft, Andreas R and Manuel M Buitrago (Dec. 2005). "Stages of motor skill learning." en. *Mol Neurobiol* 32.3, pp. 205–216 (cit. on pp. 7, 53).

MacDougall-Shackleton, Scott A (2011). "The levels of analysis revisited." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1574, pp. 2076–2085 (cit. on p. 3).

Maggi, Silvia et al. (2022). "Tracking subject's strategies in behavioural choice experiments at trial resolution." *bioRxiv*. DOI: 10.1101/2022.08.30.505807 (cit. on p. 6).

Maier, Norman RF (1931). "Reasoning and learning." *Psychological review* 38.4, p. 332 (cit. on p. 7).

Marques, João C. et al. (Jan. 2020). "Internal state dynamics shape brainwide activity and foraging behaviour." *Nature* 577.7789, pp. 239–243. DOI: 10.1038/s41586-019-1858-z (cit. on p. 86).

Marr, David (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press (cit. on p. 3).

Meyniel, Florent et al. (2013). "Neurocomputational account of how the human brain decides when to have a break." *Proceedings of the National Academy of Sciences* 110.7, pp. 2641–2646. DOI: 10.1073/pnas.1211925110 (cit. on p. 8).

Miller, Kevin J., Matthew M. Botvinick, and Carlos D. Brody (2021). "From predictive models to cognitive models: Separable behavioral processes underlying reward learning in the rat." *bioRxiv*. DOI: 10.1101/461129 (cit. on p. 23).

Miller, Kevin J. et al. (2024). "Cognitive model discovery via disentangled RNNs." *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc. (cit. on p. 13).

Mohammadi, Zeinab et al. (2024). "Identifying the factors governing internal state switches during nonstationary sensory decision-making." *bioRxiv*. DOI: 10.1101/2024.02.02.578482 (cit. on p. 81).

Moore, Sharlen and Kishore V. Kuchibhotla (2022). "Slow or sudden: Re-interpreting the learning curve for modern systems neuroscience." *IBRO Neuroscience Reports* 13, pp. 9–14. DOI: https://doi.org/10.1016/j.ibneur.2022.05.006 (cit. on p. 7).

Nassar, Matthew R and Michael J Frank (Oct. 2016). "Taming the beast: extracting generalizable knowledge from computational models of cognition." en. *Curr Opin Behav Sci* 11, pp. 49–54 (cit. on p. 12).

Nyberg, Lars et al. (2012). "Memory aging and brain maintenance." *Trends in Cognitive Sciences* 16.5, pp. 292–305. DOI: https://doi.org/10.1016/j.tics.2012.04.005 (cit. on p. 53).

Palminteri, Stefano, Valentin Wyart, and Etienne Koechlin (2017). "The Importance of Falsification in Computational Cognitive Modeling." *Trends in Cognitive Sciences* 21.6, pp. 425–433. DOI: https://doi.org/10.1016/j.tics.2017.03.011 (cit. on p. 12).

Pan-Vazquez, Alejandro et al. (2024). "Pre-existing visual responses in a projection-defined dopamine population explain individual learning trajectories." *Current Biology* 34.22, 5349–5358.e6. DOI: https://doi.org/10.1016/j.cub.2024.09.045 (cit. on p. 84).

Papachristos, Efstathios B. and C. R. Gallistel (May 2006). "AUTOSHAPED HEAD POKING IN THE MOUSE: A QUANTITATIVE ANALYSIS OF THE LEARNING CURVE." en. *Journal of the Experimental Analysis of Behavior* 85.3, pp. 293–308. DOI: 10.1901/jeab.2006.71-05 (cit. on pp. 6, 9, 49).

Piaget, Jean (1952). *The origins of intelligence in children.* The origins of intelligence in children. New York, NY, US: W W Norton & Co, pp. 419–419. DOI: 10.1037/11494-000 (cit. on pp. 9, 82).

Piray, Payam and Nathaniel D. Daw (July 2020). "A simple model for learning in volatile environments." *PLOS Computational Biology* 16.7, pp. 1–26. DOI: 10.1371/journal.pcbi.1007963 (cit. on p. 82).

— (Aug. 2021). "Linear reinforcement learning in planning, grid fields, and cognitive control." *Nature Communications* 12.1, p. 4942. DOI: 10.1038/s41467-021-25123-3 (cit. on p. 83).

Pisupati, Sashank et al. (Jan. 2021). "Lapses in perceptual decisions reflect exploration." *eLife* 10. Ed. by Daeyeol Lee et al., e55490. DOI: 10.7554/eLife.55490 (cit. on p. 8).

Polson, Nicholas G., James G. Scott, and Jesse Windle (2013). *Bayesian inference for logistic models using Polya-Gamma latent variables* (cit. on p. 28).

Power, Alethea et al. (2022). *Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets* (cit. on p. 98).

Priestley, Luke et al. (2025). "ACTIVITY IN HUMAN DORSAL RAPHE NUCLEUS SIGNALS CHANGES IN BEHAVIOURAL POLICY." *bioRxiv*. DOI: 10.1101/2025.01.08.632066 (cit. on p. 86).

Ratcliff, Roger and Gail McKoon (2008). "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks." *Neural Computation* 20.4, pp. 873–922. DOI: 10.1162/neco.2008.12-06-420 (cit. on p. 53).

Rescorla, Robert A (1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement." *Classical conditioning, Current research and theory* 2, pp. 64–69 (cit. on p. 7).

Roy, Nicholas A. et al. (Feb. 2021). "Extracting the dynamics of behavior in sensory decision-making experiments." en. *Neuron* 109.4, 597–610.e6. DOI: 10.1016/j.neuron.2020.12.004 (cit. on pp. 9, 12, 23, 29, 36, 44, 53, 85).

Schaeffer, Rylan et al. (2020). "Reverse-engineering Recurrent Neural Network solutions to a hierarchical inference task for mice." *bioRxiv*. DOI: 10.1101/2020.06.09.142745 (cit. on p. 13).

Schultz, Wolfram, Peter Dayan, and P. Read Montague (1997). "A Neural Substrate of Prediction and Reward." *Science* 275.5306, pp. 1593–1599. DOI: 10.1126/science.275.5306.1593 (cit. on p. 82).

Sih, Andrew and Marco Del Giudice (2012). "Linking behavioural syndromes and cognition: a behavioural ecology perspective." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1603, pp. 2762–2772. DOI: 10.1098/rstb.2012.0216 (cit. on p. 9).

Skinner, B. F. (1953). *Science and human behavior.* Science and human behavior. Oxford, England: Macmillan, pp. x, 461–x, 461 (cit. on p. 3).

Smith, Anne C et al. (Jan. 2004). "Dynamic analysis of learning in behavioral experiments." en. *J Neurosci* 24.2, pp. 447–461 (cit. on p. 6).

Summerfield, Christopher and Konstantinos Tsetsos (Jan. 2015). "Do humans make good decisions?" *Trends in Cognitive Sciences* 19.1, pp. 27–34. DOI: 10.1016/j.tics.2014.11.005 (cit. on p. 68).

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: an introduction, 2nd edn. Adaptive computation and machine learning* (cit. on pp. 5, 83).

Teh, Yee Whye et al. (Dec. 2006). "Hierarchical Dirichlet Processes." en. *Journal of the American Statistical Association* 101.476, pp. 1566–1581. DOI: 10.1198/016214506000000302 (cit. on p. 11).

The International Brain Laboratory et al. (2023). "A Brain-Wide Map of Neural Activity during Complex Behaviour." *bioRxiv*. DOI: 10.1101/2023.07.04.547681 (cit. on pp. 18, 97).

The International Brain Laboratory et al. (May 2021). "Standardized and reproducible measurement of decision-making in mice." en. *eLife* 10, e63711. DOI: 10.7554/eLife.63711 (cit. on pp. 5, 6, 9, 17, 43, 57).

Thorndike, Edward Lee (1911). *Animal intelligence: Experimental studies.* On cover: The animal behavior series. Lewiston, NY, US: Macmillan Press, pp. viii, 297–viii, 297. DOI: 10.5962/bhl.title.55072 (cit. on pp. 9, 29).

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 185.4157, pp. 1124–1131. DOI: 10.1126/science.185.4157.1124 (cit. on p. 3).

Vehtari, Aki et al. (June 2021). "Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis* 16.2. DOI: 10.1214/20-ba1221 (cit. on p. 30).

Watkins, Christopher JCH and Peter Dayan (1992). "Q-learning." *Machine learning* 8, pp. 279–292 (cit. on p. 4).

Wichmann, Felix A. and N. Jeremy Hill (Nov. 2001). "The psychometric function: I. Fitting, sampling, and goodness of fit." *Perception & Psychophysics* 63.8, pp. 1293–1313. DOI: 10.3758/BF03194544 (cit. on pp. 8, 75).

Wilson, Robert C and Anne GE Collins (Nov. 2019). "Ten simple rules for the computational modeling of behavioral data." *eLife* 8. Ed. by Timothy E Behrens, e49547. DOI: 10.7554/eLife.49547 (cit. on p. 3).

Wiltschko, Alexander B. et al. (Dec. 2015). "Mapping Sub-Second Structure in Mouse Behavior." en. *Neuron* 88.6, pp. 1121–1135. DOI: 10.1016/j.neuron.2015.11.031 (cit. on pp. 11, 53).

Windle, Jesse et al. (2013). *Efficient Data Augmentation in Dynamic Models for Binary and Count Data* (cit. on p. 28).

Yao, Yuling, Aki Vehtari, and Andrew Gelman (Jan. 2022). "Stacking for Non-Mixing Bayesian Computations: The Curse and Blessing of Multimodal Posteriors." *J. Mach. Learn. Res.* 23.1 (cit. on p. 30).

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede, modified by Shervin Safavi for the purpose of his thesis, and then minimally modified by Sebastian Bruijns for this thesis. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

`http://code.google.com/p/classicthesis/`