Safe Learning in Humans and Machines

Pranav Mahajan

Linacre College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2025

Abstract

Intelligent agents, biological or artificial, face a fundamental dilemma: how to learn safely from experience when learning inevitably involves making mistakes. This requires the ability to explore environments with caution during learning (i.e., *safe exploration*) and to infer deviations from homeostatic grace, such as injury in animals or faults in robots, to reorganise behaviour appropriately (i.e., *self-preservation*).

This work first explores safe exploration through strategies that combine multiple value functions. A critical safety-efficiency trade-off is identified, arising from the conflict between instrumental control, which learns the consequences of actions, and learned defensive reflexes such as Pavlovian biases to withdraw from aversive stimuli. It is hypothesised that this trade-off can be resolved by gating Pavlovian avoidance based on outcome uncertainty, and a basic test is provided in a human approach–withdrawal virtual reality experiment. Noting the suboptimality underlying Pavlovian misbehaviour, the thesis subsequently proposes a mechanism by which the dopaminergic system could optimally compose multiple values to support efficient, safe, and stable learning.

Shifting focus from external threats to bodily integrity, the thesis next addresses the problem of self-preservation, with particular emphasis on the computational representation of injury. Post-injury homeostasis is modelled as a partially observable Markov decision process (POMDP), explaining counterintuitive behaviours such as investigating an injury despite immediate pain. This framework is used to mathematically formalise an information-restriction model of pain chronification, providing a quantitative complement to the Fear-Avoidance model. These concepts are then extended to machines: robots performing stereotypical movements can employ self-supervised learning and local learning rules to build internal models of expected sensorimotor experience, enabling fault detection and adaptive responses to unexpected deviations.

Together, this work advances the understanding of safe learning, a challenge shared by humans and machines, with implications for understanding post-injury transitions to chronic pain and the development of neuro-inspired safe AI.

Safe Learning in Humans and Machines

Pranav Mahajan

Linacre College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

This thesis is dedicated to my late grandfather
Dattatray Yashwant Mahajan,
who always believed in me and had the kindest heart.

Acknowledgements

Personal

First and foremost, I would like to thank my advisor, Prof. Ben Seymour. He took a chance on me and supported me in every way possible, offering academic guidance, financial assistance, and pastoral support. He invested a great deal of time and effort in helping me grow as an independent thinker and researcher. I also cannot thank him enough for his moral support during the tough times of my DPhil, particularly during my struggle with severe back pain and sciatica. This thesis would not have been possible without him and his invaluable, steadfast support.

My sincere thanks also go to my secondary advisor, Prof. Ioannis Havoutis, for his guidance on Chapter 7 and his continued support throughout my studies. I am deeply grateful to Prof. Peter Dayan for his invaluable advice on improving Chapter 6, for hosting me as a guest scientist in his lab at the Max Planck Institute, University of Tübingen, and for many insightful discussions. I would like to thank Prof. Sang Wan Lee for his suggestions, which greatly improved Chapter 4, and for his role in co-securing the funding that supported my work. My thanks are also due to Prof. Rafal Bogacz for allowing me to join the brainstorming meetings with his group (led by my collaborator Dr. Mufeng Tang), and for his helpful suggestions on several of my projects. I am also grateful to Prof. Michael Browning for encouraging me to consider the real-world impact of my thesis and for his overall feedback during my assessments.

I would like to thank the members of the Seymour lab—Dr. Suyi Zhang, Dr. Shuangyi Tong, Dr. Charlie Yan, Sarah Schreiber, Mattan Pelah, RuoHan Liu, Dr. Danielle Hewitt, Tianjin Ed Li, Junyu Ren, Amanda Wall, Dr. Andrew Segerdahl, Dr. Rachel Crockett, Dr. Lin Qiu, and Dr. Wako Yoshida—for their helpful comments on my work at multiple stages and for making this DPhil a memorable experience. I would like to thank Dr. Leen Van Broeck from NDCN Graduate Studies for checking up on me during my struggle with chronic pain.

Finally, I would like to thank my partner, Mitsuki Mori, for her constant support and for tolerating my many eureka moments. My gratitude extends to my aunt, Prajakta Mahajan, for her assistance in securing the education loan that gave me the peace of mind to pursue this DPhil. I thank my grandparents, who have always believed in me, and especially my parents, Madhura and Parag Mahajan, who have

always put my career before their own and done everything possible to ensure I received the best education. None of this would have been possible without them.

Institutional

I am grateful for the financial support for this research from the Wellcome Trust (214251/Z/18/Z), IITP (MSIT 2019-0-01371), and JSPS (22H04998). This work was also partly supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316). All of this funding was secured by Prof. Ben Seymour; the grant from IITP was co-secured with Prof. Sang Wan Lee.

The views expressed are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care.

I am also grateful to have been hosted by the Wellcome Centre for Integrative Neuroimaging (WIN), now the Oxford Centre for Integrative Neuroimaging (OxCIN), which is supported by the Wellcome Trust (203139/Z/16/Z and 203139/A/16/Z).

Publications

Publications included in this thesis. During my DPhil, I have disseminated several of the results I report in this thesis:

- Chapter 2: Mahajan, P. & Seymour, B. (2025) Forward and reverse engineering the pain system: from computational neuroscience to neuro-engineering. *PAIN*.

- Chapter 4: Mahajan, P., Tong, S., Lee, S. W., & Seymour, B. (2025). Balancing safety and efficiency in human decision making. *eLife*.

  This article is indexed as a version of record (VoR), with editorial assessment as *important* work with *convincing* results, and included in the eLife digest. Early versions of this work were presented at the Computational and Systems Neuroscience Conference (COSYNE 2022) as an extended abstract.

- Chapter 5: Mahajan, P. & Seymour, B. (2025) Composing the value signal for dopamine-mediated learning *bioRxiv*.

  This manuscript is in preparation for a journal submission at the time of writing this thesis and was previously presented at the Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM 2025) as an extended abstract: Optimal composition of multiple value functions for dopamine-mediated efficient, safe and stable learning.

- Chapter 6: Mahajan, P., Dayan, P., & Seymour, B. (2026). Homeostasis After Injury: How Intertwined Inference and Control Underpin Post-Injury Pain and Behaviour. *PLOS Computational Biology*.

  This work was also presented at the Computational Psychiatry conference (2025) and the Cognitive Computational Neuroscience conference (CCN 2025) as an extended abstract.

- Chapter 7: Mahajan, P., Tang, M., Li, T., Havoutis, I., & Seymour, B. (2025). Neural Associative Skill Memories for safer robotics and modelling human sensorimotor repertoires. *Neural Computation*.

  This work was also presented at the International Workshop on Active Inference (IWAI 2024) as an extended abstract (Spotlight).

The literature review in Chapter 3 may be prepared into a review paper at a later stage.

Publications excluded from this thesis. During my DPhil, I have published other papers which are excluded from this thesis:

- Mancini, F., Mahajan, P., Guttesen, A. Á. V., Onysk, J., Scholtes, I., Shenker, N., ... & Seymour, B. (2025). Enhanced behavioural and neural sensitivity to punishments in chronic pain and fatigue. *Brain*, 148(6), 2151-2162.
- Palod, V., Mahajan, P., Baths, V., & Gutkin, B. (2025). Discounting and Drug Seeking in Biological Hierarchical Reinforcement Learning. *Cognitive Computational Neuroscience conference (CCN 2025)* (full paper).
- Mahajan, P., Baths, V., & Gutkin, B. (2023). Doing what's not wanted: Conflict in incentives and misallocation of behavioural control can lead to drug-seeking despite adverse outcomes. *Addiction Neuroscience*, 8, 100115.

Statements

Statement of previous acceptance or concurrent submission of thesis for the degree. No part of my thesis has been accepted or is currently being submitted for any degree, diploma or certificate or other qualification in this University or elsewhere.

Statement of Intellectual Contribution. The research presented in this thesis was carried out by me under the supervision of Prof. Ben Seymour and Prof. Ioannis Havoutis. All work presented herein is my own, except where otherwise indicated. Portions of this thesis include material from publications arising during my DPhil, some of which involved collaborations with other researchers.

- Chapters 2, 4, 5, 6, and 7 were developed by me under the guidance of Prof. Ben Seymour.
- The virtual reality experiments in Chapter 4 were designed and implemented by me, building upon the virtual reality experimental platform originally developed by Dr. Shuangyi Tong for the lab. Dr. Tong also assisted with data collection for the additional VR maze task described in Appendix A (not part of the main thesis chapters).
- Chapter 6 benefited from guidance from Prof. Peter Dayan, particularly in the restructuring and rewriting the manuscript.
- Chapter 7 was further developed through discussions with Prof. Ioannis Havoutis, Dr. Mufeng Tang, and Tianjin Ed Li.

Statement of length of thesis. The length of the thesis is estimated to be approximately 40,000 words or less, which meets the department's limit of 50,000 words.

Abstract

Intelligent agents, biological or artificial, face a fundamental dilemma: how to learn safely from experience when learning inevitably involves making mistakes. This requires the ability to explore environments with caution during learning (i.e., *safe exploration*) and to infer deviations from homeostatic grace, such as injury in animals or faults in robots, to reorganise behaviour appropriately (i.e., *self-preservation*).

This work first explores safe exploration through strategies that combine multiple value functions. A critical safety-efficiency trade-off is identified, arising from the conflict between instrumental control, which learns the consequences of actions, and learned defensive reflexes such as Pavlovian biases to withdraw from aversive stimuli. It is hypothesised that this trade-off can be resolved by gating Pavlovian avoidance based on outcome uncertainty, and a basic test is provided in a human approach–withdrawal virtual reality experiment. Noting the suboptimality underlying Pavlovian misbehaviour, the thesis subsequently proposes a mechanism by which the dopaminergic system could optimally compose multiple values to support efficient, safe, and stable learning.

Shifting focus from external threats to bodily integrity, the thesis next addresses the problem of self-preservation, with particular emphasis on the computational representation of injury. Post-injury homeostasis is modelled as a partially observable Markov decision process (POMDP), explaining counterintuitive behaviours such as investigating an injury despite immediate pain. This framework is used to mathematically formalise an information-restriction model of pain chronification, providing a quantitative complement to the Fear-Avoidance model. These concepts are then extended to machines: robots performing stereotypical movements can employ self-supervised learning and local learning rules to build internal models of expected sensorimotor experience, enabling fault detection and adaptive responses to unexpected deviations.

Together, this work advances the understanding of safe learning, a challenge shared by humans and machines, with implications for understanding post-injury transitions to chronic pain and the development of neuro-inspired safe AI.

Contents

List of Figures

List of Abbreviations

ACL . . . . . Anterior Cruciate Ligament.

AI . . . . . Artificial Intelligence.

Anxa1 . . . . . Annexin A1.

APE . . . . . Action Prediction Error.

ASM . . . . . Associative Skill Memories.

BART . . . . . Balloon Analogue Risk Task.

BDT . . . . . Bayesian Decision Theory.

BPTT . . . . . Backpropagation Through Time.

CBT . . . . . Cognitive Behavioural Therapies.

CMDP . . . . Constrained Markov Decision Process.

CSC . . . . . Complete Serial Compound model.

CVaR . . . . . Conditional Value at Risk.

D1 . . . . . Direct pathway (referring to striatal pathways).

D2 . . . . . Indirect pathway (referring to striatal pathways).

DA . . . . . Dopamine.

DCD . . . . . Directional Compensatory Deviation.

dlPFC . . . . . dorsolateral Prefrontal Cortex.

DMP . . . . . Dynamic Movement Primitives.

DR . . . . . Default Representation.

DS-5 . . . . . Constant current stimulator.

EEG . . . . . Electroencephalogram.

EMG . . . . . Electromyography.

FPR . . . . . False Positive Rate.

GSR . . . . . Galvanic Skin Response.

HI . . . . . High Information gain.

HMM . . . . . Hidden Markov Model.

HR . . . . . Heart Rate.

KL . . . . . Kullback-Leibler.

LI . . . . . Low Information gain.

LOOIC . . . . Leave-one-out Information Criteria.

MB . . . . . Model-Based.

MDP . . . . . Markov Decision Process.

MF . . . . . Model-Free.

MO . . . . . Multi-Objective.

MTRNN . . . Multiple Timescales Recurrent Neural Network.

nCVaR . . . . nested Conditional Value at Risk.

Neural ASM . Neural Associative Skill Memory.

NeuroAI . . . A research field combining Neuroscience and Artificial Intelligence.

PAL . . . . . Pavlovian Avoidance Learning.

pCVaR . . . . precommitted Conditional Value at Risk.

POMDP . . . Partially Observable Markov Decision Process.

PPE . . . . . Punishment Prediction Errors.

PWLC . . . . Piecewise Linear and Convex.

RL . . . . . Reinforcement Learning.

RLDDM . . . Reinforcement Learning Diffusion Decision-Making.

RNN . . . . . Recurrent Neural Network.

RPE . . . . . Reward Prediction Error.

RW . . . . . Rescorla-Wagner.

S-to-M . . . . Sensory-to-Motor.

SM-to-SM . . Sensorimotor-to-Sensorimotor.

SNL . . . . . Substantia Nigra pars lateralis.

SR . . . . . Successor Representation.

SSDR . . . . . Species Specific Defensive Reactions.

TD . . . . . Temporal Difference.

TDRL . . . . . Temporal Difference Reinforcement Learning.

TD-RPE . . . Temporal Difference-Reward Prediction Error.

TPE . . . . . Threat Prediction Error.

tPC . . . . . temporal Predictive Coding.

TS . . . . . Tail of the Striatum.

tDCS . . . . . transcranial Direct Current Stimulation.

Vglut2 . . . . . Vesicular glutamate transporter 2.

VR . . . . . Virtual Reality.

VTA . . . . . Ventral Tegmental Area.

WAIC . . . . . Watanabe-Aikake Information Criterion.

WASP . . . . . Surface electrode for electrodermal stimulation.

Through chances various, through all vicissitudes, we make our way....

— *Aeneid*

1 Introduction

Contents

1.1 The Survivor's Dilemma

The notion of 'learning safely' appears paradoxical at first glance. Learning from experience necessarily involves making mistakes, yet some mistakes, such as severe injury or death, can be so catastrophic that they preclude further learning altogether. This raises a fundamental question for any intelligent agent, biological or artificial: how to learn from experience while minimising the risk of potentially irreversible harm?

Addressing this dilemma requires two seemingly complementary capabilities. The first is *safe exploration*: the ability to act cautiously in uncertain environments to gather information while avoiding external threats. The second is *self-preservation*: the capacity to detect, represent, and respond to internal threats, such as injury

in an animal or a component fault in a robot. While safe exploration concerns external dangers, self-preservation focuses on maintaining internal integrity.

Throughout this thesis, these two pillars form the core of what we term *safe learning*. They represent a fundamental challenge for survival that has shaped the evolution of biological control systems and is now a frontier in the development of robust, autonomous artificial intelligence. This work argues for a bidirectional Neuro-AI approach: studying the sophisticated solutions found in nature, particularly through the lens of pain neuroscience, can yield computational principles for safer AI, while computational modelling can, in turn, refine our understanding of pain systems in biological organisms.

1.2 Perspectives on Safety: From AI to Neuroscience

The problem of safe learning is approached differently across disciplines. In artificial intelligence, research has focused primarily on safe exploration, designing agents that balance exploration and exploitation while avoiding costly states (García et al., 2015). For robotic systems, or 'physical intelligences', the problem is compounded: they must confront not only external hazards but also internal failures, necessitating built-in mechanisms for self-preservation, a problem involving decision-making under uncertainty. A robot's ability to monitor its own structural integrity and adapt to damage is often critical for success in high-stakes environments like the DARPA Robotics Challenge[1].

Economists on the other hand, are less concerned with bodily integrity. In behavioural and neuro-economics, decision-making under uncertainty is often viewed through the lens of bounded rationality (Simon et al., 1972) and prospect theory (Kahneman, 1979). While these frameworks have advanced our understanding of

[1]DARPA Robotics Challenge (DRC) was a prize competition funded by the US Defense Advanced Research Projects Agency. It aimed to develop semi-autonomous ground robots that could do "complex tasks in dangerous, degraded, human-engineered environments." It included tasks such as travelling dismounted across rubble, removing rubble, climbing industrial ladders etc, which couldn't be solved reasonably well by a robot without maintaining structural integrity and accounting for uncertainty.

phenomena like loss-aversion (Litovsky et al., 2022) and risk-aversion (Christopoulos et al., 2009; Zhang et al., 2014), they typically collapse gains and losses into a single scalar learning signal. Such treatments may fail to capture the deep qualitative distinctions between appetitive and aversive outcomes, particularly when the latter involves a threat to bodily integrity.

Pain neuroscience offers a more biologically grounded and computationally nuanced perspective. As reviewed by Seymour (2019), control systems for responding to harm are multi-layered, spanning rapid spinal reflexes (Ellrich and Hopf, 1996), hardwired innate responses (Bolles, 1970; Fanselow and Lester, 2013), learned Pavlovian withdrawal (Bolles, 1972; Mackintosh, 1983), and more sophisticated instrumental (Overmier, 1979; Mackintosh, 1983) and model-based control (Tolman, 1948; Dayan and Daw, 2008). Crucially, the motivational components of the pain system do not merely mirror reward circuits in the negative domain; they have evolved specialised, highly asymmetric mechanisms to prioritise survival (Seymour, 2019; Zimmerman et al., 2025). The first part of this thesis builds on this insight, using pain neuroscience to model safe exploration as a multi-objective control problem (also referred to as multi-attribute decision making).

Turning from external threats to bodily integrity, the problem of representing and acting upon bodily states like injury is poorly understood. While computational models of homeostasis exist (Keramati and Gutkin, 2014), they often assume direct access to physiological states like hunger or temperature. However, the brain typically operates with incomplete and noisy information about its own body (Seth, 2013; Seymour and Mancini, 2020; Allen et al., 2022). This thesis explores the hypothesis that such uncertainty is central to post-injury decision-making and the transition to chronic pain (Seymour et al., 2023b), and also in translating these ideas to fault-detection in neuro-inspired artificial systems.

1.3 Aims and Core Hypotheses

This thesis advances our understanding of safe learning by addressing two central hypotheses, which form the basis for Parts I and II of the work:

1. **Safe exploration is achieved through the flexible composition of multiple value functions.** We hypothesise that the brain's layered control architecture is a sophisticated solution to the safety-efficiency dilemma. First, we propose that seemingly suboptimal Pavlovian reflexes are not simply a vestige of evolution. They provide a safety layer at the cost of efficiency in accruing rewards, by allowing innate responses to be activated in advance of a harmful stimulus, offering the chance to prepare for, reduce, or even completely avoid it (Bolles, 1972; Mackintosh, 1983). Further, the safety-efficiency trade-off arising between this system and instrumental control is resolved by flexibly gating Pavlovian influence based on outcome uncertainty.

   Second, noting the suboptimality that can still arise from composing control systems of varying capabilities, we further hypothesise that the dopaminergic system (DA) implements a mechanism for the *optimal* composition of multiple values (of roughly similar capability but optimising different objectives). We achieve this by redefining DA's normative objective to be the optimisation of returns augmented by a policy-dependent penalty term. This single alteration not only unifies disparate findings challenging the temporal difference-reward prediction error (TD-RPE) hypothesis (Schultz et al., 1997) but also enables learning that is efficient, safe and stable. Safe learning is mediated via flexible expression of pessimistic value initialisation (e.g. avoiding a novel object), whereas efficient acquisition of multiple rewards is mediated via off-policy fast adaptation to changing priorities (non-stationary rewards).

2. **Self-preservation is a problem of inference and control under uncertainty.** We hypothesise that the brain represents and responds to injury not as a simple sensation, but as a problem of partially observable state estimation to drive optimal behaviour. We propose that post-injury behaviour can be formally modelled as a Partially Observable Markov Decision Process (POMDP), providing a computational account for counter-intuitive behaviours such as rubbing or probing an injured area so as to gain information and

reduce uncertainty, and it accounts for the high propensity of an injury to transition into a pathological, chronic pain state via information restriction (Seymour et al., 2023b).

In a subsequent chapter, we further hypothesise that these principles of self-monitoring can be translated to machines, where self-supervised predictive coding can allow robots to form an 'injury state' to detect faults and adapt behaviour (though our claims and simulations here are limited to robots performing stereotypical movements, and as we build upon the framework of Associative Skill Memories (Pastor et al., 2012, 2013), using learning from demonstrations).

1.4 Thesis Overview and Contributions

The chapters of this thesis are organised into three parts. After two background chapters, Part I addresses safe exploration, and Part II addresses self-preservation.

Background

Chapter 2: A Computational Approach to Pain

This chapter provides a perspective on the need for computational approaches to understand the multi-level complexity of pain. It reviews how reinforcement learning, control theory, and Bayesian inference can formalise concepts of pain prediction and avoidance. The central contribution is to frame the synergy between theory and experiment as a process of 'forward and reverse engineering', which provides a conceptual foundation for the rest of the thesis.

Chapter 3: A Review of Safe Learning

This chapter reviews the broader interdisciplinary literature on safe learning from both artificial intelligence and neuroscience. It leverages the taxonomy from computer science (García et al., 2015) to structure the review, highlighting how computational concepts have been used to explain avoidance and anxious behaviour (AI $\rightarrow$ Neuro). Further, the chapter particularly highlights how ideas from

neuroscience can help develop safer intelligent agents in instances, where computer scientists have previously used heuristic approaches (Neuro → AI). This situates the specific contributions of the thesis within the wider field of safe learning and equips the reader with the necessary background concepts to follow the rest of the thesis.

Part I: Safe exploration

Chapter 4: Gating Pavlovian Fear with Uncertainty

This chapter identifies and investigates the safety-efficiency dilemma that arises from the conflict between instrumental control and Pavlovian reflexes. We show that the Pavlovian system, rather than being simply maladaptive, can be flexibly arbitrated based on outcome uncertainty to help in safe exploration. The contribution is a model that promotes both safe and efficient learning, which we test in simulation and validate in a human approach-withdrawal virtual reality (VR) experiment, demonstrating a sophisticated role for the fear system in shaping exploratory behaviour.

Chapter 5: Optimal Composition of Multiple Values

Noting the suboptimality that can arise from composing control systems of varying complexity, this chapter proposes a normative framework for how the brain might optimally compose multiple value functions. The contribution is a novel reformulation of the dopaminergic system's objective, optimising not just cumulative rewards, but returns augmented by a penalty term that quantifies deviations from a default behavioural policy. We unify disparate and conflicting observations of striatal dopamine signals (Menegas et al., 2018; Akiti et al., 2022; Greenstreet et al., 2025), parsimoniously explaining phenomena related to safety, efficiency, and stability in learning.

Part II: Self-preservation

Chapter 6: A POMDP Model of Injury

This chapter addresses the problem of representing internal states such as injury. Homeostasis involves tracking deviations in internal state from homeostatic grace

to re-organise behaviour. However, the brain may not have direct access to internal states, but requires inferring them from noisy observations. The central contribution of this chapter is to formalise post-injury homeostasis as a Partially Observable Markov Decision Process (POMDP). This framework provides a quantitative explanation for counterintuitive behaviours, such as investigating an injury despite the pain, and offers a mechanistic account of how information restriction can lead to the transition from acute to chronic pain. This computational account complements the qualitative account of transition to chronic pain provided by the influential Fear-Avoidance model (Vlaeyen and Linton, 2000; Vlaeyen et al., 2016).

Chapter 7: Self-Supervised Fault Detection in Robots

This chapter extends these principles of self-preservation to neuro-robotics. The contribution is the Neural Associative Skill Memory (Neural ASM), a framework that uses self-supervised predictive coding networks (using local learning rules) to build an internal model of expected sensorimotor experience. We demonstrate that this biologically plausible architecture can unify skill learning, fault detection, and adaptive control, enabling a robot to develop an analogue of an 'injury state' to respond to unexpected deviations. Minor faults can be reactively corrected on the fly, whereas in the case of major faults, the robot can halt its operations and request for assistance. The proposed model is limited to learning from demonstrations of robots performing stereotypical movements, but inspires future goal-directed methods. Additionally, the computational model further explains how contextual inference underlies skill memory separation and expression (Sheahan et al., 2016).

Finally the last chapter (**Chapter 8**) discusses the significance of thesis contributions, limitations and future work.

Part I and II chapters start with a brief *prelude* paragraph, often to include additional motivations for the work which may not be included in the main text of the chapter or the corresponding paper.

1.5 Scope and Impact

This thesis addresses the broad question of safe learning by focusing on two specific, foundational mechanisms, though they form a fraction of the myriad safe behaviours that keep us alive. The study of computations underlying other safe behaviours such as zero-shot escapes, model-based risk aversion, prey-predator interactions, vicarious learning from observations etc, though essential for our survival are beyond the scope of this thesis.

The contributions of this thesis have two main (tempered) implications. First, in clinical neuroscience, this work aims to provide a new way of thinking about the transition to chronic pain. A reasonable contribution can be to complement previous qualitative theories by making certain concepts computationally explicit. In doing so, the computational frameworks developed here have the potential to inspire novel diagnostic and therapeutic approaches. For instance, this has been particularly highlighted in the eLife digest accompanying Chapter 4.

Second, in artificial intelligence, this work contributes to the development of safer and more robust autonomous agents. The questions asked in this thesis are central to understanding the 'mind' and the computational processes underlying intelligence (Sutton et al., 2022). By borrowing meta-level insights from neuroscience (Hassabis et al., 2017) on how to handle threat, uncertainty, and internal state monitoring, this research opens avenues for novel, bio-inspired approaches to safe AI, focusing on scalable principles rather than direct mimicry, aligning with 'The Bitter Lesson' (Sutton, 2019)[2] .

Together, this work advances our understanding of safe learning, a challenge shared by humans and machines, with implications for both the study of the mind and the engineering of intelligent systems.

[2]The Bitter Lesson is an influential essay by Prof. Sutton which posits that making systems that worked the way the researchers thought their own minds worked—trying to put that knowledge in their systems—has proven ultimately to be counterproductive and a colossal waste of researcher's time. One should instead draw inspirations from nature and builds systems that learn and build new knowledge from its own experience, the same way humans do rather than bake in the domain knowledge that humans have gathered. This approach scales gracefully with compute.

Is there anything more intimate or more internal than pain?

— René Descartes, Meditations on First Philosophy

2

Forward and reverse engineering the pain system

Contents

2.1 Why do we need a computational approach to understanding pain?

Theories play an important role in driving neuroscience research by shaping hypothesis generation and constraining the search space for experiments. Computational models bridge theories and observable behaviour, serving as mathematical structures that represent observed data (Levenstein et al., 2023). They sit in direct contrast to a purely data-driven approach, which aims to discover structure in data alone (Gershman, 2021). Thus, computational models offer an intermediary between behavioural theories (e.g., psychology) and physiological accounts (e.g., molecular,

cellular, or circuit neuroscience).

Pain can be studied through the lens of its evolutionary and phenomenological components, suggesting a framework based on its adaptive function operating across different spatiotemporal scales. Reaching a common understanding of the data from dramatically different modalities and methodologies requires a multi-level approach, which is afforded by different types of computational models providing different explanations, namely, descriptive, mechanistic and normative explanations. Each of these play distinct roles in building a multi-level account of neural and behavioural phenomena. That is, whenever a new phenomenon is observed, it is often first explained descriptively delineating what problem is the brain solving, for example see (Cecchi et al., 2012). This is often followed by mechanistic and normative (or interpretative) explanations delineating how is the brain solving it and why is the brain solving this problem (Dayan and Abbott, 2001). We structure the rest of this chapter around the past, present and future of computational modelling in pain research.

2.2 The advent of computational frameworks for pain

Computational models of pain trace back to early models of learning and conditioning, with roots in classical animal learning theory and its foundational paradigms of Pavlovian and instrumental conditioning (Mackintosh, 1983). By the late 1990s, theorists developed computational descriptions of these learning phenomena to provide unifying explanations of a variety of behavioural and neural data: most notably the reinforcement learning model of dopamine and reward (Schultz et al., 1997). Reinforcement learning (RL) is a framework for learning a behavioural policy that maximises cumulative reward (or minimise punishments) over time. This approach was soon extended to explain simple predictive pain learning (Seymour et al., 2004), aligning with fear and aversive conditioning (Johansen et al., 2010), and affective-motivational theories of pain (Fields, 2006, 2018). Central to these models was the concept of the prediction error: the difference between what

was predicted and what actually occurred, which works as the central teaching signal in these models. RL showed how prediction errors could solve difficult prediction problems (Sutton et al., 1998) across a broad range of learning phenomena involving credit assignment, optimal action learning, and the architecture of pain avoidance learning. Notably, computational parametric human neuroimaging showed that structures such as the striatum and ventromedial prefrontal cortex that had traditionally been thought of as reward-focused, were clearly also involved in pain (Seymour et al., 2005).

In parallel, models of sensory perception, particularly Bayesian perceptual inference offered insights into how prior beliefs shape pain experience, especially in placebo and nocebo effects (Büchel et al., 2014; Hoskin et al., 2019; Jepma et al., 2018; Kim et al., 2024; Seymour et al., 2013; Tabor et al., 2017; Wiech, 2016). These showed how expectations alter pain through perceptual biases, offering a normative account of the sensory-discriminative aspects of pain. However, such perceptual accounts differ from motivational models of endogenous analgesia studied in animals, where pain modulation reflects a flexible, value-based decision that balances competing goals like pain avoidance and reward seeking (Dum and Herz, 1984; Fields, 2006). These motivational models draw from the idea of value shaping, in which the perceived value of outcomes is not fixed but flexibly constructed to optimise decisions on-the-fly, paralleling theories in behavioural economics (Vlaeyen and Crombez, 2020). This approach indeed explained a number of types of endogenous analgesia beyond simple perceptual biases; showing that pain was shaped by both perceptual and motivational factors, which was also implied by the anatomical complexity of descending pathways (Bannister and Dickenson, 2017).

2.3 Theory considerations

Here we outline a conceptual map of computational approaches relevant to pain research, offering readers multiple entry points for future modelling efforts. At its core, the brain (and pain system) must solve a difficult control problem under uncertainty — selecting protective actions based on beliefs about the state of

the body and world. One influential way to formalise this is by viewing it as a partially observable Markov decision process (POMDP) (Kaelbling et al., 1998; Mahajan et al., 2025a), for which Bayesian Decision Theory provides one principled solution (Dayan and Daw, 2008), and aids as an organising map for understanding the landscape of various computational models. It requires inferring the states or beliefs over states from observations, which is then used to learn and direct optimal behaviour with respect to some utilities.

Models focusing primarily on state inference, without control, typically employ Bayesian models of perception (Büchel et al., 2014; Tabor et al., 2017) or special cases of POMDPs such as drift-diffusion models (Wiech, 2016). These reveal how pain perception is shaped by expectations and their modification through learning, and help quantify such effects (Wiech, 2016; Zaman et al., 2020). Recent work extends them to account for biased pain perception (e.g., confirmation bias) (Jepma et al., 2018) and to model sub-optimal learning and information integration in chronic pain (Vlaeyen, 2015).

Models emphasising control, while assuming full state observability, simplify the problem to a Markov decision process (MDP) and use reinforcement learning to solve it (Sutton et al., 1998). Leveraging methods from safe reinforcement learning (García et al., 2015), these models reveal how pain-related control systems guide cautious and self-preserving behaviour (Seymour et al., 2023b), inform safe exploration (Mahajan et al., 2024) and are being adapted to model maladaptive avoidance (Ball and Gunaydin, 2022) and higher punishment sensitivity (Mancini et al., 2025) in chronic pain.

When both inference and control are necessary, full POMDP solutions are required. Some use explicit models of the environment to infer beliefs and plan in belief space (belief MDPs) (Lovejoy, 1991) while others use function approximators like recurrent neural networks to learn belief-like representations through reinforcement learning (Hennig et al., 2023). These Bayesian decision theoretic models reveal how protective behaviours may obscure evidence of recovery, promoting persistent injury

beliefs and transitions to chronic pain - a hypothesis formalised as the information restriction hypothesis (Mahajan et al., 2025a; Seymour et al., 2023b).

Early frameworks such as the Helmholtz machine (Dayan et al., 1995) and predictive coding models (Friston, 2005) further provide a recipe for learning and inference in neural networks, informing recent theories of plasticity and back-propagation in the brain (Whittington and Bogacz, 2019) and implementation-level accounts of placebo analgesia (Büchel et al., 2014). Related approaches unified planning and inference, as in Active Inference, where utilities are replaced with preference priors and the agent performs Bayes-optimal behaviour (Sajid et al., 2021). An algebraic mapping between BDT approaches and active inference approaches is present in control-as-inference (Millidge et al., 2020), KL-control (Broek et al., 2012) or soft Bellman equations (Ziebart, 2010).

Framing the pain system's challenge as a POMDP allows researchers to map different aspects of pain to components such as belief states, value functions, predictive state representations, expected utilities, prediction errors, or Bayesian surprise, facilitating clearer and more testable hypotheses (Coll et al., 2024; Hoskin et al., 2019; Ishikawa et al., 2025; Mahajan et al., 2025a; Tabor et al., 2017) (Fig. 2.1).

2.4 Forward and reverse engineering

In pain research, forward engineering builds models from first principles to simulate behaviour and generate predictions, while reverse engineering infers mechanisms by testing specific hypotheses on existing systems (e.g., through experiments). In practice, one can test competing models against behavioural or neural data using techniques like Bayesian model comparison (Stephan et al., 2009). These findings can then inform forward-engineered models that explain multiple observations across datasets. Such models simulate neural and behavioural responses in diverse paradigms, generating predictions that serve as testable hypotheses. Better data analysis methods and experiments that clearly separate one computation improve reverse engineering attempts, whereas better theories guide what and where to search for and improve forward engineering attempts (Jonas and Kording, 2017).

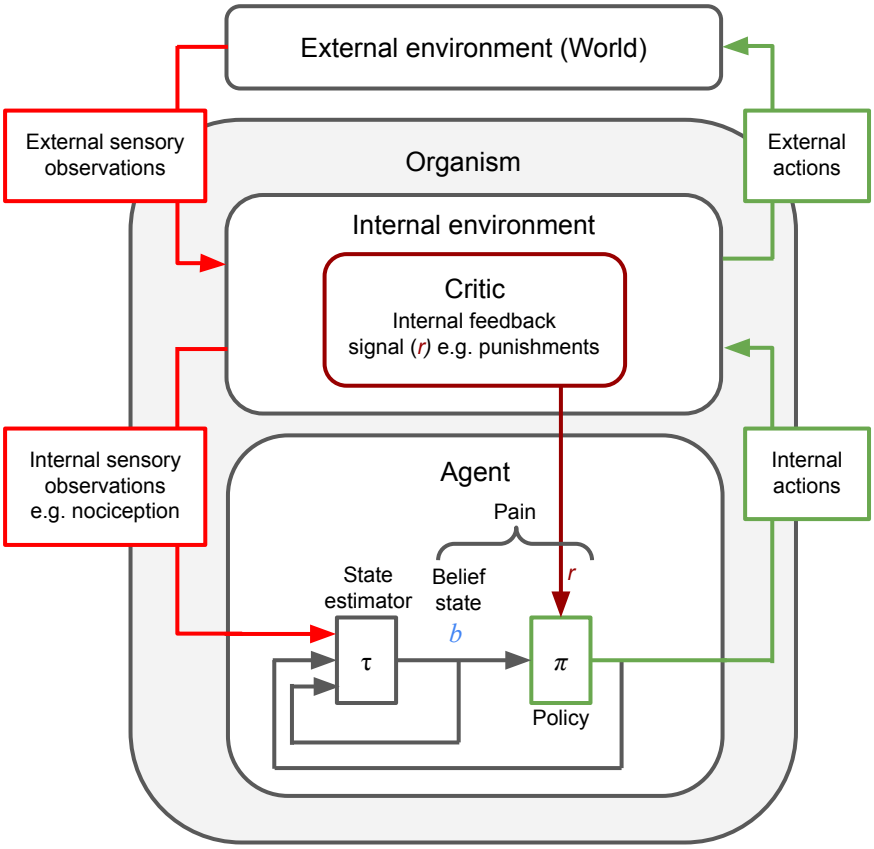


Figure 2.1: An illustration of computations underlying pain: An internal environment POMDP which is solved using Bayesian Decision Theory (BDT) by the agent, necessitating the elements of pain in the mind. The concept of an internal environment, including a critic for providing an internal feedback signal follows from Singh et al. (2009), with the only difference being that we treat the internal environment as a POMDP rather than an MDP. The sensory processing of pain is captured via the Bayesian filtering of nociceptive and other relevant sensory observations. The motivational aspect of pain is captured through the precise punishment signal generated and used for temporal credit assignment. Both play a role in affective value learning and subsequent action policy to drive protective behaviour (e.g. pain avoidance).

An example is provided in Chapter 4, where we studied how humans combine Pavlovian and instrumental responses to minimise pain/harm (Mahajan et al., 2024). Forward simulations showed that adding Pavlovian fear biases to withdraw enhanced safe learning but at the cost of efficiency in reward acquisition. This led to a normative hypothesis: uncertainty-gated Pavlovian responses could balance safety and efficiency effectively, as demonstrated in simulations. We then tested this prediction in a behavioural task using computational modelling. This illustrates the interplay between the constructivist model-building and the reductionist hypothesis

testing through carefully designed experiments.

2.5 Clinical applications and neuro-engineering

This approach offers promise for clinical translation. First, forward-engineered models of pain chronification (Mahajan et al., 2025a) provide principled hypotheses for predicting outcomes and computomics (Lee and Seymour, 2019) or computational phenotyping (Patzelt et al., 2018). This can support stratification with more robust outcome-prediction biomarkers, with fewer data than brute-force machine learning.

Second, forward-engineered models can be used to design personalised task controllers — interactive agents that guide a patient's learning process in cognitive behavioural therapies (CBT)-like settings (Lee et al., 2024a). Framed as a two-player game, the controller interacts with a forward-engineered model to best alter the patient's generative model by guiding the learning process toward desired outcomes such as improved motivation, reduced avoidance, or resilience to stress. This builds on ideas from curriculum learning (Bengio et al., 2009) and shaping (Krueger and Dayan, 2009).

Third, biases in pain and injury perception may affect post-injury behaviour (Vlaeyen, 2015), which could contribute to phenomena like boom-bust cycles in chronic pain (Antcliff et al., 2016; Moseley, 2003). Personalised models of temporal patterns in activity and pain perception could enhance activity pacing interventions by offering more targeted routine guidance.

Fourth, information-rich models beget information-rich solutions, such as functionally targeted neuro-technologies (Denison and Morrell, 2022). This dovetails with an understanding of how generative processes lead to biomarkers, allowing closed-loop technologies that tune the intervention to the outcome. Ultimately, systems engineering approaches can integrate multiple sensing and interventional technologies (alongside drugs), utilising cognitive and behavioural platforms, yielding holistic therapies that exploit the combined strength of multiple individual approaches.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— *Johann Wolfgang von Goethe*

3

An overview of safe learning and its connections to neuroscience

Contents

3.1 Introduction

In a foundational review, García et al. (2015) provided a principled taxonomy of safe reinforcement learning (RL) methods, distinguishing between strategies that (1) modify the **optimisation criterion** to embed safety directly into the learning objective, and (2) modify the **exploration process** to prevent dangerous actions. This chapter adopts and extends their structure, connecting it to findings in behavioural and systems neuroscience. We show how modifying the optimisation

criterion, as done in many machine learning approaches, has significantly informed our models of animal behaviour and neural function. Conversely, theoretical and experimental neuroscience offers valuable strategies for how agents modify their exploration processes to ensure safe exploration. These include the use of initialisation, innate priors, models of interoception, meta-control mechanisms that regulate early learning and so on.

Such computational approaches are increasingly relevant to neurobehavioural modelling and human computational psychiatry (Huys et al., 2015), particularly in domains involving fear, anxiety, panic, avoidance, aggression, and rumination (Perusini and Fanselow, 2015) and in applying principles of computational psychiatry to computers or intelligent machines (Schulz and Dayan, 2020).

Significant prior work has mapped these computations along the threat-imminence continuum preceding a predatory encounter, varying with the spatial and temporal proximity of danger (Fanselow and Lester, 2013; Mobbs et al., 2020). This review chapter does not explicitly focus on prey-predator interactions, though they contribute a significant portion of safe behaviours. This review instead focuses on learning from generic external harms (treated as losses or punishments) and the complex continuum of safe behaviours that unfold after trauma, such as an injury. This post-trauma domain, in contrast to the well-studied threat-imminence continuum, represents a significant and underexplored area for computational neuroscience.

3.2 Learning through reinforcement

Reinforcement learning (RL) provides a mathematical framework for agents to learn how to act in uncertain environments, using feedback from rewards or punishments. At its core, the agent observes a state, takes an action, and receives a reward and a new state from the environment. Over time, the agent aims to learn a policy, a mapping from states to actions, that maximises its cumulative (discounted) reward, also called the *return* (Sutton and Barto, 2018).

Formally, these problems are often modelled as Markov Decision Processes (MDPs), where each action taken in a state leads probabilistically to a new state

and a reward. The *return* from time t is typically defined as the discounted sum of future rewards: $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, where r_{t+k+1} is the reward received upon taking an action k timesteps after t and $\gamma \in [0, 1]$ is a discount factor that shrinks the value of distant rewards (Sutton and Barto, 2018). Central to RL is the *value function*, which quantifies the expected return from a state (or state-action pair) under a given policy. These value functions guide the agent's decisions, and the methods for learning them reveal a fundamental spectrum of control strategies with deep parallels in neuroscience.

At one end of this spectrum are *model-free (MF)* methods, such as TD-learning (Sutton, 1988). TD-learning updates a state-value function, $V(s)$, based on the temporal difference between successive value estimates under a behavioural policy. The update is driven by the TD-error, δ_t:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{3.1}$$

$$V(s_t) \leftarrow V(s_t) + \alpha \delta_t \tag{3.2}$$

where α is a learning rate. These TD-errors are equivalent to the 'reward prediction error' (RPE), a foundational concept in computational neuroscience strongly linked to phasic dopamine signals in the ventral tegmental area (VTA) (Schultz et al., 1997). As an on-policy algorithm, TD-learning estimates the expected return under the current behavioural policy. However, this means it often learns suboptimal values under a policy that still explores. The limitations of this approach are discussed further in Chapter 5.

These limitations of on-policy algorithms are often solved by off-policy algorithms, which learn values under a different target policy while collecting data under a potentially sub-optimal behavioural policy. A canonical example of an off-policy algorithm is the Q-learning algorithm (Watkins, 1989), which learns state-action values (referred to as Q-values, e.g. $Q(s, a)$) under a greedy target policy. This means the Q-values are updated under the assumption that the agent will choose the greedy action (i.e. the action with the maximum Q-value) at the next state, while computing the TD-error. Mathematically, this means after taking action a_t

in state s_t and observing reward r_{t+1} and next state s_{t+1}, the Q-value is updated using a TD error that assumes a greedy action will be taken in the next state:

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \tag{3.3}$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta_t \tag{3.4}$$

where α is the learning rate. This algorithm is known to converge to optimal state-action values (Watkins and Dayan, 1992).

At the other end are *model-based (MB)* approaches, which build an explicit model of the environment's transition probabilities ($T(s'|s, a)$) and reward structure ($R(s, a)$). This model allows the agent to plan by simulating future trajectories, making it more sample-efficient and flexible. The dichotomy between MF and MB control parallels the distinction between 'habitual' and 'goal-directed' behaviours in psychology (Dickinson and Balleine, 2002), where the two systems often compete for behavioural control, with arbitration guided by uncertainty (Daw et al., 2005).

Many architectures seek a middle ground. *Hybrid approaches*, such as Dyna (Sutton, 1991), integrate both strategies: a learned model generates simulated experiences to train MF value learning, a process recently mapped to the function of hippocampal replays (Mattar and Daw, 2018). The Successor Representation (SR) offers another compromise, learning predictive relationships between states to capture model-based-like flexibility using simple MF learning rules (Dayan, 1993; Momennejad et al., 2017). As we will explore in Chapter 5, related multi-objective RL architectures also manage to provide SR-like flexibility when it comes to fast adaptation or instant revaluation (Millidge et al., 2024a).

However, not all biological control neatly fits this categorisation. Even simpler mechanisms, such as hardwired innate reflexes and learned reflexes such as Pavlovian withdrawal biases, form a crucial first line of defense, contributing to safety long before instrumental values are learned (Seymour, 2019; Mahajan et al., 2024). At the same time, more sophisticated strategies like *episodic control*, a potential hippocampal contribution, may dominate the earliest stages of learning by leveraging

memory of specific past events, before gradually ceding control to MB and eventually MF systems (Lengyel and Dayan, 2007).

The existence of these multiple controllers points to a key organisational principle: meta-control, or arbitration between simpler, faster systems and more complex, deliberative ones (Lee and Seymour, 2019), echoing Kahneman's "thinking, fast and slow" (Kahneman, 2011). Various forms of uncertainty are thought to drive this arbitration, guiding the switch between MF and MB control (Daw et al., 2005; Lee et al., 2014). In Chapter 4, we will propose a arbitration mechanism where outcome uncertainty mediates the balance between a Pavlovian fear system and an instrumental avoidance system.

A compelling example application of these RL frameworks to safe learning is the organisation of defensive behaviour under threat.Mobbs et al. (2020) map these varied computational strategies onto the *threat imminence continuum* (Fanselow and Lester, 2013), which posits that different defensive systems are recruited based on the perceived proximity of danger. This model provides a neuro-computational account for distinct forms of anxiety. For instance, distant, potential threats may engage deliberative, model-based planning, corresponding to *intermittent anxiety* (contemplating future dangers) or *anticipatory anxiety* (adopting precautionary behaviours). As a threat becomes more proximal, control may shift towards faster, more reactive systems. During *encounter anxiety*, where a threat is present but may not have detected the agent, a mixture of cautious instrumental and Pavlovian systems might dominate. Finally, in a *circa-strike* scenario, where danger is immediate and unavoidable, behaviour is likely governed by hardwired, reflexive responses akin to a simple, fast-acting policy, culminating in panic or active escape. Each mode of anxiety can thus be interpreted as the activation of one or more RL regimes, flexibly tuned to the imminence of the threat. This dynamic interplay between reactive and deliberative control is central to biological survival and a primary goal for the design of safe artificial agents.

3.3 Learning through other means and possible forms of innate knowledge

Beyond learning from direct reinforcement, an agent's behaviour is profoundly shaped by innate structures and prior knowledge. In his foundational thesis, Watkins (1989) presciently outlined six categories of innate 'knowledge' that constrain and guide learning, beyond the well-known contribution of inventing Q-learning and connecting RL to Markov decision processes (MDPs). This section adopts his framework to explore these alternative learning mechanisms, from hard-wired reflexes and subjective reward systems to the powerful influence of learning from demonstration and vicarious social experience. The six types of innate characteristics that shape learning are (Watkins, 1989, Chapter 8):

1. physical capacities
2. subjective reward systems
3. methods of representation and approximation
4. initial policies, value functions, and models
5. bounds and constraints on policies, value functions and models
6. tendencies to experiment (or investigate)

The contributions of this thesis systematically engage with five of these six categories:

- **Subjective reward systems (2):** Explored in Chapters 4 and 5 through the lens of multi-attribute rewards and punishments.
- **Methods of representation (3):** A cross-cutting concern that influences all models presented.
- **Initial policies, values, and models (4):** Addressed via Pavlovian biases in Chapter 4 and through models learned from demonstration in Chapter 7.

- **Bounds and constraints on policies (5):** Leveraged in Chapter 5 to encourage stable learning by weakly constraining policies near a known default policy.

- **Tendencies to experiment (6):** Chapter 6 provides a normative account of this, explaining information-seeking behaviours like probing an injury.

While this thesis does not explore variation in physical capacities (1), it is an important factor, as an agent with well-adapted physical traits (e.g. a bird with a well-adapted beak) may learn a skill more quickly than one without.

Learning from demonstrations provides a powerful alternative to trial-and-error learning. An agent may learn an expert's policy directly via behavioural cloning (Havoutis and Calinon, 2019), or use demonstrations to bootstrap RL algorithms through pre-training and value initialisation (Hester et al., 2018). In Chapter 7, this thesis presents a method for learning an energy-based model of sensorimotor repertoires from demonstrations. These can function as associative skill memories (Pastor et al., 2013; Mahajan et al., 2025b), which can be used for abnormality detection—a process analogous to injury detection—or even to construct primary rewards from the resultant energy functional, similar to drive-reduction theories (Keramati and Gutkin, 2014; Hulme et al., 2019).

Finally, behaviour need not be innate or learned; it can be *innately learned* (Watkins, 1989, Chapter 8). Vicarious social learning is a leading contributor to safe behaviour, as it is far more beneficial to learn from observing others' mistakes than to experience the consequences oneself. Such social learning can be implemented through various mechanisms, including value-free policy updates from observed actions (Burke et al., 2010), model-free value updates from observed outcomes (Burke et al., 2010), inverse RL to infer others' preferences (Collette et al., 2017) or model-based updates to state-transition models. While this thesis focuses on single-agent experiential learning, it is important to acknowledge that vicarious learning represents a powerful mechanism for safety in multi-agent contexts.

3.4 A taxonomy of safe reinforcement learning methods

Having established the foundations of RL and the role of innate knowledge, we now turn to the core of safe RL. A central contribution of García et al. (2015) is a structured taxonomy that provides a principled way to organise safe RL approaches. They are broadly divided into two complementary categories:

1. **Modifying the optimisation criterion:** These methods embed safety directly into the learning objective itself, for instance by incorporating risk, worst-case outcomes, or explicit constraints.

2. **Modifying the exploration process:** These methods control how an agent gathers experience during learning, aiming to prevent dangerous actions before they are taken. This includes leveraging priors, demonstrations, teacher advice or any ad-hoc heuristics designed by the engineer.

This review expands the latter category to prioritise insights from neuroscience, an aspect particularly overlooked in previous computer science reviews.

While conceptually distinct, these two axes are often intertwined. The following two sections will explore each category in detail, connecting the formalisms from computer science with their analogues in behavioural and systems neuroscience.

3.5 Modifying the optimisation criterion, with applications to behavioural neuroscience

The first major class of safe RL methods involves altering the learning objective to account for safety. Rather than simply maximising expected return, these approaches embed additional considerations such as risk terms, constraints, or worst-case outcomes directly into the optimisation landscape. In this section, alongside introducing different approaches to modifying the optimisiation criterion, we also review its application to behavioural neuroscience (also referred to as AI $\rightarrow$ Neuro).

Worst-case criteria

These methods prioritise robustness by maximising performance under the worst outcomes. For example, worst-case $\hat{Q}$-*learning* (Heger, 1994) optimises over worst-case returns, while β-*pessimistic Q-learning* (Gaskett, 2003) interpolates between best and worst-case scenarios:

$$Q_\beta(s_t, a_t) \leftarrow Q_\beta(s_t, a_t) + \alpha\Big[r_{t+1} + \gamma V_\beta(s_{t+1}) - Q_\beta(s_t, a_t)\Big] \tag{3.5}$$

where the interpolated value function is given by:

$$V_\beta(s_{t+1}) = (1-\beta)\max_a Q(s_{t+1}, a) + \beta \min_a Q(s_{t+1}, a) \tag{3.6}$$

The worst-case principle can also be extended to *model-based* settings to address uncertainty in the transition dynamics (parameter uncertainty). Robust MDPs (Tamar et al., 2013) assume the true model lies within an uncertainty set P, and define the robust value function as:

$$V^*(s) = \max_\pi \min_{p \in P} V^{\pi,p}(s) \tag{3.7}$$

where the agent maximises policy performance under adversarial environment dynamics.

Applications to behavioural neuroscience: Worst-case criteria provide a simple yet powerful lens for modelling maladaptive avoidance in anxiety disorders. Zorowitz et al. (2020) use β-*pessimistic Q-learning* (Gaskett, 2003) to formalise how pessimistic beliefs about future control can lead to pathological behaviour. They contrast this with standard Q-learning, where the assumption of optimal future action ($\max_a Q$) implies that the potential for future avoidance is protective. In their pessimistic model, a high β parameter represents a belief that future avoidance will fail. This single pessimistic assumption leads to an excessive propagation of fear to situations distant from the actual threat, providing a computational account for exaggerated threat appraisals, fear generalisation, and persistent avoidance. Zorowitz and colleagues demonstrate the model's explanatory power across several

paradigms, including approach-avoidance conflict and aversive pruning in planning (Huys et al., 2012). This framework further attempts to connect the computational mechanism of pessimism to the development of anxiety and depression, linking it to maladaptive beliefs about self-efficacy and control.

Risk-sensitive criteria

Broadly, there are two ways to incorporate risk-sensitivity in the optimisation criteria: (I) using a weighted sum of return and risk and (II) using conditional value at risk (CVaR).

(I) Weighted sum of return and risk

These methods augment the reward-maximising objective with a risk penalty, weighted by a hyperparameter β that balances the trade-off between return and caution:

$$\max_{\pi} \mathbb{E}_{\pi}[G] - \beta \cdot \omega(G) \tag{3.8}$$

Here, G is the return and $\omega(G)$ denotes a risk measure. $\beta > 0$ implies risk-aversion, whereas $\beta < 0$ implies risk-seeking preference. Depending on the algorithm, this can be instantiated as:

- the *variance of returns* (Howard and Matheson, 1972; Borkar, 2002), penalising uncertain outcomes. Here, the variance of returns term arises from the Taylor expansion of exponential utility functions, to the first order.

- *temporal-difference (TD) errors* (Mihatsch and Neuneier, 2002). Here, positive and negative prediction errors have asymmetric effects on learning – a concept utilised to design neuroscientific experiments (Niv et al., 2012).

- or the *probability of failure or termination in undesirable states* (Geibel and Wysotzki, 2005).

Applications to behavioural neuroscience: Risk-sensitive criteria, particularly those based on asymmetric impacts of prediction errors, provide a compelling account of how the brain learns about variance from experience. Niv et al. (2012) tested whether the neural correlates of reinforcement learning are sensitive to risk by having participants learn the values of stimuli with equal means but different variances. They compared a standard TD model, a model with non-linear utility, and a risk-sensitive TD (RSTD) model in which positive and negative prediction errors are weighted by different learning rates (η^+ and η^-). The RSTD model provided the best fit to subjects' choices, demonstrating that risk aversion correlated with a greater learning rate for negative prediction errors ($\eta^- > \eta^+$), and vice versa for risk-seeking. Crucially, they showed a tight neurometric-psychometric coupling: BOLD signals in the nucleus accumbens, which encoded prediction errors, could be used to extract learned stimulus values. The difference in these neurally-derived values for risky versus sure options directly correlated with subjects' behavioural risk preferences. This suggests that risk sensitivity is not just a feature of outcome evaluation but is integral to the learning process itself, implemented via asymmetric updating from positive and negative feedback.

However, defining risk purely as the variance of returns has a key limitation: it treats positive and negative uncertainty as equally undesirable. A more precise formulation of risk, particularly for safety, should focus specifically on the lower tail of the return distribution, a concept addressed by methods such as Conditional Value at Risk (CVaR).

(II) Conditional Value at Risk (CVaR)

CVaR (Morimura et al., 2012; Tamar et al., 2015) targets expected returns under the lower tail of the distribution of returns:

$$\text{CVaR}_\alpha(G) = \mathbb{E}[G|G \leq v_\alpha(G)] \tag{3.9}$$

where $v_\alpha(G)$ is the α-quantile of the return G, also known as the value-at-risk. Conceptually, by optimising CVaR of returns we can formalise an agent's tendency

to focus on the 'tail events' or the set of worst-case possibilities, rather than the full spectrum of outcomes.

When applied to sequential risk, CVaR comes in two flavours: precommitted CVaR (pCVaR) and nested CVaR (nCVaR). pCVaR, precommits to a level of risk at the very first choice; and nCVaR, which re-applies the same risk level at every step in a nested manner (Gagne and Dayan, 2021).

Applications to behavioural neuroscience: CVaR provides a formal method for modelling risk-sensitive behaviours that prioritise avoiding the worst-case outcomes, a hallmark of anxious phenotypes. Gagne and Dayan (2021) applied a CVaR-based model to a dataset of human choices in the two-step sequential decision task, which is traditionally used to study MB-MF interactions. They found that a substantial minority of participants were significantly risk-averse, and that this risk aversion could account for behaviours previously attributed to other cognitive mechanisms like perseveration or stickiness in actions. This suggests that a failure to account for tail-risk aversion can lead to a mischaracterisation of the underlying learning processes. Extending this, Gagne and Dayan (2022) connect CVaR-optimal policies to pathological avoidance behaviours and link risk-sensitive offline planning building upon Mattar and Daw (2018)'s theory of prioritised replay to the cognitive processes of worry and rumination. In animal behaviour, Shen and Dayan (2024) used a Bayes-adaptive MDP with the nCVaR objective to model individual differences in how mice explore a novel, potentially threatening, object. They showed that the spectrum of behaviour from 'brave' to 'timid' could be captured by variation in the CVaR risk-sensitivity parameter, α, along with the animal's prior beliefs about potential hazards. Together, these studies demonstrate that CVaR is a powerful tool for formally characterising individual differences across a range of risk-sensitive behaviours, from adaptive caution to maladaptive avoidance, in both humans and other animals. Perhaps of most translational importance, recently (Sui et al., 2023), formulate decision-making in the Balloon Analogue Risk Task (BART) (Lejuez et al., 2002), as a risk-sensitive exploration in Bayes-adaptive MDPs. Early

results show that the structure of stochasticity in the BART is such that pCVaR is more risk-averse than nCVaR in a single trial, for the same nominal risk.

Constrained criteria

Constrained MDPs (CMDPs) (Altman, 1999) maximise return subject to some expected cost constraints (c_t):

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t} \gamma^t r_t\right] \quad \text{s.t.} \quad \mathbb{E}_{\pi}\left[\sum_{t} \gamma^t c_t\right] \leq d \tag{3.10}$$

Used in Constrained Policy Optimisation (Achiam et al., 2017) and safety layer methods (Dalal et al., 2018).

Applications to behavioural neuroscience: While CMDPs are primarily an engineering framework, they offer a useful conceptual lens for understanding how biological agents manage competing objectives. Instead of solving a single, complex constrained optimisation problem, the brain may employ separate, interacting control systems. For instance, the model proposed in Chapter 4 of this thesis (Mahajan et al., 2024) can be interpreted in this light. In this framework, the instrumental system learns to maximise rewards, while a distinct Pavlovian fear system acts as a dedicated controller to (weakly) minimise constraint violations, such as cumulative pain. This multi-system architecture, where a specialised fear system enforces a soft constraint on the behaviour of a reward-seeking system, represents a neurally plausible mechanism for approximating the solutions to constrained optimisation problems.

3.6 Modifying the exploration process, with insights from behavioural neuroscience

The class of algorithms modifying the exploration process control how the agent explores during learning, aiming to prevent dangerous actions or states from being visited during learning. In this section, alongside introducing different approaches used by computer scientists to modifying the exploration process, we also review

how insights from behavioural neuroscience can inform these modifications (also referred to as Neuro → AI).

Risk-directed exploration

A clever approach to safe exploration is to bias the agent towards more predictable, or 'controllable', regions of the environment. Gehring and Precup (2013) propose a method to achieve this by defining a state-action pair's controllability, $C(s, a)$, as the negative of the expected absolute TD-error, $C(s, a) = -\mathbb{E}[|\delta_t|]$. A running estimate of this score is learned via a stochastic approximation:

$$C(s_t, a_t) \leftarrow C(s_t, a_t) - \alpha'(|\delta_t| + C(s_t, a_t)) \tag{3.11}$$

where $\alpha' = \beta\alpha$ is a learning rate for $C(s, a)$ with $\beta < 1$ and α the learning rate for Q-values. This controllability score is then used as an additive bonus (or as a malus) during action selection, $Q(s, a) + \omega C(s, a)$. A positive weight ω encourages the agent to seek out states with high controllability (i.e., low prediction error variance), effectively biasing exploration towards 'safer', more predictable parts of the state space.

Insights from behavioural neuroscience: This principle of using prediction error variance to guide safe exploration has a strong parallel in neuroscience. In Chapter 4 of this thesis, we propose that a loosely similar yet complementary computation (Mahajan et al., 2024). Pearce-Hall associability (Krueger and Dayan, 2009), serves as a meta-control signal that arbitrates between instrumental and Pavlovian control. Associability is a running average of absolute prediction errors and tracks outcome uncertainty over time. This state and action-independent measure contrasts with (Gehring and Precup, 2013)'s $C(s, a)$ which is a measure for state-action pairs. However, instead of being an additive bonus to a single controller's values, Pearce-Hall associability dynamically adjusts the influence of a specialised Pavlovian fear system. High outcome uncertainty (high associability) increases the weight of the Pavlovian system, enforcing cautious withdrawal behaviours. As the

agent learns and uncertainty resolves, control is ceded to the more efficient instrumental system. This represents a neurally plausible mechanism for implementing risk-directed exploration, where uncertainty signals linked to the amygdala and pgACC (Zhang et al., 2016, 2018) arbitrate between distinct, specialised controllers.

Providing initial knowledge

The most elementary method for biasing exploration is to choose an initialisation based on prior knowledge. This is particularly effective for ensuring safety in the early stages of learning, before the agent has gathered enough experience to make accurate value estimates. One common approach is to bootstrap the learning algorithm from a finite set of demonstrations provided by a teacher, termed apprenticeship learning (Maire and Bulitko, 2005). This biases the agent towards relevant regions of the state-action space from the outset, reducing the time spent in potentially dangerous random exploration. Another way is to meta learn a value initialisation from a range of environments (Wang et al., 2016).

Insights from behavioural neuroscience: This engineering principle of safe initialisation has a deep biological analogue in the form of innate priors and evolutionarily-shaped value functions. Animals are not born as blank slates; they possess innate behavioural repertoires and value systems that guide their first interactions with the world. For example, neophobia—the fear of novel stimuli—can be understood as a pessimistic value initialisation that promotes cautious avoidance. This aversive value initialisation has been linked to dopamine signals in the tail of the striatum (TS), which is hypothesised to function as a separate threat-learning system (Menegas et al., 2018; Akiti et al., 2022). However, a single, fixed pessimistic value cannot explain the dynamic approach-retreat bouts observed in animal exploration. As we argue in Chapter 5, a more flexible mechanism is needed. We propose that behaviour is driven by the dynamic composition of multiple value functions with different initialisations (e.g., one aversive, one neutral). Gating the influence of these values by an inferred threat-belief state allows the agent

to flexibly express innate avoidance, reproducing the approach-retreat dynamics and the observation of threat prediction errors only at the start of a retreat from the novel object. This architecture offers a potential mapping to the opponent functions of direct and indirect pathways within the TS (Tsutsui-Kimura et al., 2025) and unifies different forms of value-shaping bonuses under a single normative framework (Kakade and Dayan, 2002).

Dedicated internal state and internal models

For physically embodied agents like robots, safety depends not only on the external world but also on the agent's own internal integrity. This necessitates mechanisms for self-monitoring and fault detection. In control theory and robotics, this is often addressed through model-based fault detection, which involves creating a forward model of the system's expected dynamics to predict expected sensor readings based on motor commands. A fault is detected when a significant discrepancy, or 'residual', emerges between the model's predictions and the actual sensor readings (Pezzato et al., 2020). A related approach, Associative Skill Memories (ASMs), learns a generative model of expected sensorimotor observations for different skills, often from demonstration (Pastor et al., 2012, 2013). An abnormality or fault can then be detected as a deviation from the learned manifold of expected experience, which can be used to trigger adaptive behaviours.

Insights from behavioural neuroscience: This engineering approach has a profound biological parallel in the concept of *interoception*—the brain's sense of its body's internal physiological state. The brain maintains sophisticated predictive models of bodily states, a process central to theories of homeostasis and allostasis. As explored in Chapter 6, the internal states are often partially observable, which can explain costly information-seeking actions (e.g., probing an injury) as rational attempts to reduce uncertainty about the body's true state to prevent larger future losses. Furthermore, as shown in Chapter 7, the principles of self-supervised predictive coding used to implement (neural) ASMs provide a

plausible computational mechanism for how the brain might learn models of expected sensorimotor feedback. Furthermore, this approach, which can be implemented with local learning rules, is formally related to a Kalman filter that tracks the mean but not the variance of the hidden state (Millidge et al., 2024b). Thus, the brain's mechanisms for monitoring its internal state offer a rich source of inspiration for building more robust, self-preserving artificial agents (Lee et al., 2023).

Lastly, we note that computer scientists have also used other methods such as teacher advice methods. Teacher advice involves interactive feedback or advisory control during training allows the agent to query or defer to a teacher, often based on uncertainty or disagreement across candidate policies. Though important, we exclude these methods from the current review as it focuses on single agent experiential learning.

3.7 Conclusion

This chapter has provided a bridge between the engineering principles of safe reinforcement learning and their analogues in behavioural and systems neuroscience. We have seen how the formalisms of safe RL, which categorise strategies into those that modify the optimisation criterion and those that modify the exploration process, map onto distinct but complementary biological solutions. AI provides a powerful language for describing what an agent should optimise for safety, through concepts like worst-case criteria and risk sensitivity. Neuroscience, in turn, provides a rich catalogue of mechanisms for how agents achieve safety in practice, through innate priors, dedicated internal models, and flexible meta-control. Having established this foundational landscape, the thesis will now transition from reviewing existing frameworks to presenting novel contributions. We begin in Part I by identifying the safety-efficiency dilemma and proposing a new model for how the brain balances these competing demands.

The ship is safest when it's in port, but that's not what ships were built for.

— *Paulo Coelho, The Pilgrimage*

4

Balancing safety and efficiency in human decision-making

Contents

4.1 Prelude

In the late 1960s, classic negative automaintenance experiments demonstrated that pigeons would persistently peck a lit key associated with food, even when this very action caused the reward to be withheld – a striking example of Pavlovian misbehaviour (Williams and Williams, 1969). This highlighted a now well-established principle: the brain does not use a single monolithic system for driving behaviour, but rather a modular control system with multiple controllers of varying capabilities and complexities for both appetitive and aversive learning (Seymour, 2019). A key normative logic for being endowed with controllers of varying complexity is the

trade-off between reaction time and the quality of a decision. While planning may yield a better action, it takes time; sometimes it is better to act fast than to act well, a niche filled by simpler, reflexive systems (Sutton et al., 2022). This further also closely links to the bias-variance trade-off in machine learning (Dorfman and Gershman, 2019). According to the Occam's razor, if the flexibility afforded by the complex system (with high variance) is not useful, then a simpler system (with high bias) will be favoured and this is observed in captured in many cost-benefit arbitration schemes (Kool et al., 2017; Mahajan et al., 2023). This backdrop begs the question: are these Pavlovian biases, which can clearly lead to suboptimal outcomes, simply a vestige of evolution, or do they play a functional, normative role in keeping us safe? We started our investigation with this motivating question.

4.2 Introduction

Humans and animals inhabit a complex and dynamic world where they need to find essential rewards such as food, water and shelter, whilst avoiding a multitude of threats and dangers which can cause injury, disability or even death. This illustrates a tension at the heart of learning and decision-making systems: on the one hand one wants to minimise environmental interactions required to learn to acquire rewards (be sample efficient), but on the other hand, it is important not to accrue excessive damage in the process – which is particularly important if you only get one chance at life. This safety-efficiency dilemma is related to the exploration-exploitation dilemma, in which the long-term benefits of information acquisition are balanced against the short-term costs of avoiding otherwise valuable options. Most solutions to the exploration-exploitation dilemma consider things only from the point of view of a single currency of reward, and hence, early losses can be overcome by later gains. Thus, many engineering solutions involve transitioning from exploratory strategies to more exploitative strategies over time as an agent gets more familiar with the environment. However, such solutions could be insufficient if some outcomes are incommensurable with others; for instance, damage accrues to the point that cannot be overcome, or worse still, leads to system failure 'death'

before you ever get the chance to benefit through exploitation, emphasising the need for safe (early) exploration. Safe learning (Garcıa and Fernández, 2015) is an emerging topic in artificial intelligence and robotics, with the advent of adaptive autonomous control systems that learn primarily from experience: for example, robots intended to explore the world without damaging or destroying themselves (or others) - the same concern animals and humans have.

A biological solution to this problem may be to have distinct systems for learning, for instance, having Pavlovian reward and punishment systems in addition to an instrumental system, which can then be integrated together to make a decision (Elfwing and Seymour, 2017; Bach and Dayan, 2017). A dissociable punishment system could then allow, for example, setting a lower bound on losses which must not be crossed during early learning. The brain seems likely to adopt a strategy like this since we know that Pavlovian fear processes influence instrumental reward acquisition processes (e.g. in paradigms such as conditioned suppression (Kamin et al., 1963) and Pavlovian-instrumental transfer (Talmi et al., 2008; Prévost et al., 2012)). However, it is not clear if this exists as a static system, with a constant Pavlovian influence over instrumental decisions, or a flexible system in which the Pavlovian influence is gated by information or experience. Computationally, it implies a multi-attribute architecture involving modular systems that separately learn different components of feedback (rewards, punishments) with the responses or actions to each then combined.

In this chapter we ask two central questions i) whether it is computationally (normatively) adaptive to have a flexible system that titrates the influence of 'fear' based on uncertainty i.e. reduces the impact of fear after exploration and ii) whether there is any evidence that humans use this sort of flexible meta-control strategy. We first describe a computational model of how Pavlovian (state-based) responses shape instrumental (action-based control) processes and show how this translates to a multi-attribute reinforcement learning (RL) framework at an algorithmic level. We propose how Pavlovian-instrumental transfer may be flexibly guided by an estimate of outcome uncertainty (Bach and Dolan, 2012; Dorfman and Gershman,

2019) - which effectively acts as a measure of uncontrollability. We use Pearce-Hall associability (Krugel et al., 2009), which is an implementationally simple and direct measure of uncertainty that has been shown to correlate well with both fear behaviour (skin conductance) and brain (amygdala) activity in fear learning studies (Li et al., 2011; Zhang et al., 2016, 2018). Below, we demonstrate the safety-efficiency trade-off in a range of simulation environments and show how it can be solved with a flexible Pavlovian fear bias. Consequently, we then test basic experimental predictions of the model in a virtual reality-based approach-withdrawal task involving pain, which builds upon previous Go-No Go studies studying Pavlovian-Instrumental transfer (Guitart-Masip et al., 2012; Cavanagh et al., 2013). The virtual-reality approach confers a greater ecological validity and the immersive nature may contribute better fear conditioning, making it easier to distinguish the aversive components.

4.3 Pavlovian Avoidance Learning (PAL) model

Our model consists of a Pavlovian punishment (fear) learning system and an integrated instrumental learning system (Fig. 4.1). The standard (rational) reinforcement learning system is modelled as the instrumental learning system. The additional Pavlovian fear system biases the withdrawal actions to aid in safe exploration, in line with our hypothesis. The Pavlovian system learns punishment expectations for each stimulus/state, with the corresponding Pavlovian responses manifest as action propensities to withdraw. For simplicity, we don't include a Pavlovian reward system, or other types of Pavlovian fear response (Bolles, 1970). The instrumental system learns combined reward and punishment expectations for each stimulus-action or state-action pair and also converts these into action propensities. Both systems learn using a basic temporal difference updating rule (or in instances, its special case, the Rescorla-Wagner rule). The ultimate decision that the system reaches is based on integrating these two action propensities, according to a linear weight, ω. Below we consider fixed and flexible implementations of this parameter, and test whether a flexible ω confers an advantage. We implement the

flexible ω using Pearce-Hall associability (see equation 4.15 in Methods). The Pearce-Hall associability maintains a running average of absolute temporal difference errors (δ) as per equation 4.14. This acts as a crude but easy-to-compute metric for outcome uncertainty which gates the influence of the Pavlovian fear system, in line with our hypothesis. This implies that the higher the outcome uncertainty, as is the case in early exploration, the more cautious our agent will be, resulting in safer exploration. For simulations, we use standard grid-world-like environments, which provide a didactic tool for understanding Pavlovian-Instrumental interactions (Dayan et al., 2006). Since Pavlovian biases influence not only choices but also reaction times, we extend our model to reinforcement learning diffusion decision-making (RLDDM) models (Pedersen et al., 2017; Fontanesi et al., 2019; Fengler et al., 2022).

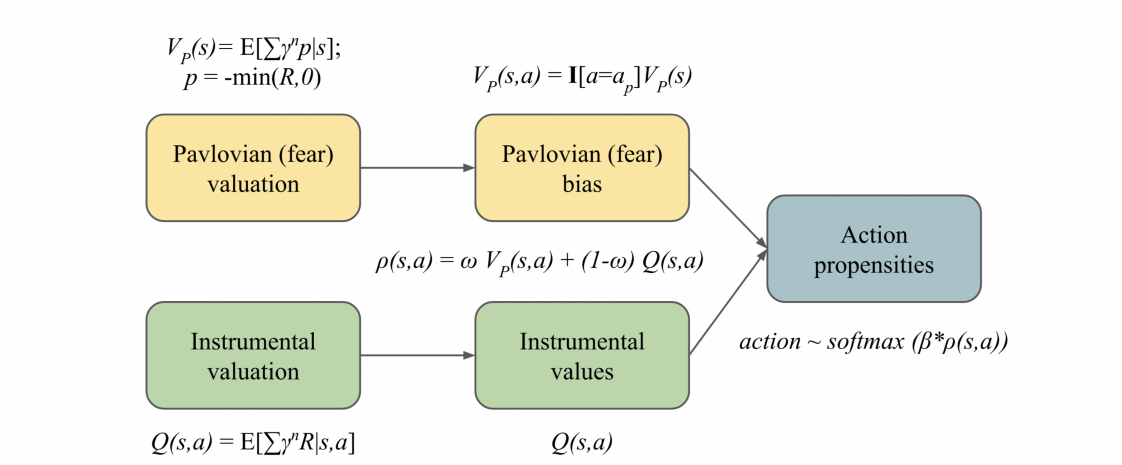


Figure 4.1: Pavlovian and Instrumental valuations are combined to arrive at action propensities used for (softmax) action selection. The Pavlovian bias influences protective behaviours through safer (Boltzmann) exploration and the arbitration between the Pavlovian and Instrumental systems is performed using the parameter ω. Here R denotes the feedback signal which can take both positive values (in the case of rewards) and negative values (in the case of punishments). Please see Methods for technical details; notations for the illustration follow Dorfman and Gershman (2019).

4.4 Results

Experiment 1: A simulated flexible fear-commissioning model balances safety and efficiency

We consider a simple fully-observable grid-world environment with stochastic state transitions and fixed starting state and fixed rewarding goal state. Fig. 4.2A illustrates how the misalignment of Pavlovian bias and instrumental action can lead to a safety-efficiency dilemma. The Pavlovian action is assumed to be an evolutionarily acquired simple withdrawal response Fig. 4.2B, and the Pavlovian state value is learned during the episode and shapes instrumental policy and value acquisition (Fig. 4.2C). Figure 4.2C shows value plots for the instrumental policy with and without a Pavlovian bias. All plots show values and policy at the end of 1000 episodes of learning. These heatmaps denote value i.e. the expectation of cumulative long-term rewards R (including any punishments) and the arrows show the policy i.e. actions that maximize this value. Additionally, the learned punishment value V_p of the Pavlovian bias is also shown along with the PAL policy. The PAL value function and policy shown in Fig. 4.2C utilizes the flexible ω scheme utilized below.

Fig. 4.2D plots cumulative pain accrued over multiple episodes during learning and is our measure of safety. Fig. 4.2E plots cumulative steps or environment interactions over episodes and is our measure of sample efficiency. Here, sampling efficiency is represented by the total number of environment interactions or samples required to reach the rewarding goal which terminates the episode. Simply, if an agent requires more samples to reinforce and acquire the rewarding goal, it is less efficient.

The simulation results with a fixed Pavlovian influence (Fig. 4.2D, 4.2E) show that adding a Pavlovian fear system to the instrumental system makes it safer in the sense that it achieves the goal of solving the environment while accruing lesser cumulative pain over episodes. However, we observe that as the influence of the Pavlovian fear system increases, with an increase in ω, it achieves safety at the expense of sample efficiency (within reasonable bounds such as until $\omega = 0.5$). Whereas under very high Pavlovian fear influence ($\omega = 0.9$), the agent loses sight

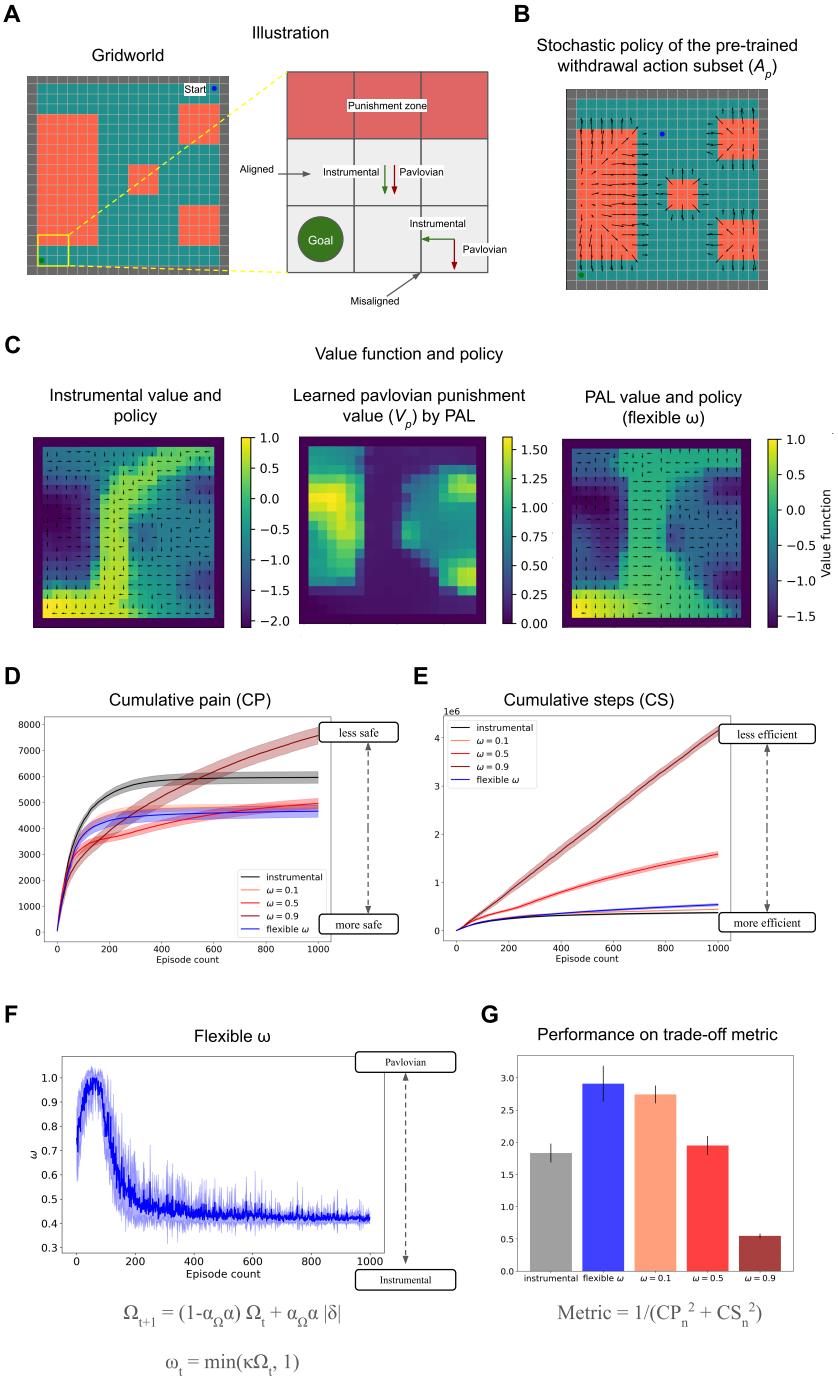


Figure 4.2: (A) Grid world environment with starting state in the top-right corner and rewarding goal state ($R = +1$) in the bottom-left corner, and the red states are painful ($R = -0.1$). The grid world layout follows Gehring and Precup (2013). The inset provides a didactic example of misalignment between Pavlovian bias and Instrumental action. (B) Stochastic policy of pre-trained withdrawal action subset A_p, which is biased with Pavlovian punishment values in the PAL agent. (Caption continued on next page)

Figure 4.2: (Continued from previous page.) (C) The learned instrumental values and Pavlovian fear bias V_p(heatmap) and policy are learned by the instrumental and flexible ω agent by the end of the learning duration. The value functions plotted are computed in an on-policy manner. For purposes of a cleaner visualisation, although the policy is stochastic, the arrows represent the greedy policy (representing the most valued action in each state). The learned policies for all agents are stochastic; therefore, for instance, even if the arrows for the PAL agent point either left or right in the corridor, showing that it prioritises avoiding the dangerous regions, the stochastic policy allows it to move down towards the rewarding goal. (D) Cumulative pain accrued by fixed and flexible ω agents whilst learning over 1000 episodes as a measure of safety, averaged over 10 runs. (E) Cumulative steps required to reach the fixed goal by fixed and flexible ω agents whilst learning over 1000 episodes as a measure of sample efficiency, averaged over 10 runs (F) Plot of flexibly modulated ω arbitration parameter over the learning duration, averaged over 10 runs. This shows a transition from a higher Pavlovian bias to a more instrumental agent over episodes as learning about the environment reduces uncertainty (G) Comparison of different agents using a trade-off metric and to be used only for didactic purposes (using equation 4.16 and more details in Methods).

of the rewarding goal and performs poorly in terms of both safety and efficiency as the episode doesn't terminate until it finds the goal.

However, the flexible omega policy (with $\alpha_\Omega = 0.6$ and $\kappa = 6.5$) achieves safety almost comparable to $\omega = 0.5$ (which is the safest fixed ω policy amongst $\omega = 0.1, 0.5, 0.9$ at a much higher efficiency than $\omega = 0.5, 0.9$, thus improving the safety-efficiency trade-offs (Fig. 4.2F). In this way, PAL model encourages cautious exploration early on when uncertainty is higher and reduces the Pavlovian biases as the uncertainty is resolved (Fig. 4.2F). The flexible ω value at convergence depends on the environment statistics: transition probabilities and reward/punishment magnitudes. We utilise a simple linear scaling of associability clipped at 1 to arrive at arbitrator ω (equation 4.15) instead of another alternative such as sigmoid to avoid additional unnecessary meta parameters (i.e. bias shift) to be tuned. In this environment, the value at convergence is $\omega = 0.42$, due to some irreducible uncertainty in state transitions (10% chance of incorrect transition). The differences in learned instrumental value functions between PAL and a purely instrumental agent are visible in (Fig. 4.2C) showing how the Pavlovian bias sculpts instrumental value acquisition.

In the appendix, we provide additional simulations that show the robustness of these results with respect to metaparameters α_Ω and κ (Appendix A.1), environments in which the reward locations vary (Appendix A.2), and other grid-world environments (Appendix A.3).

Experiment 2: Constant Pavlovian bias introduces sampling asymmetry and affects instrumental value propagation

Observing the differences in the on-policy value functions with and without the Pavlovian influence (Fig. 4.2C) prompted us to further tease apart the effect of constant Pavlovian bias on sampling asymmetry, and consequent differences in instrumental value discovery and value propagation through the states. We investigated how different fixed values of ω can lead to sampling asymmetry, which refers to exploration where certain states are visited or sampled unevenly compared to others. In this set of results, we wish to qualitatively tease apart the role of a Pavlovian bias in shaping and sculpting the instrumental value and also provide more insight into the resulting safety-efficiency trade-off. Having shown the benefits of a flexible ω in the previous section, here we only vary the fixed ω to illustrate the effect of a constant bias and are not concerned with the flexible bias in this experiment.

We tested agents with different fixed ω in two simulated environments: (1) A T-maze and (2) a three-route task. The T-maze task environment(Fig. 4.3A) has asymmetric rewards ($R = +0.1$ on the left, whereas $R = +1$ on the right). However, the agent will have to walk through a patch of painful states to reach the larger goal on the right, even the safest path will incur a damage of at least $R = -0.5$ or worse. Taking discounting into account, the goal on the right is marginally better than the one on the left and the instrumental agent achieves both of the goals nearly an equal number of times (Fig. 4.3B). Comparing the instrumental agent with other agents in Fig. 4.3C shows diminished positive (reward) value propagation leading to the $R = 1$ goal on the right as the constant Pavlovian bias increases, showing how such sampling asymmetry can prevent value discovery of states leading to $R = 1$ goal. The safety efficiency trade-off can also be observed

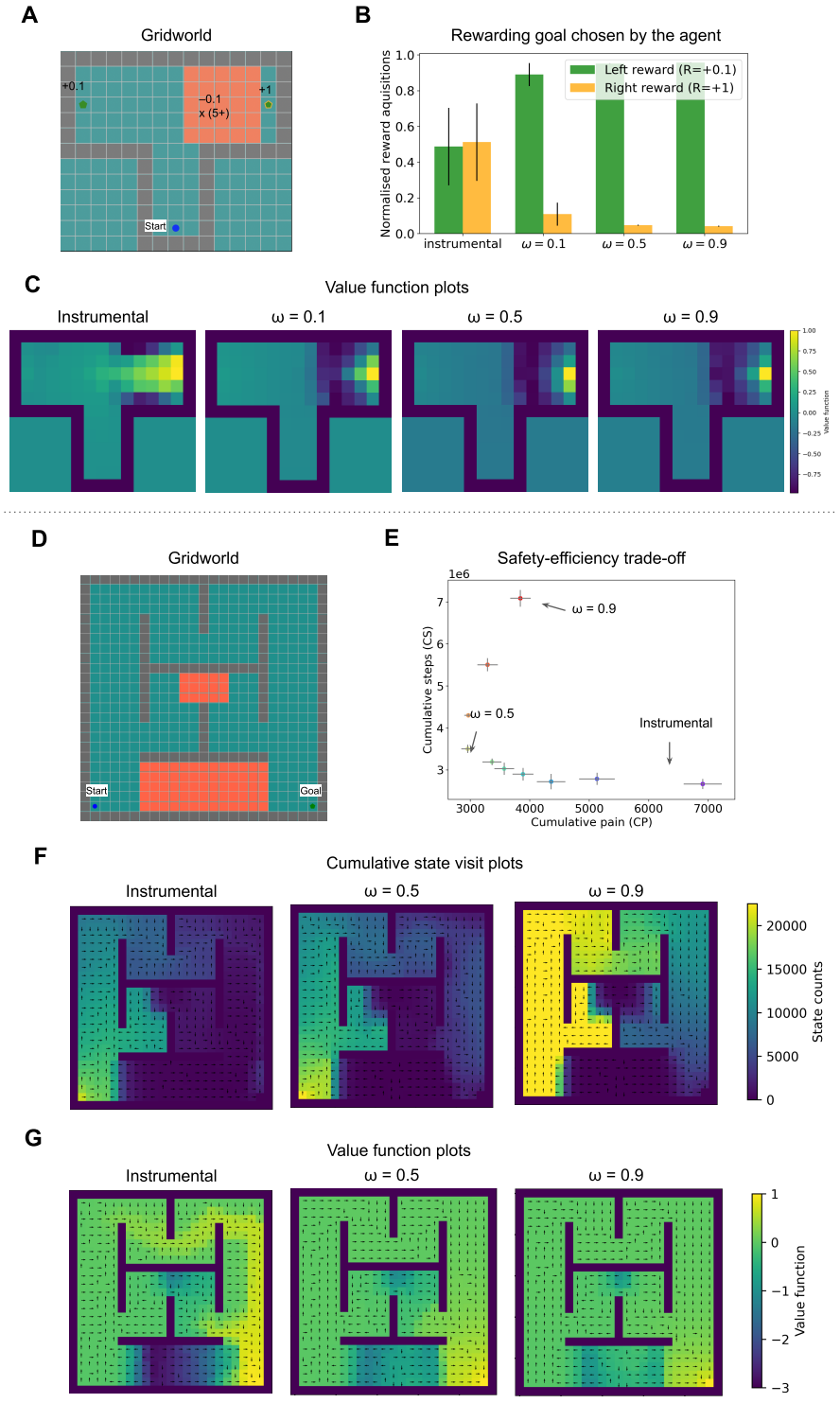


Figure 4.3: (A) T-maze grid world environment with annotated rewards and punishments. (B) Proportion of the rewarding goal chosen by the agent (C) Value function plots for $\omega = 0, 0.1, 0.5, 0.9$ shows diminished value propagation from the reward on the right (Caption continued on next page.)

Figure 4.3: (Continued from previous page.) (D) Grid world environment with three routes with varying pain (E) Cumulative steps required to reach the goal vs cumulative pain accrued by fixed ω agents ranging from $\omega = 0$ to $\omega = 0.9$ (F) State visit count plots for $\omega = 0, 0.5, 0.9$ i.e. instrumental and constant Pavlovian bias agents. (F)Value function plots for $\omega = 0, 0.5, 0.9$. Even though the learned policies are stochastic, the plotted arrows represent the greedy policy, denoting the most valued action in every state.

through Fig. 4.3B. This illustrates one of the main tenets of our model - that having a Pavlovian fear system ensures a separate 'un-erasable' fear/punishment memory which makes the agent more avoidant to punishments. This is helpful in softly ensuring an upper bound on losses, by (conservatively) foregoing decisions resulting in immediate loss, but followed by much larger rewards. This is where the safety-efficiency trade-off marks a clear distinction from the exploration-exploitation trade-off, in which earlier losses can be overcome by gains later on.

The three-route task simulation includes three routes with varying degrees of punishments (Fig. 4.3D), inspired by previous manipulandum tasks (Meulders et al., 2016; van Vliet et al., 2020, 2021; Glogan et al., 2021). We observe that increasing the constant Pavlovian bias up until $\omega = 0.7$ leads to increased safety (Fig. 4.3E). Beyond $\omega = 0.7$, a high fixed Pavlovian bias may incur unnecessarily high cumulative pain and steps as its reward value propagation is diminished (Fig. 4.3G) and attempts to restrict itself to pain-free states (Fig. 4.3F) whilst searching for reward (despite stochastic transitions which may lead to slightly more painful encounters in the long run). Comparing the cumulative state visit plots of Fig. 4.3F, the instrumental agent with an agent with high constant Pavlovian bias $\omega = 0.9$, we observe that the latter showed an increased sampling of the states on the longest route with no punishments. Comparing the value function plots (Fig. 4.3G), we observe that a high constant Pavlovian bias impairs the value propagation (it is more diffused) of the rewarding goal in comparison to an instrumental agent. Such high levels of constant Pavlovian bias can be a model of maladaptive anxious behaviour.

In conclusion, the simulations with this environment show that the Pavlovian fear system can assist in avoidance acquisition, however a constant Pavlovian bias

depending on the degree of bias, leads to sampling asymmetry and impaired value propagation.

Appendix A.3 includes the performance comparison of agents with a suitable flexible ω and with fixed ω values on the three-route task. Appendix A.4 shows the results of a human experiment with subjects navigating a 3-route virtual reality maze similar to the simulated one.

Experiment 3: Human approach-withdrawal conditioning is modulated by outcome uncertainty

Our first experiment showed the benefit of having a outcome uncertainty-based flexible ω arbitration scheme in balancing safety and efficiency, in a series of grid worlds. In this next experiment, we aimed to find basic evidence that humans employ such a flexible fear commisioning scheme. This is not intended as an exhaustive test of all predictions of the model, but to show in principle that there are situations in which a flexible, rather than fixed pavlovian influence, provides a good fit to real behavioural data. In line with our grid world simulations, we expected a Pavlovian bias in choices, but in addition to it, we also expected a Pavlovian bias in reaction times.

Fig. 4.4 describes the trial protocol (Fig.4.4A), block protocol (Fig.4.4B) and experimental setup (Fig. 4.4C). We conducted a VR-based approach-avoidance task (28 healthy subjects, of which 14 females and average age 27.96 years) inspired by previous Go-No Go task studies for isolating Pavlovian bias, especially its contributions to misbehaviour (Guitart-Masip et al., 2012; Cavanagh et al., 2013; Mkrtchian et al., 2017a,b; Dorfman and Gershman, 2019; Gershman et al., 2021). The subjects goal was to make a correct approach or withdrawal decision to avoid pain, with four different cues associated with different probabilities of neutral or painful outcomes. We expected the Pavlovian misbehaviour to cause incorrect withdrawal choices for cues where the correct response would be to approach. And in terms of reaction times, we expected the bias to slow down correct approach responses and speed up correct withdrawal responses. We explicitly attempted

to change the outcome uncertainty or controllability, in a similar way to previous demonstrations Dorfman and Gershman (2019), but with controllability changing *within* the task. To do this, we set up two of the four cues to be uncontrollable in the first half (i.e. outcome is painful 50% of the times regardless of the choice), but which then become controllable in the second half (i.e. the correct choice will avoid the pain 80% of the times). We anticipated that the Pavlovian bias in choice and reaction times would be modulated along with the change in uncontrollability. The virtual reality environment improves ecological validity(Parsons, 2015) and introduces gamification, which is known to also improve reliability of studies(Sailer et al., 2017; Kucina et al., 2023; Zorowitz et al., 2023), which is important in attempts to uncover potentially subtle biases.

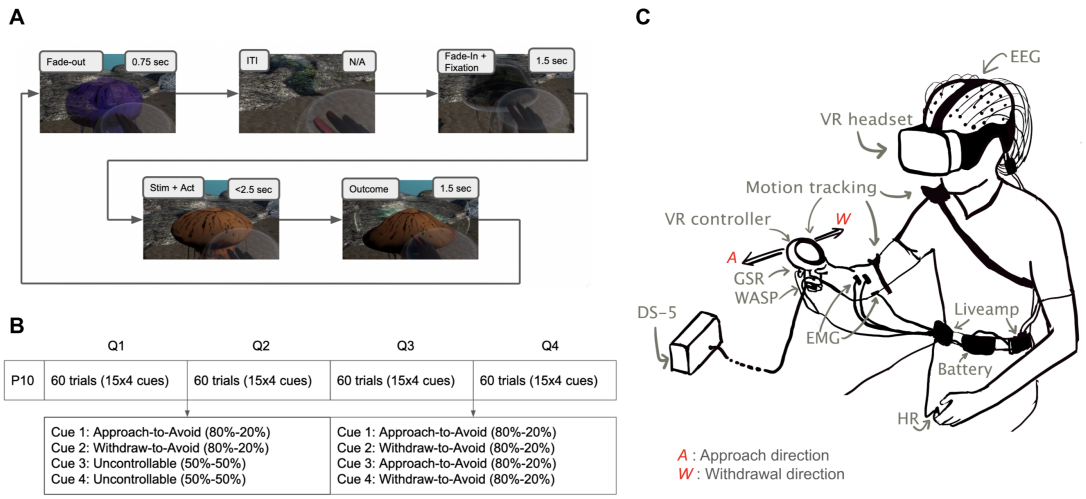


Figure 4.4: (A) Trial protocol: The participant is expected to take either an approach action (touch the jellyfish) or withdrawal action (withdraw the hand towards oneself) within the next 2.5 seconds once the jellyfish changes colour. The participant was requested to bring their hand at the centre of a bubble located halfway between the participant and the jellyfish to initiate the next trial where a new jellyfish would emerge. [Supplementary video] (B) Block protocol: First half of the trials had two uncontrollable cues and two controllable cues, and the second half had all controllable cues with aforementioned contingencies. The main experiment 240 trials were preceded by 10 practice trials which do not count towards the results. (C) Illustration of experimental setup VR: Virtual Reality, WASP: Surface electrode for electrodermal stimulation, DS-5: Constant current stimulator, GSR: galvanic skin response sensors, HR: Heart rate sensor, EMG: Electromyography sensors, EEG: Electroencephalogram electrodes, Liveamp: Wireless amplifier for mobile EEG.

We observe that all subjects learn to solve the task well and solve it better than chance (i.e. lesser than 120 shocks in 240 trials). Out of 240 trials, they receive 88.96 shocks on average (std. deviation = 12.62). We first attempted to test our hypotheses using behavioural metrics of Pavlovian withdrawal bias in choices and reaction times. However, our behavioural choice-based metrics cannot distinguish a random exploratory action from Pavlovian misbehaviour. Further, it cannot account for effects of a non-Pavlovian baseline bias b. Thus, we did not find any statistically significant result due to noisy behavioural metrics, results and more information provided in Appendix A.5.

We next aimed to test our hypotheses by model comparison of RL models (Fig.4.5A) and RLDDM models Fig.4.5E) which guides our results below. We used a hierarchical Bayesian estimation of model parameters, to increase the reliability across tasks. We found that the baseline action bias b, instrumental learning and the Pavlovian withdrawal bias competed for behavioural control, as observed in previous studies (Guitart-Masip et al., 2012; Cavanagh et al., 2013) (parameter distribution plots in Appendix A.6). However, unlike previous studies that have treated Pavlovian bias as fixed, we found that the flexible Pavlovian bias better explained the behavioural data, please see Fig.4.5 (B) and (F).

Similar to Guitart-Masip et al. (2012), the simple Rescorla-Wagner learning (RW) model represented the base model. RW+bias includes a baseline bias b that can take any positive or negative value, positive value denoting a baseline bias for approach and negative denoting a baseline bias for withdrawal. From group-level and subject-level parameter distribution plots (Appendix A.6) we observe that this baseline bias is for approach for most subjects. This is in line with previous studies(Guitart-Masip et al., 2012; Cavanagh et al., 2013) and as suggested by our data showing a significant baseline difference in the number of approaches and withdrawal actions across all subjects and trials (Appendix A.5). Note that here, this baseline bias is not learned as it is with a Pavlovian bias. RW+bias+Pavlovian(fixed) model includes a fixed Pavlovian bias and is most similar to models by Guitart-Masip et al. (2012); Cavanagh et al. (2013), which also used reward and punishment sensitivities

for the instrumental learning but did not scale the instrumental values by $(1-\omega)$ as done in our model. Our models do not have reward and punishment sensitivities. From group-level and subject-level parameter distribution plots (Appendix A.6), we observe that the distribution of fixed ω is significantly positive and non-zero. RW+bias+Pavlovian(flexible) model includes a flexible Pavlovian bias as per our proposed associability-based arbitration scheme. We found that the flexible ω model fits significantly better than the fixed ω model (Fig. 4.5B) i.e. flexible ω model has the lowest Leave-one-out information criteria (LOOIC) score amongst models compared. By comparing incremental improvements in LOOIC, we observe that adding the baseline bias term leads most improvement in model fit, followed by changing the fixed ω to a flexible ω scheme. Here, we plot LOOIC for model comparison, but Appendix A.7 include both LOOIC and WAIC scores, showing the same result. Further, it can be seen that ω tracks associability, which decreases over the trials (Fig. 4.5C) (which also resembles Fig. 4.2E). Fig. 4.5D shows a plot comparing the number of approaches (normalized to 1), aggregated over all subjects and trials by cue types for data and the best fitting model predictions. We observe qualitatively that the subjects learn to perform the correct actions for each cue and that the model predictions qualitatively reproduce the data.

We then extend the model-fitting to also incorporate reaction times, using an RLDDM (reinforcement learning drift diffusion model) (Pedersen et al., 2017; Fontanesi et al., 2019; Desch et al., 2022). The propensities calculated using the RL model are now used as drift rates in the DDM and the reaction times are calculated using a weiner distribution for a diffusion-to-bound process. Thus the drift rate is proportional to the difference in propensities between approach and withdrawal action. Since Pavlovian bias is also dependent on punishment value, similar to instrumental values, we included the Pavlovian bias and the baseline bias in the drift rate. Thus, the best propensity for an action in choice selection in RL models, drives the drift rate in our RLDDM models. We found that the RLDDM replicates the results for model-fitting (Fig. 4.5E) and flexible ω (Fig. 4.5F). Fig. 4.5H shows the distribution for approach and withdrawal reaction times

(RTs) aggregated over all subjects, over all trials, in data and model predictions. The data shows that the withdrawal RTs are slightly faster than approach RTs (Fig. 4.5H). The model captures the shape of RT distributions, but not the difference between approaches and withdrawals. This is because we kept the starting point constant and equal to half the threshold (making it an equidistant starting point for both approach and avoidance actions). However, models with variable starting points may fit the data better and may reproduce the difference in approach and withdrawal reaction times (Ratcliff et al., 2018).

Appendix A.5 includes behavioural results for the experiment data. Appendix A.6 includes group-level and subject-level (hierarchically fitted) model parameter distributions. Appendix A.7 mention model parameters with LOOIC and WAIC values for all RL models and RLDDM models. We observed that all Rhat values were strictly less than 1.05 (most parameters were less than 1.01 and generally close to 1), indicating that the models converged.

4.5 Discussion

In summary, this paper shows that addition of a fear-learning system, implemented as a Pavlovian controller in a multi-attribute RL architecture, improves safe exploratory decision-making with little cost of sample efficiency. Employing a flexible arbitration scheme where Pavlovian responses are gated by outcome uncertainty (Bach and Dolan, 2012) provides a neurally plausible approach to solving the safety-efficiency dilemma. Our experimental results support the hypothesis of such a flexible fear commissioning scheme and suggest that inflexible Pavlovian bias can explain certain aspects of maladaptive 'anxious' behaviour (please see Appendix A.8). This can be helpful in making novel predictions in clinical conditions, including maladaptive persistent avoidance in chronic pain in which it may be difficult to 'unlearn' an injury.

Broadly, our model sits amidst with the landscape of safe reinforcement learning (RL) (Garcıa and Fernández, 2015). In principle, it can be viewed through the lens of constrained Markov decision processes (Altman, 1999), where the Pavlovian fear system is dedicated towards keeping constraint violations at a minimum.

In the realm of safe learning, there exists a dichotomy: one can either apply computer science-driven approaches to model human and animal behaviour, as seen

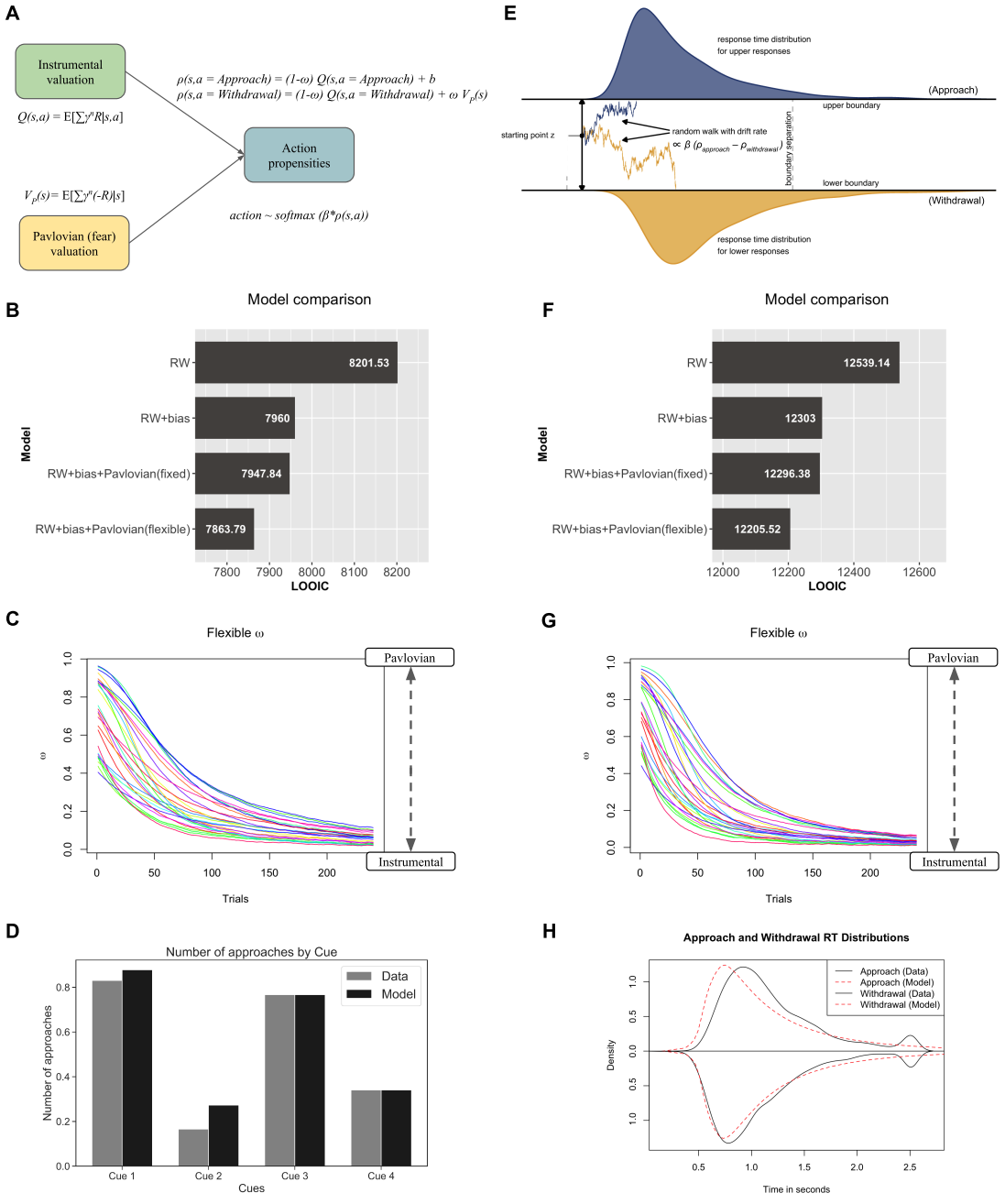


Figure 4.5: Left panels show choice model fit results using RL models. The right panels show choice and reaction times model fit results using RLDDM models. (A) Simplified RL model from Fig. 4.1 for the Approach-Withdrawal task (B) Model comparison shows that the model with flexible Pavlovian bias fits best to choices in terms of LOOIC (C) Flexible ω from the RL model over 240 trials for 28 participants. (D) Number of approaches aggregated over all subjects and all trials in data and model predictions by the RL model with flexible ω - normalized to 1. (Caption continued on next page.)

Figure 4.5: (Continued from previous page.) (E) Simplified illustration of RLDDM for the Approach-Withdrawal task, where the baseline bias b and the Pavlovian bias $\omega V_p(s)$ is also included in the drift rate (The base figure is reproduced from Desch et al. (2022) with modifications) (F) Model comparison shows that the model with flexible Pavlovian bias fits best to choices and reaction times in terms of LOOIC (G) Flexible ω from the RLDDM over 240 trials for 28 participants (H) Distribution of approach and withdrawal reaction times (RT) aggregated over all subjects and trials in data and model predictions by the RLDDM with flexible ω. The bump in RTs at 2.5 seconds is because of timeout (inactive approaches and withdrawals, please see Appendix A.5)

in optimizing worst-case scenarios (Heger, 1994) and employing beta-pessimistic Q-learning (Gaskett, 2003) for modelling anxious behaviour (Zorowitz et al., 2020), or opt for neuro-inspired algorithms and demonstrate their utility in safe learning. Our model falls into the latter category, draws inspiration from the extensive literature on Pavlovian-Instrumental interactions (Mowrer, 1951, 1960; Kamin et al., 1963; Brown and Jenkins, 1968; Mackintosh, 1983; Talmi et al., 2008; Maia, 2010; Huys et al., 2012; Prévost et al., 2012), fear conditioning (LaBar et al., 1998) and punishment-specific prediction errors (Pessiglione et al., 2006; Seymour et al., 2007, 2012; Roy et al., 2014; Berg et al., 2014; Elfwing and Seymour, 2017; Watabe-Uchida and Uchida, 2018), and elucidates a safety-efficiency trade-off. Classical theories of avoidance such as two-factor theory (Mowrer, 1951), and indeed actor-critic models (Maia, 2010) intrinsically invoke Pavlovian mechanisms in control, although primarily to act as a teaching signal for instrumental control as opposed to directly biasing action propensities such as in our case or (Dayan et al., 2006). Recent studies in computer science, particularly those employing policy optimization (gradient-based) reinforcement learning under CMDPs (Altman, 1999), have also observed a similar safety-efficiency trade-off (Moskovitz et al., 2023). Additionally, the fundamental trade-off demonstrated by Fei et al. (2020) between risk-sensitivity (with exponential utility) and sample-efficiency in positive rewards aligns with our perspective on the safety-efficiency trade-off, especially when broadening our definition of safety beyond cumulative pain to include risk considerations. Safety-efficiency trade-offs may also have a close relation with maladaptive avoidance (Ball and Gunaydin, 2022) often measured in clinical anxiety, and our work provides insights into the

maladaptive avoidance via the heightened threat appraisal pathway. Similar safe exploration behaviour in choices could be achievable using a risk-sensitive criterion such as conditional value at risk (CVaR) that relies only on the instrumental systems without needing a Pavlovian system. However, these work in different ways. Embracing the decision-theoretic psychiatry framework by Huys et al. (2015), which attempts to categorise dysfunctions as the agent either solving the problem with a wrong solution or solving the wrong problem correctly, or solving the right problem correctly but in an unfortunate or wrong environment, then we see the following. CVaR provides the correct solution to the wrong problem (an objective that only maximises the lower tail of the distribution of outcomes). In contrast, the Pavlovian bias provides the wrong solution to the correct problem (normative objective) (Huys et al., 2015). Further, approaches such as CVaR might not be the best approach to capture the Pavlovian withdrawal bias effect we find in reaction times.

When it comes to our experiments, both the simulation and VR experiment models are related and derived from the same theoretical framework, maintaining an algebraic mapping. They differ only in task-specific adaptations i.e. differ in action sets and differ in temporal difference learning rules - multi-step decisions in the grid world vs. Rescorla-Wagner rule for single-step decisions in the VR task. This is also true for Dayan et al. (2006), who bridge Pavlovian bias in a Go-No Go task (negative auto-maintenance pecking task) and a grid world task. A further minor difference between the simulation and VR experiment models is the use of a baseline bias in the human experiment's RL and the RLDDM model, where we also model reaction times with drift rates which is not a behaviour often simulated in the grid world simulations. As mentioned previously, we use the grid world tasks for didactic purposes, similar to Dayan et al. (2006) and common to test-beds for algorithms in reinforcement learning (Sutton et al., 1998). The main focus of our work is on Pavlovian fear bias in safe exploration and learning, rather than on its role in complex navigational decisions. Future work can focus on capturing more sophisticated safe behaviours, such as escapes (Evans et al., 2019;

Sporrer et al., 2023) and model-based planning, which span different aspects of the threat-imminence continuum (Mobbs et al., 2020).

In our simulation experiments, we assume the coexistence of the Pavlovian fear system and the instrumental system to demonstrate the emergent safety-efficiency trade-off from their interaction. It is possible that similar behaviours could be modelled using an instrumental system alone, with higher punishment sensitivity, therefore, we do not argue for the necessity for the Pavlovian fear system here. Instead, the Pavlovian fear system itself could be a potential biologically plausible implementation of punishment sensitivity. Unlike punishment sensitivity (scaling of the punishments), which has not been robustly mapped to neural substrates in fMRI studies; the neural substrates for the Pavlovian fear system are well known (e.g., the limbic loop and amygdala, further see Appendix A.9). Additionally, Pavlovian fear system provides a separate punishment memory that cannot be erased by greater rewards like (Elfwing and Seymour, 2017; Wang et al., 2018b). This fundamental point can be observed in our simple T-maze simulations, where the Pavlovian fear system encourages avoidance behaviour and the agent chooses the smaller reward instead of the greater reward. We next discuss the plausibility of pre-training to select the hardwired actions. In the human experiment, the withdrawal action is straightforwardly biased, as noted, while in the grid world, we assume a hardwired encoding of withdrawal actions for each state/grid. This innate encoding of withdrawal actions could be represented in the dPAG (Kim et al., 2013). We implement this bias using pre-training, which we assume would be a product of evolution. Alternatively, this could be interpreted as deriving from an appropriate value initialization where the gradient over initialized values determines the action bias. Such aversive value initialization, driving avoidance of novel and threatening stimuli, has been observed in the tail of the striatum in mice, which is hypothesised to function as a Pavlovian fear/threat learning system (Menegas et al., 2018).

We illustrate that a high Pavlovian impetus is characterized by reduced sample efficiency in learning, worsened/weakened (instrumental) value propagation and impervious rigidity in the policy, and misbehaviour due misalignment of bias

with the instrumental action. This way it also further promotes short term safer smaller rewards opposed long term higher rewards, echoing the idea of Pavlovian pruning of decision trees(Huys et al., 2012). The idea of alignment between the Pavlovian and instrumental actions leading to harm-avoiding safe behaviours and misalignment being the root of maladaptive behaviours was proposed by Mkrtchian et al. (2017b) through a Go-No Go task with human subjects and the threat of shock responsible for the Pavlovian instrumental transfer. Recently Yamamori et al. (2023), have developed a restless bandit-based approach-avoidance task to capture anxiety-related avoidance, by using the ratio of reward and punishment sensitivities as a computational measure of approach-avoidance conflict. We show in our simulations that misalignment can also lead to safe behaviours, but at the cost of efficiency. But having a flexible fear commissioning alleviates majority of Pavlovian misbehaviour and in turn making the agent more cautious in the face of uncertainty and catastrophe, contrasting with 'optimism bias' observed in humans (Sharot, 2011). A limitation of our work would be that we do not model the endogenous modulation of pain and stress induced analgesia which may have the opposite effect of the proposed uncertainty-based fear commissioning scheme. A limitation of our VR experiment is that we only consider uncertainty decrease from first half to second half. This was motivated to make it similar to the grid world simulations as well as to help with behavioural tests (Appendix A.5), as this would keep all of the reducible and irreducible uncertainty in the first half and none in the second half. However, a stringent test would also require a balanced case, where the outcomes of cues 3 and 4 are more certain in first half and more uncertain in the second half, or consider differentiating uncertainty and volatility.

While our flexible ω scheme, rooted in associability, shares motivation with Dorfman and Gershman (2019) to track uncontrollability, our approach differs. Unlike Dorfman and Gershman (2019), which employs a Bayesian arbitrator emphasizing the most useful predictor (Pavlovian or Instrumental), our Pearce-Hall associability-based measure provides a direct and separate controllability assessment. This distinction allows our measure to scale effectively to complex tasks, including

gridworld environments, maintaining stability throughout experiments. In contrast, the measure by Dorfman and Gershman (2019) exhibits notable variability, even when the controllability of the cue-outcome pair remains constant throughout the task. Previous fMRI studies have associated associability signals with the amygdala (Zhang et al., 2016) and pgACC (Zhang et al., 2018). Additionally, outcome uncertainty computation could be possibly performed within the basal ganglia using scaled prediction errors (Mikhael and Bogacz, 2016; Moeller et al., 2022) and is encoded in the firing rates of orbitofrontal cortex neurons and possibly in slow ramping activity in dopaminergic midbrain neurons (Fiorillo et al., 2003; O'Neill and Schultz, 2010; Bach and Dolan, 2012). Associability as a measure of outcome uncertainty, though very practical and useful on an implementational level, cannot distinguish between various kinds of uncertainties. Further, future work can help differentiate between controllability and predictability; (Ligneul et al., 2022) suggest that controllability and not predictability may arbitrate the dominance of Pavlovian versus instrumental control. Future work could also use a formal account of uncertainty, which could fit the fear-conditioned skin-conductance response better than Pearce-Hall associability (Tzovara et al., 2018). Additionally, Cavanagh et al. (2013) demonstrated that theta-band oscillatory power in the frontal cortex tracks and overrides Pavlovian biases, later suggesting its connection to inferred controllability (Gershman et al., 2021). Notably, Kim et al. (2023) revealed that upregulation of the dorsolateral prefrontal cortex (dlPFC) through anodal transcranial direct current stimulation (tDCS) induces behavioral suppression or changes in Pavlovian bias in the punishment domain, implying a causal role of the dlPFC in Pavlovian-Instrumental arbitration.

An natural clinical application (Fullana et al., 2020) of this model is towards mechanistic models of anxiety and chronic pain. Quite simply, both have been considered as reflecting excessive Pavlovian punishment learning systems. In the case of anxiety disorder, this equates a strong influence of Pavlovian control with subjectively experienced anxiety symptomatology, leading to excessively defensive behaviour and avoidance of anxiogenic environments (Norton and Paulus, 2017). In

the case of chronic pain, the idea is that failure to overcome a Pavlovian incentive to avoid moving results in failure to discover that pain escape is possible (the fear avoidance model) (Vlaeyen and Linton, 2000). In both cases, the pathological state can be considered a failure to turn down the Pavlovian system when the environment becomes more predictable (i.e. less uncertain). This illustrates a subtle distinction between existing theories that simply propose a constant excess Pavlovian influence, from the possibility they might result from a deficit in the flexible commission of Pavlovian control. This distinction can be therefore be experimentally tested in clinical studies. Furthermore, accruing evidence also indicates a role of excessive Pavlovian punishment learning in models depression (Nord et al., 2018; Huys et al., 2016), suggesting that this may be a common mechanistic factor in comorbidity between chronic pain, anxiety and depression. Recent experiments and perspectives also suggest apsychological mechanism of how avoidance in humans can lead to growth of anxiety (increased belief of threats)(Urcelay, 2024). A key distinctive prediction of our model for an intervention is that we should help patient groups reduce Pavlovian bias not by training to reduce the bias, but rather by attempting to make the arbitration more flexible. This could potentially be done via some sort of controllability discrimination paradigm, i.e. helping distinguish between what is controllable and what is not - this is something also found in stoicism-based approaches to cognitive behavioural therapy (CBT) (Turk and Rudy, 1992; Thorn and Dixon, 2007).

In conclusion, we outline how the Pavlovian fear system provides an important and computationally precise mechanism to shape or sculpt instrumental decision-making. This role for the Pavlovian fear system extends its utility far beyond merely representing the evolutionary vestiges of a primitive defence system, as sometimes portrayed. This opens avenues for future research in basic science of safe self-preserving behaviour (including in artificial systems), and clinical applications for mechanistic models of anxiety and chronic pain.

4.6 Materials and Methods

Instrumental learning and Pavlovian fear learning

We consider a standard reinforcement learning setting in an environment containing reward and punishments (pain). In each time step t, the agent observes a state s_t and selects an action a_t according to its stochastic policy $\pi_t(s_t, a_t)$ (i.e., the probability of selecting action $a_t = a$ in state $s_t = s$). The environment then makes a state transition from the current state s_t to the next state s_{t+1} and the agent receives a scalar reward $R_{t+1} \in (-\infty, +\infty)$. This represents that the scalar reward includes both positive rewards and negative rewards or punishments. We use the standard notation used by Sutton and Barto (2018).

In the Instrumental system, we define the value of taking action a in state s under a policy π, denoted as the action-value function $Q^\pi(s, a)$, as the expected return starting from s, taking the action a, and thereafter following policy π:

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a \right] \tag{4.1}$$

where R_{t+k+1} is the scalar reward received k time steps in the future, when evaluating the Q-values at timesteps t. The discount factor is γ and the reward k timesteps into the future is discounted by γ^k. The optimal action-value function is defined as $Q^*(s, a) = \max_\pi Q^\pi(s, a)$. Note this are purely instrumental Q-values and do not include the Pavlovian bias.

In addition to the instrumental system, we define a Pavlovian fear (i.e. punishment/pain) system over-and-above the instrumental system which makes it safer. The Pavlovian fear systems aims to increase the impetus of the pain-avoidance actions that minimize pain. For that we split the standard reward R and only extract the punishment feedback signal $p \geq 0$:

$$p = -\min(R, 0), \tag{4.2}$$

We can similarly define a Pavlovian reward system trained on $max(R, 0)$, however that's not relevant to the questions of this study so we will only focus

on the arbitration between the instrumental (state-action based) model-free and Pavlovian (state-based) fear system. And we define the pain state-value $V_p(s)$ of the Pavlovian fear system as follows :

$$V_p^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k p_{t+k+1} | s_t = s\right], \tag{4.3}$$

The subset of actions with the Pavlovian bias A_p are arrived at using a pretraining in the same environment with only punishments and random starting points. $V_p(s)$ then bias this pretrained subset of actions A_p according to equation 4.13.

Here onwards, we will drop the time subscript for simplicity and write value update equations considering the state transitions from s to s'. The Pavlovian fear state-value functions are updated as follows:

$$V_p(s) := V_p(s) + \alpha(p + \gamma V_p(s') - V_p(s)) \tag{4.4}$$

The instrumental value function for qualitative value plots is updated in an on-policy manner as follows (but is not used in the PAL algorithm):

$$V(s) := V(s) + \alpha(R + \gamma V(s') - V(s)) \tag{4.5}$$

And the Instrumental action-value functions are updated as follows:

$$Q(s,a) := Q(s,a) + \alpha(\delta) \tag{4.6}$$

where α is the learning rate and while using off-policy Q-learning (sarsamax) algorithm, the TD-errors are calculated as follows:

$$\delta = R + \gamma Q(s', \text{argmax}_{a'}(Q(s', a'))) - Q(s,a) \tag{4.7}$$

The equations above are valid for a general case and are used in grid-world simulations. For model-fitting purposes for the VR Approach-Withdrawal task, there is no next state s', thus the equations reduce to a simpler form of the Rescorla-Wagner learning rule.

Action selection

Let A be the action set. In the purely instrumental case, propensities $\rho(s, a)$ of actions $a \in A$ in state s are the advantages of taking action a in state s:

$$\rho(s, a) = Q(s, a) \tag{4.8}$$

And thus using softmax action selection with a Boltzmann distribution, the stochastic policy $\pi(a|s)$ (probability of taking action a in state s) as follows:

$$\pi(a|s) = \frac{e^{(\rho(s,a)/\tau)}}{\sum_{a' \in A} e^{(\rho(s,a')/\tau)}} \tag{4.9}$$

where τ is the temperature that controls the trade-off between exploration and exploitation. For gridworld simulations, we use hyperbolic annealing of the temperature, where the temperature decreases after every episode i:

$$\tau(i) = \frac{\tau_0}{1 + \tau_k i} \tag{4.10}$$

Here τ_0 is the initial temperature and τ_k controls the rate of annealing. This is to ensure the policy converges in large state spaces like a gridworld and follows previous studies (Elfwing and Seymour, 2017; Wang et al., 2018b). For model-fitting of the VR Approach-Withdrawal task we do not anneal it and keep it as a free parameter (inverse temperature $\beta = 1/\tau$) to be fitted to each participant, this is also consistent with previous literature (Guitart-Masip et al., 2012; Cavanagh et al., 2013; Dorfman and Gershman, 2019; Gershman et al., 2021) and several other works modelling Go-No Go tasks.

In case of a Pavlovian pain/fear system, let A_p be the subset of actions in state s which has the Pavlovian pain urges or impetus associated with it. These are usually a small set of species specific defensive reactions (SSDR). In the VR Approach-Withdrawal task we assume or rather propose it is the bias to withdraw from potentially harmful stimuli (in our case jellyfish). For the purpose of the gridworld simulations, these can either be hardcoded geographical controller moving away from harmful states or through a pretrained value-based controller (Dayan et al., 2006). This work does not delve into the evolutionary acquisition of these biases, but

one can derive the action subset A_p from evolutionarily acquired value initializations which may also help avoid novel stimuli and is a direction for future work.

Thus after adding a Pavlovian fear system over and above the instrumental system, the propensities for actions are modified as follows:

$$\rho(s, a_n) = (1 - \omega)Q(s, a_n); \text{ where } a_n \in A_n = A \setminus A_p. \tag{4.11}$$

$$\rho(s, a_p) = (1 - \omega)Q(s, a_p) + \omega(V_p(s)); \text{ where } a_p \in A_p \tag{4.12}$$

The same can be compactly written as mentioned in the illustration (Fig. 4.1):

$$\rho(s, a) = (1 - \omega)Q(s, a) + \omega(V_p(s, a)); \text{ where } V_p(s, a) = \mathbb{I}[a = a_p]V_p(s) \tag{4.13}$$

where ω is the parameter responsible for Pavlovian-Instrumental transfer. These equations are constructed following the preceding framework by Dayan et al. (2006) which laid out the foundation for interplay between Pavlovian reward system and the instrumental system. $\mathbb{I}[\cdot] = 1 \forall a_p \in A_p$ and $\mathbb{I}[\cdot] = 0 \forall a_n \in A_n = A \setminus A_p$ following the succinct vectorised notation by (Dorfman and Gershman, 2019).

We have refered to this algorithm as the Pavlovian Avoidance Learning (PAL) algorithm in this paper. The equations above assume only Pavlovian fear system in addition to the an instrumental system, and the given equations would vary depending on if we add a Pavlovian reward system too. After this modification, the action selection probabilities are calculate in similar fashion as described in equation 4.9.

Uncertainty based modulation of ω

We further modulate the parameter ω which is responsible for Pavlovian-instrumental transfer using perceived uncertainty in rewards. We use Pearce-Hall associability for this uncertainty estimation based on unsigned prediction errors (Krugel et al., 2009; Zhang et al., 2016, 2018). We maintain a running average of absolute TD-errors δ (equation 4.7) at each state using the following update rule:

$$\Omega_{t+1} = (1 - \alpha_\Omega * \alpha)\Omega_t + \alpha_\Omega * \alpha \mid \delta \mid \tag{4.14}$$

where Ω is the absolute TD-error estimator, α is the learning rate for $Q(s, a)$ and $V(s)$ values as mentioned earlier and $\alpha_\Omega \in [0, 1]$ is the scalar multiplier for the learning rate used for running average of TD-error. To obtain parameter $\omega \in [0, 1]$ from this absolute TD-error estimator $\Omega \in [0, \infty)$, we scale it linearly using scalar κ and clip it between $[0, 1]$ as follows:

$$\omega_t = \min(\kappa\Omega_t, 1) \tag{4.15}$$

We note that the range values Ω takes largely depends on the underlying reward function in the environment and α_Ω. Thus we choose a suitable value of κ for α_Ω using gridsearch in each gridworld environment simulation to ensure that the Pavlovian system dominates in cases of high uncertainty and that the instrumental system starts to take control as uncertainty reduces. We aim to show that this flexible ω scheme is a viable candidate for arbitration between the two systems and addresses the safety-efficiency dilemma wherever it arises. The initial associability Ω_0 is set to 0 in grid-world simulations as there is no principled way to set it. In the case of model-fitting for the VR Approach-Withdrawal task, Ω_0, κ and α_Ω are set as free parameters fitted to each participant and instead of the TD-errors, we have the Rescorla-Wagner rule equivalent - punishment prediction errors (PPE) without any next state s'.

Gridworld simulation details

We consider a series of painful grid-world based navigational tasks including moderate sources of pain (more variations in appendix with catastrophic and dynamic sources of pain). In the grid-worlds, the goal is to navigate from starting position (in blue) to goal position (in green) while avoiding the static moderately painful states (in red). The agent receives a positive reward of 1 for reaching the goal and pain of 0.1 for moderately painful states (red). The pain is encoded as a negative reward of -0.1 in the case of standard RL. Four actions move the agent one step north, south, east, or west (or choose not to move, allowed only in certain environments). If the agent hits a wall, it stays and remains in its

current state. All simulation environments have the following stochastic transition probabilities: 0.9 probability of correct (desired) state transition, whereas with 0.05 probability, the agent's state transitions to the position perpendicular to action taken (right or left). We test the PAL algorithm for varying $\omega = 0.1, 0.5, 0.9$ and for uncertainty-based modulation of flexible ω and compare the performance with standard instrumental policy (Q-learning). The following meta parameters are fixed for all our tabular grid world simulations - learning rate $\alpha = 0.1$, discount factor $\gamma = 0.99$, temperature annealing meta parameters $\tau_0 = 1, \tau_k = 0.025$. The meta parameters α_Ω and κ are tuned using grid-search on the safety-efficiency trade-off metrics for each environment. This is necessary, as different environments have different underlying reward distributions leading to different distributions of TD-errors, thus its running average needs to appropriately scaled to map it to $\omega \in [0, 1]$. Due to these meta parameter tuning, the claim in the simulation experiments is a modest one that there exists a α_Ω and κ that mitigate the trade-off as opposed to the trade-off is mitigated by every possible combination of α_Ω and κ. This resembles the model-fitting procedure in other experimental tasks, where α_Ω and κ are fit in a hierarchical Bayesian manner, suggesting that humans perform this tuning to varying degree to the best of their ability. The Q-tables and V_p-tables are initialized with zeros. Plots are averaged over 10 runs with different seed values.

We quantify safety using cumulative pain accrued by the agent, and sample efficiency using the cumulative steps (or environment interactions or samples) taken by the agent across all the episodes in the learning process. The lesser is the cumulative pain accrued over episodes, the safer is the learning; and the lesser the cumulative steps (or environment interactions), the more efficient is the learning in terms of reward seeking and task completion in each episode. Furthermore, we also construct a trade-off metric to measure how well the safety-efficiency trade-off is improved. We define the safety-efficiency trade-off metrics as follows which is maximised when both cumulative pain and cumulative steps are independently minimised:

$$\text{Trade-off metric} = \frac{1}{CP_n^2 + CS_n^2} \tag{4.16}$$

where CP_n and CS_n are cumulative pain and cumulative steps normalized by dividing the maximum cumulative pain and steps achieved (usually by fixed $\omega = 0$ or $\omega = 0.9$) in that run. We acknowledge that this normalization can make the metric favour improvements in either safety or efficiency unequally to an extent, as it weighs the improvements in safety or efficiency relative to worst performance in each of them. This metric can be further weight-adjusted to give more priority to either CP or CS as required but we don't do that in the current instance. Thus this metric should only be used as didactic tool and not an absolute metric of performance, and one should instead draw conclusions by observing at the cumulative pain accrued and steps taken over multiple episodes.

Approach-Withdrawal conditioning task: experimental design

Participant recruitment and process

30 adults participated in the experiment (15 females, 15 males; age: min=18, max=60, mean=30.5, standard deviation=12.44). Healthy participants from ages 18-60 were allowed to participate in the study (pre-established inclusion criteria). All subjects provided written informed consent for the experiment, which was approved by the local ethics board - University of Oxford Central University Research Ethics Committee (CUREC2 R58778/RE002). One participant withdrew and did not complete the study and one participant turned out to be a fibromyalgia patient upon arrival, thus were excluded. The rest of the 28 healthy subjects' (14 female, average age 27.96 years) data was used for the analysis.

Participants filled a short demographic form upon arrival, followed by pain tolerance calibration procedure, followed by putting on all of the sensors, followed by a re-calibration of their pain tolerance before starting the practice session and the main experiment. All of this was usually completed within 2 hours and participants were paid £30 for their participation (and were adequately compensated for any

unexpected overtime and reasonable travel reimbursements). They were free to withdraw from the experiment at any time.

Trial protocol

We use a trial-based approach-withdrawal task, however the subjects had complete control over when to start the next trial. Each trial consisted of four events: a choice to initiate the trial, a coloured jellyfish cue, an approach or withdrawal motor response and a probabilistic outcome. The timeline is displayed in Fig. 1. In each trial, subjects will initiate the trial by bringing in their hand inside a hovering bubble in front of them. Then a jellyfish will emerge and fade-in (gradually decreasing transparency) within the next 0.5 seconds and then stay in front of the subject for another 1 second, making the total fixation segment 1.5 seconds long. Throughout the fixation segment, the jellyfish colour will remain greyish-black. After this fixation segment terminates, the jellyfish takes one of the four colours with associated pain outcome contingencies. This is the stimulus phase and the subject is required to perform either an approach or a withdrawal response within the next two seconds. Approach response involved reaching out their hand and touching the jellyfish, whereas the withdrawal response involved withdrawing the hand away from the jellyfish and towards one ownself. The subjects practised these two actions in the practice session before the main experiment and were instructed to perform either of these two actions. The stimulus ended as soon as an action was successfully completed and was followed by the probabilistic outcome phase. In the rare case that the two seconds time window completed before the subject could successfully perform either of these two actions, then for the purpose of the probabilistic outcome segment, the action was decided based on the hand-distance from the jellyfish (i.e. whether it was closer to an approach or a withdrawal action). The possible outcomes were either a painful electric shock (along with some shock animation visualisations around the jellyfish) or a neutral outcome (along with bubble animations from the jellyfish). The outcomes were presented depending on the action taken and the contingencies for each cue, as per shown in Fig 4.4. After the outcome segment

which lasted for 1.5 seconds, the jellyfish proceeded to fade-out (become more transparent gradually) for the next 0.75 seconds and then the subject could start the next trial by again bringing their hand within the bubble in front of them.

Subjects were instructed to try to keep their hand inside the bubble during this fixation segment and only move the hand after the jellyfish changes colour. The bubble was placed halfway between the subject and the jellyfish and was placed slightly to the right for right-handed and slightly to the left for left-handed subjects. The subjects performed the task with their dominant hand.

Block protocol

Prior to the main task and the practice session, we perform a calibration of the intensity of pain stimulation used for the experiment according to each individual's pain tolerance. To do this, we start with the minimum stimulation value and gradually increase the value using the 'staircase' procedure. We will record a "threshold" value (typically rated as 3/10 on Likert scale), which is identified as the participant first reports pain sensation. We will record a second "maximum" value, which the participant reports as the maximum pain sensation that the participant would be comfortable to tolerate for the complete experiment (typically rated 8/10 on the Likert scale). We then use 80% of that maximum value for stimulation throughout the experiment.

Before the main task, the subjects had to go through a short practice session to get acquainted with approach and withdrawal motions and the speed requirements. Subjects had one attempt at each of the two actions with no painful outcomes and no timeouts followed by a short practice session with two jellyfish (5 trials each, randomised) and with 80% painful outcome contingencies for approach and withdrawal respectively. They were informed as to which of these two jellyfish likes to be touched and which does not, during the practice session but not for the main experiment. The colours of the jellyfish for the practice session were different from that used for the main experiment. The four colours of the jellyfish cues for the main experiment were chosen so as to be colourblind-friendly. The

main experiment had a total of 240 trials, 60 trials for each of the four jellyfish which was balanced across each quarter of the block i.e. 15 trials per jellyfish per quarter block. The jellyfish 1 was the approach-to-avoid type and jellyfish 2 was the withdraw-to-avoid type throughout 240 trials. The jellyfish 3 was uncontrollable for the first half of the block (first 120 trials) and then was approach-to-avoid type for the rest of the block. The jellyfish 4 was uncontrollable for the first half of the block (first 120 trials) and then was the withdraw-to-avoid type for the rest of the block. Approach-to-avoid type means that outcome would be a neutral outcome 80% of the times (and shock, 20% of the times) if the 'correct' approach action was performed or else the outcome would be a shock 80% of the times (and neutral, 20% of the times) if the 'incorrect' withdrawal action was performed. Withdraw-to-avoid type means that outcome would be a neutral outcome 80% of the times (and shock, 20% of the times) if the 'correct' withdraw action was performed or else the outcome would be a shock 80% of the times (and neutral, 20% of the times) if the 'incorrect' approach action was performed. Uncontrollable type means that the outcome could be shock or neutral with 50% probability each, regardless of the actions performed. After each quarter of the block, the subjects were informed of their progress through the block with a 10 second rest.

Analysis

The choices and reaction times were extracted and used for model-fitting. The EEG, EMG and skin conductance data was acquired but found to be too corrupted by movement artefacts and noise to allow reliable analysis.

Hierarchical bayesian model-fitting choices and reaction times

For both RL model-fitting to choices and RLDDM model-fitting to choices and reaction times, we built 4 models each: RW (i.e. Rescorla-Wagner learning rule model), RW+bias (i.e. RW model with baseline bias), RW+bias+Pavlovian(fixed) (i.e. RW+bias model with a fixed Pavlovian withdrawal bias) and RW+bias+Pavlovian(flexible)

(i.e. RW+bias model with a flexible Pavlovian withdrawal bias) similar to Guitart-Masip et al. (2012).

The action selection for RL models was performed using a softmax as per equation 4.9 with free parameter $\beta = 1/\tau$ and $\beta > 0$. The learning rule for RW models was:

$$Q(s,a) := Q(s,a) + \alpha(R - Q(s,a)) \tag{4.17}$$

where $R = -1$ in case of electric shocks or $R = 0$ in case of neutral outcome. Punishment p can be defined as per equation 4.2 and thus $p = 1$ in case of electric shocks or $p = 0$ and the Pavlovian punishment value is calculated as per:

$$V_p(s) := V_p(s) + \alpha(p - V_p(s)) \tag{4.18}$$

$\alpha > 0$ is the learning rate and fitted as a free-parameter and note that here V_p is always positive.

For RW+bias model,

$$\rho(s,a) = Q(s,a) + b; \text{ if } a = \text{Approach} \tag{4.19}$$

$$\rho(s,a) = Q(s,a); \text{ else} \tag{4.20}$$

Here $b \in (-\infty, +\infty)$ is the baseline bias, which if positive represents a baseline approach bias and if negative represents baseline negative bias and is not Pavlovian in nature.

For RW+bias+Pavlovian(fixed) and RW+bias+Pavlovian(flexible) models,

$$\rho(s,a) = (1-\omega)Q(s,a) + b; \text{ if } a = \text{Approach} \tag{4.21}$$

$$\rho(s,a) = (1-\omega)Q(s,a) + \omega(V_p(s)); \text{ if } a = \text{Withdrawal} \tag{4.22}$$

Here $\omega \in [0,1]$ is a free parameter for the RW+bias+Pavlovian(fixed) model. ω is not a free parameter for the RW+bias+Pavlovian(flexible) model, but computed as per equations 4.14 and 4.15 with free parameters Ω_0 (initial associability), κ (scaling factor for ω) and α_Ω (learning rate multiplier for associability).

For RLDDM models, it is assumed that within a trial, the evidence is accumulated using a drift-diffusion process with parameters drift rate (v), non-decision time

(ndt), threshold and starting point. Non-decision time and threshold were kept as free parameters and the starting point was kept constant and equal to half the threshold (making it equally likely starting point for both approach and avoidance actions). The drift rate v was set according to the difference in action propensities between the choices as follows.

$$v = \rho(s, a = \text{Approach}) - \rho(s, a = \text{Withdrawal}) \tag{4.23}$$

Thus, the baseline bias and the Pavlovian biases were also included in the drift rate.

For model-fitting we used a hierarchical Bayesian modelling approach, all models were fit using Stan. They were fit using both custom code in PyStan as well as using the hBayesDM package (Ahn et al., 2017) and final plots of group-level and subject-level parameter distributions were generated using the plotting functions in hBayesDM. Four parallel chains were run for all models. To assess the predictive accuracy of the models, we computed the leave one out information criterion (LOOIC) and Watanabe-Aikake information criterion (WAIC) (Vehtari et al., 2017).

Software and hardware setup

We used the HTC Vive Pro Eye for the virtual reality (VR) with Alienware PC setup and the experiment was designed in Unity (game engine). The pain stimulator used was DS5 with WASP electrodes for the VR approach-withdrawal task and Silver-Silver Chloride (Ag/AgCl) Cup Electrodes for the VR maze task. We also collected glavanic skin response (GSR), heart rate(HR), electromyography(EMG) signals, wireless EEG using Brainproducts LiveAmp and Vive tracker movement signals and eye-tracking inside the VR headset.

The pain stimulator electrodes were attached on the ring finger, between the ring and the middle finger. The GSR sensors were attached on the middle and the index fingers and the EMG sensors were attached on the brachioradialis muscle of the active hand used in the task with the ground electrode on the elbow. Heart rate sensor was attached to the index finger of the opposite hand.

The task is ... not so much to see what no one has yet seen; but to think what nobody has yet thought, about that which everybody sees.

— *Erwin Schrödinger*

5

Optimal composition of multiple values for dopamine-mediated efficient, safe and stable learning

Contents

5.1 Prelude

The previous chapter explored the safety-efficiency dilemma as a conflict between control systems of differing complexity, suggesting the possibility that the trade-off could be mitigated by composing controllers of similar capabilities. This, however, raises a more subtle question: how does the brain combine these controllers when they are optimising different, potentially conflicting objectives? This is the essence of multi-attribute decision-making, a challenge animals must constantly solve when balancing competing needs in changing environments (Dulberg et al., 2023; Millidge

et al., 2024a). The biological substrate for such an architecture is evident in the rich heterogeneity of dopaminergic signals, which show distinct tuning to different reward attributes, punishments, and threats (Lak et al., 2014; Watabe-Uchida and Uchida, 2018). Indeed, the dopamine system is directly implicated in safety, with signals in the tail of the striatum (TS) hypothesised to encode an innate, pessimistic value initialisation that promotes cautious avoidance of novel stimuli (Menegas et al., 2018; Akiti et al., 2022). This points to a critical computational challenge that this chapter addresses: if the brain learns multiple, parallel value functions, how are they optimally composed to ensure that learning is not only efficient and safe, but also stable and reliable?

5.2 Introduction

Almost three decades ago, Schultz et al. (1997) proposed that midbrain dopamine (DA) neuron phasic activity encodes temporal difference errors (TDEs). This fundamental idea, leveraging the temporal difference (TD) learning algorithm (Sutton, 1988), suggested the brain could assign credit in terms of expected future reward using temporally successive predictions. Numerous experiments have since substantiated this relationship within the TD reinforcement learning (TDRL) framework (Niv and Schoenbaum, 2008; Glimcher, 2011; Eshel et al., 2015; Watabe-Uchida et al., 2017; Zhang et al., 2025). A core motivation was that engineered systems employ similar algorithms to optimise actions in complex environments, mirroring challenges faced by animals (Schultz et al., 1997). However, three key challenges currently impede TDRL research from fully realising this core motivation.

The first challenge arises because typical experiments and computational models, utilising single-attribute rewards (e.g., juice) and monolithic value functions, inadequately capture the multi-objective nature of challenges animals face in real life. That is, they constantly need to simultaneously satisfy distinct, often conflicting objectives under changing priorities, a kind of non-stationary rewards. This necessitates either multiple value functions (Dulberg et al., 2023; Millidge et al., 2024a) or alternative efficient representations like the successor representation (Dayan, 1993; Gershman,

2018; Gardner et al., 2018), as standard RL approaches struggle with instant revaluation under non-stationary rewards (Dulberg et al., 2023; Millidge et al., 2024a; Padakandla et al., 2020). This need for fast adaptation to multiple objectives (often encoded by different rewarding attributes), under shifting priorities, is ubiquitous in homeostasis (Robinson and Berridge, 2013; Keramati and Gutkin, 2014; Richman et al., 2023) and extends to human cognitive tasks (Tomov et al., 2021). Neural evidence further suggests dopaminergic circuits projecting to different targets may encode multiple value functions corresponding to various reward modalities (e.g., food, juice, water (Schultz, 2000; Carelli et al., 2000; Enomoto et al., 2011; Lak et al., 2014; Takahashi et al., 2017)), valence (e.g., threats versus rewards (Watabe-Uchida and Uchida, 2018)), substance type (Chang et al., 1998), or even abstract features and contexts (Babayan et al., 2018; Gershman and Uchida, 2019; Takahashi et al., 2023), but this isn't captured by the standard TDRL framework.

The second challenge concerns the TD-learning rule used by Schultz et al. (1997), which learns values under the current behavioural policy (on-policy algorithms in RL), as opposed to values under an optimal policy (off-policy algorithms). A key implication is that the TD-learning algorithm learns values which are estimates of the so-called future returns, a long-run measure of cumulative rewards, assuming the agent continues to act using its current behavioural policy. While a core motivation was to learn optimal actions (Schultz et al., 1997), on-policy algorithms often learn suboptimal values under an often-exploring, suboptimal policy. This compromise, well-noted in traditional single-objective RL (Sutton and Barto, 2018) and usually dealt with by tuning exploratory noise, becomes particularly problematic in multi-objective RL, where the behavioural policy can change drastically with shifts in needs or non-stationarity (Dulberg et al., 2023).

The third challenge stems from widespread dopamine heterogeneity, which calls into question the physiological basis of a single, broadcasted scalar reward prediction error (RPE). While a final scalar variable is computationally necessary to guide choice, the notion of a monolithic RPE signal is increasingly at odds with

observations of diverse dopamine responses, both between and within striatal subregions. Furthermore, extending the classical model Schultz et al. (1997) to multiple objectives (Millidge et al., 2024a) only partially explains this diversity (e.g. between different DA targets), leaving more perplexing forms of heterogeneity unresolved.

Nowhere is this challenge more apparent than in the tail of the striatum (TS), whose normative role is hotly debated. Evidence suggests TS-projecting neurons encode threat prediction errors (TPEs) to guide avoidance, even without explicit aversive reinforcement (Menegas et al., 2017, 2018; Akiti et al., 2022; Tsutsui-Kimura et al., 2025), with direct and indirect pathway heterogeneity (Tsutsui-Kimura et al., 2025). Others posit that TS-projecting neurons encode action prediction errors (APEs) (Greenstreet et al., 2025), which have been linked to "value-free" habits (Miller et al., 2019) and are difficult to integrate into a standard value-based TDRL framework, as they are often modelled as a separate ad-hoc component (Miller et al., 2019; Lindsey and Litwin-Kumar, 2022; Bogacz, 2020). Reconciling these conflicting views is a central problem; recent unifying accounts, for instance, often highlight the TS's role in stimulus-associated salience predictions but tend to disregard APEs (Green et al., 2024). Thus, a parsimonious account that integrates the normative roles of both TPEs and APEs into the TDRL framework remains elusive.

While distinct, the first two challenges - the need for multi-objective learning and the pitfalls of on-policy methods - converge on a single, fundamental question of optimal composition of multiple values. The brain clearly possesses mechanisms for learning multiple values, but how can they be reliably combined to produce a coherent and optimal policy? Ideally, this "recipe" for combining values should satisfy two key properties: optimality and compositionality. Optimality, the bedrock of modern reinforcement learning (Bellman, 1952), simply means choosing the best possible action (Todorov, 2009b). Compositionality, conversely, is the formal principle of constructing solutions to complex problems from a set of modular components, for instance by allowing a multi-attribute task to be represented by a basis set of value functions tuned to individual reward dimensions. Crucially, this structure allows policies for novel priority landscapes to be constructed by

flexibly recombining these components, obviating the need to learn each new configuration de novo.

As noted previously (Zhang et al., 2025), the choice of learning algorithm is not a mere technicality but a foundational choice dictating the system's computational objectives. Recent proposals often overlook this, foregoing either optimality (Millidge et al., 2024a) or composability (Dulberg et al., 2023). On-policy methods (e.g., SARSA), for instance, suffer from learning interference, especially under shifting contextual priorities. When the current context prioritises one reward, the agent's policy becomes biased; because learning is tied to this policy, the valuation of alternative rewards is not learned in isolation but is corrupted by being evaluated through the lens of the current trajectory. In contrast, off-policy methods (e.g., Q-learning) fail because of the non-linearity of the Bellman optimality (max) operator, which is not additive. This creates an unrealistic assumption: the agent is treated as perfectly rational during valuation (identifying the single best future action) yet as boundedly rational during action selection (making stochastic choices). This inconsistency between an ideal valuation and a bounded action-selection corrupts the composition of different reward values, revealing the need for an internally consistent framework for decision making.

To resolve this, we adopt a normative framework from control engineering (Kappen, 2005; Todorov, 2006, 2009b; Dvijotham and Todorov, 2012), broadly termed linear RL (Piray and Daw, 2021). We propose that the dopamine system's objective is not merely to optimise cumulative reward, but to optimise returns augmented by a penalty for deviating from a default policy. This single modification provides a principled solution to the optimal composition problem (Todorov, 2009a). By enforcing a consistent assumption of bounded rationality throughout both valuation and action selection, it resolves the paradoxical logic of standard algorithms, allowing multiple values to be robustly combined.

Remarkably, the same principle that ensures optimal composition also provides a unified normative account of dopamine heterogeneity. The framework's parallel architecture for outcome-specific prediction errors supports efficient learning across

multiple rewards and rapid adaptation. Furthermore, composing values from different initialisations can explain within-target heterogeneity and threat prediction errors (TPEs) and how the flexible expression of such innate values can drive safe learning. Most strikingly, the penalty term itself manifests computationally as an action prediction error (APE), revealing its function in promoting stable learning by conferring a value on controllability. We substantiate these claims in our Results, which use a didactic example and simulations to demonstrate each of these respective normative advantages.

5.3 Theory sketch

The model utilises the linearly-solvable Markov Decision Process (MDP) framework (Kappen, 2005; Todorov, 2006, 2009b; Dvijotham and Todorov, 2012), also known as or entropy-regularised or KL-regularised RL. Distinct from the original control theoretic formulations (Todorov, 2006, 2009b), we focus on temporal difference learning rules for solving this and extend it to learning values for multiple objectives or reward types, leveraging inherent compositionality theory afforded by the linear MDP framework (Todorov, 2009a).

At Marr's computational level (Marr, 2010), our agent optimises returns augmented by an additional KL divergence term ($D_{\mathrm{KL}}(\pi(\cdot|s_t)\|\pi^d(\cdot|s_t))$) that penalises proportional to how much the current behavioural policy ($\pi(\cdot|s_t)$) deviates from some default policy ($\pi^d(\cdot|s_t)$)) at each timestep:

$$G_t = \sum_{k=0}^{\infty} \gamma^k (r_{t+k+1} - \tau D_{\mathrm{KL}}(\pi(\cdot|s_t)\|\pi^d(\cdot|s_t)))$$

where G_t is the return from timestep t onwards and r_{t+k+1}, is the reward at the k timesteps after t, following standard notation from Sutton and Barto (2018). The KL divergence term acts as a relative-entropy regularisation. If π^d is uniform, this encourages random exploration (Max Entropy RL); if π^d slowly tracks the learned policy π, it promotes choice perseveration or soft-habit formation by penalising deviations from recently taken actions (Miller et al., 2019; Piray

and Daw, 2021; Gershman, 2020; Thorndike, 1900). Lastly, we note that entropy-regularised RL is mathematically equivalent to planning/control as probabilistic inference (Levine, 2018).

At the algorithmic level, we implement this using soft Q-learning (Haarnoja et al., 2017), an off-policy temporal difference (TD) method, adapted here for the general relative-entropy objective. Q-values are updated via:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t,$$

where α is the learning rate, and δ_t is the reward prediction error at timestep t, defined as:

$$\delta_t = r_{t+1} + \gamma V_Q(s_{t+1}) - Q(s_t, a_t),$$

where the soft state-value is given by $V_Q(s) = \tau \log \mathbb{E}_{a\sim\pi^d} \exp(Q_\pi(s, a)/\tau)$(see Methods, Eq. 5.9). $V_Q(s)$ replaces the $\max_a Q(s, a)$ operation in the Bellman optimality equation, which is often (biologically) intractable in large or continuous action spaces, with a more computationally feasible expectation, with other properties favouring optimal composition of multiple values. Notably, for this one-step algorithm, the KL-divergence from the objective does not appear in δ_t because the action a_t has already been taken (Ziebart, 2010; Haarnoja et al., 2017; Schulman et al., 2017); however, it re-emerges in our derivations of multi-step extensions (Section 5.4). Actions are selected using a Boltzmann policy derived from the Q-values, which is optimal under this regularised objective (Methods, Eq. 5.8).

This soft Q-learning rule is applied independently to learn multiple Q-functions for different objectives (r_i), which are then composed using a weighted softmax or weighted summation operation f, with weight vector $\mathbf{w}$ (Theorem 1, Methods and Appendix B.2). Note, taking an action at timestep t results in rewards $r_{i,t+1}$, for the i-th reward dimension, but we will omit the time subscript from here for convenience. The reward weighting $\mathbf{w}$ is determined by either internal homeostatic priorities (e.g. hunger, thirst, salt deprivation (Robinson and Berridge, 2013; Millidge et al., 2024a)

or injury belief state (Mahajan et al., 2025a)) or inferred task-based rules (Babayan et al., 2018; Gershman and Uchida, 2019) or survival priorities (e.g. threat inference).

$$r_c(s,a) = f(r_1, r_2, \ldots, r_n; \mathbf{w})$$

Importantly, the compositionality of optimal control laws in linear MDPs (Todorov, 2009a) allows us to decompose the multi-objective RL problem, into individual value components $Q_i(s,a)$ optimising independent objectives r_i, whilst ensuring that combining these individual values, using the same composition f, ensures the resultant policy optimises the intended reward composition r_c on the multi-dimensional reward manifold.

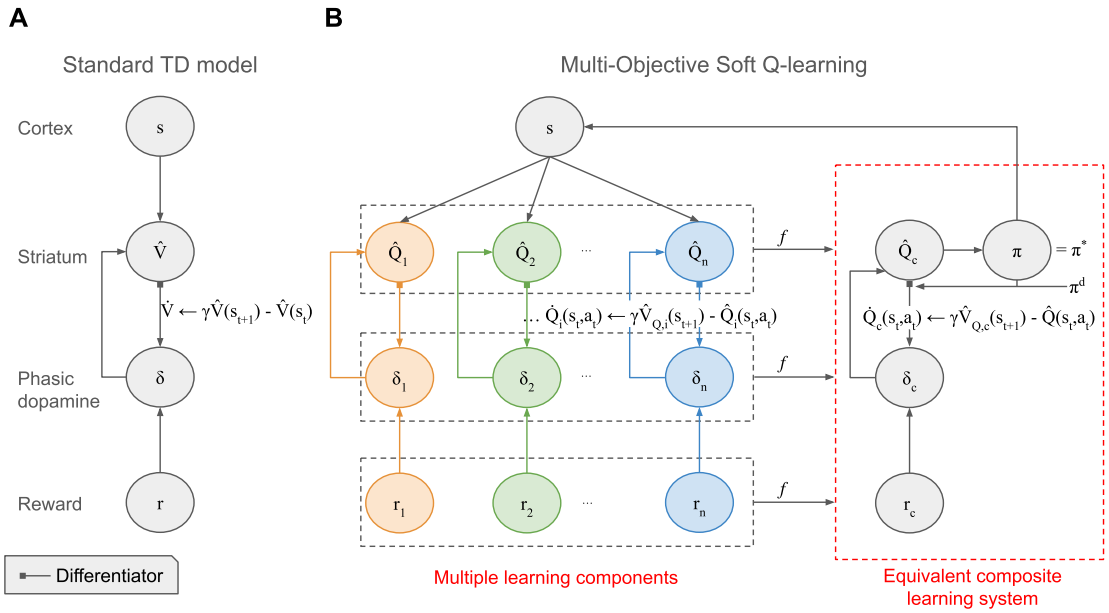


Figure 5.1: Proposed neural implementation for optimally composable multi-objective RL. (A) Conventional TD learning model (Schultz et al., 1997): Dopamine neurons compute a scalar reward prediction error (RPE) that drives learning of a single state value function in the striatum. Extensions feature parallel systems for different outcomes, generating vectors of state values and RPEs (Millidge et al., 2024a). (B) Our model: Parallel soft Q-learning modules learn distinct state-action values (Q-values) for multiple or shared outcomes within the linearly solvable MDP framework (Todorov, 2006). These Q-values are optimally composed (f) to guide behaviour that optimises a weighted combination of rewards. This architecture supports parallelisation across different dopaminergic targets and allows for within-target heterogeneity through differing value initialisations (even without immediate outcomes). A key desideratum is access to the default policy π^d by the TD error computing differentiator (Schultz et al., 1997) for computing the soft state values V_Q.

At Marr's implementation level, our model posits parallel soft Q-learning mechanisms to account for heterogeneity in dopaminergic responses, both between and within distinct neural targets (Fig. 6.1). Learning separate, composable Q-values for different reward attributes (e.g., appetitive, aversive) allows for efficient acquisition of multi-attribute rewards. It also enables the representation of dedicated threat or punishment values essential for safe behaviour, preventing them from being overridden by unrelated positive rewards (Elfwing and Seymour, 2017; Mahajan et al., 2024; Dulberg and Cohen, 2025). However, within a single dopaminergic target (e.g. the tail of the striatum), we propose that heterogeneity can also arise from multiple value functions sharing a common outcome signal but differing in their initialisations, potentially reflecting priors for different contexts, associated with different priorities. Such value initialisations can drive behaviour and explain dopaminergic responses even without explicit outcomes (Kakade and Dayan, 2002; Dayan, 2022; Akiti et al., 2022), a concept addressed in Section 5.4.

We first illustrate the principles of optimal value composition with a didactic simulation, building on Todorov (2009a), while situating and comparing with contemporary biological multi-objective RL models, then proceed to more concrete experiments.

5.4 Results

Optimal composition of multiple value functions for reliable optimisation

This is a didactic analysis and simulation demonstrating our claim that recently proposed methods for learning and combining multiple values either forego optimality (Millidge et al., 2024a) or composability (Dulberg et al., 2023). A recipe for optimal (or near-optimal) composition allows a system to combine several independent components to reliably optimise the resultant composite reward function. Put simply, it is a strategy to get several selfish value functions to behave well together.

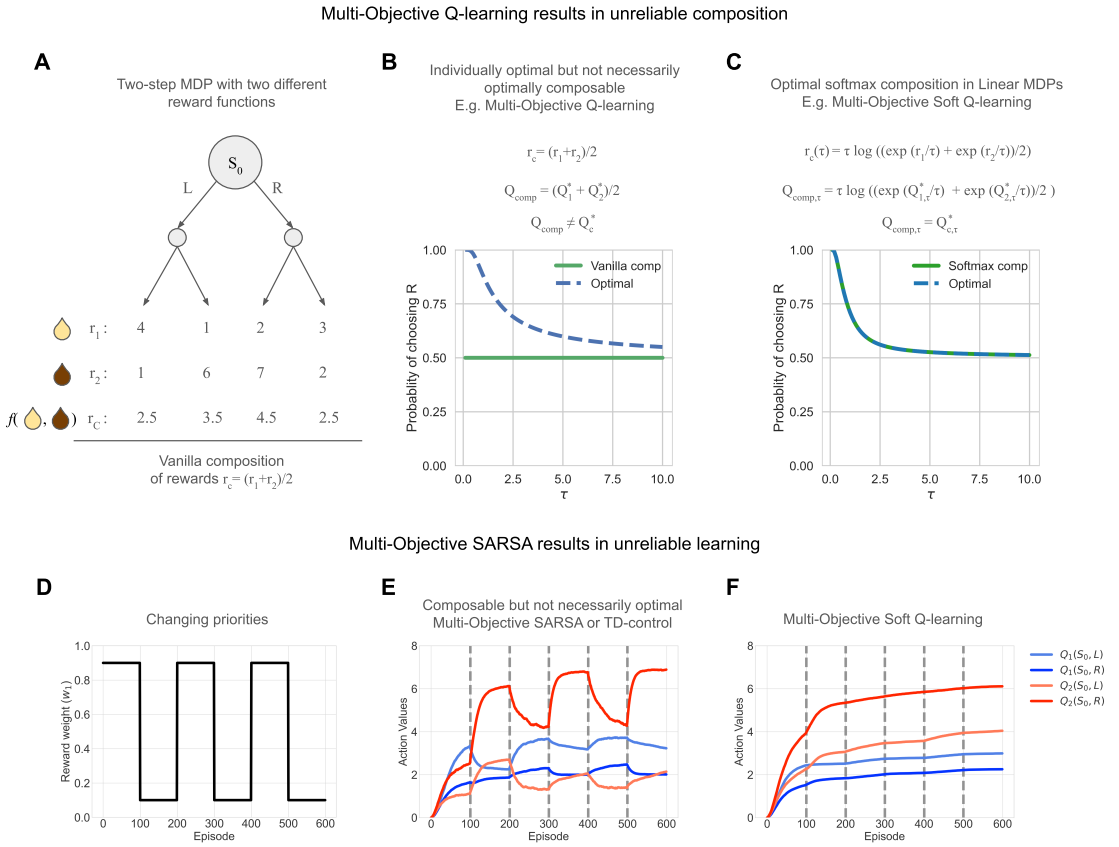


Figure 5.2: Demonstration of the reliable and optimal composition of values in linear MDP. (A) Two-step MDP with two (diverging) reward functions, say Juice 1 (r_1) and Juice 2 (r_2). (B) Action selection probabilities under sub-optimal additive composition of Q-values in MDPs (Q_{comp}) deviate from optimal behaviour mandated by Q^*_c optimising r_c. (C) Action selection probabilities under optimal softmax composition in Linear MDPs (Q_{comp}) match optimal behaviour mandated by Q^*_c optimising r_c. Weights set to $w_1 = w_2 = 0.5$ (equal priority for both rewards) and action probabilities plotted for a range of τ. (D) Changing reward weight w_1, denoting the change in priorities and $w_2 = 1 - w_1$. (E) Action values of multi-objective (MO) SARSA show unstable and unreliable learning over episodes. (F) Action values of MO Soft Q-learning show reliable and stable learning over episodes. Note that the focus is on value learning over episodes is stable or has interference; action values from (E) and (F) cannot be compared as the objective functions are different.

The general approach in prior works in multi-objective RL often linearly decomposes a composite reward r_c into attributes $(r_1, r_2, \dots)$. Separate value functions $(Q_1, Q_2, \dots)$ are then learned, each optimising returns for a single attribute. These individual value functions are subsequently linearly combined (using the same weights as the reward decomposition) to form a composite value function

Q_{comp} that guides a unified policy (Russell and Zimdars, 2003; Dulberg et al., 2023; Millidge et al., 2024a). However, depending on the specific value learning algorithm, this can lead to a critical trade-off: either the individual values Q_i are optimal for their respective rewards r_i but their composition Q_{comp} is not optimal for r_c (i.e., $Q_{comp} \neq Q_c^*$), or the composition is well-defined but the individual values Q_i themselves are sub-optimal (i.e., $Q_i \neq Q_i^*$).

To illustrate the first issue - individually optimal values that compose sub-optimally, we consider the off-policy, multi-objective Q-learning, devised by Russell and Zimdars (2003) and utilised by Dulberg et al. (2023). This method learns optimal Q_i for each r_i using standard Q-learning (using equation 5.10) and then additively combines them (referred to as Vanilla comp). In a two-step MDP with two reward functions r_1, r_2 (Fig. 5.2A), if we equally weight rewards ($w_1 = w_2 = 0.5$) to get r_c, the (undiscounted) optimal Q-values for r_c in the starting state S_0 are $Q_c^*(S_0, a = L) = 3.5$ and $Q_c^*(S_0, a = R) = 4.5$. However, the additively composed Q-values are $Q_{comp}(S_0, a = L) = Q_{comp}(S_0, a = R) = 5$. Consequently, an agent using Q_{comp} with a softmax policy chooses actions L and R with equal probability, irrespective of the temperature τ, deviating from the optimal behaviour for r_c (Fig. 5.2B). This sub-optimality arises from the non-linearity of the max operator in the Bellman optimality equation (see Methods, Section 5.6).

To make this concrete, one can view this as multi-attribute decision making problem with two juices of similar utility per millilitre consumption. The left path from S_0 leads to maximum consumption of Juice 1 (r_1) whereas the right path leads to maximum Juice 2 (r_2). The abstract result above — where the flawed multi-objective Q-learning algorithm computes both left and right actions at S_0 as equally valuable even when the path leading to $r_c = 4.5$ is objectively better — demonstrates how the inconsistent rationality assumption in standard Q-learning leads to suboptimal choices. Here r_c is an internally constructed composite reward based on priorities w_1 and w_2 and external rewards r_1 and r_2.

In contrast, our approach using soft Q-learning within the linear MDP framework allows for a reliable composition. Here, rewards $r_{c,\tau}$, a composite reward function

composed from individual rewards r_1 and r_2, and Q-values $Q_{comp,\tau}$ are functions of τ, which controls the influence of a default policy π^d (here, uniform, though its specific choice does not alter this result). As shown in Fig. 5.2C, the resulting composed policy reliably optimises the target reward composition, achieving $Q_{comp,\tau} = Q^*_{c,\tau}$ (Theorem 1, Methods Section 5.6). This demonstrates optimal value composition (Todorov, 2009a). For completeness, we also simulated a viable alternative - the additive reward composition within this linear MDP (Appendix Fig. B.1 and text), which also outperforms MO Q-learning additive composition in MDPs, along with theoretical guarantees (Haarnoja et al., 2018).

On the other hand, works that employ on-policy algorithms, i.e. TD(0) (Schultz et al., 1997), its multi-objective extension (Millidge et al., 2024a) and extensions to control (SARSA), are known to not reliably learn optimal policies, especially from sub-optimal trajectories (see Appendix Fig. B.2).

However, the limitations of on-policy learning become particularly acute in multi-objective scenarios with dynamically shifting priorities. This is because policies optimal to one value component are bound to be sub-optimal for other value components, but values for all components are learnt under a common behavioural policy. Here, we highlight a critical form of interference in multi-objective (MO) on-policy algorithms such as MO SARSA (or vanilla Reward Bases (Millidge et al., 2024a)): as priorities shift (Fig. 5.2D), the ensuing changes in the global behaviour policy directly impact the valuation of individual components (Appendix B.1 illustrates different policies under different priorities). Because on-policy value updates (e.g., in MO SARSA) depend on the current policy π (i.e., $V_{i,\pi}(s')$), actions taken to optimise one reward modality can lead to unintended revaluation and even unlearning of values for other modalities (Fig. 5.2E). In our juice example, this models a task where the rules frequently switch, making only one juice the prioritised or most rewarding option at a time (Fig. 5.2D). This could also be implemented via changes in homeostatic needs, although these typically occur on a slower timescale. The unstable value estimates demonstrate how an on-policy agent struggles to adapt: its learned value for 'Juice 1' becomes corrupted while it pursues 'Juice 2', preventing

a flexible switch when the rules change. In contrast, our off-policy multi-objective (MO) soft Q-learning framework effectively mitigates this interference, ensuring stable and accurate learning of all value components (Fig. 5.2F). These divergent learning dynamics — instability in on-policy versus stability in off-policy approaches under shifting priorities — offer clear, experimentally testable predictions.

In summary, this section demonstrates that MO soft Q-learning achieves a reliable and optimal composition of multiple values. This contrasts with MO SARSA, which suffers from unstable value learning under shifting priorities, and standard MO Q-learning, which can yield sub-optimal compositions. It is important to note that several of the diverging predictions arise in MDPs with two or more steps sharing some common states, whereas several experiments (Lak et al., 2014) and subsequent modelling (Millidge et al., 2024a) are limited to one-step tasks where all learning rules reduce to the Rescorla-Wagner rule and do not yield diverging predictions. These diverging predictions can be used to tease apart strategies used in multi-attribute decision making.

Efficient learning and off-policy fast adaptation

Efficient adaptation to non-stationary rewards is a hallmark of intelligence. While full model-based RL offers a solution, it is computationally costly. Part model-free solutions rely on either (i) composing independent value functions for different (pre-defined) reward types, as discussed previously (Dulberg et al., 2023; Millidge et al., 2024a), or (ii) learning efficient representations such as the successor representation (SR) (Dayan, 1993), which allow rapid revaluation of policies when reward functions change.

Crucially, in the case of on-policy model-free algorithms, Millidge et al. (2024a) show that a combination of TD(0) learning rule for each of the reward types is akin to a compressed SR with rewards tuned only to relevant dimensions. An equivalent relationship in off-policy algorithms is lacking. The default representation (DR), an off-policy counterpart to the SR derived from linear MDPs (Piray and Daw, 2021; Todorov, 2006, 2009b), overcomes the on-policy limitations of the SR and

offers such a path. We first establish a theoretical link between our multi-objective (MO) Soft Q-learning and the DR, and then proceed to empirically demonstrate its superior performance over on-policy alternatives.

Theoretical result: Relationship to the default representation (DR). MO Soft Q-learning learns values equivalent to those learned by a compressed DR tuned to only relevant (predefined) reward dimensions (Methods, Section 5.6). It further provides two benefits: First, MO Soft Q-learning scales linearly with state space size (assuming a fixed, smaller number of reward dimensions), whereas the full DR scales quadratically. Second, unlike the SR, the DR lacks an efficient TD learning algorithm and typically requires matrix inversions for its computation, which are biologically implausible. MO Soft Q-learning provides a TD-based mechanism to learn DR-like values for the relevant pre-defined reward dimensions.

Simulation result: MO Soft Q-learning enables superior adaptation to shifting priorities. We empirically tested these advantages in a four-room grid world where an agent pursued one of three goals, with priorities shifting every 1000 episodes (Fig. 5.3A; see Methods for more details). This task, though standard, is more complex than some prior grid worlds (Dulberg et al., 2023; Millidge et al., 2024a), by introducing walls.

MO Soft Q-learning outperformed SR and other MO TD algorithms in total rewards accrued (Fig. 5.3B), demonstrating superior adaptation. The policy-dependence of SR was particularly detrimental during priority shifts requiring substantial re-planning (Russek et al., 2017). For instance, when needing to switch back to a previously learned goal after extensive training on another, the SR agent took orders of magnitude more steps than MO Soft Q-learning or even MO SARSA (Fig. 5.3C, episodes 3000, 4000, 5000). MO Soft Q-learning, by maintaining relatively stable, independent values for each reward, could immediately leverage the appropriate value function. The adaptation rates over episodes further illustrate these differences (Fig. 5.3D). These findings not only highlight SR's limitations in dynamic environments but also offer differentiating testable predictions against SR-based models of dopamine (Gardner et al., 2018) and other MO TD approaches

like Reward Bases (Millidge et al., 2024a), previously unidentified in Millidge et al. (2024a).

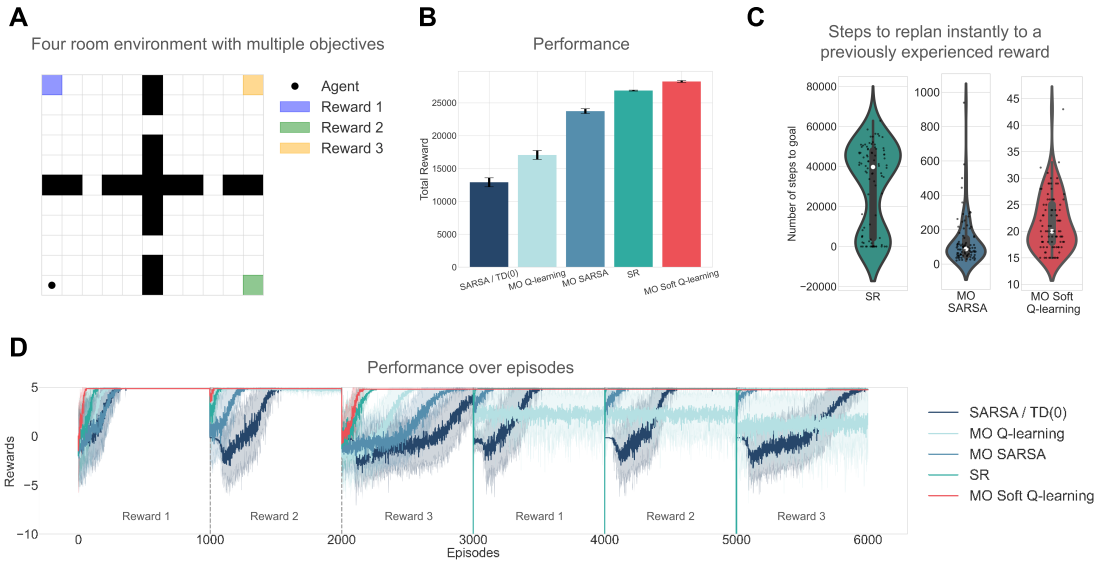


Figure 5.3: Demonstration of efficient learning and fast adaption to changing priorities in a four-room environment. (A) Four-room environment with changing priorities between three rewards every 1000 episodes, and the agent starts in the same starting position (bottom-left corner, akin to a shelter). Rewards for goals: +5 points, step cost: -0.01. The episode terminates only on reaching one of the rewarding goals. Meta-parameter $\tau = 0.5$ to allow all algorithms to converge to an efficient path to the goal. (B) Multi-objective (MO) Soft Q-learning algorithm performs the best amongst comparisons in terms of total rewards accrued, highlighting its fast adaptation capabilities. (C) Steps to the goal on episodes 3000, 4000 and 5000 are plotted for different algorithms to test replanning to a previously experienced reward. SR performs the worst at replanning, requiring substantial policy re-evaluation, while MO Soft Q-learning performs the best amongst comparisons. (D) Performance over episodes shows different rates of adaptation for different algorithms upon a change of priorities. Do note, SR accrues many losses on episodes 3000, 4000 and 5000, but the plot is truncated -10 average rewards for visualisation.

Lastly, we observe that off-policy algorithms continue to propagate optimal Q-values for all components, while collecting data under different policies (priorities), unlike on-policy algorithms (Appendix Fig. B.4). However, MO Q-learning requires commensurate (heuristic-based) temperature annealing to get the most benefits, as also seen in Dulberg et al. (2023), whereas MO soft Q-learning manages this trade-off without such explicit annealing. Such a multi-objective RL solution for efficient learning can be mapped to different DA targets, responsible for different reward

bases (for example, see Millidge et al. (2024a)). Next, we show how the same recipe can be extended to model and explain certain heterogeneities within DA targets.

Safe learning and explaining novelty responses in TS

Beyond efficient adaptation to changing rewards, relying on an aversive value initialisation that can be flexibly expressed can be useful in generating safe behaviours, such as avoiding potential threats. Recent experiments (Watabe-Uchida and Uchida, 2018; Menegas et al., 2018; Akiti et al., 2022; Tsutsui-Kimura et al., 2025) propose that the tail of the striatum (TS) encodes initial threat predictions for novel stimuli, contributing to avoidance, and are updated by dopamine-mediated threat prediction errors (TPEs). We revisit an idiosyncrasy in the observed data, propose a model qualitatively explaining some of the findings and discuss implications for pathway-dependent heterogeneity in the TS.

Simulation result: Threat belief-gated value composition reproduces approach-retreat dynamics and TS dopamine signals. Studies in mice reveal that interactions with a novel object involve approach-retreat bouts, importantly, accompanied by TS dopamine activity during retreat but not during approach (Menegas et al., 2018; Akiti et al., 2022) (Figs. 5.4A, B). These studies model TS dopamine with aversive prediction errors and model this phenomenon using a complete serial compound (CSC) model (Akiti et al., 2022), which incorporates value initialisation, which is equivalent to potential-based reward shaping bonus (Wiewiora, 2003) (Fig. 5.4C). The CSC-based model, albeit tremendously helpful in describing the TS responses, relies on arbitrary thresholds for engage-avoid decisions. Further, it cannot produce approach-retreat bout behaviour in space, and therefore, it is unclear if reward shaping-based modelling of TS responses extends directly to 2D state-action spaces (e.g. Ng et al. (1999)). Further, a shaping bonus is a non-distorting bonus (Kakade and Dayan, 2002; Ng et al., 1999), which would mandate that the net effect of any cycle of states (such as an approach-retreat bout) is zero, a condition not adequately tested in CSC, which is unidirectional in space and time with no cycles.

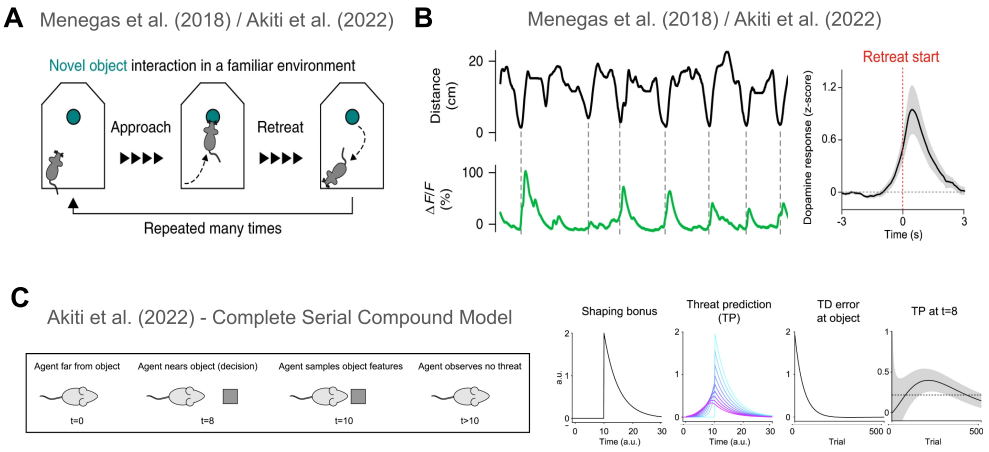


Figure 5.4: A summary of experimental findings of the role of TS in threat prediction. (A) Illustration of the novel object task and the approach-retreat behaviour. (B) Approach-retreat behaviour observed in the distance to the novel object and the TS responses observed on retreat start. (C) Complete Serial Compound (CSC) model of TS responses proposed by Akiti et al. (2022), uses value initialisation, which is equivalent to a potential-based reward shaping. CSC is a temporal representation where the trial (or an episode) is broken down into a sequence of states, each corresponding to a distinct moment in time ($t = 0, 1, 2, ...$). The agent starts at a state corresponding to time $t = 0$, and at each time step, the agent moves on to the next state (unidirectionally in time). State $t = 8$ in their model corresponds to a state where the agent can either choose to continue the trial (equivalent to engage with the novel object) or terminate the trial (equivalent to avoiding). State $t = 10$ is when the agent samples the novel object, and this state has a high threat value initialisation (shaping bonus). For states $t > 10$, the value initialisation is gradually decayed. These correspond to states when the agent observes no threat after sampling the object. The X-axis of the Shaping bonus and Threat prediction plots represents time (states in the CSC model). The threat prediction plot further shows the values learned over trials by their model for all states (blue - early trials, pink - late trials). TD error and the threat prediction at state $t = 8$ are further plotted over 500 learning trials. This model lacks spatial aspects; therefore, it cannot produce approach-retreat bouts and forces a task which inherently doesn't have any trials into a trial-based CSC learning framework. All figures are retrieved and minimally adapted from Menegas et al. (2018) with permission from Springer Nature and from Akiti et al. (2022) under CC-BY-NC-ND 4.0 license with permission from Elsevier.

To address these limitations, we simulate a grid world environment without explicit rewards or punishments (Fig. 5.5A). The agent receives threatening observations o_t from a Bernoulli process: if the agent is in the vicinity of the novel object (demarcated by the grey area), $p(o_t = 1|\text{threatened}) = 0.9$, otherwise outside the grey area, $p(o_t = 1|\text{not-threatened}) = 0.1$. Using these observations, the agent infers a threat belief state b_t via a Bayesian Beta posterior. Two value functions with different initialisations drive behaviour: $V_{not-threatened}$ (neutral initialisation) and $V_{threatened}$ (aversive initialisation at the novel object) (Fig. 5.5B). Both values share the common outcome signal for learning, here zero since there is no external outcome. These values are dynamically combined into a composed value V_{comp} using softmax composition, weighted by the belief state: $V_{threatened}$ and $(1 - b_t)$ weights $V_{not-threatened}$. However, all of the results would also hold true for additive composition, and with TD-learning or SARSA with two different initialisations, we simply use soft maximum for consistency throughout the paper.

We find that the proposed model and the baseline model both produce approach-retreat bout behaviour. In 10 simulation runs, each running for 10^5 timesteps, we find that the proposed model produces 336.1 bouts on average (std. deviation: 10.4), whereas the baseline model produces 12.4 bouts on average (std. deviation: 9.05). In both models, the approach to the novel object is a result of stochastic actions that lead to the novel object state. The number of approach-retreat bouts is higher in the proposed model due to the belief-state dynamics, allowing a higher contribution of the $V_{not-threatened}$ value function. Adding a third value function responsible for seeking the novel object (not implemented by the TS) may increase the number of approach-retreat bouts.

The proposed model reproduces TS dopaminergic responses during retreat, where the threat prediction (TP $= -V_{comp}$) aligns with experimental observations (Menegas et al., 2018) (Fig. 5.5C&D). Importantly, using only $V_{threatened}$ (analogous to PBRS in 2D space) fails to reproduce these responses in a grid world setting, as it generates high TP during both approach and retreat (Fig. 5.5F&G).

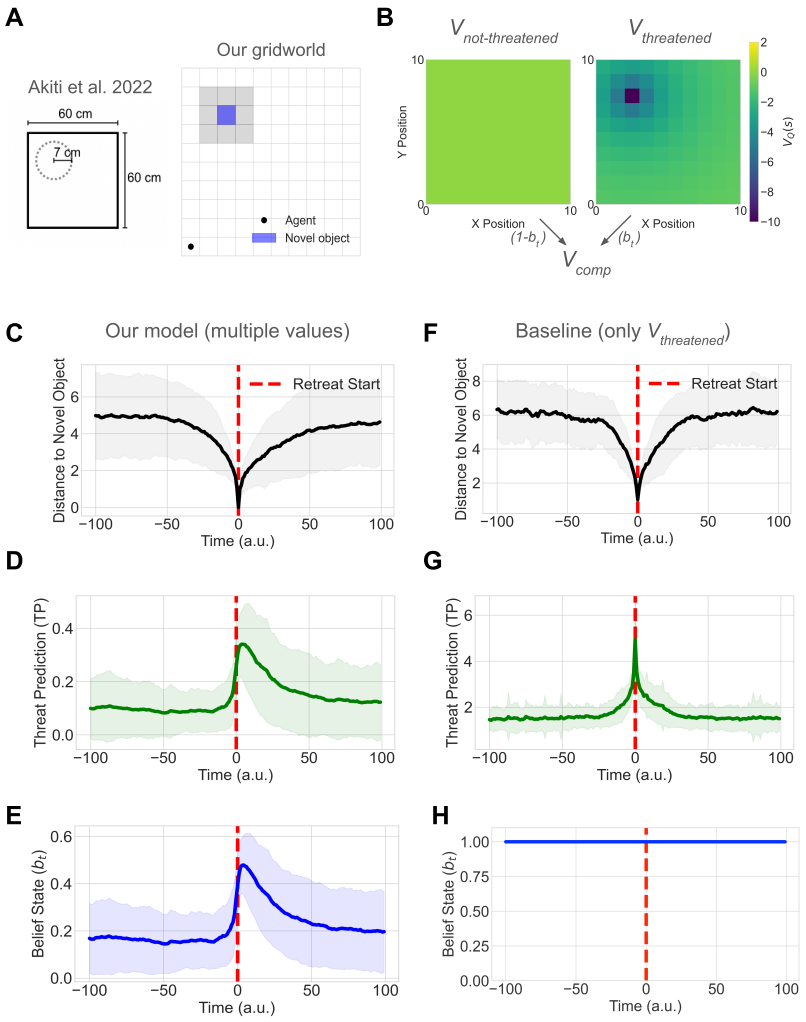


Figure 5.5: Threat belief-gated value composition reproduces approach-retreat dynamics and TS dopamine signals.(A) Grid-world environment analogous to Akiti et al. (2022). (B) Two value initialisations used to produce the observed results, gated by the threat belief state b_t. (C) Distance to novel object shows approach-retreat bout behaviour (D) Threat Prediction (TP) is analogous to the TS dopamine activity during retreat but not during approach. The relative magnitude of the composite Threat Prediction response depends on the nature of composition - soft maximum or additive, here we use soft maximum for consistency in the paper (E) The threat belief state b_t over time that gates the contribution of the two values. (F) When using only the $V_{threatened}$, the approach-retreat bout behaviour is also observed in distance to novel object (albeit we observe fewer bouts) (G) Threat Prediction (TP) is observed during approach and retreat phases unlike the TS dopamine activity and (H) Using only $V_{threatened}$ is akin to setting the $b_t = 1$.

The dynamics of the belief state (Fig. 5.5E) introduce the asymmetry in time. Low belief state during approach results in low Threat Prediction values (more dominated by $V_{not-threatened}$), whereas higher belief states during retreat result in high Threat Prediction values (more dominated by $V_{threatened}$). This cannot be produced using a single value initialisation over spatial states. Akiti et al. (2022) reproduce it using a single value initialisation in a CSC model (where states do not correspond to space but rather time), by having different value initialisation before and after the time corresponding to the state when the agent reaches the novel object (Fig 5.4C).

Our results make a testable prediction that gating aversive value based on a threat belief state (Fig. 5.5E), akin to a context-dependent switch, is a normative mechanism for adaptive safe behaviour. Similar asymmetric TP responses can also be generated by non-Bayesian, switch-like dynamics modulating the two values (Appendix Fig. B.5). We did not include explicit curiosity-based exploration rewards for simplicity of modelling TS responses, but their inclusion in the behavioural model could increase agent tendencies to approach the novel object. Such a dynamic composition of multiple value initialisation also provides one way of unifying the distorting novelty bonuses and non-distorting shaping bonuses (Kakade and Dayan, 2002), i.e. under fixed weights, it acts like a shaping bonus and under varying weights it produces novelty-like bonuses.

The belief-gated, dual-value architecture further offers a potential neural implementation for TS function, where $V_{threatened}$ and $V_{not-threatened}$ could potentially map onto direct (D1) and indirect (D2) TS pathways, respectively (Appendix Fig. B.6). This generates several testable predictions: First, in a task involving threat-reward conflicts, this predicts opposing effects on avoidance behaviour, with D1 promoting avoidance and D2 suppressing it. Second, ablating D1 TS neurons in the same task is predicted to reduce avoidance, while D2 TS ablation should increase it. Third, suppose the inferred threat belief b_t covaries with observable threat cues like object size or whether it is moving. In that case, the composite TS dopamine responses should scale accordingly with those features of the potentially

threatening novel object. Fourth, it predicts distinct learning dynamics: with repeated exposure to the novel object, the D1-associated $V_{threatened}$ should diminish, reducing avoidance, while the D2-associated $V_{not-threatened}$ might remain stable. Fifth, in a one-step task, a punishing outcome with a magnitude falling between the initial $V_{threatened}$ and $V_{not-threatened}$ values should elicit negative threat prediction errors (TPEs) in D1-projecting SNL neurons and positive TPEs in D2-projecting SNL neurons (Menegas et al., 2018); such asymmetric TPEs could potentially encode outcome uncertainty (Mikhael and Bogacz, 2016). Recent work by Tsutsui-Kimura et al. (2025) in a threat-reward conflict task provides evidence supporting the first three predictions: opposing roles, ablation effects, and modulation by threat salience. Their findings on learning dynamics were mixed (non-significant D1 decrease, significant D2 increase), and the prediction of asymmetric TPEs remains to be directly tested, to our knowledge. Our model thus offers a novel explanation for within-target dopaminergic heterogeneity based on differentially initialised, composable values sharing a common outcome; distinct from feature-specific heterogeneity accounts (Lee et al., 2024b).

Stable learning and reconciling conflicting views on TS function

While multiple value initialisations can account for threat prediction in the tail of striatum (TS), recent findings also implicate TS dopamine in encoding action prediction errors (APEs) (Greenstreet et al., 2025) that support perseverative behaviours or habits (Miller et al., 2019; Bogacz, 2020; Gershman, 2020). This raises the question of how these seemingly distinct TPE and APE signals can be reconciled within a unified normative framework and what computational purpose such APE-like signals serve.

The linear MDP framework offers a parsimonious explanation. The default policy π^d, so far assumed uniform, can be slowly updated to track the agent's learned policy π, thereby encouraging perseveration (Piray and Daw, 2021). We find it is worthwhile to mine the analogy from RPEs for $Q(s, a)$ and $V(s)$ to APEs

- whether there exist analogous state-dependent aggregate APEs? We propose that the KL divergence term, $D_{\mathrm{KL}}(\pi(\cdot|s_t)\|\pi^d(\cdot|s_t))$, within the regularised objective itself normatively accounts for an aggregate, state-dependent APE. Although this KL term is implicit in one-step soft Q-learning TD errors (as action a_t is already chosen), it becomes explicit in multi-step formulations, which motivates our theoretical results.

Theoretical result: Multi-step soft Q-learning reveals an explicit APE-like term. We derive novel, multi-step extensions: N-step soft Q-learning and soft Q(λ) with eligibility traces (Methods Section 5.6; derivations in Appendix B.3 and B.4). Note, these are fully off-policy as well, using Tree Backup operations, which simplify under Boltzmann policy. For these, the TD error for $Q(s_t, a_t)$ can be expressed in relation to state-value changes as: $\delta_t = [r_{t+1} + \gamma V_Q(s_{t+1}) - V_Q(s_t)] - \tau \mathrm{KL}_t$, where $\mathrm{KL}_t = D_{\mathrm{KL}}(\pi(\cdot|s_t)\|\pi^d(\cdot|s_t))$ is the action policy divergence term. We interpret this KL_t term as a normative, scalar APE signal. Thus, the Q-value update effectively becomes $\delta_t = \mathrm{RPE}_t - \tau \mathrm{APE}_t$. Given that TS-projecting neurons in SNL encode magnitude of aversive prediction errors $-\delta_t$ (Menegas et al., 2018; Akiti et al., 2022), and do not respond to positive rewards (Watabe-Uchida and Uchida, 2018; Greenstreet et al., 2025), their activity in tasks without explicit aversive stimuli could be primarily driven by this APE-like KL term (or APE_t).

Simulation result: KL-divergence dynamics in multi-step Soft Q-learning partly mimic APEs and support unified TS function. To test this, we simulated a multi-step two-choice task (loosely) inspired by Greenstreet et al. (2025), where rewards depended probabilistically on an observable context (Fig. 5.6A, B). Here, the agent starts in the state S_0 and takes a series of primitive actions a_L or a_R to get closer to either of the terminal states S_L or S_R. This setup makes the two-choice bandit tasks more granular by dividing them into simpler actions. Greenstreet et al. (2025) showed that TS dopamine decreased over trials in an APE-like manner (Fig. 5.6C). Greenstreet et al. (2025) modelled the data using a rectified version of APEs proposed by Miller et al. (2019). Since APEs are action-specific, they converted them to a scalar value by rectifying, i.e. choosing the positive value of the two APEs for the two actions, which corresponds to the APE

in the chosen action. The update rule for the default policy in our model is similar to that of Miller et al. (2019) and exactly the same as that of Piray and Daw (2021). Similar to Greenstreet et al. (2025), we plot the APEs used to update the default policy, and they qualitatively reproduce the data (Fig. 5.6D), but do not directly appear in the TD-errors. However, the KL_t term in our soft $Q(\lambda)$ model's TD errors also decreased across episodes (Fig. 5.6E) and qualitatively matches the decrease in TS responses. The KL term starts from zero as both behavioural and default policies are uniform distributions at the beginning, and therefore the initial ascent in the KL term doesn't match the data, however a sufficiently low $\tau < 0.3$, keeps it steep. This minor difference in the two conceptions of APEs, both of which can be found in our model, arises because the KL term is calculated using the deviation of the current behavioural policy distribution from the default policy distribution. Whereas Miller et al. (2019)'s APE is calculated from a single sample of behavioural policy (i.e. the chosen action) and is used to update the default policy in our model. Lastly, the decrease in the KL term also reflects the default policy π^d gradually aligning with the learned contextual policies π, effectively encoding soft habits (Fig. 5.6F). This result offers a unifying perspective: TS dopamine could reflect a composite signal where TPEs dominate in threat-relevant contexts, while APEs (the KL_t term) dominate when behaviour stabilises around a default/habitual policy.

Lastly, we note that KL term implements the computational logic of the APEs implemented by the TS by capturing the divergence between the entire behavioural policy from the default policy, across all actions. Observations regarding choice bias contralateral to the TS hemisphere are captured in the updating of the default policy for individual discrete actions (Piray and Daw, 2021), which is similar to value-free updates (Greenstreet et al., 2025; Miller et al., 2019).

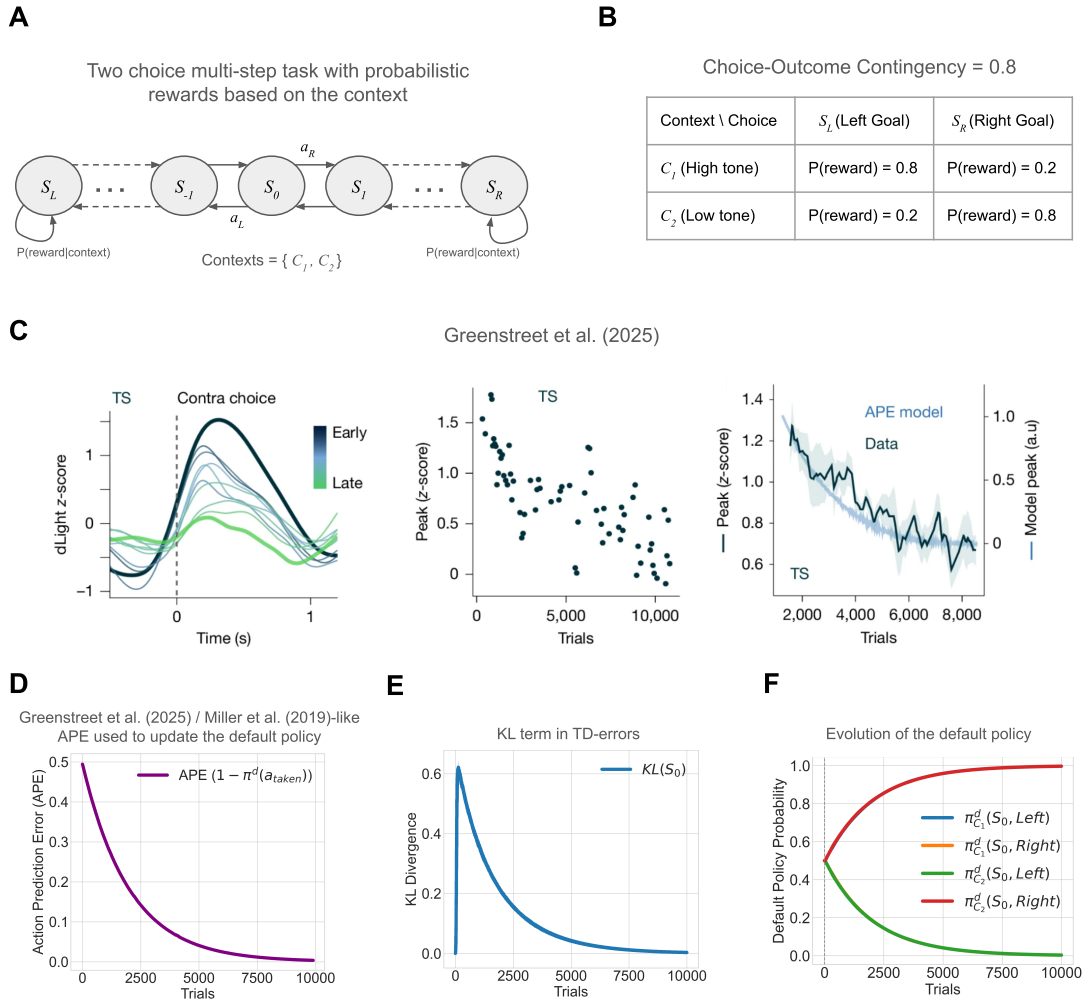


Figure 5.6: KL-divergence dynamics in multi-step Soft Q-learning mimic APEs and support unified TS function.(A) Two-choice multi-step task with probabilistic rewards based on the context (two possible contexts, treated as fully observable states). (B) Choice-Outcome Contingencies are dependent on the context, leading to different correct choices based on the context. (C) Evidence for APEs, figures adapted from Greenstreet et al. (2025) under CC-BY 4.0 license. (Figures were cropped and combined) (D) In our simulation experiment, an action prediction error similar to Greenstreet et al. (2025) or Miller et al. (2019) is used to update the default policy. It qualitatively matches the gradual decrease in TS-projecting responses, but it doesn't directly appear in the TD-error for learning values from the rewards. (E) The decrease (but not the initial increase) in the KL divergence term in soft Q(λ) TD-errors qualitatively recapitulates the decrease in APE-like TS responses for $\tau = 0.1$. Here showing the KL divergence term for the middle starting state (S_0), using $\lambda = 0.5$. (F) The evolution of the default policy shows acquisitions of soft habits.

Simulation result: Perseverative bias confers a value on stability against uncontrollability. What is the normative benefit of such APE-driven perseveration or stickiness in choices? We hypothesised that the bias towards π^d (soft

habits) promotes learning stability during periods of environmental uncontrollability. We tested this by introducing sporadic episodes where choice-outcome contingencies were degraded (Fig. 5.7A). During these uncontrollable periods, agents using soft Q-learning or soft Q(λ) exhibited perseveration of prior choices and limited unlearning of the behavioural policy, due to the influence of the slowly updated default policy (Fig. 5.7B, C). This resulted in better overall reward accrual compared to standard Q-learning (which lacks this perseverative bias) across a range of τ values (Fig. 5.7D).

Thus, the APE-like KL term in our framework not only arises normatively, in a way that would scale to any action space (unlike other models which explicitly include an APE term (Miller et al., 2019; Lindsey and Litwin-Kumar, 2022)), but also confers a functional advantage of stable learning by promoting conservative policy updates. Unlike (Greenstreet et al., 2025), it is not included as an additional ad-hoc controller, but normatively shapes the policy through the objective function, reconciling with the TDRL view of dopamine. Distinct from Miller et al. (2019), these "soft habits" are not entirely value-free but rather accounted for in the value itself, through the augmented RL objective and confer a value on controllability.

Similar to our model, Bogacz (2020) proposes a model where APEs arise normatively from a Bayesian analysis and the APEs encode deviations from an action prior. This bears similarities to the KL divergence between the behavioural policy and the default policy in our framework of Linear RL. The default policy in Linear RL corresponds to the passive dynamics in early control theoretic formulations of linear MDPs (Todorov, 2009b), which were considered as a prior, when viewed under control as an inference (Todorov, 2008). In Bogacz (2020), the action is inferred by iteratively minimising the Free Energy which involves a term responsible for minimising the difference between the action that is to be chosen and the mean of the habit prior (if the action distributions are Gaussians). In our model, if we assume Gaussian action distributions with equal variance for behavioural and default policies, then the KL term can be decomposed into two terms: a policy entropy term (an entropy maximisation bonus) and a Bogacz (2020)-like squared action prediction term (further details in Appendix B.5).

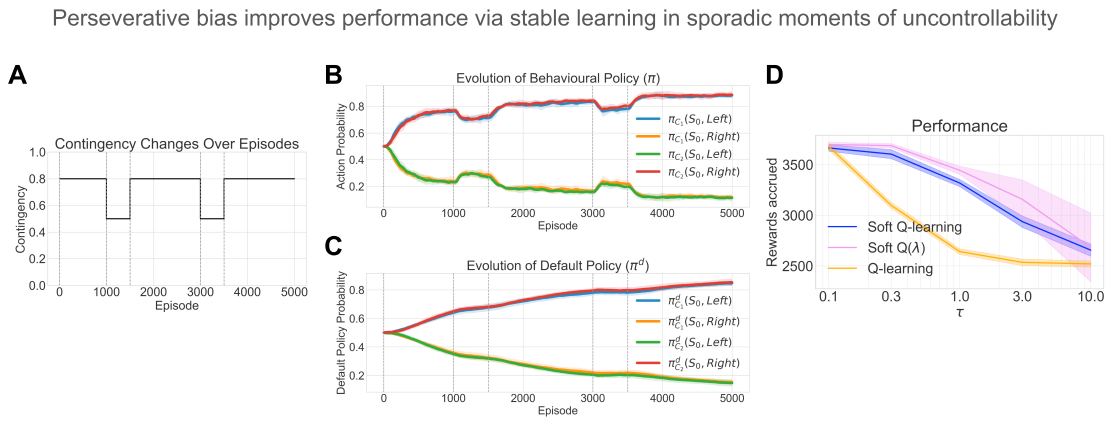


Figure 5.7: Perseverative bias confers a value on stability against uncontrollability. (A) Spurious contingency degradation to test the role of perseverative bias (B) Evolution of behavioural policy shows stickiness in actions (C) Evolution of the default policy. Both B & C are plotted for the soft Q-learning model with $\tau = 0.3$. (D) Soft Q-learning and Soft Q(λ) outperform Q-learning for a range of τ.

5.5 Discussion

In this work, we address the fundamental challenge of reliably optimising multiple objectives in reinforcement learning and propose a model that outperforms existing multi-objective methods. It invites a reconsideration of the dopaminergic system's computational goals: shifting from the classical view of cumulative discounted reward maximisation (Schultz et al., 1997) towards optimising returns augmented by a KL-penalty in policies deviating from a default policy. This reframing not only yields distinct functional advantages but also generates novel testable predictions in efficient, safe and stable learning. In doing so, we resolve a major outstanding puzzle about how to reconcile conflicting views of the role of TS-projecting dopamine in threat prediction errors (TPEs) and action prediction errors APEs).

This approach was designed to directly address several limitations of the standard model, specifically the challenge of flexibly pursuing multiple, often conflicting, rewards without the learning process becoming unstable or inefficient—a known issue for classic on-policy methods. Our model's core innovation is to augment the classic reward prediction error with a regularisation term that imposes a 'cost' for deviating from a default policy. This single conceptual shift yields several powerful

benefits; for instance, it confers a 'value on controllability,' which promotes stable learning. The framework further allows innate priors (e.g., for threat avoidance) to be flexibly expressed for safe behaviour. Crucially, this framework provides a clear normative function for the APE-like signals in the TS, uniting them with TPEs and RPEs under a single, coherent objective and thereby distinguishing our model from other multi-objective RL approaches.

At an algorithmic level, this work addresses how to ensure (often individually selfish) value functions cooperate effectively to drive reliable behaviour, despite constantly competing for control. This contrasts with "delegation" approaches (Dayan and Hinton, 1992; Dietterich, 2000; Parr and Russell, 1997), where only one value function controls, thus avoiding this problem. Regarding our optimal composition results, previous multi-objective TD(0) or SARSA approaches that scale on-policy prediction errors with weights (e.g. state-dependent RPE modulation (Millidge et al., 2024a; van Swieten and Bogacz, 2020) or feature-specific weight updates (Gershman and Uchida, 2019; Lee et al., 2024b)) may slightly ameliorate unreliable learning or policy interference, but do not resolve it (Appendix Fig. B.3). This excludes trivial conditions where only one value component is active, preventing interference. A better alternative might be to use importance sampling, which also ensures off-policy learning. However, it incurs higher update variance and cannot explain action prediction errors. Hence, we primarily utilise the Tree-Backup approach in our derivations for novel multi-step extensions of soft Q-learning (Appendix B.3 and B.4).

Off-policy learning is a prominent theme in this work, as it demonstrably improves performance over on-policy multi-objective RL algorithms and the successor representation, while also relating directly to its off-policy counterpart (Piray and Daw, 2021). The brain may implement off-policy learning for several reasons: First, it prevents interference and unlearning between multiple values amidst changing priorities. Second, it facilitates learning amidst motor noise and competition from distributed control systems, like the motor cortex and cerebellum (Lindsey and Litwin-Kumar, 2022; Lindsey et al., 2024). Third, on-policy algorithms, such as the

successor representation, exhibit strong policy dependence where goal information contaminates the state map, hindering flexible transfer (Russek et al., 2017; Lehnert et al., 2017), a problem solved by off-policy alternatives (Piray and Daw, 2021). Fourth, the ability of episodic memories to utilise cached values (Sadacca et al., 2016; Krausz et al., 2023) or stale behavioural data for performance improvements points towards an underlying off-policy mechanism, akin to its necessity in deep Q-learning's episodic replay buffers (Mnih et al., 2015). However, finding strong neural and behavioural evidence for interference and unintended unlearning between two or more value systems (for different rewards) under changing priorities in a two-step task similar to Fig. 5.2D-F, would falsify our hypothesis of phasic dopamine performing off-policy multi-objective RL and find evidence for on-policy multi-objective RL (e.g. (Millidge et al., 2024a)). Rapid change in priorities could be potentially implemented as a task rule that needs to be inferred.

Our framework synthesises several threads of the evolving dopamine story, which has progressively expanded from a simple scalar reward signal (Schultz et al., 1997) to a multifaceted control signal. The model naturally accommodates aversive and threat prediction errors (Matsumoto and Hikosaka, 2009; Watabe-Uchida and Uchida, 2018; Menegas et al., 2018; Akiti et al., 2022), aligning with multi-threaded and outcome-specific prediction error views (Takahashi et al., 2023; Millidge et al., 2024a), and highlights their role in safe learning. This approach complements the feature-specific vector RPE model (Lee et al., 2024b) while remaining compatible with it. Further, we propose an alternative account of within-target dopaminergic heterogeneity based on differentially initialised, composable values sharing a common outcome, distinct from that of Lee et al. (2024b). Further, we find that such value compositions with different initialisations unify (reward-distorting) novelty bonuses and (non-distorting) shaping bonuses in TDRL (Kakade and Dayan, 2002), previously used to explain early observations of novelty responses in phasic dopamine (Ljungberg et al., 1992; Horvitz et al., 1997). Our results (Fig. 5.3C) also highlight the behavioural inefficiencies of the SR model in replanning after overtraining (Russek et al., 2017). Indeed, SR models of dopamine often need to treat reward as

a feature to explain RPEs as a form of state prediction error (Gardner et al., 2018), as Lee et al. (2024b) shows they otherwise fail to consistently respond to rewards. This highlights the benefit of outcome-specific models such as ours (and that of Millidge et al., 2024a), which efficiently achieve values comparable to an SR/DR model (Dayan, 1993; Piray and Daw, 2021) when tuned to a subset of reward types. The introduction of belief-state weighted value composition bears resemblance to other contextual learning models (Gershman and Uchida, 2019; Babayan et al., 2018); though not a focus of this work, this mechanism could be extended in the future to account for beliefs about reward timing (Takahashi et al., 2023) or state uncertainty, potentially explaining ramping effects (Howe et al., 2013; Kim et al., 2020; Mikhael et al., 2022). Our derivation of eligibility traces allows the model to flexibly incorporate the presence or absence of a gradual shift in dopamine responses, with $\lambda = 0$ and higher learning rate or higher λ and lower learning rate, respectively (Amo et al., 2022; Amo, 2024). Finally, the soft Bellman value function links our work to a kind of risk-sensitive control (KL-control) due to its use of exponential utilities (Howard and Matheson, 1972; Borkar, 2002; Dvijotham and Todorov, 2012), as it inherently considers a spectrum of action values, not just the mean, with the temperature parameter τ modulating this sensitivity. Both our approach and distributional RL go beyond a single average value, but they do so differently, and future work is needed to compare with full distributional RL in outcomes (Dabney et al., 2020) and time (Tano et al., 2020; Sousa et al., 2025; Masset et al., 2025).

This work is the first, to our knowledge, to mathematically unify the conflicting views about the tail of the striatum's (TS) phasic dopamine. However, previous work has attempted to unify the general role of dopamine in learning and action inference (Bogacz, 2020). Our proposal goes beyond the (non-mathematical but very helpful) unifying hypothesis that TS shifts attention to orient or avoid (Green et al., 2024) by acknowledging the action prediction errors and their normative implications for stable learning. Biologically, one possibility is that threat and movement-related activity are encoded by separate dopaminergic subpopulations. In support of this, it has recently been reported that threat and acceleration-related dopamine

responses are encoded in separate genetic subpopulations that both project to the TS, expressing Slc17a6 (also known as Vglut2) and Anxa1, respectively (Azcorra et al., 2023; La Manno et al., 2016; Poulin et al., 2014, 2018; Greenstreet et al., 2025).

Our work introduces the concept of an inferred threat belief state (potentially cortical), demonstrating how it can titrate the balance between multiple values to guide flexible avoidance. This aligns with a performance effect, rather than a learning effect, in models of striatal direct (D1) and indirect (D2) pathway balance (Collins and Frank, 2014). Our hypothesis, mapping different value initialisations to TS D1 and D2 pathways, yields testable predictions (Tsutsui-Kimura et al., 2025). Furthermore, the flexible expression of innate values can reconcile associative and non-associative fear conditioning accounts (Zambetti et al., 2022), promoting more adaptable safe behaviour than previous models based on outcome uncertainty (Mahajan et al., 2024). Crucially, we show that the same computations for efficient multi-reward acquisition can, with minor modifications, form a modular instrumental system for avoidance, unlike Pavlovian misbehaviour (Mahajan et al., 2024). Lastly, our approach extends beyond existing models (Akiti et al., 2022) by demonstrating approach-retreat bouts with associated TPEs, complementing risk-sensitive model-based RL efforts in modelling cautious behaviours (Shen and Dayan, 2024).

In terms of limitations; first, though broadly applicable to decision-making under changing priorities, when applied to homeostatic priorities, it cannot explain physiological state-dependent modulation of prediction errors (Cone et al., 2016). This is a limitation common to several vanilla multi-objective RL approaches, which learn equally from all reward types, at all times. Millidge et al. (2024a) address this issue by introducing a variant of the Reward Basis model, which captures state-dependent RPE by scaling prediction error with the reward weight. We find that this model empirically doesn't converge to optimal values in our task with changing priorities (Appendix Fig. B.3), and future work is needed to tease apart the strengths and weaknesses of this variant. Homeostatically regulated reinforcement learning (HRRL) (Keramati and Gutkin, 2014) derives rewards from changes in a convex multi-dimensional drive function, and this normatively accounts for the

scaling of prediction errors based on the extent of deviations from homeostatic setpoints. However, Dulberg et al. (2023) show that a modular multi-objective RL system outperforms HRRL in tasks with changing rewards/priorities.

Second, despite its advantages, soft maximum composition optimises an objective different from a simple weighted sum of utilities. While this is not necessarily an issue for modelling homeostasis, where value representation is debated (Keramati and Gutkin, 2014; Dayan, 2022; Dulberg et al., 2023), and may even account for inhibitory effects of irrelevant drives (Keramati and Gutkin, 2014; Dickinson and Balleine, 2002), we find it can fit poorly to human behaviour when participants explicitly maximise weighted sums of utilities or show generalisation in task structure rather than in values (Tomov et al., 2021) (see Appendix Fig. B.7). Third, while there are compelling reasons for the brain to implement off-policy learning (Lindsey and Litwin-Kumar, 2022; Lindsey et al., 2024; Greenstreet et al., 2025), some early work (Morris et al., 2006; Niv et al., 2006) suggested phasic dopamine implements on-policy algorithms like SARSA. However, those experiments involved overtrained monkeys where no learning occurred, making observed TD error signals potentially epiphenomenal. Our framework could partly explain these differences in action values as the KL divergence from the overtrained default policy (APEs). Furthermore, the observed use of cached values in computing TDEs (Sadacca et al., 2016; Krausz et al., 2023) better aligns with off-policy rather than on-policy TD learning.

In offering a novel normative framework for multi-objective reinforcement learning, this paper re-conceptualises the computational role of striatal dopamine. Our findings demonstrate how the brain might achieve efficient, safe, and stable learning, simultaneously reconciling disparate experimental observations and generating testable predictions for future neuroscientific inquiry.

5.6 Methods

Reinforcement learning in MDPs

Let the environment be a Markov Decision Process, where at time $t = 0, 1, 2, ...$, the agent is in state $s_t \in \mathcal{S}$ and takes action $a_t \in \mathcal{A}$ and receives the next state $s_{t+1} \in \mathcal{S}$

and the reward $r_{t+1} = r(s_t, a_t) \in \mathcal{R}$ giving rise to trajectories $s_0, a_0, r_1, s_1, a_1, r_2, ...$ The dynamics of MDP are given by the conditional probability $p(s', r|s, a) \doteq \Pr(s_t = s', r_t = r|s_{t-1} = s, a_{t-1} = a)$.

The discounted return at time t is given by $G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, where $\gamma \in [0, 1]$. Policy $\pi(a|s)$ is a mapping from states to the probabilities of choosing each possible action. The value function of a state s under the policy π is is the expected return when starting in s and following π thereafter, which is formalized as $V_\pi \doteq \mathbb{E}_\pi[G_t|s_t = s], \forall s \in \mathcal{S}$. Similarly, the value of taking action a in state s and following policy π thereafter is given by the Q-value or the action-value function, $Q_\pi(s, a) \doteq \mathbb{E}_\pi[G_t|s_t = s, a_t = a]$.

The Bellman equation of a value function v_π is a fundamental property in reinforcement learning expressing the recursive relationship between a value of state and the value of its possible successor states.

$$V_\pi(s) \doteq \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma V_\pi(s')], \forall s \in \mathcal{S} \tag{5.1}$$

Since value functions define a partial ordering over policies, there exists at least one optimal policy π^* that is better than all policies, where a policy $\pi \geq \pi'$ if and only if $V_\pi(s) \geq V_{\pi'}(s), \forall s \in \mathcal{S}$. The optimal state-value function is $V^*(s) \doteq \max_\pi V_\pi(s), \forall s \in \mathcal{S}$. Similarly, the optimal action-value function is $Q^*(s, a) \doteq \max_\pi Q_\pi(s, a) = \mathbb{E}[r_{t+1} + V^*(s_{t+1})|s_t = s, a_t = a]$. Once we have the optimal action-values, one can simply perform actions greedily to get the optimal policy $\pi^* = [\mathcal{G}Q^*](s) = \arg\max_a Q^*(s, a)$.

The recursive Bellman equations can also be written for the value function under the optimal policy, referred to as the Bellman optimality equations:

$$V^*(s) = \max_a \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma V^*(s')] \tag{5.2}$$

Entropy-regularised reinforcement learning in Linear MDPs

Entropy-regularised RL (Todorov, 2006, 2009b; Van Niekerk et al., 2019) augments the reward function with a term that penalises deviating from some default policy π^d, essentially making "soft" assumptions about the future policy (in the form of a stochastic action distribution). When π^d is an uniform policy, this reduces to max entropy reinforcement learning (Ziebart, 2010; Haarnoja et al., 2017). The expected reward on taking action a_t in state s_t is given by $\mathbb{E}_{a_t\sim\pi}[r(s_t,a_t) - \tau D_{\mathrm{KL}}(\pi(\cdot|s_t)\|\pi^d(\cdot|s_t))]$, which can be further compactly written as $\mathbb{E}_{a_t\sim\pi}[r_{t+1} - \tau\mathrm{KL}(s_t)]$. Here, τ is the scalar temperature parameter, and $\mathrm{KL}(s_t)$ is the Kullback-Leibler divergence between the current policy π and a default policy π^d in state s_t. Thus, the entropy-augmented return is $G_t = \sum_{k=0}^{\infty}\gamma^k(r_{t+k+1} - \tau\mathrm{KL}(s_{t+k}))$.

The value function definitions under a policy π at any timestep t based on the entropy-augmented returns are as follows,

$$V_\pi(s) \doteq \mathbb{E}_\pi[G_t|s_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k(r_{t+k+1} - \tau\mathrm{KL}(s_{t+k}))\middle| s_t = s\right] \tag{5.3}$$

$$Q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|s_t = s, a_t = a] = \mathbb{E}_\pi\left[r_{t+1} + \sum_{k=1}^{\infty}\gamma^k(r_{t+k+1} - \tau\mathrm{KL}(s_{t+k}))\middle| s_t = s, a_t = a\right] \tag{5.4}$$

Note that this Q-function does not include the first KL penalty term ($\mathrm{KL}(s_t)$), as it does not depend on action action a_t which has already been chosen (Ziebart, 2010; Haarnoja et al., 2017; Schulman et al., 2017). This gives the following relationship which holds for all policies π.

$$V_\pi(s) = \mathbb{E}_{a\sim\pi}[Q_\pi(s,a)] - \tau\mathrm{KL}(s) \tag{5.5}$$

The Bellman equation and the Bellman optimality equation are as follows:

$$V_\pi(s) \doteq \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r - \tau\mathrm{KL}(s) + \gamma V_\pi(s')] \tag{5.6}$$

$$V^*(s) = \max_a \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r - \tau\mathrm{KL}(s) + \gamma V^*(s')] \tag{5.7}$$

Note, unlike the greedy (deterministic) policy $[\mathcal{G}Q](s) = \arg\max_a Q(s,a)$ in standard RL, the greedy (stochastic) policy in entropy-regularised RL is the Boltzmann policy ($\pi^{\mathcal{B}}_Q$).

$$\pi^{\mathcal{B}}_Q(\cdot|s) = [\mathcal{G}Q](s) = \frac{\pi^d(a|s)\exp(Q(s,a)/\tau)}{\sum_{\mathcal{A}}\exp(Q_\pi(s,a')/\tau)\pi^d(a'|s)} \tag{5.8}$$

Prior work (Todorov, 2006, 2009b; Haarnoja et al., 2017; Van Niekerk et al., 2019) shows that this Boltzmann policy holds the two properties: (1) it is the optimal policy ($\pi^* = \pi^{\mathcal{B}}_{Q^*}$) i.e. it uniquely solves the Bellman optimality equations and (2) under the Boltzmann policy, the Bellman equation is equivalent to the "soft" Bellman equation, thus the value function $V_{\pi^{\mathcal{B}}_Q}(s) = V_Q(s)$, essentially performing a soft maximum operation over Q-values. These known results can be verified easily and for completeness purposes, we provide an intuitive explanation in Appendix B.1.

$$\begin{aligned} V_Q(s) &= \tau \log \mathbb{E}_{a\sim\pi^d} \exp(Q_\pi(s,a)/\tau) \\ &= \tau \log \sum_{\mathcal{A}} \exp(Q_\pi(s,a)/\tau)\pi^d(a|s) \end{aligned} \tag{5.9}$$

Note, this log-sum-exp performs a soft maximum because, $\max\{x_1,...,x_n\} \leq \text{softmax}(x_1,...,x_n) \leq \max\{x_1,...,x_n\} + \log(n)$.

Multi-objective reinforcement learning and optimal composition in Linear MDPs

Having discussed reinforcement learning (RL) in MDPs and Linear MDPs with single-attribute rewards, we now focus on multi-objective RL, which concerns multiple rewarding attributes $\mathbf{r} = [r_1, r_2, \ldots, r_n]$. The objective is to maximise a cumulative discounted return of a reward function composed of these attributes:

$$r_c(s,a) = f(r_1, r_2, \ldots, r_n; \mathbf{w}),$$

where $\mathbf{w}$ is a set of non-negative parameters that weight each attribute, satisfying $\sum w_i = 1$. We address the problem of *optimal compositions*: determining how the reward function should be composed of multiple attributes to motivate meaningful

behaviour and how to compose value functions to ensure the resulting policy acts optimally with respect to the composed reward function.

As shown in Results section 5.4, a simple linear composition of Q-values, such as $w_1 Q_1^*(s,a) + w_2 Q_2^*(s,a) + \cdots + w_n Q_n^*(s,a)$, may not maximise the composed reward function

$$r_c(s,a) = w_1 r_1(s,a) + w_2 r_2(s,a) + \cdots + w_n r_n(s,a)$$

due to the non-linearity introduced by the *max* operation in the Bellman optimality equation in MDPs:

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma V^*(s')] \\ &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma \max_{a'} Q^*(s',a')]. \end{aligned} \tag{5.10}$$

However, in linear MDPs (Kappen, 2005; Todorov, 2006, 2009a,b), which replace the *max* operation over Q-values by a soft-maximum V_Q with respect to the default policy (see equation 5.9), optimal (softmax) and near-optimal (additive) compositions are possible. The Bellman optimality equation for Q-values in linear MDPs is as follows:

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma V^*(s')] \\ &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma \mathbb{E}_{a\sim \pi_Q^{\mathcal{B}}}[Q(s',a')] - \tau D_{\mathrm{KL}}[\pi_Q^{\mathcal{B}} \| \pi^d](s')] \\ &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}[r + \gamma V_Q(s')]. \end{aligned} \tag{5.11}$$

Theorem 1 (Optimal Softmax Composition)**.** *(Todorov, 2009a; Van Niekerk et al., 2019)*

Let $Q_{i,\tau}^*(s,a)$ be the optimal entropy-regularized Q-functions for individual rewards $r_i(s,a)$.

Then the reward function for the composed task is given by the log-sum-exp (soft maximum) of the individual reward functions:

$$r_c(s,a) = \tau \log \left(\sum_{i=1}^{n} w_i \exp\left(\frac{r_i(s,a)}{\tau} \right) \right), \tag{5.12}$$

where τ is a temperature parameter.

The optimal Q-function $Q^*_{c,\tau}(s,a)$ for the composed task is equal to the composition of Q-values $Q_{comp,\tau}(s,a)$ given by:

$$Q^*_{c,\tau}(s,a) = Q_{comp,\tau}(s,a) = \tau \log \left(\sum_{i=1}^{n} w_i \exp \left(\frac{Q^*_{i,\tau}(s,a)}{\tau} \right) \right), \tag{5.13}$$

where $Q^*_{i,\tau}$ are the optimal Q-functions for the individual tasks.

The theorem for additive composition in Linear MDPs is provided in the Appendix B.2.

Off-policy model-free learning algorithms in Linear MDPs

Model-free algorithms do not assume a probabilistic model about state transitions and rewards but instead learn value functions through reward prediction errors, which can be implemented by phasic dopamine signals in the striatum. We focus on online algorithms, such as soft Q-learning (Haarnoja et al., 2017), which update values continuously during episodes rather than waiting until the end, unlike offline algorithms like Z-learning (Todorov, 2006), a Monte Carlo control algorithm. We further also prefer off-policy algorithms like Soft Q-learning and our subsequent extensions.

Soft Q-learning (One-Step)

We adopt soft Q-learning and extend it from the maximum entropy formulation to a relative entropy formulation. The Q-value update equation is given by:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta_t, \tag{5.14}$$

where α is the learning rate, and δ_t is the reward prediction error at timestep t, defined as:

$$\delta_t = r_{t+1} + \gamma V_Q(s_{t+1}) - Q(s_t, a_t), \tag{5.15}$$

where V_Q is given by equation 5.9. Deep RL implementations inspired by Mnih et al. (2015) may use a separate target network (e.g., $\underline{Q}$, resulting in $V_{\underline{Q}}$) to construct the loss function, which we exclude here for simplicity.

Multi-step Soft Q-learning

The KL term does not arise in the TD-error of Q-values of soft Q-learning, as the action is already chosen. Alternatively, the KL-term arises in TD-error for just learning state values. One way to introduce the KL term (hypothesised to play the role of APE in this chapter) into the TD-error by extending it to multi-step learning. The motivation for this section is to derive an eligibility traces solution which has the KL term in the TD-error. The eligibility traces solution is more biologically plausible than n-step learning, which requires storing "n" intermediate prediction errors; however, the process to derive the eligibility traces solution requires first deriving the n-step learning solution.

This section presents novel update rules for multi-step extensions of soft Q-learning, where the agent learns from multiple steps rather than the most immediate step. Under the assumption that the state action values are approximately unchanging (Sutton and Barto, 2018), we can write the update rule for the N-step soft Q learning and its extension with eligibility traces, soft Q(λ) using TD-errors.

When following the Boltzmann policy, the N-step soft Q-learning is simply,

$$Q_{t+n}(s_t, a_t) \leftarrow Q_{t+n-1}(s_t, a_t) + \alpha \left(\sum_{k=t}^{min(T-1, t+n-1)} \gamma^k \delta_k \right). \tag{5.16}$$

Where T is the time step at which the episode terminated, Q_{t+n} denotes Q-value accessed or updated at timestep $t+n$ and the TD-errors are defined as follows. For $k = t$, the TD-error is given by equation 5.15. For $k > t$, it includes the KL divergence term and is given by,

$$\delta_k = r_{k+1} - \tau \mathrm{KL}_k + \gamma V_Q(s_{k+1}) - V_Q(s_k) \tag{5.17}$$

The KL term is the KL divergence between the behavioural policy and the default policy, both of which are softmax distributions over discrete choices in our

simulations. Therefore, this is computed as $\mathrm{KL}(s) = \sum_a \pi(a|s) \log\left(\frac{\pi(a|s)}{\pi^d(a|s)}\right)$. An alternative plausible way to compute the KL term, under the Boltzmann policy ($\pi_Q^{\mathcal{B}}$) is by exploiting the relationship from equation 5.5 and re-arranging it. Under Boltzmann policy, $V_{\pi_Q^{\mathcal{B}}}(s) = V_Q(s)$ (equation 5.9), therefore:

$$\tau \mathrm{KL}(s) = \mathbb{E}_{a\sim\pi}[Q_\pi(s,a)] - V_Q(s) \tag{5.18}$$

If the agent's behavioural policy that is not Boltzmann, we need a truly off-policy update rule. If the agent has access to the behavioural policy, then it can use importance sampling (detailed derivation provided in Appendix B.3). However, this can lead to higher variance in the updates and requires access to the behavioural policy. Therefore we derive an alternative method using Tree Backup which does not require knowing the behavioural policy. The update rule is as follows,

$$Q_{t+n}(s_t,a_t) \leftarrow Q_{t+n-1}(s_t,a_t) + \alpha \left(\sum_{k=t}^{min(T-1,t+n-1)} \delta_k \prod_{i=t+1}^{k} \gamma \pi_Q^{\mathcal{B}}(a_i|s_i) \right). \tag{5.19}$$

We next extend these methods to incorporate eligibility traces. Under Boltzmann policy, the Q-value update rule is,

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha \delta_t e_t(s,a) \;\; \forall s,a \tag{5.20}$$

and eligibility traces are updated as follows (in the tabular setting),

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) + 1, & \text{if } (s,a) = (s_t,a_t), \\ \gamma\lambda e_{t-1}(s,a), & \text{otherwise,} \end{cases} \tag{5.21}$$

The TD-error (δ_t) is the same as equation 5.17 (except substitute k with t). Note, this algorithm is entirely online.

For a full off-policy Soft Q(λ), we build upon the Tree Backup approach. The Q-value update rule and the TD-errors remain the same, but the eligibility trace updates are adjusted to include the target policy $\pi_Q^{\mathcal{B}}$,

$$e_t(s,a) = \begin{cases} \gamma\lambda \pi_Q^{\mathcal{B}}(a_t|s_t) e_{t-1}(s,a) + 1, & \text{if } (s,a) = (s_t,a_t), \\ \gamma\lambda \pi_Q^{\mathcal{B}}(a_t|s_t) e_{t-1}(s,a), & \text{otherwise,} \end{cases} \tag{5.22}$$

All detailed derivations for N-step soft Q-learning and Soft Q(λ) are provided in Appendix B.3 and B.4, respectively.

Relationship to the default representation (DR)

The optimal values $V^*(s)$ are calculated using the default representation as follows (Piray and Daw, 2021; Todorov, 2006, 2009b) (in a model-based fashion):

$$\exp(\mathbf{V}^*/\tau) = \mathbf{MP}\ \exp(\mathbf{r}/\tau) \tag{5.23}$$

where, $\mathbf{V}^*$ is the vector of optimal values at nonterminal states, $\mathbf{r}$ is the vector of rewards at terminal states, $\mathbf{P}$ is the one-step transition probabilities $\mathbf{T}_{NT}$ from non-terminal states to terminal states under the default policy π^d and $\mathbf{M}$ is the DR matrix defined as $\mathbf{M} = (\text{diag}(\exp(-\mathbf{r}_N)/\tau) - \mathbf{T}_{NN})^{-1}$, where $\mathbf{r}_N$ is vector of rewards at non-terminal states and $\mathbf{T}_{NN}$ is the one-step transition probabilities between non-terminal states under the default policy π^d. Further, (Piray and Daw, 2021) extend and define a more general version of the DR matrix D over all states (not just terminal states), as $\mathbf{D} = (\text{diag}(\exp(-\mathbf{r}_A)/\tau) - \mathbf{T})^{-1}$, where $\mathbf{r}_A$ is vector of rewards over all states and $\mathbf{T}$ are transition probabilities over all states under the default policy π^d. Here, $\mathbf{M}$ is a sub-block of $\mathbf{D}$ and $\mathbf{T}_{NT}$ and $\mathbf{T}_{NN}$ are sub-blocks of matrix $\mathbf{T}$. We observe that in both cases, it requires storing and/or learning the one-step transitions $\mathbf{T}$ under the default policy over all states, resulting in a $S \times S$ memory cost, where S is the size of the state space. Therefore, the memory cost of the DR scales quadratically with the size of the state space.

We propose our model (composition of multiple soft Q-learning rules) is intuitively akin to a compressed DR tuned to only relevant reward dimensions and converges to the same. To show this, we decompose the reward vector at terminal states, by assuming it to be composed using the optimal softmax composition (Todorov, 2009a), $\mathbf{r} = \tau \log\left(\sum_{i=1}^{n} w_i \exp\left(\frac{\mathbf{r}_i}{\tau}\right)\right)$. Therefore, the vector of optimal values using the DR will be:

$$\begin{aligned}\exp(\mathbf{V}^*/\tau) &= \mathbf{MP}\left(\sum_{i=1}^{N} w_i \exp\left(\frac{\mathbf{r}_i}{\tau}\right)\right) \\ &= \sum_{i=1}^{N} w_i \left(\mathbf{MP}\exp\left(\frac{\mathbf{r}_i}{\tau}\right)\right) \\ &= \sum_{i=1}^{N} w_i \exp(\mathbf{V}_i^*/\tau)\end{aligned} \tag{5.24}$$

Here, the vector of optimal values for each reward decomposition can be computed using soft Q-learning and consumes memory cost of S (state-space size). Therefore, for N reward decompositions, our method consumes a memory cost of $S \times N$. We believe the necessary reward decompositions would often be much lesser than the total number of states, $N << S$. Therefore, our method scales linearly with the size of the state space, whilst achieving the same optimal values as the DR for a predefined reward basis.

Simulation parameters

Simulations in Fig. 5.2D,E and Appendix Fig. B.2 and B.3 use learning rate $\alpha = 0.1$, discount rate $\gamma = 1$, temperature $\tau = 1$ (unless explicitly varied) and are averaged over 100 runs. Simulations on efficient learning (Fig. 5.3) and safe learning (Fig. 5.5) use $\alpha = 0.1$, $\gamma = 0.99$, $\tau = 0.5$ for all algorithms. Fig. 5.3 results were averaged over 30 runs. For Fig. 5.5, we ran the simulations for 10^5 steps and then averaged over the time points that the agent was closest to the novel object and inside the grey zone to capture retreat start. In the safe learning experiment, we used a Bayesian approach to infer the belief state b_t, representing the probability of a "threatened" context c_t ($c_t = 1$ for "threatened" and $c_t = 0$ for "not threatened"). The belief state dynamically adjusted the weights $w_{\text{threatened}} = b_t$ for $Q_{\text{threatened}}$ and $w_{\text{not-threatened}} = 1 - b_t$ for $Q_{\text{not-threatened}}$.

Observations o_t were modelled as samples from a Bernoulli distribution, with likelihoods:

$$p(o_t = 1|c_t = 1) = 0.9, \quad p(o_t = 0|c_t = 1) = 0.1, \tag{5.25}$$

$$p(o_t = 1|c_t = 0) = 0.1, \quad p(o_t = 0|c_t = 0) = 0.9. \tag{5.26}$$

We assume the true context c_t depends on the agent's position. If the agent was in the vicinity of the novel object (demarcated as the grey states), the true context was $c_t = 1$ ("threatened"); otherwise, $c_t = 0$ ("not threatened"). Using these observations, the posterior distribution over c_t was modelled using a Beta distribution with parameters α_t and β_t. These parameters were updated iteratively with a decay term $\zeta = 0.1$ as follows:

$$\alpha_t = (1-\zeta)\alpha_{t-1} + o_t, \quad \beta_t = (1-\zeta)\beta_{t-1} + (1-o_t). \tag{5.27}$$

Finally, the belief state b_t was computed as the mean of the Beta posterior distribution:

$$b_t = \frac{\alpha_t}{\alpha_t + \beta_t}. \tag{5.28}$$

The stable learning simulations slowly update the default policy, given by a delta rule (Piray and Daw, 2021; Miller et al., 2019) upon taking action a in state s: $\hat{\pi}^d(a|s) \leftarrow \pi^d(a|s) + \alpha_d(1 - \pi^d(a|s))$ followed by normalising $\hat{\pi}^d(\cdot|s)$ to get the updated default policy $\pi^d(\cdot|s)$. Fig. 5.6, 5.7 use $\gamma = 0.99$, $\alpha = 0.1$ and $\alpha_d = 0.001$ (τ and λ mentioned in Figure captions, wherever applicable) and results are averaged over 10 runs.

Man is not worried by real problems so much as by his imagined anxieties about real problems.

— *Epictetus*

6

Homeostasis after injury: How intertwined inference and control underpin post-injury pain and behaviour

Contents

6.1 Prelude

The preceding part of this thesis addressed safe exploration, the challenge of learning and acting in the face of external threats. We now turn to the seemingly complementary problem of self-preservation, which concerns an agent's internal integrity - be it a physiological injury in an animal or a fault in a robot. This links directly to the biological principles of homeostasis (Davies, 2016) and allostasis (Sterling, 2012), where maintaining a stable internal milieu is as critical to survival as avoiding external harm. The core computational challenge, which this part of

the thesis will address, is that this internal state is rarely fully known; it must be inferred from noisy and incomplete signals. The following chapters will formalise this as a problem of partially observable control, offering a framework to understand not only post-injury behaviour and its transition to chronic pain but also potentially the design of more robust, self-aware machines.

6.2 Introduction

Injuries are common across the lifespan of many species and often lead to a period of vulnerability and reduced functionality whilst healing occurs. In both ecological and laboratory studies, characteristic behavioural changes such as increased anxiety and reduced activity are also often observed. They are considered to reflect adaptive changes and altered motivational priorities appropriate for safe recovery (Wall, 1979; Bolles and Fanselow, 1980; Williams, 2019). These changes are notoriously accompanied, and indeed potentially mediated, by forms of pain. Noting that organising appropriate behaviour is, by definition, a problem of control (Sutton and Barto, 2018), Seymour et al. (2023a) offered a substantially new perspective on tonic pain. Here, we centre Seymour et al. (2023a) in its natural reinforcement learning context, and provide the first concrete computational realisation of these ideas.

The critical innovation in Seymour et al. (2023a) was to note that the brain suffers inevitable uncertainties about the nature and extent of the injury and the status of the recovery process. They suggested that the brain continually integrates multisensory and physiological inputs to infer, and thereby represent, the uncertain injury state. This representation would then be tied to the appropriate choice of action or inaction and generate internal signals interpreted as pain. This framework extends Bayesian models of pain perception (Seymour et al., 2013; Büchel et al., 2014; Wiech, 2016) to address not only inferential aspects (for instance, expectancy effects) but also control under uncertainty. Crucially, they proposed that protective behaviours might restrict access to informative signals about recovery, potentially leading to persistent or maladaptive injury beliefs and contributing to chronic pain.

We formalise the suggestion of Seymour et al. (2023a) as a partially observable Markov decision process (POMDP; Kaelbling et al. (1998)). This situates injury, recovery, and pain within the explanatory framework of neural reinforcement learning (Sutton and Barto, 2018; Dayan and Daw, 2008), where the solution is a policy mapping observations to actions or inaction. In doing so, our work addresses several gaps in the literature: it (i) unifies Bayesian inference and control approaches in pain research (Seymour and Mancini, 2020), (ii) mathematically formalises in a minimal model, how information restriction about whether injury has resolved, may drive chronic pain states (Seymour et al., 2023a), and (iii) makes explicit how the value of information influences pain-related behaviour (Seymour, 2019).

We focus on the belief about the injury state constructed by the brain, proposing that this belief underlies the experience of tonic pain and modulates phasic pain responses. While this belief is informed by ascending sensory signals—particularly sustained nociceptor firing and hyperexcitability—such inputs, especially those from small-diameter, unmyelinated C-fibres (Debanne, 2004), may be unreliable or imprecise, especially during the later stages of healing. Accurate injury state estimation therefore, requires integrating multiple sources of information, including autonomic and physiological signals, exteroceptive cues such as vision (Höfle et al., 2010), and prior expectations (Mancini et al., 2022; Seymour and Mancini, 2020; Seymour et al., 2023a), accumulated over time to construct a coherent, control-relevant internal representation (Büchel et al., 2014; Wiech, 2016).

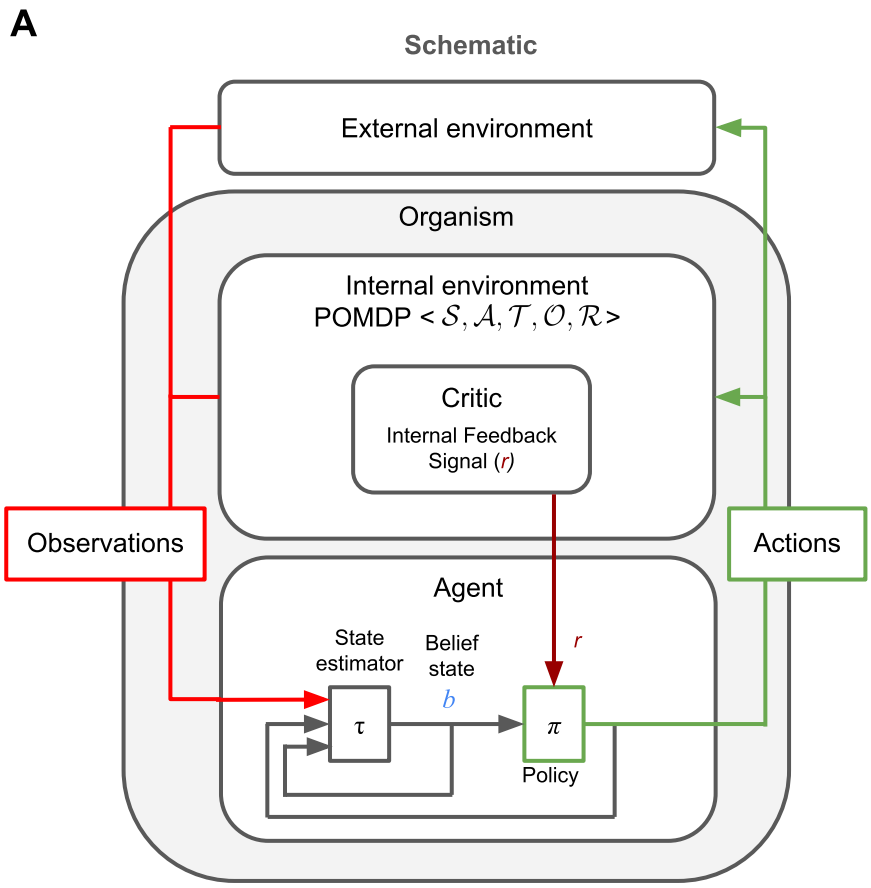
Although our simulations are inevitably limited, they show how tonic pain, as the belief about the state of the injury in the first instance, is a critical, internally constructed and interpreted signal which has broadcast properties, reorganising macroscopic patterns of behaviour. In our simulation results, we will dwell on a particular facet of the injury POMDP, which is the potentially large cost of the conventional policy solution to uncertainty, namely, information gathering to determine which options are better, or at least less worse. This cost could accrue if, for instance, the only way to learn about the current state of the injury is to attempt to use the injured body part. We examine such counter-intuitive actions

of injury investigation (e.g. rubbing injured areas conventionally interpreted in the context of Gate Control Theory) in the light of our theory, and also suggest how the costs might lead to the development of pathological states of chronic pain, even when the periphery is apparently no longer presenting nociceptive input (Fitzcharles et al., 2021).

We proceed by building a simple POMDP model with just a handful of states and actions, and studying the properties of its optimal policy for different parameter values and starting states. The POMDP may seem rather abstract; we therefore tie it as closely as possible to phenomena observed in the context of pain.

6.3 Theory sketch

We propose that the brain treats injury in terms of a partially observable Markov Decision Problem (POMDP; Fig. 6.1A) (Drake, 1962; Astrom et al., 1965; Sondik, 1971; Kaelbling et al., 1998). A POMDP comprises states $\mathcal{S}$, actions $\mathcal{A}$, transitions $\mathcal{T}$, observations $\mathcal{O}$ and utilities $\mathcal{R}$; we make substantial simplifications in our construction of all of these. The decision-theoretic task for the brain is then to determine a policy which maps the history of observations to the probability of taking an action to maximise the so-called future *return*, which is a long-run measure of cumulative utility. The general Bayesian Decision Theoretic framework is further described in the Methods section.



B

Simulation Action Set

a_{act}	Commit to a demanding activity e.g. get groceries	
$a_{r\&r}$	Not committing to the activity e.g. rest at home	
a_{que}	Investigating the injury e.g. probing the injured area	
a_{nul}	Avoiding to disturb the injury (omitted in some experiments)	

Figure 6.1: (A) Schematic of the injury POMDP, with an internal environment generating observations and conferring utilities and an internal agent inferring a belief state (Kaelbling et al., 1998) to choose optimal actions. (B) The POMDP action set used for simulations.

For the purpose of our simulations, we construct a simplified injury POMDP describing the following situation: A patient contemplating whether or not to commit to a demanding activity (e.g. getting groceries), whilst being uncertain about their injury. The true state characterises the current circumstance of the

injury. We consider just two possible states $s_t \in \{0, 1\}$ for healthy or injured at time t. However, crucially, the brain lacks full information about the state – interoception is incomplete, as well as being noisy. The agent thus only has a so-called *belief* state $b_t \in [0, 1]$, which is a probabilistic distribution over all states (Kaelbling et al., 1998). The actions include everything that could be done; however, we make the radical simplification of considering just three to four actions: a_{act}, which involves anything physically demanding but can collect resources; $a_{\text{r\&r}}$, which allows time for recovery and recuperation; and a_{que}, which involves assessing the injury, for instance trying to walk on a recently broken leg (Fig. 6.1B). We also allow for a null action a_{nul} that does nothing (we omit this from a few experiments for simplicity). Transitions specify how states change, probabilistically, based on the actions – in general, with recovery and worsening of the injury, although here, we restrict ourselves to the case that the state actually remains constant (and so drop the time index for s). Observations include all the exteroceptive and interoceptive information available to the brain about the injury, requiring multisensory integration, but we reduce it to a single observation channel for simplicity in our simulations. Investigatory actions, such as probing the injured area, might provide both types of information.

Utilities are central to determining the optimal policy and are a contentious aspect of the POMDP. In standard RL, utilities are externally provided by the environment (e.g., points in a video game). However, in nature, animals must infer or construct affective consequences from observations alone. This has inspired work on homeostatic reinforcement learning (Keramati and Gutkin, 2014) and intrinsic rewards (Chentanez et al., 2004; Barto and Simsek, 2005; Singh et al., 2009; Dayan, 2022). Here, as a simplification, we assume the agent has an intrinsic reinforcement function $r(s, a)$, defining immediate affective consequences of action a in true state s. For instance, this function might be large and negative/positive for activity/rest actions when injured (the former as a convenient proxy for the very long-run costs of incurring extra damage); small and negative for investigating while injured (due to potential harm); and negative for resting while uninjured, reflecting unmet resource needs (again, as a proxy for the very long run effects

of homeostatic deterioration). Since the agent only knows its belief b_t about s, the expected utility is internally constructed.

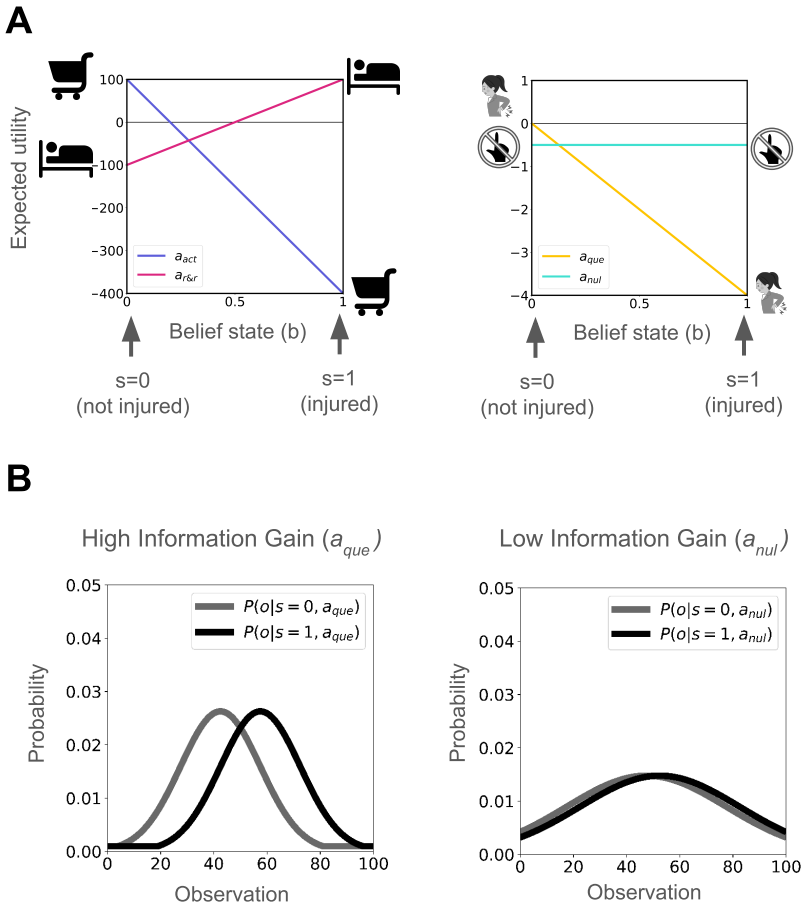


Figure 6.2: Injury POMDP: (A) Expected utilities of actions a_{act}, $a_{\text{r\&r}}$, a_{que} and a_{nul} in belief space. Utilities: when injured: $r(s = 1, a_{\text{act}}) = -400, r(s = 1, a_{\text{r\&r}}) = +100, r(s = 1, a_{\text{que}}) = -4$; when uninjured: $r(s = 0, a_{\text{act}}) = +100, r(s = 0, a_{\text{r\&r}}) = -100, r(s = 0, a_{\text{que}}) = 0$. $r(a_{\text{nul}}) = -0.5$ regardless of the internal state as a minor opportunity cost. (B) Observation processes for actions a_{que} and a_{nul}. (Further details in Methods - model simulation)

We illustrate the injury POMDP in a simplified setting where the agent chooses a single, final, choice of a_{act} or $a_{\text{r\&r}}$, but can choose to a_{que} along the way if it is sufficiently uncertain about the injury state. The true state remains fixed throughout the decision-making episode, and the return is the cumulative utility of the trajectory of choices. The one-step expected utilities in the belief space of the injury POMDP used in our simulations are shown in Fig. 6.2A. Note, that these expected utilities

differ from long-run return for actions a_{que} and a_{nul}. Choosing a_{que} and a_{nul} also results in sampling observations from distributions shown in Fig. 6.2B. We see a_{que} provides more discriminating or discerning observations about the internal state and is therefore more informative than a_{nul}. In our results, we often report state-action values $Q(b_t, a)$, which represent the expected long-run return for belief state b_t and action a. The optimal action in any b_t is the one with the highest Q value.

For concreteness, as a substantial simplification, we associate the belief state b_t with tonic pain – higher belief in being injured corresponds to greater pain. This pain becomes chronic if the agent fails to act or gather evidence to revise its belief. We link the expected negative reinforcement from a_{que} to phasic pain caused by injury investigation during the episode. This expectation averages over the belief state (see Methods), providing a mechanism for precisely tuning the feedback (Seymour, 2019) in case of correct inferences, but is also susceptible to incorrect or underinformed inferences.

6.4 Results

Normative consequences

Why do we investigate injury despite it being painful?

Our model provides a normative explanation for the behaviour as to why we investigate our injury despite it being painful. This is difficult to explain through a simple inferential or a control-theoretic approach alone.

In this instance, the feedback for a_{que} when injured is $r(s = 1, a_{\text{que}}) = -4$, denoting the negative consequence of probing the injury when injured. However, it also results in sampling observations using an observation process that can help infer the internal state more accurately. For this simulation demonstrating the value of information gained from injury investigation, we omit a_{nul} for simplicity, but including it does not affect the results.

When starting from an uncertain belief state of $b_0 = 0.5$, the agent progressively samples observations by choosing a_{que} at the cost of some (self-constructed) phasic pain. This is because there is a value to the information that can be gained given

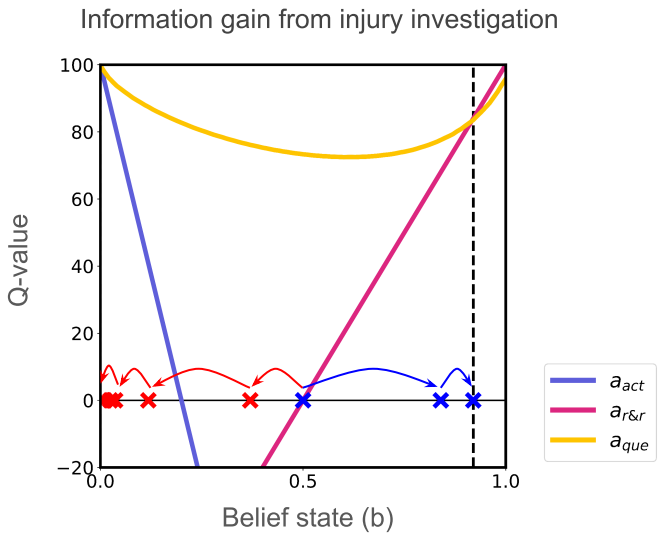


Figure 6.3: Action values after value iteration. Red arrows show belief updates under true state $s = 0$ (not injured), whereas blue arrows show belief updates under true state $s = 0$ (injured). The agent takes multiple a_{que} actions, reaching either belief thresholds, where either choosing a_{act} or $a_{\text{r\&r}}$ is more valuable than a_{que} and thus terminating the episode.

this uncertainty (see Fig. 6.3). With every sample, the agent accumulates evidence about the internal state and updates its belief until it can commit to either action a_{act} or $a_{\text{r\&r}}$ (Fig. 6.3, red and blue arrows). The belief state acts as the context for driving optimal behaviour. When the true internal state is not injured $s = 0$, the agent updates its beliefs (red arrows) and chooses the optimal decision a_{act} in the end. When the true internal state is injured $s = 1$, the agent updates its beliefs (blue arrows) and chooses the optimal decision $a_{\text{r\&r}}$ in the end. In conclusion, this demonstrates agent choosing injury investigating/probing actions despite associated with costs so as to accrue enough evidence to take the optimal decision. Note, actions investigating an injury extend beyond “rubbing the injured area" (explained by Gate Control Theory as "touch inhibits pain") to include non-contact explorations like bending a painful back or moving a painful joint etc.

Trade-off between information gain and phasic pain:

Consider the example of assessing your back after an injury. Mild discomfort from actions like stretching (e.g., "pins and needles") can provide valuable information

about the injury's state, outweighing the discomfort. In contrast, persistent tonic pain from bending offers little clarity about healing, as the sensory feedback may not distinguish between injury or recovery, making such actions less valuable.

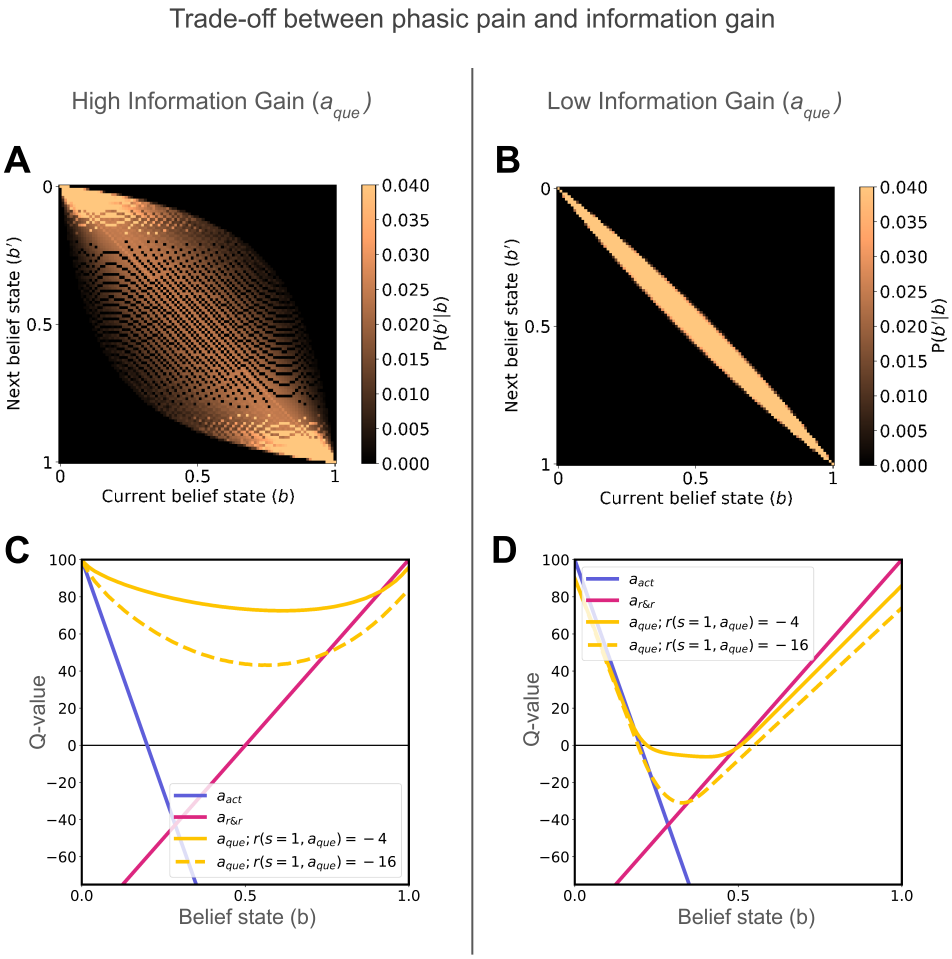


Figure 6.4: Didactic demonstration of a 2×2 factorial design. (A, B) Belief transitions from current belief state (b) to subsequent belief state (b') reflect the influence of high vs. low informativeness of the observation process on belief updating. (C, D) Action value plots show how phasic pain ($r(s = 1, a_{\text{que}}) = -4$ vs. $r(s = 1, a_{\text{que}}) = -16$) and high vs low information gain influence a_{que}. Higher phasic pain reduces action value, while higher information gain increases value when the expected cost remains constant.

We explored this trade-off between information gain and phasic pain using a 2×2 factorial simulation experiment. We manipulated the information gain by altering the observation process for a_{que}, using the two observation processes from Fig. 6.2B. In this simulation, we omit a_{nul} for simplicity of our didactic demonstration.

We find higher information gain of the observation process results in greater belief updates (Fig. 6.4A), and thus confers a higher action value of a_{que} across various uncertain belief states (Fig. 6.4C). This contrasts with less informative observation process which results in small belief updates (Fig. 6.4B) and lower value (Fig. 6.4D). Further, when observation process is the same, higher cost of a_{que} decreases the value of choosing a_{que} (Fig. 6.4C,D). This dynamic illustrates the trade-off between exploratory behaviour and pain avoidance. This experiment also highlights a key facet of our framework: the informativeness of observations and the motivational (phasic) punishment used for credit assignment are disentangled/treated independently. In conclusion, the trade-off between information gain and phasic pain determines the value of an action in our model.

Dysfunctional consequences

Having illustrated the normative consequences of the model, we will next consider how it might predict pathways to post-injury chronic pain. We describe two such pathways: (1) Information restriction and (2) Aberrant priors.

Information restriction:

A person recovering from a back injury might reasonably avoid movement due to fear of pain. This would hinder them from gathering valuable information that could signal the resolution of the injury. Such a Fear-Avoidance (Crombez et al., 2012) route to pain chronification has been elaborated in the path-dependent (Dayan et al., 2020) suggestion of the information restriction hypothesis (Seymour et al., 2023a).

To investigate this route to pain chronification, we simulated the injury POMDP in which the true internal state is no longer injured ($s = 0$), as in the late stages of recovery. In this scenario, the action to investigate the injury (a_{que}) is nevertheless slightly painful when $b_t > 0$, due to the uncertainty of the injury. Action a_{que} is more informative than a_{nul}, using the observation processes shown in Fig. 6.2B. The action a_{nul} has no cost associated with it, but is less informative than a_{que}.

The difference from $a_{\text{r\&r}}$ is that it is not a terminating action - the agent could subsequently change its mind and execute one of the three other action.

We varied the the cost of a_{que} from $r(s = 1, a_{\text{que}}) = -4$ to $r(s = 1, a_{\text{que}}) = -16$, while keeping the observation processes fixed to explore different action values where information gain exceeds phasic pain and vice versa respectively (Fig. 6.5A and B). We set the initial belief at the mid-point between the thresholds for choosing actions a_{act} and $a_{\text{r\&r}}$. When the information gained from a_{que} outweighs its associated pain, the individual's beliefs adaptively update, correctly resolving the hidden state to $b_t = 0$, leading ultimately to the optimal action ($a_{\text{r\&r}}$) over time (Fig. 6.5C, averaged belief state trajectories). However, when the phasic pain of a_{que} exceeds the information gain, it is less valued than a_{nul} and a_{nul} is preferred. Due to the less informative nature of a_{nul}, the beliefs do not change greatly. This leads to information restriction, as observed in belief state trajectories averaged over all simulation episodes (Fig. 6.5D). Thus, we can observe a prolonged incorrect belief that the injury persists, leading to slower recovery or persistent injury misinference, which in our model correlates with increased tonic pain (as per eqn. 6.11). Additionally, this behaviour results in more suboptimal avoidance behaviour - we see an increase in the agent choosing $a_{\text{r\&r}}$ (Fig. 6.5D, right). This simulation illustrates how restricted information due to pain avoidance can contribute to the prolonged incorrect belief that the injury persists, providing a computational account for the chronic pain pathway described by (Seymour et al., 2023a).

Aberrant priors:

We next study the adverse effects of having a strongly held incorrect belief about the severity of an injury, e.g. an aberrant prior. A prior (belief at the start of the decision-making episode) that underestimates the injury leads to further unnecessary harm, whereas a prior that overestimates the injury leads to over-protective behaviour. Both of our simulations use the same Q-values as Fig. 6.5A.

In the first experiment, we use an underestimating prior when the true state is injured ($s = 1$). With aberrant starting belief states closer to zero (but greater than

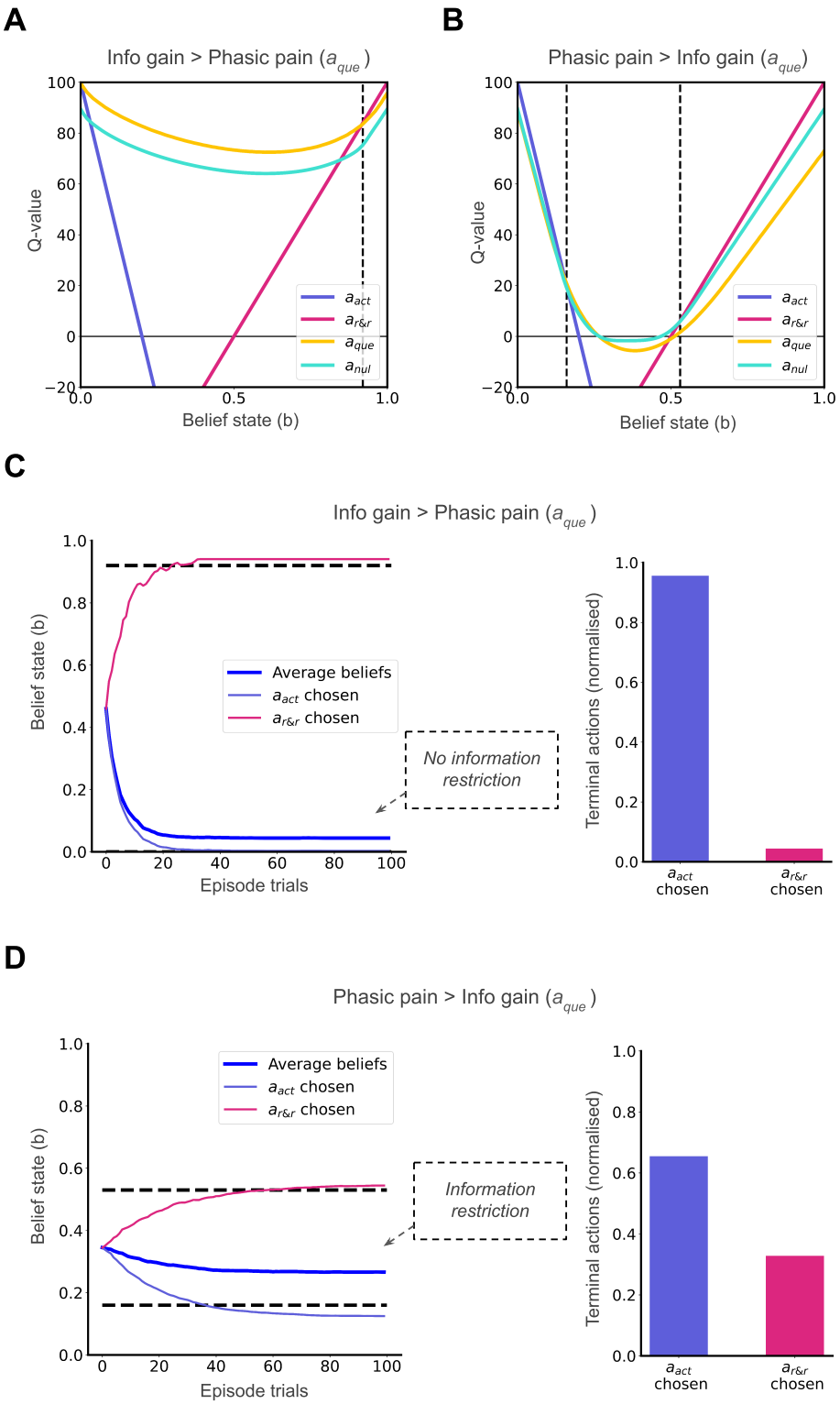


Figure 6.5: Two possible scenarios along the recovery where the internal state is $s = 0$: (A) Q-values when information gain from a_{que} outweighs phasic pain costs; (B) Q-values when phasic pain costs outweigh information gain. (Caption continued on next page.)

Figure 6.5: (Continued from previous page.) (C) Faster adaptive hidden state resolution is seen when more informative a_{que} is chosen over less informative a_{nul} (D) Maladaptive information restriction occurs when less informative a_{nul} is chosen over more informative a_{que}. Belief state trajectories are averaged over 1000 simulation episodes (dark blue) or only the episodes ending in a_{act} (violet) and $a_{\text{r\&r}}$ (pink). 'Episode trials' represent steps within a decision-making episode.

zero, as this is a decision threshold for a_{act}), we observe that the agent chooses a_{que} over a_{nul}. This results in greater average cumulative expected costs (interpreted as cumulative phasic pain) accrued during the episode in the interest of accurate state inference and subsequent optimal decision-making (Fig. 6.6A). These excess costs decrease as the starting belief becomes closer to $b_0 = 1$, i.e., being better aligned with the true state (Fig. 6.6A). This demonstrates a pathway for excess phasic pain through injury investigation due to maladaptive priors underestimating an injury when the true state is injured.

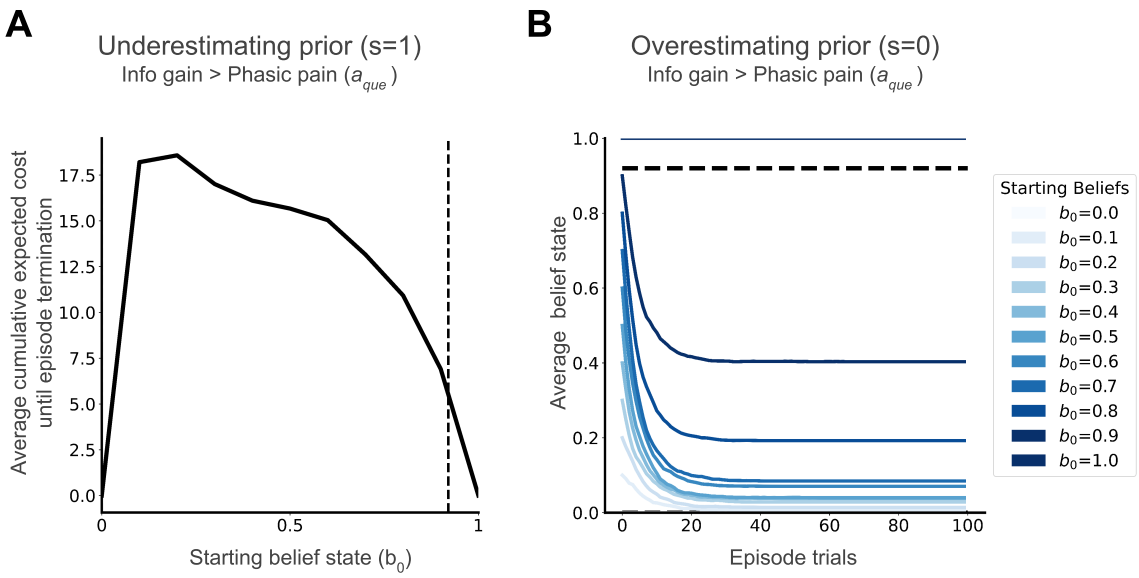


Figure 6.6: (A) Underestimation of injury when injured ($s = 1$), modelled as start belief states closer to 0, increases average phasic pain when injured. (B) Overestimation of injury when not injured ($s = 0$), modelled as start belief states closer to 1 but short of the decision threshold, can lead to prolonged incorrect beliefs, as seen in mean belief state trajectories. Belief state trajectories are averaged over 1000 simulation episodes.

In our second experiment, we use an overestimating prior when the true state

is not injured ($s = 0$). If the starting belief state is aberrantly closer to 1 (itself a possible consequence of probability distortions in the context of risk aversion; Gagne and Dayan, 2022), then it contributes to prolonged incorrect belief that the injury persists, seen in belief trajectories in Fig. 6.6B. Such prolonged incorrect belief decreases as the starting belief becomes closer to $b_0 = 0$, i.e., again being better aligned with the true state (Fig. 6.6B). This demonstrates a pathway for persistent tonic and potentially chronic pain, due to maladaptive priors overestimating the injury when the true state is recovered.

6.5 Discussion

We present a theoretical framework for understanding the computational logic of a dedicated homeostatic state for injury (Seymour et al., 2023a). Central to this framework is the notion that internal states are partially observable and require inference. This perspective provides a normative explanation for behaviours such as probing an injury despite immediate phasic pain, as these actions prioritise information acquisition. Additionally, we identify two broad "fault lines" that can lead to suboptimal behaviour and chronic pain, offering insights into how the homeostatic injury system can go awry.

Our framework builds on prior models of homeostatic motivation (Chentanez et al., 2004; Barto and Simsek, 2005; Singh et al., 2009) in which reinforcement signals are generated within the agent from observations, rather than being provided externally. However, we highlight the necessity of inferring internal states based on noisy observations and thus partial observability.

Our results align with aspects of the Fear-Avoidance model of pain chronification (Vlaeyen and Linton, 2000; Vlaeyen et al., 2016), offers a rich psychological framework grounded in fear learning and patient experience, positing that fear of movement leads to avoidance and physical deconditioning, thereby reducing opportunities to update beliefs about injury resolution. We offer the complementary approach (see Appendix C.1 Figure) of formalising information restriction and avoidance dynamics, providing a normative, computational, inference-based account that quantifies when

and why avoidance could sustain maladaptive beliefs, consistent with theories of ex-consequentia reasoning (Arntz et al., 1995; van Vliet et al., 2018; Meulders, 2019). Thus our model makes a precise, testable prediction about how avoidance behaviour (or its absence) can lead to persistent chronic pain (or recovery), respectively. This information-restriction pathway can be assessed in terms of stagnated belief updates about injury resolution. In contrast, the Fear-Avoidance model remains largely qualitative. While it clearly links avoidance to persistent fear, and lack of avoidance to fear extinction — drawing on insights from fear conditioning — its connection to chronic pain is more indirect. It relies on the narrative that fear of pain or re-injury discourages activity, leading to physical deconditioning of muscles and disability and, ultimately, persistent pain (Vlaeyen and Linton, 2000). Crucially, it thereby lacks such a formal account of how injury-related beliefs are inferred or updated.

Our simulations with aberrant priors, particularly those reflecting overestimation, also align with self-fulfilling prophecy and catastrophisation models (Jepma et al., 2018; Flink et al., 2013), which highlight how catastrophic thinking enhances attentional demand and hinders disengagement from pain (Van Damme et al., 2002, 2004).

Looking ahead, the "fault lines" we identify can be categorised as the agent: solving the "wrong problem," solving the correct problem with the "wrong solution," or solving the correct problem correctly in an unfortunate, "wrong environment" (Huys et al., 2015). For example, aberrant priors illustrate solving the wrong problem, while maladaptive behaviours, such as excessive avoidance due to Pavlovian biases (Mahajan et al., 2024) or habitual traits (Ball and Gunaydin, 2022), represent wrong solutions that could lead to information restriction. A wrong environment may yield priors or utilities that were once adaptive but have become maladaptive, creating scenarios that appear to involve the wrong problem or solution. This categorisation provides a theoretical framework for linking current behavioural theories of chronic pain under a unified computational perspective, akin to computational nosology (Mathys et al., 2016; Friston et al., 2017).

Our model is particularly relevant to clinical chronic pain conditions involving ambiguous symptoms, such as sciatica after lumbar radiculopathy. For instance, it

may explain observations with sciatica muscle weakness patients, where baseline muscle weakness (a worse symptom) is counter-intuitively associated with improved outcomes (Konstantinou et al., 2018; Vroomen et al., 2002). Recovery of muscle strength could provide objective and unambiguous observations thereby providing high-information feedback for accurate injury inference and self-improvement. Similarly, our model could potentially explain phenomena such as boom-bust activity cycles (Moseley, 2003; Antcliff et al., 2016) via faulty injury inference feeding into maladaptive behaviour. Potential interventions could involve improving injury inference, for example, through external cues or reducing maladaptive behaviour by activity pacing. Moreover, our framework may offer insights into the positive outcomes of Pain Reprocessing Therapy (Gordon and Ziv, 2021), particularly the role of somatic tracking and guided touch (Kim et al., 2022; McParlin et al., 2022).

The model is just a first step. In particular, we did not include transitions in internal states, and thus actual recovery or exacerbation of injuries. Incorporating this could allow us to specify multiple 'activity' actions on a spectrum of intensity, with varying probabilities for worsening one's internal state, differential costs for acting more quickly, intensely or vigorously, and different times of completion. This will allow us to relate our model to notions of activity pacing, i.e. choosing appropriate intensity and timing, and also theories of vigour (Niv et al., 2005, 2007). We rather arbitrarily generated the utility terms in our simulations; in the future, we plan to derive them from both physiological costs and the opportunity costs associated with the threats to homeostasis that forced inaction imposes. We have yet to model concrete neural or behavioural data (although we have identified ACL-tear as an ideal test case). More subtly, there are uncertainties about the exact modeling construct of tonic pain. Here, we modelled it directly using the belief over an injury state, which is in line with previous lines of work treating pain in terms of Bayesian inference. However, if an experiment is able to demonstrate that actions associated with information gain reduce tonic (as well as phasic) pain, as in endogenous analgesia (Basbaum and Fields, 1978; Bannister and Dickenson, 2017) then it would deem value over belief states $V(b)$ as a more accurate correlate

of tonic pain. Alternatively, one may also choose to represent tonic pain and its multiple attributes as a predictive state representation, as in general value functions (Sutton et al., 2011). Each of these various approaches implies a different meaning to tonic pain and its role in guiding behaviour.

The core contribution of this work is at the computational level, rather than an algorithmic one (Marr, 2010). We use a model-based approach in this work, which is sufficient to account for the proposed phenomena. Model-free approaches, such as the actor-critic framework (Maia, 2010; Moutoussis et al., 2008), may be able to reproduce the results, provided one uses recurrent neural networks (RNNs) as function approximators. These can learn to represent belief information correctly as long as the RNN capacity is sufficient (Hennig et al., 2023). However, a large amount of interaction data under different scenarios is typically required to train these RNNs and allow them to operationalise the value of information gain, e.g. meta-learning (Wang et al., 2016, 2018a). This data requirement presents a challenge, as taking precarious actions (e.g., repeatedly acting when severely injured) risks worsening the injury or, in the worst case, causing death, which would preclude learning altogether. Therefore, it is likely that we are endowed with innate models that are either fine-tuned through experience or used to guide model-free learning. Note further that in the case of actor-critic models, only the critic units would be expected to represent the value of information gain.

Finally, and perhaps of greatest translational importance, chronic pain likely arises from complex interactions between learning processes, internal and external environments, and behavioural feedback loops. By offering a computational framework that spans from phasic to chronic pain, our account provides a mathematical foundation for making theories of this condition explicit.

6.6 Methods

Bayesian decision theoretic approach to homeostasis

We build upon the Bayesian Decision Theoretic (BDT) framework (Dayan and Daw, 2008; Huys et al., 2015) utilising POMDPs (Drake, 1962; Astrom et al.,

1965; Sondik, 1971; Kaelbling et al., 1998; Ross et al., 2008) and extend it to the problem of homeostasis.

Our internal environment POMDP is formally represented as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, \mathcal{R} \rangle$, where: $\mathcal{S}$ is the set of all internal states (s). $\mathcal{A}$ is the set of all possible actions (a). $\mathcal{T}$ is the transition function, where $\mathcal{T}(s, a, s') = P(s'|s, a)$ is the probability of ending in s' by performing action a in state s. $\mathcal{O}$ is the set of all observations (o), where the observation function is $\mathcal{O}(s', a, o) = P(o|a, s')$ is the probability of observing o if action a is performed and resulting state is s'. $\mathcal{R}$ are the utilities, where $r(s, a)$ is the internal feedback signal obtained by taking action a in internal state s.

At any time t, the agent does not necessarily have access to its complete internal state s_t, but has access to observations o_t from the internal environment, which provide incomplete (noisy) information about the internal state. The agent therefore infers a belief state b_t over internal state space $\mathcal{S}$ (Astrom et al., 1965). The belief state is defined as the posterior probability density of being in each internal state s, given the complete history h_t of actions and observations at any time t, $h_t = \{a_0, o_1, ...o_{t-1}, a_{t-1}, o_t\}$ and initial belief b_0

$$b_t(s) = P(s_t = s|h_t, b_0) \tag{6.1}$$

The belief state b_t is a sufficient statistic for the history h_t (Smallwood and Sondik, 1973; Sondik, 1978). At any time t, the belief state b_t can be computed from the previous belief state b_{t-1}, using the previous action a_{t-1} and the current observation o_t using a state estimator function (τ) derived using Bayes rule (Kaelbling et al., 1998; Ross et al., 2008)

$$\begin{aligned} b_t(s') &= \tau(b_{t-1}, a_{t-1}, o_t)(s') \\ &= \frac{\mathcal{O}(s', a_{t-1}, o_t) \sum_{s \in \mathcal{S}} \mathcal{T}(s, a_{t-1}, s') b_{t-1}(s)}{P(o_t|b_{t-1}, a_{t-1})} \end{aligned} \tag{6.2}$$

where the $P(o_t|b_{t-1}, a_{t-1})$ is the marginal distribution over observations o_t (independent of s', acting as a normalising factor for b_t).This belief transition function (τ) is used to obtain belief transition plots and belief trajectories.

A critical concept for the POMDP is a stationary policy $\pi(b, a)$ which is a probability distribution over the action a that the agent will take given that its current belief is b. Define $V^{\pi}(b_t)$ as the expected discounted long-term reward following policy π as the value of this belief state:

$$V^{\pi}(b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{a_t \in \mathcal{A}} \rho(b_t, a_t)\pi(b_t, a_t)|b_0 = b\right] \tag{6.3}$$

where γ is a temporal discount function and

$$\rho(b_t, a_t) = \sum_{s \in \mathcal{S}} b_t(s) r(s, a_t) \tag{6.4}$$

is the feedback function on belief states, constructed from the original feedback signal on internal states $r(s, a)$. Then, it can be shown that the value satisfies the Bellman equation:

$$V^{\pi}(b) = \sum_{a \in \mathcal{A}} \pi(b, a)\Big[\rho(b, a) + \gamma \sum_{o \in \mathcal{O}} P(o|b, a)V^{\pi}(\tau(b, a, o))\Big] \tag{6.5}$$

where the sum over $\mathcal{O}$ is interpreted as the expected future return over the infinite horizon of executing action a, assuming the policy π is followed afterwards. It is known that there is a deterministic policy $\pi^*(b, a)$ that optimizes the value of all belief states (Ross et al., 2008)

$$\pi^*(b) = \arg\max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}} \rho(b_t, a)\pi(b_t, a)|b_0 = b\right] \tag{6.6}$$

The value function V^* of the optimal policy π^* is the fixed point of the Bellman's equation (Bellman, 1954)

$$V^*(b) = \max_{a \in \mathcal{A}}\Big[\rho(b, a) + \gamma \sum_{o \in \mathcal{O}} P(o|b, a)V^*(\tau(b, a, o))\Big] \tag{6.7}$$

and one can define a corresponding optimal Q-value function, which refers to the value of taking action a and then following the optimal policy, as

$$Q^*(b, a) = \rho(b, a) + \gamma \sum_{o \in \mathcal{O}} P(o|b, a)V^*(\tau(b, a, o)) \tag{6.8}$$

Model implementation and simulation details

We do not aim to model the entire lifetime of an agent but rather an episode in the agent's lifetime. Here, we maintain a binary hidden state $s = \{0, 1\}$ and we do not model transitions between states. Since we do not model transitions in s_t within the POMDP episode, the state estimator equation 6.2 defining the belief transitions is reduced to the following,

$$b_t(s) = \tau(b_{t-1}, a_{t-1}, o_t)(s) = \frac{\mathcal{O}(s, a_{t-1}, o_t) b_{t-1}(s)}{P(o_t | b_{t-1}, a_{t-1})} \tag{6.9}$$

This is similar to the Tiger POMDP environment by (Kaelbling et al., 1998), except that the a_{que} is also belief-dependent. Further we refer to the belief state over s as $P(s = 1) = b_t(s = 1) = b$ and $P(s = 0) = b_t(s = 0) = 1 - b$. If the POMDP has more states, the belief state provides the probability distribution of being in each of those states.

To solve the POMDP, we treat it as a belief MDP, and of the several approaches to perform value approximation (Hauskrecht, 2000; Kochenderfer et al., 2022), we utilise belief grid value iteration (Lovejoy, 1991)leading to piecewise linear and convex (PWLC) value functions. Here, we discretise the belief space in steps of 0.01 and the observation distribution, $o \in (0, 100] = \mathcal{O}$, in steps of 1 and assume the observation function to be known.

From the action set $\mathcal{A} = \{a_{\text{act}}, a_{\text{r\&r}}, a_{\text{que}}, a_{\text{nul}}\}$ described in Fig. 6.1, a_{act} and $a_{\text{r\&r}}$ result in episode termination and a_{que} and a_{nul} do not. We then used the Q values of each action in each belief state to determine the decision thresholds for choosing actions a_{act} and $a_{\text{r\&r}}$ which terminate the episode, defined as the most uncertain belief state b where the optimal action is no longer a non-terminal action.

Simulation parameters

Experiment	True state	Outcome for a_{que} when injured	Observation model for a_{que}
Injury investigation	$s = 1$	$r(s = 1, a_{\text{que}}) = -4$	$O(s, a_{\text{que}}, o) = HI$
Phasic pain - information gain trade-off (High information gain condition)	$s = 1$	$r(s = 1, a_{\text{que}}) = -4, -16$	$O(s, a_{\text{que}}, o) = HI$
Phasic pain - information gain trade-off (Low information gain condition)	$s = 1$	$r(s = 1, a_{\text{que}}) = -4, -16$	$O(s, a_{\text{que}}, o) = LI$
Information restriction (Info gain >phasic pain)	$s = 0$	$r(s = 1, a_{\text{que}}) = -4$	$O(s, a_{\text{que}}, o) = HI$, $O(s, a_{\text{nul}}, o) = LI$
Information restriction (Phasic pain >info gain)	$s = 0$	$r(s = 1, a_{\text{que}}) = -16$	$O(s, a_{\text{que}}, o) = HI$, $O(s, a_{\text{nul}}, o) = LI$
Underestimating prior	$s = 1$	$r(s = 1, a_{\text{que}}) = -4$	$O(s, a_{\text{que}}, o) = HI$, $O(s, a_{\text{nul}}, o) = LI$
Overestimating prior	$s = 0$	$r(s = 1, a_{\text{que}}) = -4$	$O(s, a_{\text{que}}, o) = HI$, $O(s, a_{\text{nul}}, o) = LI$

The utilities mentioned in Fig. 6.2 were fixed throughout, with the exception of experiments where phasic pain exceeds information gain for a_{que}, therefore, we vary $r(s = 1, a_{\text{que}}) = -4$ to $r(s = 1, a_{\text{que}}) = -16$. The observation process for a_{que} and a_{nul} were always high information gain (HI) and low information gain (LI) respectively, as shown in Fig. 6.2, with the exception of the Fig. 6.4, where we vary the information gain of a_{que}. For the HI process, the means of observation distributions for $s = 0$ and $s = 1$, were apart by 15 units and the observation distributions had the std. dev. of 15 units. For the LI process, the means of observation distributions for $s = 0$ and $s = 1$, were apart by 5 units and the observation distributions had the std. dev. of 30 units. Action a_{nul} was omitted from Normative Consequences results for simplicity but included in Dysfunctional Consequences results. The maximum trials within a POMDP episode were set to 100. Results were averaged over 1000 independent POMDP episodes for a fixed utility function. The discount factor was set to $\gamma = 1$.

Definition of phasic and tonic pain in our model simulations

In the first instance, we define phasic pain based on equation 6.4, as belief-dependent negative feedback (expected negative costs of an action), only when concerning the pain/injury systems.

$$\text{Phasic Pain} \propto \sum_{s \in \mathcal{S}} b(s)\zeta(s, a) \tag{6.10}$$

Here, $\zeta(s, a)$ is the negative rewards or punishments concerning the pain systems and excludes other kinds of punishments such as opportunity costs or loss in resources. In our simulations, we only require and report the phasic pain for action a_{que}, where the $\zeta(s, a_{\text{que}}) = r(s, a_{\text{que}})$.

We define tonic pain to be proportional to the belief-weight drive in the injury state space. This further contributes to the affective value, which is a function of the belief state $V^{\pi}(b)$.

$$\text{Tonic Pain} \propto b(s = 1) \tag{6.11}$$

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'

— *Isaac Asimov*

7

Towards an injury state for robots using Neural Associative Skill Memories

Contents

7.1 Prelude

The previous chapter formalised self-preservation using a POMDP, which relied on a given, handcrafted generative model of the injury state. While useful for exploring the computational principles of control under uncertainty, this approach leaves a critical question unanswered: how might such a model of expected bodily states be learned from experience in the first place? This chapter addresses this by translating the problem to neurorobotics. We explore how a generative model of expected sensorimotor experience can be acquired from demonstrations using self-supervised, local learning rules. Although we compromise by focusing on reactive fault detection

rather than goal-directed planning, this approach provides a biologically plausible mechanism for how an agent—biological or artificial—can build an internal model of its own body to detect abnormalities, bridging the gap between computational theory and practical implementation.

7.2 Introduction

Animals spend their lives learning, storing, and refining a repertoire of sensorimotor skills, which they must not only express adaptively but also use to detect abnormalities. To recognise an abnormal event, an animal requires an understanding of what normal actions 'feel' like. This corresponds to having a robust internal model of the body and its interactions with the world, i.e., a generative model that predicts the normal sensory consequences of actions. When reality deviates from these predictions, the resulting prediction error, or 'predictive surprise' (Friston, 2005; Clark, 2013), serves as a crucial signal. From a biological standpoint, this signal can indicate a sensorimotor conflict or an external perturbation (Shadmehr et al., 2010), or even bodily harm, such as an injury (Seymour and Mancini, 2020; Walters et al., 2023). The brain's capacity to detect such violations of its internal priors is therefore not a secondary feature but a core component of safe and self-preserving behaviour.

Inspired by this principle, roboticists have sought to equip machines with analogous capabilities through Associative Skill Memories (ASMs) (Pastor et al., 2012, 2013). The core idea is to link motor commands, often represented by dynamic movement primitives (DMPs) (Ijspeert et al., 2013), with their expected sensory feedback. This allows a robot to detect 'faults', the engineering equivalent of injuries or perturbations, and react accordingly. However, traditional ASMs have a critical architectural limitation compared to their biological counterparts: they rely on a hard-coded, dictionary-like library of skills. Each skill is learned and stored as an independent module, requiring an external mechanism to explicitly select which one to execute. Consequently, fault detection is constrained to a single skill, failing to capture the integrated nature of biological motor repertoires.

This contrast between engineered libraries and integrated biological systems highlights a key unresolved question: how can a single, unified neural network learn a repertoire of sensorimotor skills in a biologically plausible manner? Such a system would need to solve two fundamental challenges that are handled seamlessly in the brain. First, it must be able to detect abnormalities with respect to all learned skills in the repertoire, without being explicitly told which skill is being performed. Second, it must be able to use sensory cues to contextually infer which skill memory is most appropriate to express in different scenarios.

This paper addresses this question by introducing Neural Associative Skill Memories (Neural ASMs), a framework that uses a temporal predictive coding network to learn an embodied generative model of a robot's sensorimotor repertoire. We focus on how a sequential memory of sensorimotor observations can be learned from skill demonstrations, eschewing the separate problem of how optimal motor actions leading to these sequences are planned. Our central goal is to demonstrate how a single network, trained with self-supervised, local learning rules, builds upon the original concepts of Associative Skill Memories (Pastor et al., 2012, 2013). To this end, we provide basic demonstrations of three outcomes. First, the model performs fault detection across all learned skills in its repertoire. This is achieved by identifying abnormally high network energy, which is indicative of an out-of-distribution state with respect to the demonstration data. Second, it supports reactive correction by minimising proprioceptive prediction errors. Third, it facilitates the expression of different skills through contextual inference from early-stage cues, in a robotics simulation inspired by a human motor experiment (Sheahan et al., 2016). By doing so, we aim to provide a step towards safer, self-preserving robotics and offer a basic computational model for how the brain might learn, express, and monitor its own motor skills. Lastly, this study explores the foundational capabilities of a novel learning method involving local learning and inference within a temporal predictive coding framework. We demonstrate these capabilities using streamlined simulations, specifically fault detection in pick-and-place movements and perturbation compensation in point-to-point movements.

7.3 Related Work

Associative Skill Memories and Movement Primitives

The foundation of our work lies in Associative Skill Memories (ASMs) (Pastor et al., 2012, 2013), which extend Dynamic Movement Primitives (DMPs) (Ijspeert et al., 2013), a class of attractor-based models for generating stereotyped movements. The key idea of ASMs is to associate these motor primitives with their expected sensory feedback, enabling fault detection and reactive control. However, the standard ASM framework implements this using multiple handcrafted modules: an explicit skill library to store DMPs, a separate system to maintain sensory statistics for each movement, and a prediction module to find the closest match in the library (Fig. 7.1A). This modular, library-based architecture requires explicit skill selection and can only detect faults relative to one active skill at a time. Our work aims to overcome this limitation by learning a repertoire of skills within a single, unified neural network.

Predictive Coding for Sensorimotor Learning

Our model is built on the principles of predictive coding (Rao and Ballard, 1999; Friston, 2005; Clark, 2013), an influential theory in neuroscience positing that the brain continuously generates predictions of sensory input and updates its internal model based on prediction errors. This self-supervised process has recently been proposed as a biologically plausible alternative to backpropagation (Whittington and Bogacz, 2017, 2019; Song et al., 2024) and has been applied to associative memory (Salvatori et al., 2021) and temporal sequence learning (Tang et al., 2024b). In robotics, predictive coding has been used for body state perception and multisensory integration from noisy information to filter internal state (Lanillos and Cheng, 2018). However, these models typically do not learn the dynamics of different movements, which is crucial for differentiating skills. Similarly, novelty detection (akin to abnormality detection in our work) has been demonstrated in static predictive coding networks by identifying inputs that generate high prediction errors (or energy)

(Li et al., 2025), but not applied to temporal sequence learning or to the domain of neurorobotics. Our work extends these ideas by using a temporal predictive coding network to learn a dynamic model of a full sensorimotor repertoire, using the model's energy as a natural signal for fault detection across all learned skills.

Temporal Sequence Learning with Recurrent Architectures

Learning sensorimotor sequences has long been a domain of recurrent neural networks (RNNs). Influential work has used RNNs trained with backpropagation through time (BPTT) to generate complex sensorimotor sequences by learning associations between initial states and subsequent trajectories (Nishimoto and Tani, 2004; Nishimoto et al., 2008; Yamashita and Tani, 2008). A key insight from this research, particularly from Yamashita and Tani (2008), is the use of hierarchical structures with multiple timescales to capture complex temporal dependencies (MTRNNs). While powerful, these models typically rely on BPTT (Rumelhart et al., 1986), a non-local learning rule considered biologically implausible because it requires propagating error signals backwards through the network's entire temporal history.

Our model, based on temporal predictive coding (tPC) (Millidge et al., 2024b; Tang et al., 2024b), offers an alternative. Temporal predictive coding networks learn temporal dependencies using local, Hebbian-like updates that operate only between adjacent layers and time steps. This locality, while biologically plausible, is an approximation of BPTT truncated to one time step (Tang et al., 2024a). A benefit of tPC networks is that they integrate models of temporal sequence learning in the rich field of Bayesian inference. This allows us to use the notions of 'energy' of a model to formalise the concept of faults in this work. A key feature of our predictive coding framework is the iterative inference process used to converge on a hidden state at each time step. This process introduces an additional, faster timescale for online inference, distinct from the slower timescale of weight updates during learning. This mechanism shares some conceptual similarities with fast-weight RNNs (Ba et al., 2016) and error-regression in MTRNNs (Ahmadi and Tani, 2017). It further provides a candidate neural process for motor preparation,

allowing contextual cues to shape a subsequent motor plan. These kinds of iterative inference from early-stage contextual cues further align with the hypothesis of the role of motor preparatory activity in setting the initial state of a dynamical system rather than explicitly representing movement parameters (Churchland et al., 2010, 2006a,b; Cisek, 2006; Fetz, 1992).

Computational Models of Homeostasis and Injury-related Behaviours

The concept of fault detection in our model aligns with computational theories of interoception and homeostasis, the processes by which the brain senses, predicts, and regulates the body's internal state (Seth, 2013; Barrett, 2017). Our model adopts the view that the brain must infer its bodily state from multiple, often noisy, sensory signals (interoceptive, exteroceptive, and proprioceptive) to guide control (Seymour and Mancini, 2020). This contrasts with theories that assume direct access to (noise-free) internal states (Keramati and Gutkin, 2014) or rely purely on potentially noisy peripheral nociceptive signals to direct self-preserving behaviours (Walters et al., 2023). In this Bayesian inference-based view, the prediction errors arising from unexpected bodily signals drive an updated belief that the body is damaged, which in turn elicits protective behaviours (Seymour et al., 2023a; Mahajan et al., 2025a). Our model's use of prediction errors to detect deviations from a learned 'normal' repertoire provides an analogous mechanism.

While other computational models have explored goal-directed homeostatic and interoceptive control (Tschantz et al., 2022; Mahajan et al., 2025a), simulating interesting behaviours like investigating one's injury to gain information about the injury state despite associated phasic pain (Mahajan et al., 2025a), they typically rely on hand-crafted generative models or restrictive notions of homeostatic 'setpoints' (Keramati and Gutkin, 2014). This leaves two questions unanswered: (1) how can these ideas be extended to more complex robotics tasks, and (2) how can these computational-level theories be implemented with biologically plausible learning rules? Our work addresses both points by demonstrating that a generative model

can be learned directly from demonstrations using local learning rules, providing a framework for fault detection and reactive control in a neuro-robotic system, while leaving goal-directed self-preserving behaviours as future work.

7.4 Neural ASMs: A Theory Sketch

Neural Associative Skill Memories (Neural ASMs) replace the modular, library-based architecture of traditional ASMs with a single, unified network. This network is a temporal predictive coding (tPC) model (Millidge et al., 2024b; Tang et al., 2024b) that learns a generative model of sensorimotor sequences from demonstrations (Fig. 7.1). The model has a hierarchical structure analogous to a Hidden Markov Model (HMM), where hidden states (z) predict observations (x) as well as their own future states. The sensory and motor observations are concatenated at the observation layer (x), similar to previous work (Nishimoto et al., 2008; Yamashita and Tani, 2008). The entire system learns by minimising a single energy function via local, Hebbian-like updates (see Methods, equations 7.1-7.4).

The model operates in two distinct phases: memorisation and recall. During memorisation, the network learns from demonstrated sensorimotor sequences. For each step in a sequence, it infers a hidden state and updates its weights, creating an attractor in its energy landscape that corresponds to that sequence (Salvatori et al., 2021; Tang et al., 2024b). This process allows the model to store multiple skill memories within a single set of weights.

During recall, the network retrieves a stored sequence from a partial cue. This begins with cued inference, where initial sensorimotor observations (e.g., the robot's starting position and sensory state) are used to infer the corresponding hidden state (Fig. 7.1C, purple arrow). This inference process is analogous to motor preparation, where context sets the initial conditions for a movement (Churchland et al., 2010). Once the initial state is set, the network autonomously generates the predicted sensorimotor trajectory for the remainder of the skill (Fig. 7.1C, orange arrows). In this work, we use offline recall (Tang et al., 2024b), where the trajectory is generated ballistically after inferring the hidden state from a few early-stage

contextual cues. Examples of such offline processes in biology include sequences guided from working memory (Mizes et al., 2023, 2024) or mental simulation, where trajectories are generated without producing actual movements (Yamashita and Tani, 2008; Jeannerod, 1994; Tani, 1996). An alternative, online recall, would involve continuous inference using incoming sensory data at each step, allowing for dynamic switching between learned skills, but this is not explored here.

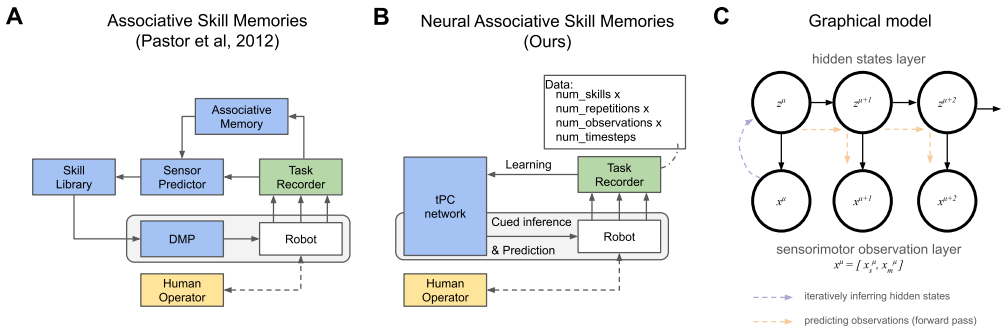


Figure 7.1: (A) Traditional ASMs (Pastor et al., 2012) use a modular, library-based architecture where each movement primitive and associated sensory statistics are stored separately. (B) Our Neural ASM approach replaces this with a single tPC network that learns multiple skill memories from demonstrations, where each demonstration is a time series of sensorimotor observations. (C) The underlying graphical model is a Hidden Markov Model (HMM), where hidden states (z) capture the dynamics and inferred context, and predict observations (x). Cued inference (purple arrow) sets the initial state from early observations, which then allows for offline prediction (orange arrows) of the sequence. While prediction can be a simple forward pass in an HMM, more complex generative models would require iterative energy minimisation.

Finally, the predicted sequence of motor commands (e.g., joint angles) from the Neural ASM serves as a high-level dynamical policy. This policy provides a reference trajectory to a low-level controller, which executes the movement and can make minor reactive adjustments online (cf. Schaal et al. (2007); Appendix D.1).

7.5 Results

Memorised skills are useful in fault detection and simple reactive correction

Having introduced Neural ASMs as a viable alternative to ASMs, we now demonstrate their core functionalities: fault detection and reactive fault correction in a simple simulation. We train a model to memorise two pick-and-place skills from demonstrations (Fig. 7.2A) and then test its ability to detect and react to simulated faults. The model uses implicit skill recognition during cued inference to retrieve the correct skill from initial observations using offline recall, unlike traditional ASMs, which require explicit selection of a movement primitive (Pastor et al., 2012, 2013).

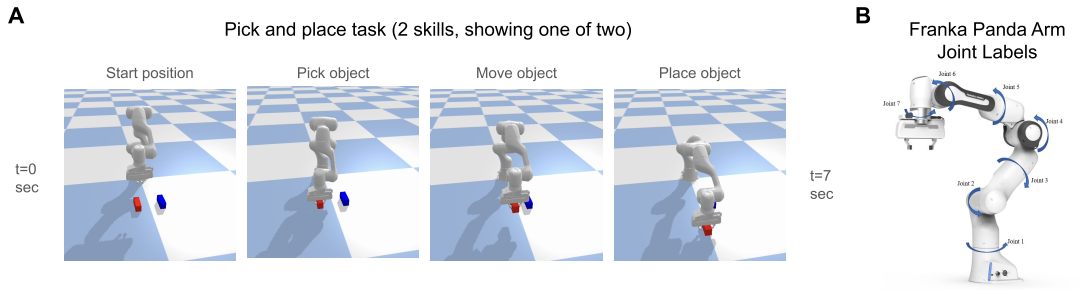


Figure 7.2: (A) Demonstration of Neural ASMs learning two pick and place skills in simulation. The sensorimotor sequences in the dataset used for learning from demonstrations are generated using predefined end-effector goals and use inverse kinematics to get joint angles. This is intended to be a proxy for teleoperation in simulation. (B) A schematic of Franka Panda arm joint labels (Adapted from Rogel et al. (2022), under CC BY 4.0 license. A title line was added.)

We first demonstrate fault detection qualitatively by measuring the network's energy (i.e., sum of squared prediction errors). Fig. 7.3 shows two examples of a joint-locking fault. In a minor fault where joint 5 gets stuck (Fig. 7.3A), the network's energy increases abnormally, successfully detecting the fault (Fig. 7.3B) using out-of-distribution detection on the network's energy. The individual prediction errors also correctly identify joint 5 as the primary location of the fault (Fig. 7.3C-E). In a more severe fault where joint 2 gets stuck and causes a collision (Fig. 7.3F), the model again detects the fault via a spike in energy (Fig. 7.3G). However, in this case, the downstream effects of the collision cause larger prediction errors in other joints

and sensors, illustrating that while the effects of a fault can be isolated, identifying the root cause remains a challenge (Fig. 7.3H-J). Please refer to the Methods section for more details on how the faults are simulated. A subtlety about Neural ASMs is that since the skill recognition is entirely implicit, the fault detection only depends on the current predictions, which in turn depend on the inferred context. This is unlike ASMs, which require knowing the explicit movement being performed to compare the observations to the corresponding signal statistics.

These qualitative examples of fault detection in different sensors are similar to Pastor et al. (2012, 2013). In addition, we perform a systematic evaluation of fault detection and isolation. For simplicity, we use percentile-based thresholding of the energy distribution during normal operation without faults to set the threshold for (out-of-distribution) fault detection. Fig. 7.3K shows the energy distribution during our pick-and-place task over 10 trials of normal operation for each skill, along with the 95th percentile threshold as an example, which is equivalent to a 5% false positive rate (FPR). We compare the performance of our Neural ASM against a baseline analogous to traditional ASMs that uses normalised Z-scores for error detection (see Methods for full details). By setting the detection threshold to correspond to a false positive rate (FPR) of 1–5%, we find that the Neural ASM correctly detects 82–83% of simulated faults, a modest improvement over the baseline's 74–82% detection rate. However, the Neural ASM demonstrates a substantial advantage in fault isolation, correctly identifying the specific joint responsible for the fault in 79.5% of cases, whereas the baseline is only able to do so in 41% of cases. This concludes our basic demonstration of systematic evaluation, which can be extended to real-world robots with more realistic faults and comprehensive comparisons with alternate benchmarks in future work.

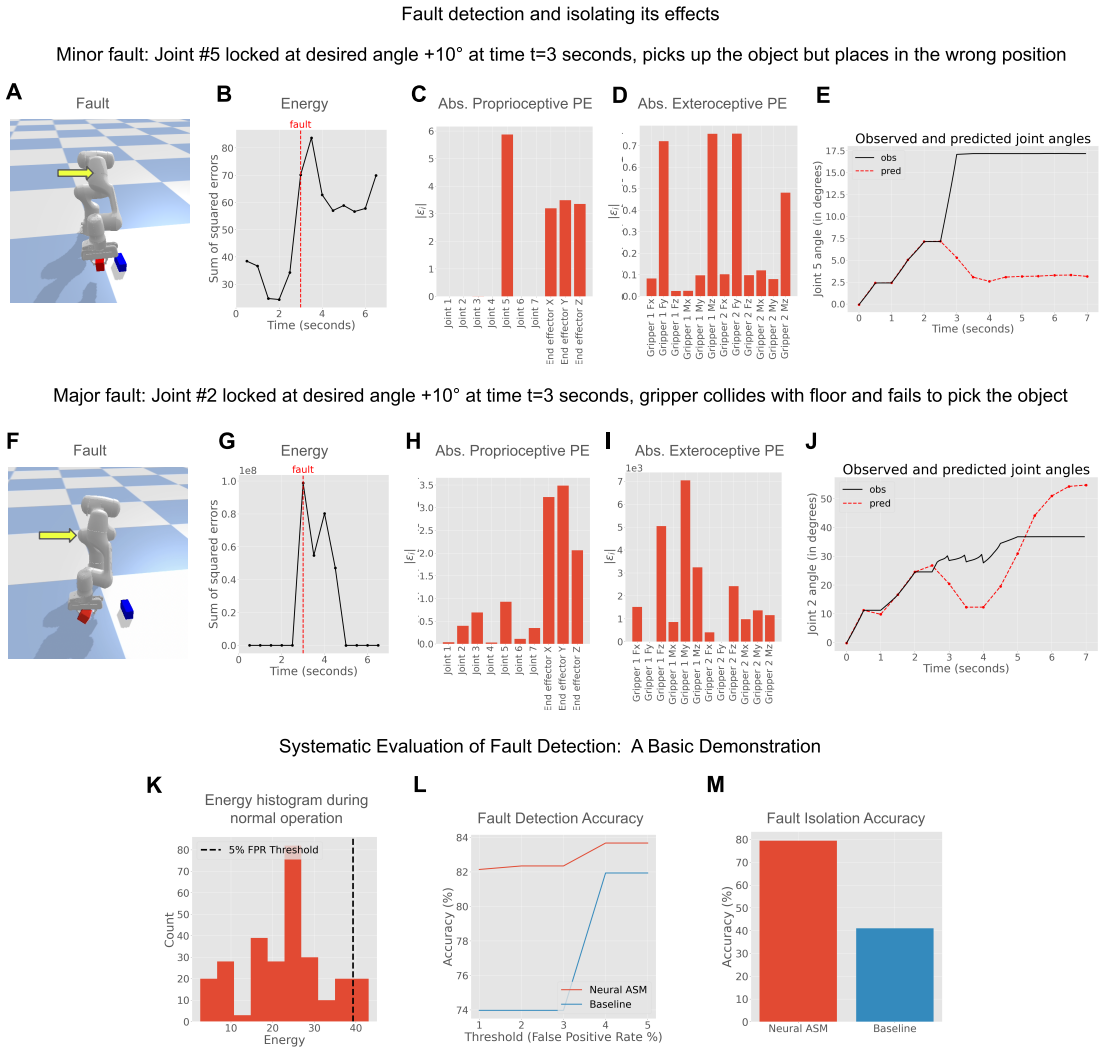


Figure 7.3: (A-E) Minor fault example with fault detection using energies and correct fault isolation using absolute prediction errors (denoted by abs. PE in figures) along with the joint angle time series. (F-J). Major fault example with fault detection using energies and incorrect fault isolation using absolute prediction errors (denoted by abs. PE in figures) along with the joint angle time series. The yellow arrows in Figure panels A and F point to the joints where the respective minor and major faults took place. The X-axis in panels C and H labels the proprioceptive sensors for joint angles (1-7, see Fig. 7.2B) and end-effector coordinates (X, Y, Z). The X-axis in panels D and I labels the exteroceptive sensors measuring X, Y and Z components of forces and torques at the two grippers. The Y-axis in panels C, D, H, and I represents the absolute prediction error (abs. PE). (K-M) A basic demonstration of the systematic evaluation of fault detection and isolation.

.

Lastly, in cases where faults can be corrected on-the-fly, Neural ASMs enable reactive correction. In ASMs, this is facilitated by DMPs, which themselves provide the reactive movement dynamics in end-effector space. In Neural ASMs, reactive

correction is modelled by minimising proprioceptive prediction errors in either end-effector or configuration (joint) space, which is used for control. We demonstrate reactive fault correction by simulating a fault caused by a falling cube colliding with the robot, leading to temporary prediction errors. By minimising proprioceptive prediction errors in configuration-space based on predicted trajectory, the low-level controller automatically corrected for this disturbance (Fig. 7.4). In this simulation experiment, almost all faults can be corrected on-the-fly unless the grip strength is too weak and the object slips out of the grip due to the collision. A more systematic evaluation will require extending Neural ASMs to real-world robots along with alternative human-like methods for fault-correction, e.g. Collins et al. (2005), please see the Discussion section for more details.

In summary, this simple setup showcases that the core aspects of ASMs can be implemented using Neural ASMs: (1) fault detection, enabling the robot to halt and seek assistance for unresolvable faults, and (2) reactive fault correction, supporting real-time adjustments for robust, fault-tolerant control.

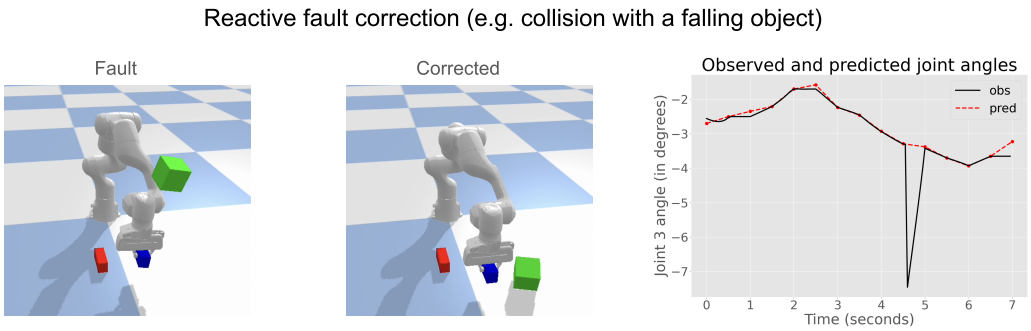


Figure 7.4: Demonstration of a fault resulting from a collision with a falling object. The fault is corrected reactively on the fly by having the low-level controller minimise proprioceptive prediction errors in joint configuration space using the proprioceptive predictions from Neural ASMs.

Contextual inference in skill memory separation and expression

Having demonstrated the utility of Neural ASMs for self-preserving robots, we now demonstrate the role of contextual inference in the separation and expression

of skill memories in our model. We utilise a robotics setup loosely inspired by Sheahan et al. (2016), who showed that motor planning of a follow-through motion, but not simply its execution, separates sensorimotor memories. However, we radically simplify the setup and eschew the optimal control that goes into arriving at the optimal trajectories, which compensate for the perturbations applied in their experiment. We rather assume that the appropriate sensorimotor sequences comprising the motor plan for each context are available in the demonstration dataset (for learning from demonstrations) and focus solely on under which conditions our model can or cannot learn and express these memories. We aim to explain and qualitatively simulate certain aspects of human motor behaviour in this robotics task. In doing so, we will also compare Neural ASM with less biologically plausible counterparts (baselines): sequence-to-sequence recurrent neural networks (RNNs) trained using backpropagation through time (BPTT). This simulation highlights the challenges shared by humans and machines that utilise neural networks to learn multiple skills, and would not usually arise if the system stored each skill independently in a library-like manner.

Our simulation is inspired by the work of Sheahan et al. (2016), who investigated how motor planning affects the separation of opposing motor skills. In their study, participants reached towards a target while counteracting one of two opposing force fields. A visual cue, available either before or during the movement, indicated which force field was active. The crucial finding was that participants only learned to separate the two motor memories, producing distinct compensatory trajectories for each field, when the contextual cue was available before the movement, allowing for motor preparation (Fig. 7.5A-B). Therefore, in 'Planning only' and 'Planning and Execution' (full follow-through) conditions, participants could separate the memories, whereas they could not do so in the execution only condition, where the cue appeared mid-movement. This highlights the critical role of contextual inference during motor preparation for separating and selecting skill memories, a principle we test with our model.

To evaluate the capabilities of our framework, we compare the Neural ASM against two standard baseline models: a sensory-to-motor (S-to-M) RNN and a sensorimotor-to-sensorimotor (SM-to-SM) RNN (Fig. 7.5D). The crucial difference between these models lies in how they process contextual information. While our Neural ASM uses an iterative inference phase to model motor preparation from early cues, the baseline RNNs are trained with backpropagation through time (BPTT) and lack this distinct inference mechanism at runtime. All models are then tested on their ability to perform offline recall, generating a full trajectory from initial cues, with the specific architectures and training procedures detailed in the Methods section.

We find that in the 'Planning only' and 'Planning and Execution' conditions, all models successfully separate the two skill memories, but fail to do so in the 'Execution only' condition, where the contextual cue is unavailable during preparation. To quantify this, we measure the Directional Compensatory Deviation (DCD), which captures the (post-exposure) maximum trajectory deviation specifically in the direction that would correctly compensate for the expected perturbation (see Methods for formal definition). As shown in Figure 7.5E, the DCD is significantly greater than zero for the planning conditions across all models, indicating successful memory separation. In contrast, the DCD is near zero for the execution-only condition, confirming the models' failure to express the correct motor plan. This result is further illustrated by the recalled trajectories in Figure 7.5F, which show distinct, opposing paths for the two contexts, qualitatively replicating the post-exposure hand paths reported by Sheahan et al. (2016). Sheahan et al. (2016) visualise their post-exposure hand path trajectories with 4 different starting points on the outside and ending in the central target, whereas we simplify when plotting our recalled motor plan in Fig. 7.5F. The same results can also be observed with alternative metrics such as mean squared error in predicted trajectories in comparison to the motor plan (please see Appendix D.3).

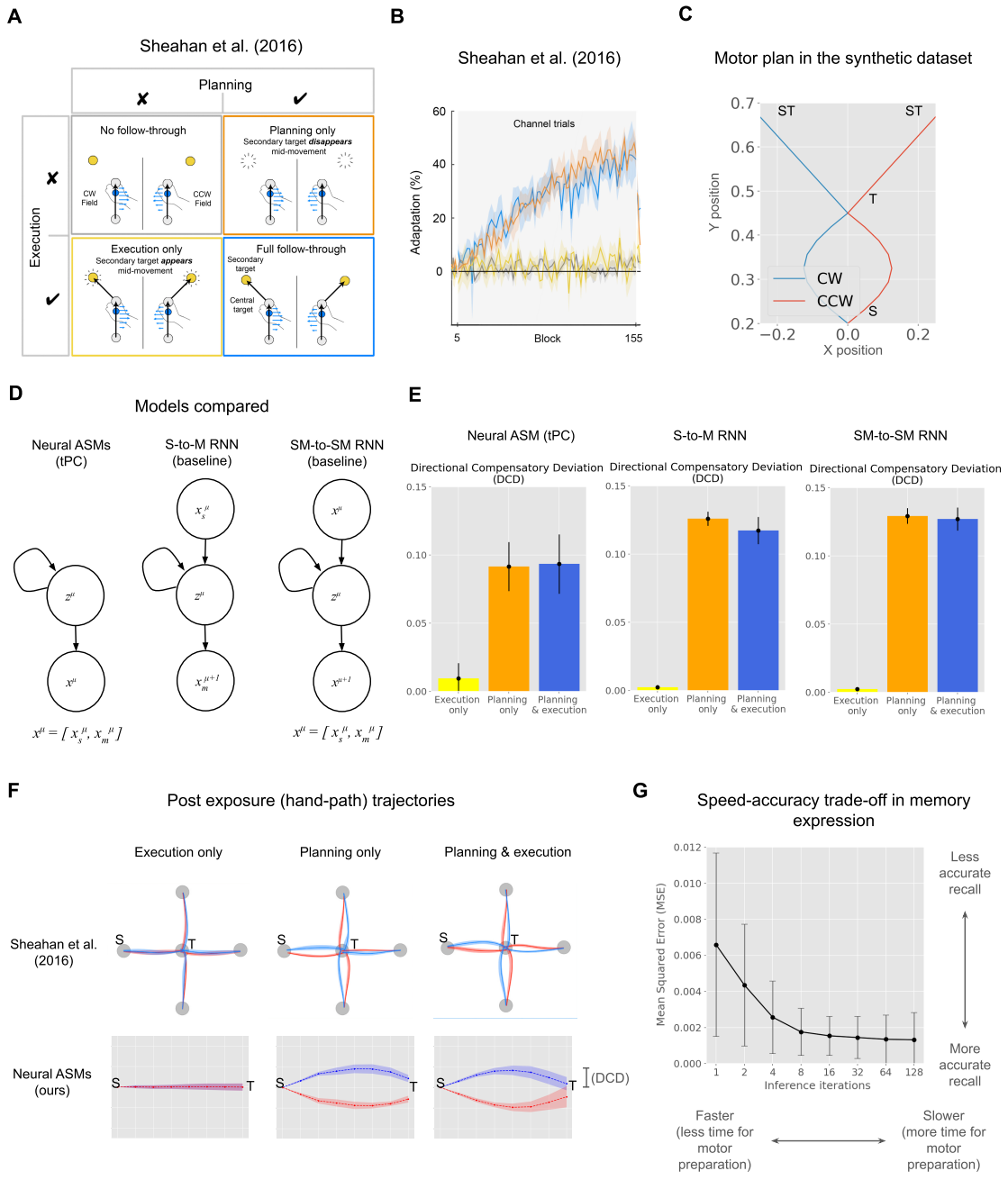


Figure 7.5: (A) Schematic from Sheahan et al. (2016) describing their experiment which inspires our robotics experiment. (Adapted under CC BY 4.0 license. Figure was cropped and a title line was added.) (B) Adaptation result from Sheahan et al. (2016)(C) Motor plan (end-effector positions) in the synthetic dataset, S: starting point, T: central target, ST: secondary target. Further details in Appendix D.2. (D) Simplified representations of our model and baseline RNNs. S-to-M RNN predicts the motor observations at the next discrete time step $(\mu+1)$ using the sensory observations at discrete time step (μ) as input. SM-to-SM predicts the sensorimotor observations at the next discrete time step $(\mu+1)$ using sensorimotor observations at the current time step (μ). (E) Results qualitatively replicated by our model, on par with baseline models. (F) Qualitative comparison of post-exposure (hand-path) trajectories (G) Speed accuracy trade-off, predicted in-memory expression. Here, demonstrated in offline skill recall.

Interestingly, our model predicts a speed-accuracy trade-off in such ballistic actions, which is not predicted by baseline RNNs (Fig. 7.5G). Unlike baseline RNNs, our model has a cued iterative inference phase akin to setting the initial condition of the hidden state from multi-sensory integration. Since we hypothesise such iterative inference of the hidden state as a potential mechanism of motor preparation, fewer inference iterations mean a shorter or constrained time of motor preparation. We find that very few iterations may be inadequate for complete hidden state inference, leading to less accurate skill recall, which improves with an increase in inference iterations. This additional timescale in our model predicts such speed-accuracy recall during skill recall or expression. A similar speed-accuracy trade-off has been observed in the expression of habitual actions in human experiments by Hardwick et al. (2019), please refer to the Discussion section for more details.

We have demonstrated that our model can capture essential aspects of the results by (Sheahan et al., 2016) on the role of contextual inference in sensorimotor memory separation and expression in our simplified robotics setup, and is qualitatively on par with RNN baselines using non-local learning rules. Further, it predicts a speed-accuracy trade-off during skill memory expression, which is not predicted by baseline RNNs.

7.6 Discussion

In this work, we introduced Neural Associative Skill Memories, a framework that leverages temporal predictive coding to learn an integrated generative model of a sensorimotor repertoire. Our results demonstrate that this approach, using biologically plausible local learning rules, can unify fault detection, reactive control, and contextual skill expression within a single network.

The model's behaviour can be understood as a form of procedural memory, capturing the automatic, stimulus-driven recall of a well-learned sequence of actions (Simor et al., 2019; Dezfouli and Balleine, 2012; Éltető et al., 2022; Mizes et al., 2023). Our main finding is that the model's energy, representing the sum of squared prediction errors, serves as a natural and effective signal for fault detection. This

aligns with the neuroscientific view of the brain as a prediction machine, where 'predictive surprise' is fundamental for adaptive behaviour. While our fault-detection performance showed only a modest improvement over a simple baseline, we argue that the key contribution is conceptual. Unlike traditional approaches that require a separate module to compare observations against the sensory statistics of an explicitly selected skill, in our framework, this capacity emerges inherently from the model's primary objective of predicting its sensorimotor stream.

Furthermore, we showed that the iterative inference process, where the model settles on a hidden state based on early cues, acts as a model of motor preparation and selects the appropriate skill memory to express. This provides a potential mechanistic account of motor preparatory activity, hypothesised to initialise a dynamical system without explicitly encoding movement parameters (Churchland et al., 2010, 2006a,b; Cisek, 2006; Fetz, 1992). Iterative inference before plasticity refers to performing multiple inference steps to settle on a stable hidden state representation of the current observation before any synaptic weight updates are made. Such iterative inference before plasticity has also recently been demonstrated to have benefits over back-propagation in some biologically plausible tasks (Song et al., 2024). This provides a computational foundation for testing whether motor preparation (in terms of iterative inference) provides benefits in skill learning.

Further, a speed-accuracy trade-off arises from the model's iterative inference mechanism, where fewer iterations (less preparation time) can lead to a less accurate initial state for recall. We note, that this trade-off does not arise in off-the-shelf recurrent neural networks. It directly links the time available for preparation (i.e., the number of inference iterations) to the accuracy of the recalled skill, a phenomenon observed in human habitual actions (Hardwick et al., 2019). When it comes to motor control, on one hand, Neural ASMs appeals to the concepts such as the equilibrium point hypothesis (Feldman and Levin, 1995) and passive motion paradigm (Mohan et al., 2019), aligning with the free energy principle (Friston et al., 2010; Friston, 2010). Whilst, on the other hand, Neural ASMs are also compatible with views by Schaal et al. (2007), where the dynamical systems

policy sits atop the optimal control system, thus providing a possible unification of both approaches to motor control (Appendix D.1).

Limitations

Our study serves as a proof of concept, and several limitations should be acknowledged. The experiments used a simplified robotics setup with a limited skill repertoire and basic evaluation metrics; future work should test the scalability, generalisation capabilities and robustness of Neural ASMs in more complex, dynamic environments. To simplify the demonstration, we simulate a limited repertoire of two pick-and-place skills, although the framework supports more. We also find that the rate of learning depends on the number of skills, consistent with observations in Howard et al. (2015) (see Appendix D.4). The reliance on learning from demonstrations means the model is not learning optimal control strategies itself, a key distinction from reinforcement learning approaches.

Furthermore, our experiment, inspired by Sheahan et al. (2016) is not an exact replica. We omit their "no follow-through" condition based on findings that no adaptation occurs from static cues alone (Gandolfo et al., 1996; Howard et al., 2012, 2013, 2015; Sheahan et al., 2016). The inability to learn to counteract perturbations from static cues is a limitation of the learning-from-demonstration paradigm itself, where optimal trajectories are provided, rather than our specific model. Nonetheless, this does not affect our central finding regarding how contextual inference supports the separation and expression of already learned skills.

On a framework level, further limitations exist. First, the model, under offline recall, relies on early-stage cues for contextual inference and cannot adapt to goal changes mid-trajectory. Online recall could be used for re-recognition or switching between alternative skills in the repertoire. This would involve inferring the hidden states using the sensorimotor observations at each time step, allowing the model to continuously infer the skill it is performing and then predict the sensorimotor observations for the next time step. However, it would not be able to perform novel goal-directed planning to change goals or to achieve certain preferred sensory

observations. Second, our model concatenates sensorimotor channels into a single input vector, whereas a more neurally plausible architecture might involve separate, interacting pathways for different modalities. Contextual selection of the correct skill would require motor and sensory observations to be connected to a shared hidden state. Third, the capacity of the tPC network is constrained by its local-in-time learning rule, which restricts its ability to learn complex, long-horizon sequences. Finally, all Associative Skill Memory models (Pastor et al., 2012, 2013) assume that skill-related movements are stereotyped; high sensory variability would make the learned predictions unreliable.

Future Work

The limitations discussed above highlight several promising directions for future research. A critical next step is to test the scalability and robustness of the Neural ASM framework in more complex, dynamic environments with a larger repertoire of skills. For robotic applications, improving reactive controllers, for instance by incorporating principles from passive dynamic machines (Collins et al., 2005), would be a valuable step towards more energy-efficient and human-like reactive behaviours.

At a systemic level, incorporating goal-directed planning would require disentangling sensory and motor modalities. The concatenated input structure, similar to (Nishimoto et al., 2008; Yamashita and Tani, 2008), would need to be replaced by a more neurally plausible architecture with separate, interacting pathways for different modalities. Future work can aim to integrate ideas on goal-directed planning to achieve a preferred sensory state similar to (Matsumoto et al., 2022). One could potentially use a sensory tPC network for inferring the internal states from sensory observations and a "model inversion" of a motor tPC network to implement goal-directed motor planning, exploiting the duality of Bayesian inference and optimal control (Doya, 2021). This would allow for more flexible behaviour, such as adapting to mid-trajectory goal changes, and is an avenue for future work. Additionally, enabling Neural ASMs to track posterior variance, e.g., via Monte Carlo Predictive Coding (Oliviers et al., 2024) would make them viable as planning-capable world

models. To overcome the limited temporal credit assignment of the current tPC network, one could investigate multi-timescale tPC architectures (Yamashita and Tani, 2008) or alternative biologically plausible approximations to BPTT using memory traces, such as e-prop (Bellec et al., 2020).

Conclusion

In conclusion, this work presents a biologically inspired framework for sensorimotor learning in robots which unifies skill learning, fault detection, and contextual expression within a single predictive coding network. Neural ASMs further offer a plausible approximation of how associative sensorimotor memories might be implemented in the brain using local learning rules. The conceptual contribution of this work provides a step towards creating more adaptive, self-preserving robots and offers a tractable computational testbed for exploring theories of biological motor control.

7.7 Methods

Model algorithm

The sequential memory in Neural ASMs is modelled using a temporal Predictive Coding (tPC) network (Millidge et al., 2024b; Tang et al., 2024b). Predictive coding models learn in a self-supervised fashion with the aim of best predicting the incoming input based on its own learned generative model. The model evaluates the actual input against its prediction by determining the difference in activity in the respective error neurons and minimises these 'errors' through adjusting neural activities and synaptic weights, which corresponds to the processes of inference and learning, respectively (Clark, 2015; Bogacz, 2017).

In mathematical terms, the task of neural ASMs can be seen as learning a sequence of sensorimotor observations (such as motor coordinates and associated sensory events like haptic feedback, etc) $(\mathbf{x}^t)_{t=0}^T$. This can be reduced to learning the dynamics in these sensorimotor observations for each skill, i.e. learning to associate each $\mathbf{x}^\mu$ with the next $\mathbf{x}^{\mu+1}$ $(\mu = 0, 1, ..., T-1)$. We use a 2-layer tPC model, whose

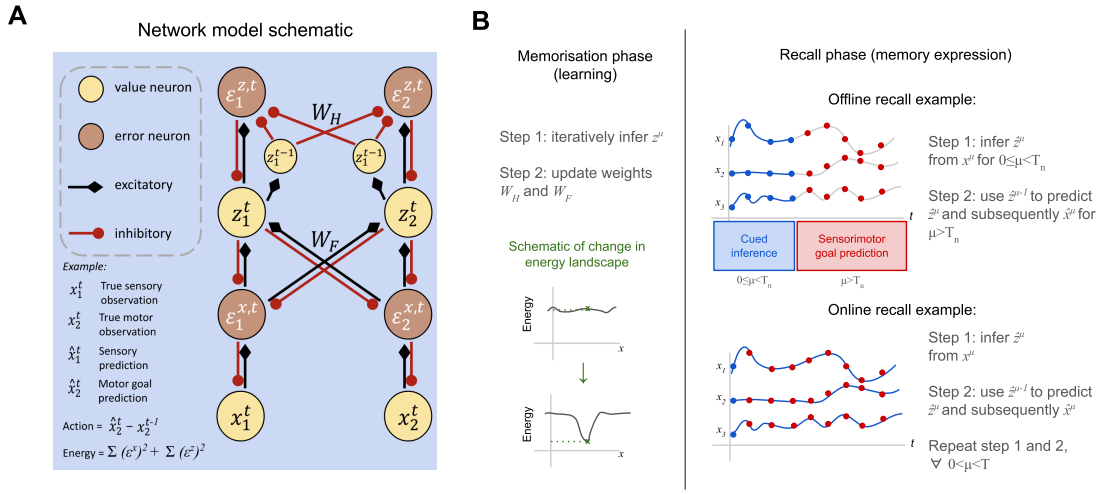


Figure 7.6: (A) Neural network implementations of temporal predictive coding (tPC) model used in Results for simulations. The network is illustrated with a single sensory and motor observation input for didactic purposes (though in reality, the network will have multiple sensorimotor observations as inputs). (B) The model learns in the memorisation phase in a self-supervised manner, and the weight updates change the energy landscape to store these memories as attractors in the energy landscape (which is crucial in recognising memorised skills). During the recall phase, the learned memories are expressed and can be either done offline (e.g. ballistic actions) or in an online manner where ground-truth observations are utilised to provide online feedback through model inversion at each step.

underlying graphical structure is that of a Hidden Markov Model (HMM). The lower layer of the tPC predicts the sensorimotor observations ($\mathbf{x}^\mu$) and the upper layer predicts the next hidden state ($\mathbf{z}^{\mu+1}$). This predictive processing account loosely models the hierarchical processing of raw sensory inputs by the neocortex, where hidden value neurons $\mathbf{z}^\mu$ models the brain's internal neural responses to the sequential sensory inputs $\mathbf{x}^\mu$ (Fig. 7.6A).

The working of the tPC model in neural ASMs can be divided into two stages: (1) memorisation and (2) recall (Fig. 7.6B). During memorisation, tPC tries to minimise the sum of squared errors at step μ, with respect to the weights and the hidden activities:

$$F_\mu(\mathbf{z}^\mu, \mathbf{W}_H, \mathbf{W}_F) = \|\mathbf{z}^\mu - \mathbf{W}_H f(\hat{\mathbf{z}}^{\mu-1})\|_2^2 + \|\mathbf{x}^\mu - \mathbf{W}_F f(\mathbf{z}^\mu)\|_2^2 \tag{7.1}$$

where $\mathbf{W}_H$ governs the temporal prediction in the hidden state, $\mathbf{W}_F$ are the weights governing predictions from $\mathbf{z}^\mu$ to $\mathbf{x}^\mu$, with $\hat{\mathbf{z}}^{\mu-1}$ being the hidden state inferred at

the previous time-step. During memorisation, the model first infers the hidden representation of the current sensorimotor observational input $\mathbf{x}^\mu$ by:

$$\dot{\mathbf{z}}^\mu \propto -\frac{\partial F_\mu(\mathbf{z}^\mu, \mathbf{W}_H, \mathbf{W}_F)}{\partial \mathbf{z}^\mu} = -\varepsilon^{\mathbf{z},\mu} + f'(\mathbf{z}^\mu) \odot \mathbf{W}_F^\top \varepsilon^{\mathbf{x},\mu} \tag{7.2}$$

where $\odot$ denotes the element-wise product between two vectors, and $\varepsilon^{\mathbf{z},\mu}$ and $\varepsilon^{\mathbf{x},\mu}$ are defined as the hidden temporal prediction error $\mathbf{z}^\mu - \mathbf{W}_H f(\hat{\mathbf{z}}^{\mu-1})$ and the top-down error $\mathbf{x}^\mu - \mathbf{W}_F f(\mathbf{z}^\mu)$, respectively. After $\mathbf{z}^\mu$ converges, $\mathbf{W}_H$ and $\mathbf{W}_F$ are updated following gradient descent on F_μ:

$$\Delta \mathbf{W}_H \propto -\frac{\partial F_\mu(\mathbf{z}^\mu, \mathbf{W}_H, \mathbf{W}_F)}{\partial \mathbf{W}_H} = \varepsilon^{\mathbf{z},\mu} f(\hat{\mathbf{z}}^{\mu-1})^\top \tag{7.3}$$

$$\Delta \mathbf{W}_F \propto -\frac{\partial F_\mu(\mathbf{z}^\mu, \mathbf{W}_H, \mathbf{W}_F)}{\partial \mathbf{W}_F} = \varepsilon^{\mathbf{x},\mu} f(\mathbf{z}^\mu)^\top \tag{7.4}$$

which are performed once for every presentation of the full sequence. Importantly, the converged $\mathbf{z}^\mu$ is then used as $\hat{\mathbf{z}}^\mu$ for the memorisation at time-step $\mu + 1$.

After memorisation (learning) is completed, the model enters the recall stage where all weights no longer change and the previously learned memories are expressed in response to certain cued input observations (also referred to as queries q). Note that during the recall phase, the observation layer has no access to the correct patterns for the complete movement. Instead, it needs to dynamically change its value to retrieve the memories to predict these sensorimotor observations. The sequential memories are recalled or expressed using the learned weights $\mathbf{W}_H$ and $\mathbf{W}_F$. The loss thus becomes:

$$F_\mu(\mathbf{z}^\mu, \hat{\mathbf{x}}^\mu) = \|\mathbf{z}^\mu - \mathbf{W}_H f(\hat{\mathbf{z}}^{\mu-1})\|_2^2 + \|\hat{\mathbf{x}}^\mu - \mathbf{W}_F f(\mathbf{z}^\mu)\|_2^2 \tag{7.5}$$

where $\hat{\mathbf{x}}^\mu$ denotes the activities of value neurons in the observation layer during recall. Both the hidden and observation layer value neurons are updated to minimise the loss. The hidden neurons will follow similar dynamics specified in Eq. 7.2, whereas the observation layer neurons are updated according to:

$$\dot{\hat{x}}^\mu \propto -\frac{\partial F_\mu(\mathbf{z}^\mu, \hat{\mathbf{x}}^\mu)}{\partial \hat{\mathbf{x}}^\mu} = -\varepsilon^{\mathbf{x},\mu} \tag{7.6}$$

and the converged $\hat{\mathbf{x}}^\mu$ is the final retrieval.

In the case of sequential memory, there are two types of recall, offline and online. In case of offline recall, first $\hat{\mathbf{z}}^{\mu}$ is iteratively inferred from T_n cued input ground-truth observations or queries $q = \mathbf{x}^{\mu}$, where $\mu = (0, 1, ..., T_n)$ (here, $0 \leq T_n < T$). Once $\hat{\mathbf{z}}^{\mu}$ is converged, it is used to recall $\hat{\mathbf{x}}^{\mu}$ for $\mu > T_n$. In case of online recall, we query the model with $q = \mathbf{x}^{\mu}$ (ground-truth), use the query to infer $\hat{\mathbf{z}}^{\mu}$, and then use $\hat{\mathbf{z}}^{\mu}$ for the recall the next time step $\mu + 1$ for $\mu = 0, 1, ..., T - 1$. This distinction is important in our results; here, we only present offline recall results for skill memory expression in ballistic movements.

Network training details

The network details for the tPC network are as follows: The number of hidden units was 256, and the number of sensorimotor observation units depended on the task. The learning rate for the weight updates was 10^{-4} and the default learning iterations were 1000 per skill, trained with a batch size of 1. The iterative inference learning rate for hidden state update was 10^{-2} with default inference iterations set to 100. The same hyperparameters were used for all simulation experiments. Kaiming uniform initialisation was used in hidden layers for all networks.

The S-to-M RNN and SM-to-SM RNN are sequence-to-sequence RNNs. They used the same number of hidden units, learning rate, learning iterations and batch size at the tPC network. They did not have iterative inference functionality, like the tPC network. The input unit size is the number of sensory observations, and the output unit size is the number of motor observations. The skill memory expression experiment trains the RNNs to predict the output sequence using inputs only from the first time step to mimic the offline recall used in tPC.

The sensorimotor sequences were Z-score normalised for each observation channel before being provided as inputs to all neural networks. Outputs were again unnormalised to original units before movement.

Experimental setup and robot simulation details

Task 1: Fault Detection and Reactive Correction

In our simulations, the Neural ASM operates at 2 Hz, guiding an underlying low-level position controller operating at 40 Hz. The demonstration dataset was generated in simulation by providing end-effector goals, from which joint angles were calculated using inverse kinematics and sensory observations were recorded. This process acts as a proxy for teleoperation. The experiment used 10 repetitions for each of two distinct pick-and-place skills. Each repetition is a sensorimotor sequence of 15 time steps (at 2 Hz), involving 25 observations (including desired joint angles, end-effector positions, gripper states, and sensed forces and torques), with realistic noise from the simulation process.

The minor fault is simulated as follows: the joint 5 overshoots the desired goal by 10° while attempting to pick up the object ($t = 3$ seconds) and gets stuck in that configuration for the remaining duration (Fig. 7.3A). The major fault is simulated as follows: the joint 2 overshoots the desired goal by 10° and gets stuck while attempting to pick up the object, which results in the arm colliding with the floor (Fig. 7.3F).

For the systematic evaluation of fault detection, we use percentile-based thresholding of the model's energy distribution during normal operation to establish a detection threshold, as illustrated in Fig. 7.3K. This method allows us to control for the false positive rate (FPR); for instance, a 95th percentile threshold corresponds to a 5% FPR. We then measure accuracy on 980 simulated fault trials, created by systematically varying the locked joint (1 to 7), the time of fault (1 to 6 seconds in 0.5s steps), and the degree of joint angle overshoot (-15° to +15° in 5° steps). Fault isolation accuracy is measured as the proportion of trials where the joint with the highest absolute prediction error matches the simulated fault location.

We compare our model's performance against a simple baseline analogous to traditional ASMs (Pastor et al., 2012), which relies on stored signal statistics. This baseline computes normalised errors for each observation channel using Z-score normalisation ($\epsilon = (x_i^t - \bar{x}_i)/\sigma_i$). The sum of squared normalised errors is then used for fault detection, and the channel with the maximum absolute normalised

error is used for fault isolation. We acknowledge that this represents one possible implementation of a statistics-based approach and that other methods may exist, and a comprehensive comparison with alternate methods is left for future work.

Task 2: Contextual Skill Expression

Inspired by the experimental paradigm of Sheahan et al. (2016), we designed a simulation to test the model's ability to separate skill memories based on contextual cues. It is important to note that our goal was not to model the learning of optimal trajectories to counteract perturbations, but rather to assess if the model could learn and express distinct, pre-defined motor plans from demonstrations under different contextual conditions. The synthetic dataset, therefore, provides these optimal compensatory trajectories directly. For each skill, the data includes end-effector positions, joint angles calculated via inverse kinematics, and a one-hot coded visual cue representing the context (see Appendix D.2 for plots). The sensory and motor sequences used in the skill memory expression are presented in Appendix D.2. We did not model the "no follow-through" condition from the original study, for reasons explained in the Discussion.

We compared our Neural ASM against two baseline recurrent neural network (RNN) models (Fig. 7.5D): a sensory-to-motor (S-to-M) RNN and a sensorimotor-to-sensorimotor (SM-to-SM) RNN, a discrete-time variant of the model used in Nishimoto et al. (2008). Unlike our tPC-based model, these baselines are trained using backpropagation through time (BPTT) and lack an iterative inference phase. To test skill expression from preparation, all models were evaluated using an offline recall procedure, generating the full trajectory based only on the inputs from the first time step. To parallel the human experiment, we trained all networks on each condition using 6 different random seeds for 1200 learning trials, and post-training recall was averaged over 24 trajectories.

To quantify memory separation, we introduce the Directional Compensatory Deviation (DCD) metric, which, unlike a simple absolute deviation, accounts for the direction of movement. DCD is defined as the maximum deviation of the

recalled end-effector trajectory perpendicular to the straight line connecting the start (S) and target (T) points, projected onto the axis of correct compensation for a given context (e.g., leftward for a clockwise field). A positive DCD value indicates that the trajectory deviates in the appropriate direction to counteract the expected force field, while a value near zero indicates a failure to express the correct motor memory. This metric serves as a direct proxy for the directional adaptation measured by Sheahan et al. (2016).

The best way to predict the future is to invent it.

— Alan Kay

8 Discussion

Contents

8.1 NeuroAI: A marriage between neuroscience and AI

Key motivations in NeuroAI research are twofold: utilising advancements from AI to generate useful hypotheses for neuroscience, and leveraging insights from neuroscience to develop more effective algorithms for AI. One could poetically say this discipline represents a marriage between neuroscience and AI. This thesis contributes to this growing NeuroAI field in the context of pain neuroscience and safe AI. At the beginning of the thesis, we saw a categorisation of main contributions into either safe exploration using multi-objective RL or self-preservation, leveraging POMDPs. Here, we present an alternative categorisation of the thesis contributions to the NeuroAI marriage.

British marriages often require four things: something old, something new, something borrowed and something blue. In this thesis, we find them as follows:

- **Something old:** The conflict between Pavlovian reflexes and instrumental control has been a foundational topic in learning theory since the late 1960s (Williams and Williams, 1969). Chapter 4 revisits this 'old' problem in the aversive domain, but reframes it through a normative lens. We argue that the Pavlovian fear system is not merely an evolutionary vestige that causes 'misbehaviour', but a functional adaptation that confers a critical safety advantage during exploration. The thesis demonstrates that this system can operate efficiently, avoiding excessive costs/penalties whilst seeking rewards, as long as its influence is carefully titrated by outcome uncertainty. Our human virtual reality experiment provided empirical support for this model, showing that a flexible, uncertainty-gated Pavlovian system best explained participants' choices and reaction times in a virtual reality approach-withdrawal task, building upon foundational Go-No Go tasks (Guitart-Masip et al., 2012; Cavanagh et al., 2013) in the literature.

- **Something new:** The story of dopamine has evolved over three decades, from a simple scalar temporal difference-reward prediction error (TD-RPE) signal to a representation of multifaceted learning signals. Chapter 5 presents a 'new' chapter in this story by providing a normative solution to the problem of optimally composing multiple, parallel value functions. By redefining the dopaminergic system's objective to optimise returns augmented by a policy deviation penalty, our framework, built on linear RL (Todorov, 2009b; Piray and Daw, 2021), achieves what previous models could not: a learning method that is simultaneously efficient, safe, stable, and optimally composable. This single theoretical shift parsimoniously unifies a range of disparate observations, including the flexible expression of innate safety priors and the reconciliation of threat and action prediction errors in the tail of the striatum (Akiti et al., 2022; Greenstreet et al., 2025).

- **Something borrowed:** In Chapter 7, we turned to the challenge of learning a generative model of the self for neurorobotics. For this, we 'borrowed' the

architecture of temporal predictive coding (Millidge et al., 2024b; Tang et al., 2024b), a framework gaining traction for its biological plausibility and reliance on local learning rules. We built upon this borrowed foundation to advance the concept of Associative Skill Memories (ASMs) (Pastor et al., 2012, 2013). The resulting Neural ASM overcomes the key limitation of traditional, library-based ASMs by allowing a single, unified network to learn a full repertoire of sensorimotor skills, enabling fault detection across all behaviours without requiring explicit skill selection.

- **Something blue:** Chapter 6 delves into the 'blue' mood of pathology. The POMDP model of homeostasis after injury provides a formal account of how an adaptive system can go awry, leading to dysfunctional consequences. Specifically, it demonstrates how information restriction (Seymour et al., 2023b), often a consequence of protective avoidance behaviours, can lead to aberrant, persistent beliefs about the body's state, providing a quantitative mechanism for the transition from acute to chronic pain. This work embeds these ideas within the field of computational psychiatry (Huys et al., 2015), characterising this pathological state as an agent either solving the wrong problem, or the right problem with the wrong solution or using the right solution in an unfortunate (wrong) environment.

8.2 Significance

The contributions of this thesis are significant in two main areas: advancing our computational understanding of pain and its transition to chronicity, and contributing to the development of safer, more robust artificial intelligence.

Ameliorating chronic pain

This work frames pain-related behaviour as a sophisticated control problem involving the interaction of multiple systems, the composition of competing objectives,

and inference under uncertainty. These perspectives offer distinct, clinically relevant insights.

First, the model of Pavlovian-instrumental arbitration in Chapter 4 suggests a novel target for intervention. Rather than focusing on extinguishing fear through exposure, a key prediction is that therapeutic efforts should aim to make the arbitration process more flexible by targeting the patient's sense of uncontrollability. This could be achieved through a controllability discrimination paradigm, helping patients distinguish between what they can and cannot control—an approach that aligns with principles of stoicism-based cognitive behavioural therapy (CBT) (Turk and Rudy, 1992; Thorn and Dixon, 2007).

Second, the POMDP model of injury in Chapter 6 provides a formal, mechanistic complement to the influential Fear-Avoidance model. It suggests that chronic pain can arise from faulty inference driven by information restriction. This has direct therapeutic implications, suggesting that interventions providing safe, unambiguous information about the body's recovered state could be highly effective. This could take the form of a guided activity-pacing app that uses wearable data to provide an unbiased estimate of the body's state, helping patients avoid maladaptive 'boom-bust' activity cycles (Moseley, 2003; Antcliff et al., 2016). Furthermore, this framework offers a potential mechanism for explaining the success of treatments like Pain Reprocessing Therapy (Gordon and Ziv, 2021), where techniques such as somatic tracking and guided touch may function to provide the very evidence needed to update a patient's aberrant prior beliefs about their body (Kim et al., 2022; McParlin et al., 2022).

Third, the work on sensorimotor learning in Chapter 7 highlights the critical role of context. Following an injury, individuals often develop new, protective movement patterns that become ingrained as part of their motor repertoire. These maladaptive memories can be highly context-sensitive and resistant to extinction if therapy is conducted in a different context. Our model suggests that therapeutic success may depend on first eliciting the specific post-injury context to ensure that

these sensorimotor memories are properly overwritten, an insight that aligns with computational theories of memory modification (Gershman et al., 2017).

Towards neuro-inspired safer and robust AI

This thesis argues that building truly autonomous agents requires a multi-faceted approach to safety that mirrors the solutions found in biology. This work contributes to a blueprint for such an approach by outlining three key principles. Safe agents should be able to: (1) manage the conflict between fast, reflexive systems and slower, deliberative ones (Chapter 4); (2) possess a principled method for optimally and stably composing multiple, potentially conflicting, goals (Chapter 5); and (3) preferably maintain an internal, generative model of the self to detect faults and maintain integrity (Chapter 7).

The principle of optimal composition is particularly relevant to current challenges in AI safety. Decomposing a complex objective into multiple, simpler value functions can create a more interpretable and controllable system, providing distinct 'control knobs' for different safety and performance criteria. This approach has recently been applied to the alignment of large language models, where compositional preference models were shown to improve generalisation and prevent reward-hacking to a greater degree than monolithic preference models (Go et al., 2023). Together, these principles argue for a more holistic, neuro-inspired approach to building safe and robust AI.

8.3 Limitations and Future Work

This thesis provides a theoretical and computational foundation for safe learning, but each component has limitations that point towards clear avenues for future research.

The work on Pavlovian-instrumental arbitration in Chapter 4 has a few limitations. The experimental design only considered a decrease in uncertainty from the first half to the second; a more stringent test would involve a balanced design that also examines the effects of increasing uncertainty and further includes conditions to dissociate it from volatility. Furthermore, the model did not account for endogenous

pain modulation, such as stress-induced analgesia, which could counteract the uncertainty-gated fear commissioning we propose. On a broader level, while this work establishes a normative role for Pavlovian withdrawal, the computational underpinnings of more sophisticated survival behaviours, such as escape, remain largely unexplored (Evans et al., 2019). Future work should investigate how compositional, seemingly model-based escape trajectories are generated under extreme time pressure, and explore the recursive theory-of-mind dynamics inherent in prey-predator interactions.

The framework for optimal composition in Chapter 5 is primarily theoretical, and its critical next step is empirical validation. Targeted experiments could test its core predictions, for instance, by using tasks with shifting reward priorities to look for the neural signatures of stable, off-policy value representation. Another key direction is to manipulate the predictability of action outcomes to dissociate the hypothesised threat prediction error (TPE) and action prediction error (APE) components of the dopamine signal. These hypotheses are bolstered by recent findings suggesting that threat and motor-related dopamine responses are encoded in genetically distinct subpopulations projecting to the TS (Azcorra et al., 2023; La Manno et al., 2016; Poulin et al., 2014, 2018; Greenstreet et al., 2025), providing a clear biological target for these investigations.

The POMDP model of injury in Chapter 6 represents a minimal, first-step formalisation of a complex biological process. A key avenue for future work is to enrich the model's dynamics by incorporating internal state transitions corresponding to recovery or exacerbation. This would allow for a more nuanced representation of activity, relating the model to theories of activity pacing and vigour (Niv et al., 2005, 2007). Furthermore, the utility terms, which were set arbitrarily in the current work, should be derived from first principles, incorporating both physiological costs and the opportunity costs of inaction (a 'Resource-Trauma POMDP'). The most critical future direction is empirical validation; the model needs to be tested against concrete neural and behavioural data from conditions such as ACL tear or sciatica recovery to refine its structure and parameters, including

the modelling construct of tonic pain. The neural substrates of such homeostatic and allostatic control may be the hypothalamus and parabrachial hub, orchestrated by neuropeptides - a direction for future work.

Finally, the Neural ASM in Chapter 7 was limited to learning reactive control from demonstrations and did not incorporate goal-directed planning. A major direction for future work is to integrate this self-monitoring system with a goal-directed decision-making architecture, such as the optimal composition framework from Chapter 5. Testing this integrated system on a physical robotic platform would be a crucial step towards creating agents that can not only perform tasks but also understand when they are 'injured' and adapt their plans accordingly.

Ultimately, the work in this thesis addresses the question of how an agent can survive. A broader, more ambitious goal for the future is to understand how an agent can thrive. This would require moving beyond the current models to embrace the challenge of continual learning, exploring how the brain adapts and acquires new skills and knowledge throughout a lifetime, in both function and dysfunction.

Appendices

A

Appendix for Chapter 4

Contents

A.1 Robustness of the associability-based ω in gridworld simulations

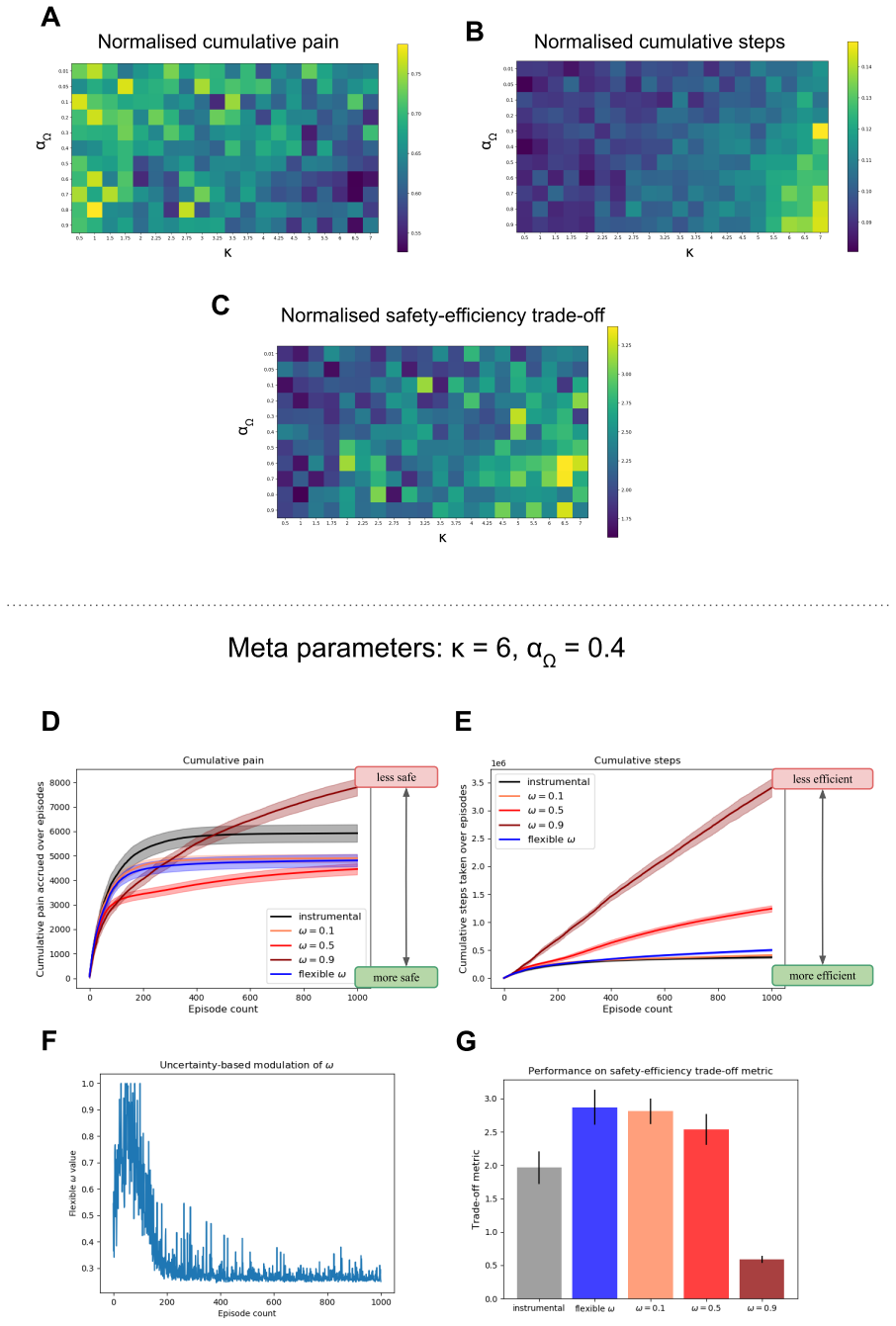


Figure A.1: This figure shows the robustness of grid search for tuning the meta parameters for the associability-based ω in grid world simulations. We show that the results hold for a range of values close to the chosen meta-parameters. (A-C) Grid search results for the environment in Fig. 4.2 for varying κ and α_Ω. (D-G) Results for another set of meta-parameters.

A.2 Flexible ω agent better adapts to reward relocation than a fixed ω agent.

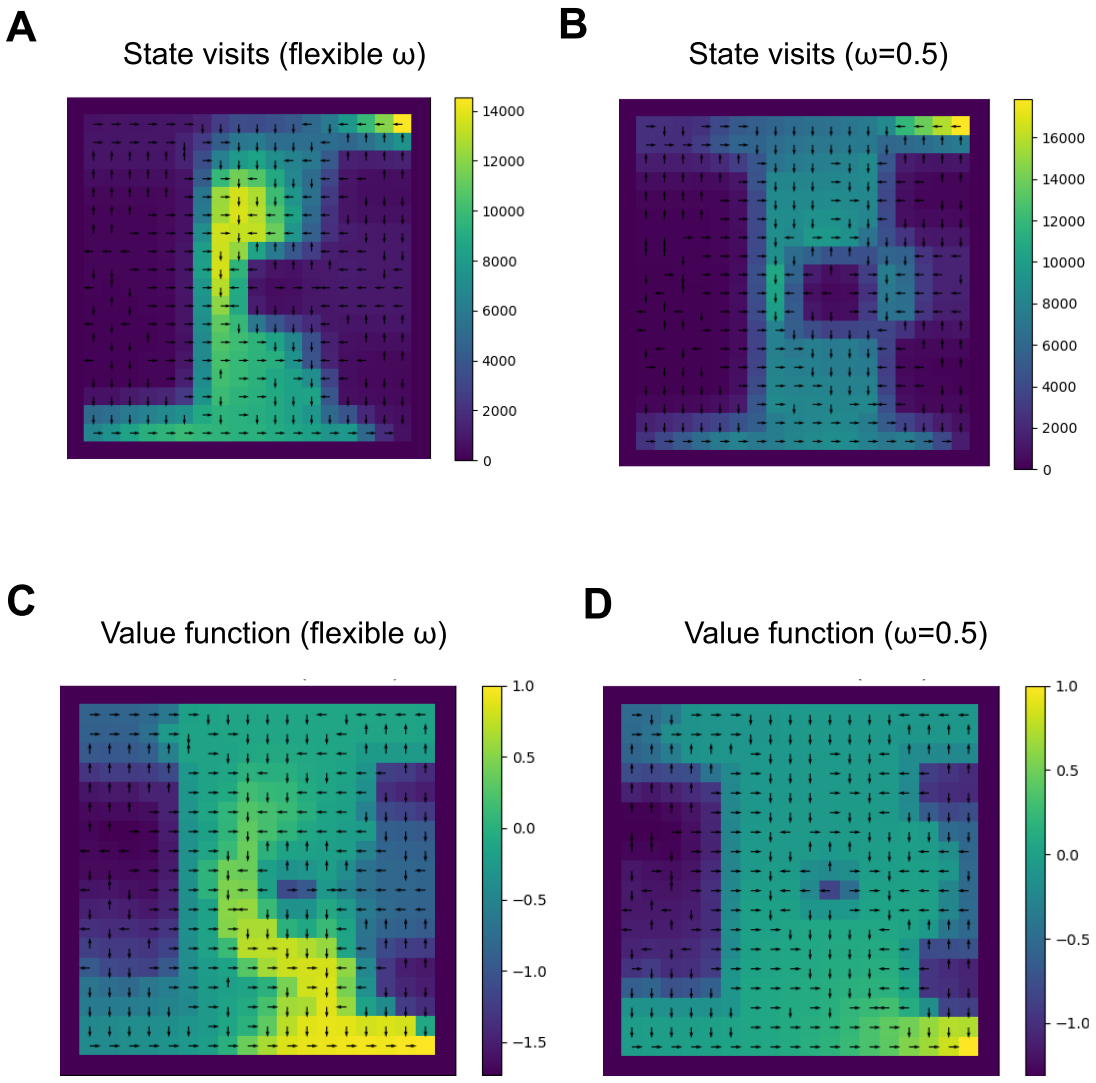


Figure A.2: This figure shows cumulative state visit plots and value function plots of the flexible ω and fixed ω agents at the end of 1000 episodes when we relocate the reward goal from the bottom left corner (Fig. 4.2) to the bottom right corner on episode 500. Comparing state visit plots A & B and comparing value function plots C & D, we observe that persistent Pavlovian influence leads to persistent rigidity while the flexible fear commissioning scheme is able to efficiently locate the goal. We observe that unlike flexible ω, constant $\omega = 0.5$ leads to diminished value propagation of the rewarding value (C & D).

A.3 Solving the safety-efficiency trade-off in a range of grid world environments

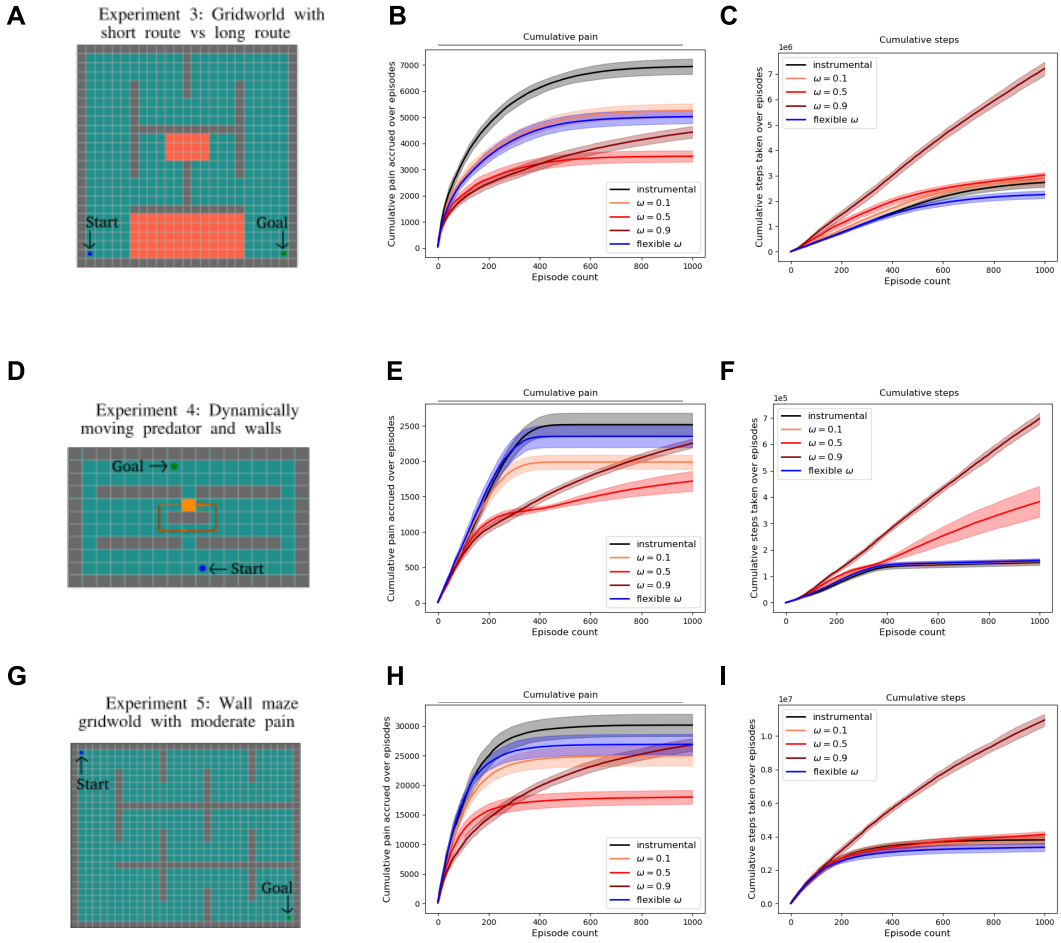


Figure A.3: In this figure, we show the performance of fixed $\omega = 0.1, 0.5, 0.9$ and flexible ω agents on a range of grid world environments, namely (A) the three-route environment from Fig. 4.3, (D) an environment with a moving predator on routine path and (G) wall maze grid world from Elfwing and Seymour (2017). Colliding with the predator results in a negative reward of -1 and catastrophic death (episode terminates). Otherwise, colliding with the walls results in moderate pain of 0.1, and the agent's state remains unchanged. The latter two are completely deterministic environments unlike the previous environments in the main paper. We show the safety-efficiency trade-off arises in these three environments as well and, there is a separate optimal fixed ω for each environment. Alternatively, there exists a flexible ω scheme for each environment that can solve the trade-off, suggesting that the brain may be calibrating ω flexibly.

A.4 Human three-route virtual reality maze results

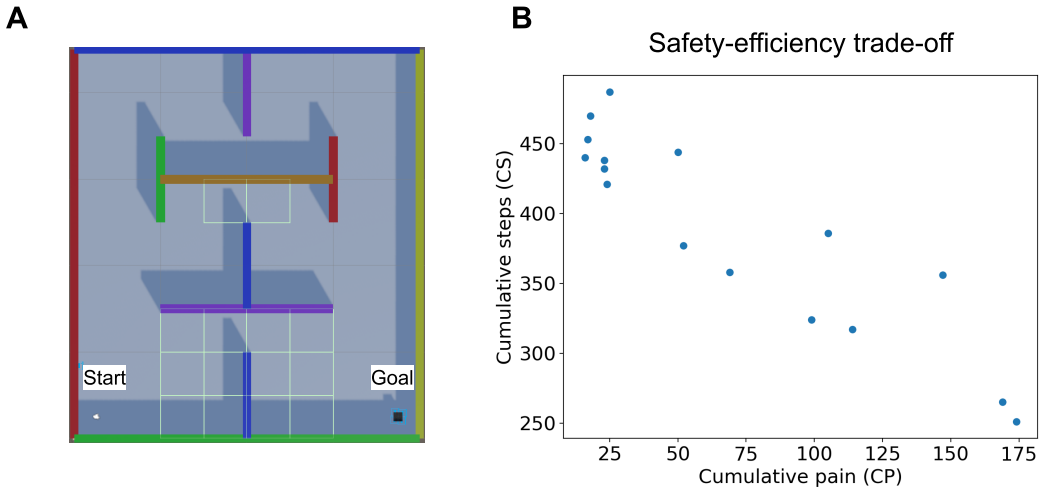


Figure A.4: (A) Top-view of virtual reality (VR) maze with painful regions annotated by highlighted borders (B) Cumulative steps required to reach the goal vs cumulative pain acquired by participants over 20 episodes in the VR maze task. In this figure, we show the results of a VR maze replicating the three-route grid world environment from simulation results, however, it had fewer states and the participants were instructed to reach the goal which was visible to them as a black cube with "GOAL" written on it. In order to move inside the maze participants had to physically rotate in the direction they wanted to move and then press a button on the joystick to move forward in the virtual space. Thus the participant did not actually walk in the physical space but did rotate up to 360 degrees in physical space. The painful regions were not known to the participants but they were aware that some regions of the maze may give them painful shocks with some unknown probability. Walking over the painful states in the VR maze, demarcated by grid borders (see A) in the potentially shocked them with 75% probability while ensuring 2 seconds of pain-free interval between two consecutive shocks. Participants were not given shocks with 100% probability as that would be too painful for participants due to the temporal summation effects of pain. The participants engaged in 20 episodes of trials and were aware of this before starting the task and were free to withdraw from the experiment at any point. 16 participants (11 female, average age 30.25 years) were recruited and were compensated adequately for their time. The pain tolerance was acquired similarly to the Approach-Withdrawal task. (B) All participant trajectories inside the maze were discretized into an 8x9 (horizontal x vertical) grid. Entering a 1x1 grid section counted incremented the cumulative steps (CS) count. Upon receiving the shocks, the cumulative pain (CP) count was incremented. CP and CS over 20 episodes were plotted against each other to observe the trade-off. A limitation of this experiment is that it reflects the constraints of the grid world and future experiments are necessary to show the trade-off in a range of environments.

A.5 Behavioural results from Approach-Withdrawal VR task

We observe a baseline approach bias through significant asymmetry in average number of cumulative approaches and withdrawals (Fig. A.5A) and we consider the few inactive approaches and inactive withdrawals due to timeout as approaches and withdrawals respectively (Fig. A.5B).

We consider a couple of model-free metrics of Pavlovian withdrawal bias prior to model fitting. The withdrawal bias metric on choices for two cues (say, cues X and cue Y) calculated as follows:

$$\begin{aligned}\text{Choice bias metric(cue X, cue Y)} = {} & \%\text{ withdrawal choices on 'cue X'} \\ & - \%\text{ approach choices on 'cue Y'}\end{aligned} \tag{A.1}$$

and the metric for withdrawal bias in reaction times is simply the subtraction of average withdrawal times from average approach times in a half (60 trials) or the quarter block (30 trials) under consideration. This choice metric is an extension of the metric used by Dorfman and Gershman (2019) to punishment bias, and it's logic is as follows. Consider Choice bias metric(cue2,cue1) - As Pavlovian withdrawals will increase %withdrawal choices i.e.(correct choice) for cue2 and decrease %approach choices i.e. (correct choice) for cue1. Similarly it'll also make sense for metric(cue4,cue3) in the second half, albeit the bias would be lesser as they will be exploiting the optimal actions. It makes less sense for (cue4,cue3) in the first half as there is no optimal action however, helps act as a control and quantify a baseline approach bias. Unfortunately, this metric cannot differentiate an action due to random exploration from an action due to Pavlovian misbehaviour, leading to noisy estimates. Further, it cannot capture baseline approach bias b at all, because the model by Dorfman and Gershman (2019) does not consider this parameter, unlike Guitart-Masip et al. (2012); Cavanagh et al. (2013). However we show that including baseline bias contributes the most to an incremental improvement in model fit.

We expect this bias to be largest in the first half with uncontrollable cues, and especially in the second quarter as opposed to first quarter, by when enough

punishment value would have been accrued for each of the cues and there would be a significant drop in random exploration. The Pavlovian withdrawal bias in choices is measured using the controllable cues 1 and 2 and thus computing the same quantity in uncontrollable cues 3 and 4 acts as control (Fig. A.5C). Likewise we also hypothesized that there will exist a Pavlovian bias in reaction times which speed up all withdrawals and slow down all approaches regardless of the cue (Fig. A.5E). For our second hypothesis, the Pavlovian bias should decrease with decrease in outcome uncertainty i.e. it would be higher in the second quarter as opposed to the fourth quarter (Fig. A.5D & F). We compare the quarters rather than first and second half to minimise the noise through random exploration in the first quarter. However, the differences we observe are not statistically significant.

In addition to these results from behavioural metrics in choices and reaction times, we further observe certain change-of-mind like patterns in motor responses. It is unclear if these are due to a Pavlovian bias or due to other factors and can be investigated in future studies.

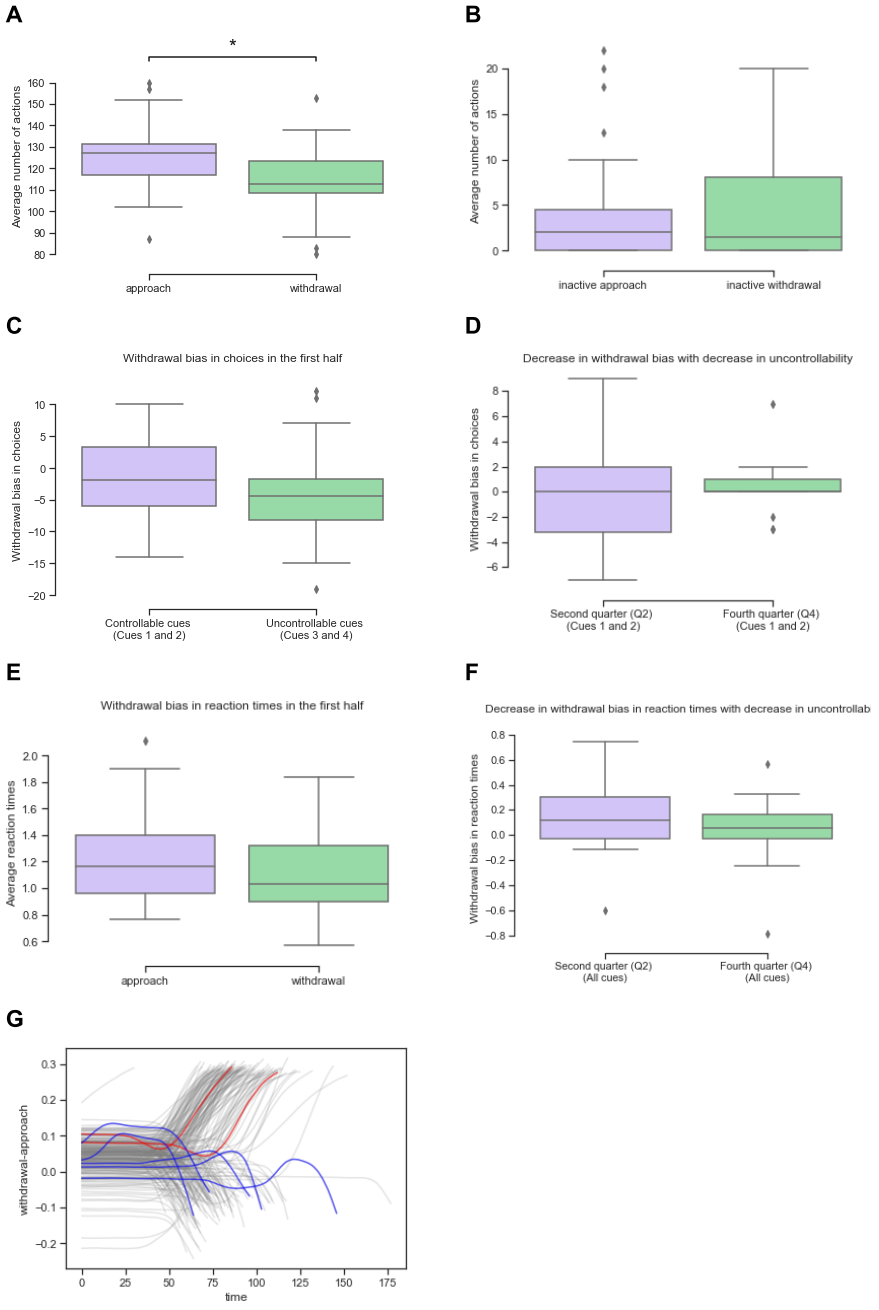


Figure A.5: (A) Asymmetry in average approach and withdrawal responses over all subjects showing a baseline approach bias - Mann-Whitney U test (statistic=597.0, p-value=0.0004)(B) Incomplete approaches and incomplete withdrawals that were counted as approaches and withdrawals, respectively. (C) Withdrawal bias in choices in the first half with uncontrollable cues - Mann-Whitney U test (statistic=492.5, p-value=0.0504) (D) Decrease in withdrawal bias in choice with decrease in uncontrollability - Mann-Whitney U test (statistic=350.5, p-value=0.7611) (E) Withdrawal bias in reaction times in the first half with uncontrollable cues - Mann-Whitney U test (statistic=475.0, p-value=0.0882) (F) Decrease in withdrawal bias in reaction times with decrease in uncontrollability - Mann-Whitney U test (statistic=456.0, p-value=0.1490) (G) Change-of-mind trials observed in motor data.

A.6 Group and subject level parameter distributions of RL and RLDDM models

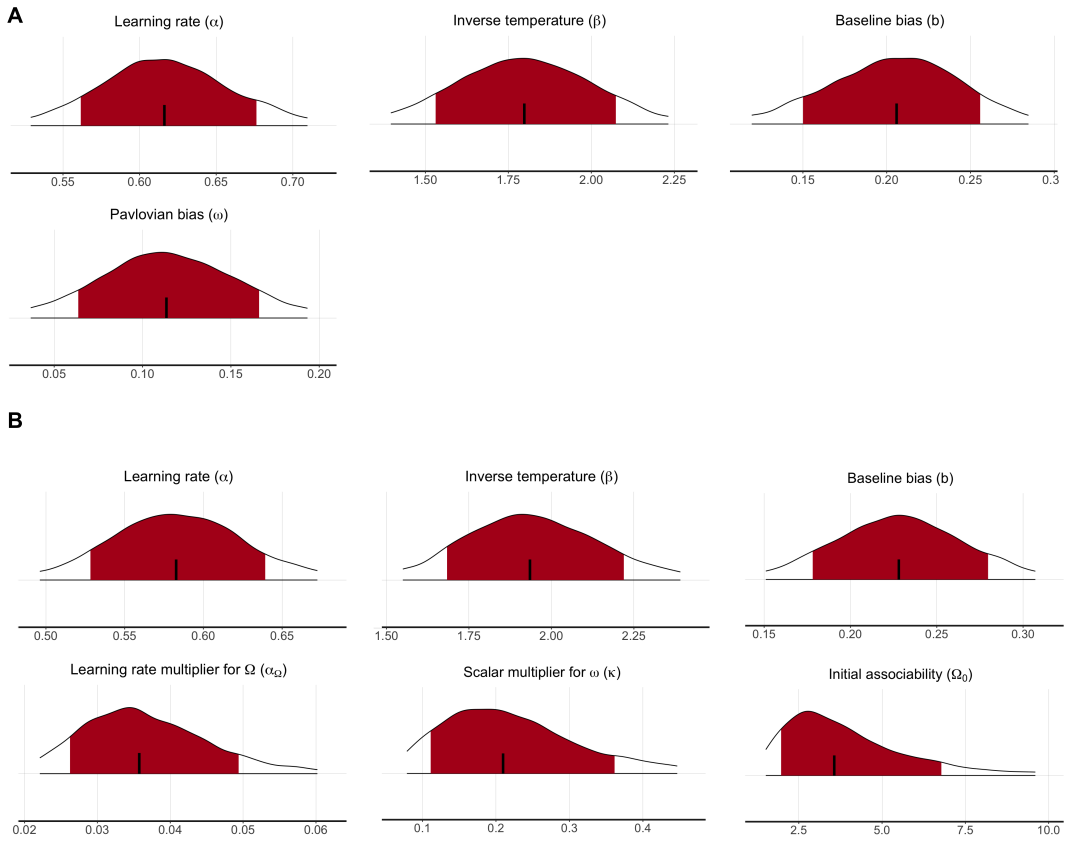


Figure A.6: Group-level parameter distributions from (A) the RL model (M3) with fixed ω and (B) the RL model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

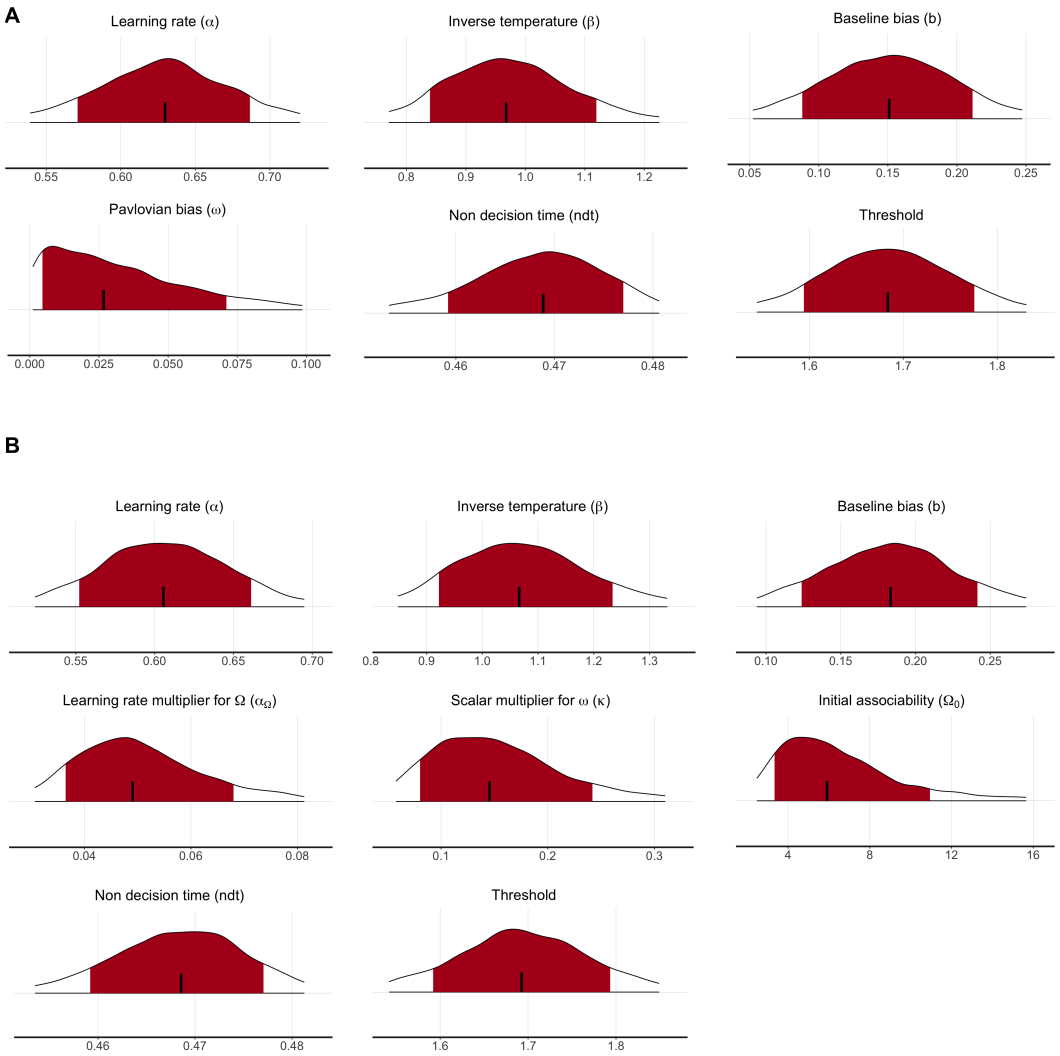


Figure A.7: Group-level parameter distributions from (A) the RLDDM model (M3) with fixed ω and (B) the RLDDM model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

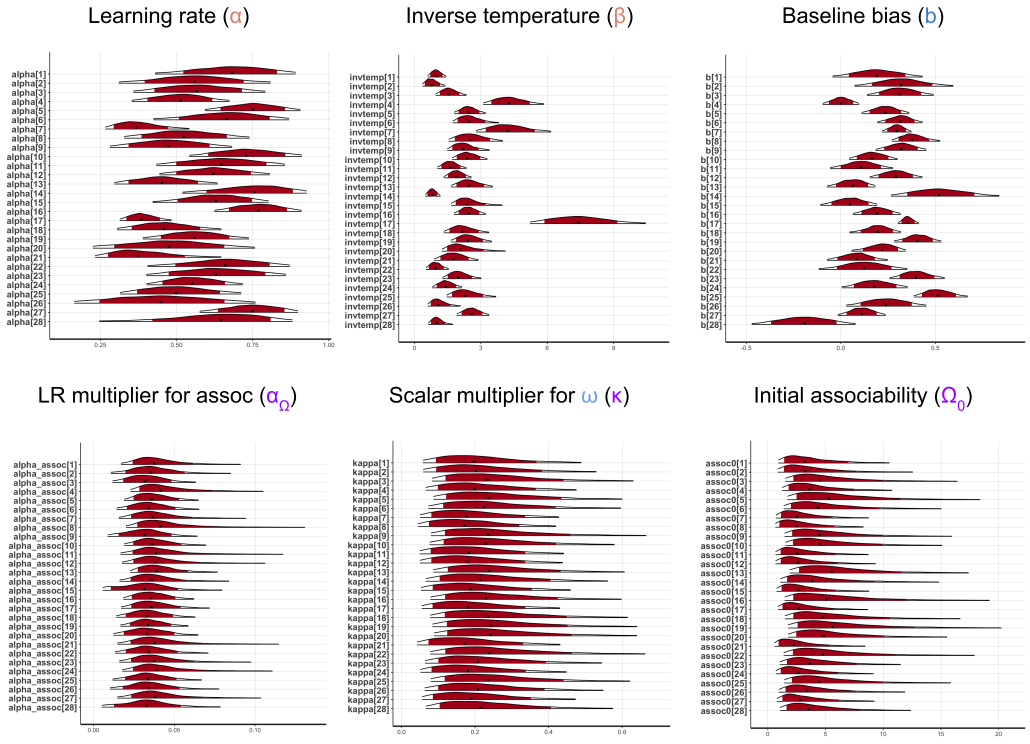


Figure A.8: Subject-level parameter distributions from (A) the RL model (M3) with fixed ω and (B) the RL model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

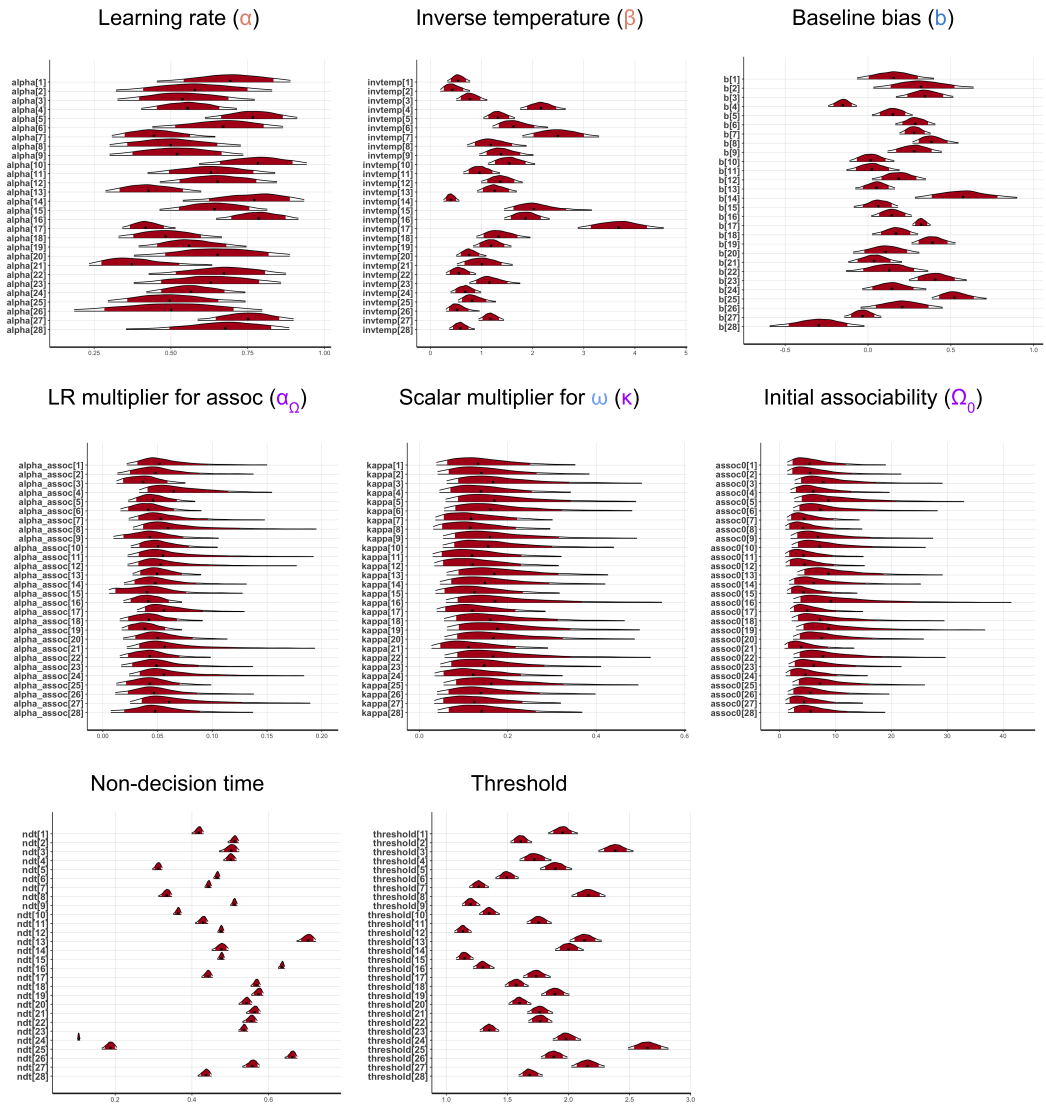


Figure A.9: Subject-level parameter distributions from (A) the RLDDM model (M3) with fixed ω and (B) the RLDDM model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

A.7 RL and RLDDM model parameters and model comparison tables

Please refer to Table 1 and Table 2.

Table A.1: Model comparison results for RL models

Model	Free Parameters	LOOIC	WAIC
M1	α, β	8201.53	8182.15
M2	α, β, b	7960.00	7926.73
M3	α, β, b, ω	7947.84	7918.20
M4	$\alpha, \beta, b, \alpha_\Omega, \kappa, \Omega_0$	7863.79	7830.18

Table A.2: Model comparison results for RLDDM model

Model	Free Parameters	LOOIC	WAIC
M1	ndt, threshold, α, β	12539.14	12495.44
M2	ndt, threshold, α, β, b	12303.00	12247.63
M3	ndt, threshold, α, β, b, ω	12296.38	12247.22
M4	ndt, threshold, $\alpha, \beta, b, \alpha_\Omega, \kappa, \Omega_0$	12205.52	12164.14

A.8 Model predictions: Adapting fear responses in a chronic pain gridworld

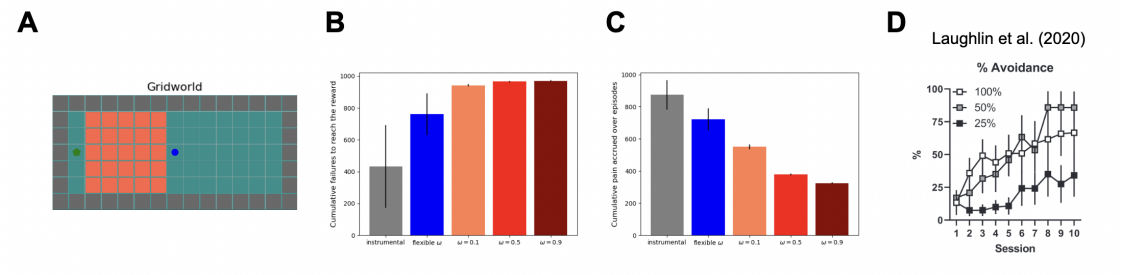


Figure A.10: Pavlovian-instrumental interactions are invoked in a popular model of chronic pain, in which excessive Pavlovian fear of movement is self-punitive in a context in which active avoidance would reduce pain (Meulders et al., 2011; Crombez et al., 2012). (A) Grid world with a start at the centre (blue) and goal at the left end (green) operationalises this. We augment the action set to include an additional "immobilise" action, to the action set, resulting in no state change and repeated rewards. An upper bound of 100 steps per episode is set; exceeding it leads to a painless death and episode restart. (B) Cumulative failures to reach the goal as a measure of efficiency. With a constant Pavlovian fear influence, the agent struggles to complete episodes, resembling effects seen in rodent models of anxiety (Laughlin et al., 2020) (C) Cumulative pain accrued as a measure of safety. In clinical terms, the agent remains stuck in a painful state, contrasting with an instrumental system that can seek and consume rewards despite pain. Flexible parameter ω ($\kappa = 3, \alpha_\Omega = 0.01$) allows the agent to overcome fear and complete episodes efficiently, demonstrating a safety-efficiency dilemma. The flexible ω policy outperforms fixed variants, emphasising the benefits of adapting fear responses for task completion. (D) Results from Laughlin et al. (2020) show that 25% of the (anxious) rats fail the signalled active avoidance task due to freezing. GIFs for different configurations: pure instrumental agent, adaptively safe agent (flexible ω) and maladaptively safe agent (constant ω) can be found here.

A.9 Neurobiology of Pavlovian contributions to bias avoidance behaviour

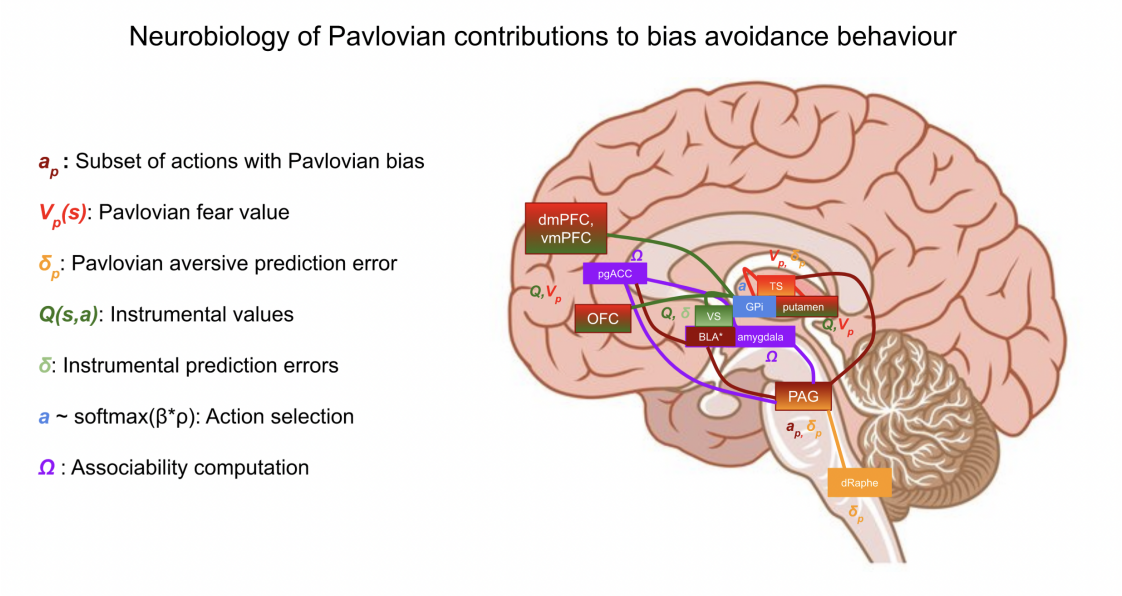


Figure A.11: An overview of neurobiological substrates for the proposed PAL model based on relevant prior literature.

B

Appendix for Chapter 5

Contents

B.1 How does the Boltzmann policy achieve the stochastic Bellman optimal policy in Linear MDPs?

We here aim to provide an intuitive explanation for known results. Consider the KL divergence between any policy π and the Boltzmann policy $\pi_Q^{\mathcal{B}}$ under some Q-values.

$$D_{\mathrm{KL}}(\pi(\cdot|s)\|\pi_Q^{\mathcal{B}}(\cdot|s)) = \mathbb{E}_{a\sim\pi}[\log\pi(a|s) - \log\pi_Q^{\mathcal{B}}(a|s)] \tag{B.1}$$

$$= \mathbb{E}_{a\sim\pi}\Big[\log\pi(a|s) - \log\pi^d(a|s) - Q(s,a)/\tau + \log\mathbb{E}_{a\sim\pi^d}[Q(s,a)/\tau]\Big] \tag{B.2}$$

$$= D_{\mathrm{KL}}[\pi\|\pi^d](s) - \mathbb{E}_{a\sim\pi}[Q(s,a)/\tau] + \log\mathbb{E}_{a\sim\pi^d}[Q(s,a)/\tau] \tag{B.3}$$

We can rearrange this equation and multiply by τ to get V_π (as per equation 5.5) on the left-hand side.

$$\mathbb{E}_{a\sim\pi}[Q(s,a)] - \tau D_{\mathrm{KL}}[\pi\|\pi^d](s) = \tau\log\mathbb{E}_{a\sim\pi^d}[Q(s,a)/\tau] - \tau D_{\mathrm{KL}}[\pi\|\pi_Q^{\mathcal{B}}](s) \tag{B.4}$$

Here, we can see that the left-hand side of the equation (i.e. $V_\pi(s)$)) is maximised with respect to π, during generalised policy iteration (GPI), when the KL term on the right-hand side is minimized (as the other term does not depend on π), and $D_{\mathrm{KL}}[\pi\|\pi_Q^{\mathcal{B}}](s)$ is minimized at $\pi = \pi_Q^{\mathcal{B}}$. After each GPI, as $D_{\mathrm{KL}}[\pi\|\pi_Q^{\mathcal{B}}](s)$ approaches zero at $\pi = \pi_Q^{\mathcal{B}}$, we observe that left-hand side of the equation is the "soft" Bellman value function $V_Q(s)$. This shows that for fixed Q-values, the Boltzmann policy is the stochastic greedy policy that maximises value. Under optimal Q-values, this greedy policy can lead to the Bellman optimal policy in linear MDPs.

During the generalised policy evaluation (GPE), these optimal Q-values can be learnt using any reinforcement learning algorithm with convergence guarantees. Repeating these generalised policy updates (GPE+GPI) will lead to the optimal policy π^* will be given by the Boltzmann policy $\pi_{Q^*}^{\mathcal{B}}$. This concludes our intuitive explanation.

B.2 Theorem for additive composition in Linear MDPs

Theorem 2 (Additive Composition)**.** *(Haarnoja et al., 2018; Van Niekerk et al., 2019)*

Let $Q^*_{1,\tau}(s,a)$ and $Q^*_{2,\tau}(s,a)$ be the optimal entropy-regularized Q-functions for two tasks with rewards $r_1(s,a)$ and $r_2(s,a)$.

Then the reward function for the composed task aimed to ensure both objectives are given by the average of the individual reward functions:

$$r_c(s,a) = \frac{r_1(s,a) + r_2(s,a)}{2}.$$

Let the composition of Q-values $Q_{comp,\tau}(s,a)$ be:

$$Q_{comp,\tau}(s,a) = \frac{Q^*_{1,\tau}(s,a) + Q^*_{2,\tau}(s,a)}{2}.$$

The optimal Q-function $Q^*_{c,\tau}(s,a)$ for the composed task is bounded by:

$$Q_{comp,\tau}(s,a) \geq Q^*_{c,\tau}(s,a) \geq Q_{comp,\tau}(s,a) - C^*_\tau(s,a),$$

where C^*_τ is a fixed point of

$$C^*_\tau = \tau \mathbb{E}_{s' \sim \rho(s,a)} \left[D_{\frac{1}{2}}(\pi^*_1(s) \| \pi^*_2(s)) + \max_{a'} C(s',a') \right],$$

where $\pi^*_i(s)$ is the optimal Boltzmann policy for task i, and $D_{\frac{1}{2}}(\cdot\|\cdot)$ is the Rényi divergence of order $\frac{1}{2}$.

B.3 Novel derivations extending Soft Q-learning to N-step soft Q-learning

In this section, we provide a detailed derivation of how soft Q-learning can be extended to N-step soft Q-learning. We will first begin with the on-policy setting, under the special case of Boltzmann policy (the stochastic optimal policy) and then extend it to a fully off-policy algorithm.

N-step Soft Q-learning (on-policy with Boltzmann policy)

N-step soft Q-learning incorporates multiple future rewards and KL penalties for deviating from the default policy, starting from the second time step onward.

The N-step return at time t, after taking an action a_t in state s_t is defined as:

$$G_{t:t+n} \doteq r_{t+1} + \gamma(r_{t+2} - \tau \mathrm{KL}_{t+1}) + \gamma^2(r_{t+3} - \tau \mathrm{KL}_{t+2}) + \ldots + \gamma^{n-1}(r_{t+n} - \tau \mathrm{KL}_{t+n-1}) + \gamma^n V_Q(s_{t+n}), \tag{B.5}$$

Note that the KL penalty terms appear only from the second timestep onward, as the cost of deviating from the default policy affects subsequent actions. If the episode terminates at timestep T, which can be less than $t+n$, then we will see next that the summation of TD-errors is appropriately truncated to $min(T-1, t+n-1)$.

We can rewrite $G_{t:t+n}$ in terms of the temporal difference (TD) error δ, by adding and subtracting $\gamma V_Q(s_{t+1})$, $\gamma^2 V_Q(s_{t+2})$, $\gamma^3 V_Q(s_{t+3})$ and so on:

$$\begin{aligned} G_{t:t+n} = (r_{t+1} + \gamma V_Q(s_{t+1})) + \gamma(r_{t+2} - \tau \mathrm{KL}_{t+1} + \gamma V_Q(s_{t+2}) - V_Q(s_{t+1})) \\ + \ldots + \gamma^{n-1}(r_{t+n} - \tau \mathrm{KL}_{t+n-1} + \gamma V_Q(s_{t+n}) - V_Q(s_{t+n-1})). \end{aligned} \tag{B.6}$$

Simplifying, we obtain:

$$\begin{aligned} G_{t:t+n} &= (r_{t+1} + \gamma V_Q(s_{t+1})) + \sum_{k=t+1}^{min(T-1,t+n-1)} \gamma^k \delta_k \\ &= Q_{t-1}(s_t, a_t) + (r_{t+1} + \gamma V_Q(s_{t+1}) - (Q_{t-1}(s_t, a_t)) + \sum_{k=t+1}^{min(T-1,t+n-1)} \gamma^k \delta_k \\ &= Q_{t-1}(s_t, a_t) + \sum_{k=t}^{min(T-1,t+n-1)} \gamma^k \delta_k \end{aligned} \tag{B.7}$$

where the TD error δ_k at each timestep is given as follows.

If $k = t$, the same as soft Q-learning:

$$\delta_t = r_{t+1} + \gamma V_Q(s_{t+1}) - Q(s_t, a_t) \tag{B.8}$$

For $k \geq t$,

$$\delta_k = r_{k+1} - \tau \mathrm{KL}_k + \gamma V_Q(s_{k+1}) - V_Q(s_k) \tag{B.9}$$

The first TD error term, $\delta_t = r_{t+1} + \gamma V_Q(s_{t+1}) - Q(s_t, a_t)$, does not include the KL penalty since it doesn't depend on the action a_t which has already been chosen (Ziebart, 2010; Haarnoja et al., 2017; Schulman et al., 2017).

Thus, the N-step soft Q-learning update rule is defined as:

$$Q_{t+n}(s_t, a_t) \leftarrow Q_{t+n-1}(s_t, a_t) + \alpha \left(G_{t:t+n} - Q_{t+n-1}(s_t, a_t) \right), \tag{B.10}$$

where α is the learning rate. The subscripts denote the timestep in the episode when the Q-value was used or updated. Note that n-step returns for $n > 1$ involve future rewards and states that are not available at the time of transition from t to $t+1$. Thus, the first Q-update of state s_t is performed at timestep $t+n$ and not t.

If the approximate action-values are unchanging, i.e. $Q_{t-1}(s_t, a_t) \simeq Q_{t+n-1}(s_t, a_t)$ (similar to Exercise 7.11 in Sutton and Barto (2018)), then we can substitute the expression for $G_{t:t+n}$ to get:

$$Q_{t+n}(s_t, a_t) \leftarrow Q_{t+n-1}(s_t, a_t) + \alpha \left(\sum_{k=t}^{min(T-1, t+n-1)} \gamma^k \delta_k \right). \tag{B.11}$$

If the approximate action values are changing, then we will have an additional term of $Q_{t-1}(s_t, a_t) - Q_{t+n-1}(s_t, a_t)$ in the update.

N-step Soft Q-learning (off-policy with importance sampling)

We can now extend this to an off-policy algorithm that learns the Boltzmann policy ($\pi_Q^{\mathcal{B}}$) as the target policy while collecting data under any behavioural policy b. Considering that soft Q-learning is akin to expected SARSA for relative-entropy regularised objective, this derivation is similar to the N-step expected SARSA derivation (Sutton and Barto, 2018).

We define the importance sampling ratio as follows (T is the last time step of the episode),

$$\rho_{t:h} = \prod_{k=t}^{min(h, T-1)} \frac{\pi_Q^{\mathcal{B}}(a_k|s_k)}{b(a_k|s_k)} \tag{B.12}$$

Now the update from the previous section can be replaced with its off-policy form,

$$Q_{t+n}(s_t, a_t) \leftarrow Q_{t+n-1}(s_t, a_t) + \alpha \rho_{t+1:t+n-1} \left(G_{t:t+n} - Q_{t+n-1}(s_t, a_t) \right), \tag{B.13}$$

$$Q_{t+n}(s_t, a_t) \leftarrow Q_{t+n-1}(s_t, a_t) + \alpha\rho_{t+1:t+n-1}\left(\sum_{k=t}^{t+n-1}\gamma^k\delta_k\right). \tag{B.14}$$

where, δ_{t+k} is defined as per equations B.8 and B.9. Note, we use $\rho_{t+1:t+n-1}$ and not $\rho_{t+1:t+n}$ as in any N-step expected SARSA such as this one, all possible actions are taken into account in the last state; the one actually taken has no effect and does not have to be corrected for (Sutton and Barto, 2018, Page 150). One can further write this recursively using per-decision importance sampling (Sutton and Barto, 2018; Precup, 2000), but it is not essential to our derivations.

N-step Soft Q-learning (off-policy with Tree Backup)

We next present N-step Soft Q-learning using the Tree Backup algorithm. N-step soft Q-learning with importance sampling only uses the expectation over actions in the last time step. Tree Backup instead uses it at every step. This provides the following advantages: (1) reduces the variance due to the importance sampling ratio, (2) an importance sampling ratio does not need to be computed, thus the behavioural policy b does not need to be stationary, Markov, or even known (De Asis et al., 2018; Precup, 2000).

We begin by writing the N-step return under the Boltzmann policy after taking action a_t in state s_t in the Tree Backup format. Note, this is the soft-Bellman optimal return regardless of the behavioural policy which chooses actions $a_t, a_{t+1}, a_{t+2}, ...$ leading to states $s_{t+1}, s_{t+2}, s_{t+3}, ...$ respectively.

$$G_{t:t+n} \doteq r_{t+1} + \gamma V_{\pi^{\mathcal{B}}_Q}(s_{t+1}) \tag{B.15}$$

Using equation 5.5, we get,

$$G_{t:t+n} \doteq r_{t+1} + \gamma\left(\sum_a \pi^{\mathcal{B}}_Q(a|s_{t+1})Q_t(s_{t+1}, a) - \tau\mathrm{KL}_{t+1}\right) \tag{B.16}$$

We can now write it in Tree-Backup format,

$$
\begin{aligned}
G_{t:t+n} \doteq\ & r_{t+1} + \gamma \sum_{a \neq a_{t+1}} \pi_Q^{\mathcal{B}}(a|s_{t+1}) Q_t(s_{t+1}, a) - \gamma\tau \mathrm{KL}_{t+1} \\
& + \gamma \pi_Q^{\mathcal{B}}(a_{t+1}|s_{t+1}) \Big(r_{t+2} - \tau \mathrm{KL}_{t+1} \\
& \quad + \gamma \sum_{a \neq a_{t+2}} \pi_Q^{\mathcal{B}}(a|s_{t+2}) Q_{t+1}(s_{t+2}, a) - \gamma\tau \mathrm{KL}_{t+2} \Big) \\
& + \gamma^2 \pi_Q^{\mathcal{B}}(a_{t+2}|s_{t+2}) \pi_Q^{\mathcal{B}}(a_{t+1}|s_{t+1}) \Big(r_{t+3} - \tau \mathrm{KL}_{t+2} \\
& \quad + \gamma \sum_{a \neq a_{t+3}} \pi_Q^{\mathcal{B}}(a|s_{t+3}) Q_{t+2}(s_{t+3}, a) - \gamma\tau \mathrm{KL}_{t+3} \Big) \\
& + \dots \\
& + \gamma^{n-1} \prod_{i=t+1}^{\min(t+n-1, T-1)} \pi_Q^{\mathcal{B}}(a_i|s_i) \Big(r_{t+n} - \mathrm{KL}_{t+n-1} \\
& \quad + \gamma \sum_{a} \pi_Q^{\mathcal{B}}(a|s_{t+n}) Q_{t+n-1}(s_{t+n}, a) - \gamma\tau \mathrm{KL}_{t+n} \Big)
\end{aligned}
\tag{B.17}
$$

This is visualised as follows: The update is from the estimated action values of the leaf nodes of the tree. The action nodes in the interior, corresponding to the actual actions taken, do not participate. Each leaf node contributes to the target with a weight proportional to its probability of occurring under the target policy.

This can now be written recursively as,

$$
G_{t:t+n} \doteq r_{t+1} + \gamma \sum_{a \neq a_{t+1}} \pi_Q^{\mathcal{B}}(a|s_{t+1}) Q_t(s_{t+1}, a) + \gamma \pi_Q^{\mathcal{B}}(a_{t+1}|s_{t+1}) (G_{t+1:t+n} - \tau \mathrm{KL}_{t+1}) \tag{B.18}
$$

Alternatively, it can also be compactly written in terms of temporal difference errors, by using the following relation from equations 5.9 and B.4:

$$
\begin{aligned}
\sum_{a \neq a_k} \pi_Q^{\mathcal{B}}(a|s_k) Q_{k-1}(s_k, a) &= \sum_{a} \pi_Q^{\mathcal{B}}(a|s_k) Q_{k-1}(s_k, a) - \pi_Q^{\mathcal{B}}(a_k|s_k) Q_{k-1}(s_k, a_k)) \\
&= V_Q(s_k) + \tau \mathrm{KL}_k - \pi_Q^{\mathcal{B}}(a_k|s_k) Q_{k-1}(s_k, a_k))
\end{aligned}
\tag{B.19}
$$

By substituting this relation in equations B.17, the $\tau \mathrm{KL}_k$ terms cancel out and we can write the Tree-Backup return in terms of TD-errors as follows:

$$
\begin{aligned}
G_{t:t+n} \doteq\ & r_{t+1} + \gamma\Big(V_Q(s_{t+1}) - \pi_Q^{\mathcal{B}}(a_{t+1}|s_{t+1})Q_t(s_{t+1}, a_{t+1})\Big) \\
& + \gamma\pi_Q^{\mathcal{B}}(a_{t+1}|s_{t+1})\Big(r_{t+2} - \tau\mathrm{KL}_{t+1} \\
& \quad + \gamma V_Q(s_{t+2}) - \gamma\pi_Q^{\mathcal{B}}(a_{t+2}|s_{t+2})Q_{t+1}(s_{t+2}, a_{t+2})\Big) \\
& + \gamma^2\pi_Q^{\mathcal{B}}(a_{t+2}|s_{t+2})\pi_Q^{\mathcal{B}}(a_{t+1}|s_{t+1})\Big(r_{t+3} - \tau\mathrm{KL}_{t+2} \\
& \quad + \gamma V_Q(s_{t+3}) - \gamma\pi_Q^{\mathcal{B}}(a_{t+3}|s_{t+3})Q_{t+2}(s_{t+3}, a_{t+3})\Big) \\
& + \dots \\
& + \gamma^{n-1}\left[\prod_{i=t+1}^{\min(t+n-1,T-1)} \pi_Q^{\mathcal{B}}(a_i|s_i)\right]\Big(r_{t+n} - \mathrm{KL}_{t+n-1} \\
& \quad + \gamma V_Q(s_{t+n}) - \gamma\pi_Q^{\mathcal{B}}(a_{t+n}|s_{t+n})Q_{t+n-1}(s_{t+n}, a_{t+n})\Big)
\end{aligned}
\tag{B.20}
$$

If we combine $r_{k+1} - \tau KL_k + V_Q(s_{k+1})$ with the last term of the (previous) k-th term, and add and subtract $Q(s_t, a_t)$ for the first term, then we have the following.

$$
\begin{aligned}
G_{t:t+n} = & Q_{t-1}(s_t, a_t) \\
& + \sum_{k=t}^{\min(T-1,t+n-1)} \left[\delta_k \prod_{i=t+1}^{k} \gamma\pi_Q^{\mathcal{B}}(a_i|s_i)\right] \\
& - \gamma^n Q_{t+n-1}(s_{t+n}, a_{t+n}) \prod_{i=t+1}^{\min(T-1,t+n-1)} \pi_Q^{\mathcal{B}}(a_i|s_i)
\end{aligned}
\tag{B.21}
$$

If the $t + n - 1 > T - 1$, that is, the last state is terminal, then we can set the last Q-term to zero, and this expression simplifies to,

$$
G_{t:t+n} = Q_{t-1}(s_t, a_t) + \sum_{k=t}^{T-1} \left[\delta_k \prod_{i=t+1}^{k} \gamma\pi_Q^{\mathcal{B}}(a_i|s_i)\right] \tag{B.22}
$$

Again, if we assume the approximate Q-values are unchanging (similar to Exercise 7.11 in Sutton and Barto (2018)), then this gives us our Q-update equation as follows,

$$
Q_{t+n}(s_t, a_t) \leftarrow Q_{t+n-1}(s_t, a_t) + \alpha\left(\sum_{k=t}^{min(T-1,t+n-1)} \delta_k \prod_{i=t+1}^{k} \gamma\pi_Q^{\mathcal{B}}(a_i|s_i)\right). \tag{B.23}
$$

where δ_k are defined as per equations B.8 and B.9. These updates lead to the estimation of off-policy multi-step returns under any behavioural policy, without knowing the behavioural policy.

Note, that if one starts the Tree Backup derivation with $V_Q(s_t + 1)$ instead of $V_{\pi^{\mathcal{B}}_Q}(s_{t+1})$, then this leads to an alternate equivalent derivation in terms of the default policy instead of the Boltzmann policy (which requires calculating TD-errors under the default policy as well). We think this alternate derivation is less relevant as the agent is the target policy for the agent is the soft-Bellman optimal Boltzmann policy; therefore, we focus on the derivation in terms of the Boltzmann policy.

This concludes our novel derivations of off-policy N-step extensions of Soft Q-learning, using either importance sampling or Tree-Backup. One may further aspire to unify these two multi-step off-policy methods, as done in the standard RL setting by De Asis et al. (2018), but it is not essential to the current work and is left as future work.

B.4 Novel derivations extending N-step soft Q-learning to an elegant algorithm with eligibility traces

Soft Q(λ) (on-policy with Boltzmann policy)

Here, we build upon the N-step Soft Q-learning results to develop Soft Q(λ), a solution using eligibility traces.

We define a λ-return, which is the weighted summation of n-step returns (Sutton and Barto, 2018).

$$G^{\lambda}_t = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t:t+n} \tag{B.24}$$

To simplify the derivation, we define the Boltzmann backup operator following Schulman et al. (2017),

$$\begin{aligned}[\mathcal{T}_{\pi^{\mathcal{B}}_Q}Q](s,a) &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}\left[r + \gamma\tau\log\mathbb{E}_{a'\sim\pi^d}[\exp(Q(s',a')/\tau)]\right] \\ &= \mathbb{E}_{(s',r)\sim p(s',r|s,a)}\left[r + \gamma V_Q(s')\right]\end{aligned} \tag{B.25}$$

We can now define the SARSA(λ) version of this backup operator under the Boltzmann policy, $[\mathcal{T}_{\pi^{\mathcal{B}}_{\mathcal{Q}},\lambda}\mathcal{Q}](s,a)$, as follows.

$$G_t^\lambda = [\mathcal{T}_{\pi^{\mathcal{B}}_{Q},\lambda}Q] = (1-\lambda)(1+\lambda\mathcal{T}_{\pi^{\mathcal{B}}_{Q}} + (\lambda\mathcal{T}_{\pi^{\mathcal{B}}_{Q}})^2 + ...)\mathcal{T}_{\pi^{\mathcal{B}}_{Q}}Q \tag{B.26}$$

Based on n-step methods, we can derive it to be,

$$G_t^\lambda = [\mathcal{T}_{\pi^{\mathcal{B}}_{Q},\lambda}Q](s,a) = Q(s,a) + \mathbb{E}\left[\sum_{k=t}^{\infty}(\gamma\lambda)^k\delta_k\right] \tag{B.27}$$

where,

$$\delta_k = r_{k+1} - \tau\mathrm{KL}_k + \gamma V_Q(s_{k+1}) - V_Q(s_k) \tag{B.28}$$

The update rule using G_t^λ , with a forward-view but offline algorithm is,

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha(G_t^\lambda - Q_t(s,a)) \tag{B.29}$$

This can be approximated using a backwards view (SARSA(λ)-like) online algorithm under the Boltzmann policy, with eligibility traces (e_t) and the TD-errors as mentioned above in equation B.28 (δ_t).

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha\delta_t e_t(s,a) \;\; \forall s,a \tag{B.30}$$

and eligibility traces are updated as follows (in the tabular setting),

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) + 1, & \text{if } (s,a) = (s_t,a_t), \\ \gamma\lambda e_{t-1}(s,a), & \text{otherwise,} \end{cases} \tag{B.31}$$

Soft Q(λ) (off-policy with Tree Backup)

We next extend the algorithm to a full off-policy algorithm, developing upon the n-step method using the Tree Backup algorithm.

$$G_t^\lambda \approx Q(s,a) + \left[\sum_{k=t}^{\infty}\delta_k \sum_{i=t+1}^{k}\gamma_i\lambda_i\pi_Q^{\mathcal{B}}(a_i|s_i)\right] \tag{B.32}$$

Which gives us an online off-policy soft Q(λ) algorithm, similar to the previous one, but the eligibility trace update is adjusted with the target policy $\pi_Q^{\mathcal{B}}$,

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha\delta_t e_t(s,a) \;\; \forall s,a \tag{B.33}$$

where,

$$\delta_t = r_{t+1} - \tau \mathrm{KL}_t + \gamma V_Q(s_{t+1}) - V_Q(s_t) \tag{B.34}$$

and,

$$e_t(s,a) = \begin{cases} \gamma\lambda\pi_Q^{\mathcal{B}}(a_t|s_t)e_{t-1}(s,a) + 1, & \text{if } (s,a) = (s_t,a_t), \\ \gamma\lambda\pi_Q^{\mathcal{B}}(a_t|s_t)e_{t-1}(s,a), & \text{otherwise,} \end{cases} \tag{B.35}$$

This concludes our derivation of a basic online off-policy Soft Q(λ) algorithm. Such algorithms can be extended to (1) function approximation, (2) a more "true" online algorithm and (3) more stable algorithms following Chapter 12 in Sutton and Barto (2018).

B.5 Connection between the KL term and Bogacz (2020) model APE

In this section, we formally establish the connection between the Kullback-Leibler (KL) divergence penalty in our Linear RL framework and the action prediction error (APE) derived in the Bayesian inference model of Bogacz (2020).

We begin with the standard KL divergence between the current behavioural policy $\pi(\cdot|s_t)$ and the default policy (or prior) $\pi_d(\cdot|s_t)$, which is defined as an expectation over the actions drawn from the behavioural policy:

$$D_{\mathrm{KL}}(\pi(\cdot|s_t) \parallel \pi_d(\cdot|s_t)) = \mathbb{E}_{a\sim\pi}\left[\log \pi(a|s_t) - \log \pi_d(a|s_t)\right] \tag{B.36}$$

To relate this to Bogacz (2020), we assume that both policies are parameterized as normal distributions with equal variance. Specifically, let the behavioural policy be $\pi(a|s_t) = \mathcal{N}(a;\mu^\pi,\sigma^2)$ and the default policy be $\pi_d(a|s_t) = \mathcal{N}(a;\mu^d,\sigma^2)$, where μ^d is the mean of the default action prior in state s_t.

Under these assumptions, we can decompose the KL divergence into two distinct components: a negative entropy term and an expected cross-entropy term:

$$D_{\mathrm{KL}}(\pi \parallel \pi_d) = \underbrace{\mathbb{E}_{a\sim\pi}[\log \pi(a|s_t)]}_{-\mathcal{H}(\pi)} - \underbrace{\mathbb{E}_{a\sim\pi}[\log \pi_d(a|s_t)]}_{\text{Prior matching term}} \tag{B.37}$$

By substituting the Gaussian probability density function into the prior matching term, we obtain:

$$\begin{aligned} -\mathbb{E}_{a\sim\pi}[\log \pi_d(a|s_t)] &= -\mathbb{E}_{a\sim\pi}\left[-\frac{(a-\mu^d)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \\ &= \mathbb{E}_{a\sim\pi}\left[\frac{(a-\mu^d)^2}{2\sigma^2}\right] + C \end{aligned} \tag{B.38}$$

where C is a constant dependent on the variance. This formulation reveals that the KL divergence penalizes the expected squared deviation of the behavioural policy's actions from the prior mean μ^d, balanced against the entropy of the behavioural policy.

Sample-based Approximation In practice, algorithms such as soft actor-critic (SAC) often rely on a single-sample Monte Carlo estimate of the KL penalty evaluated at the executed action $a_t \sim \pi(\cdot|s_t)$. Dropping the expectation yields the sampled penalty $\hat{D}_{\mathrm{KL}}$:

$$\hat{D}_{\mathrm{KL}} = \log \pi(a_t|s_t) + \frac{1}{2\sigma^2}(a_t - \mu^d)^2 + C \tag{B.39}$$

This single-sample approximation directly exposes the mathematical equivalence to the Bogacz (2020) model. Bogacz (2020) defines the habit prediction error (APE) simply as the deviation of the chosen action from the expected habitual action: $\delta_h = a_t - \mu^d$.

Consequently, the sample-based KL penalty embedded within our regularized reward function incorporates the exact squared Action Prediction Error of the Bogacz model:

$$\hat{D}_{\mathrm{KL}} \propto \log \pi(a_t|s_t) + \frac{1}{2}\delta_h^2 \tag{B.40}$$

This demonstrates that penalizing the KL divergence from a passive prior in Linear RL inherently applies a quadratic penalty based on the Bogacz (2020) APE (δ_h), alongside an entropy maximization bonus.

B.6 Additional Results and Figures

Additional info accompanying Fig. B.1: Haarnoja et al. (2018) shows that simple additive composition of Q-functions (Q_{comp}) never overestimates Q_c^* by more than the divergence of the constituent policies. Here, constant C^* is the "value" of an adversarial policy that seeks to maximise this divergence (Theorem 2, Appendix B.2). For $\tau = 0$, the behaviour of such additive composition resembles standard Q-learning (Fig. 5.2B), but, for $\tau > 0$, there is a weak bias towards the optimal action $a = R$, even with adversarially chosen rewards r_1, r_2 designed for maximal divergence. To our knowledge, equivalent performance guarantees are unavailable for standard multi-objective Q-learning. While mathematically distinct from a simple weighted sum, the weighted soft maximum approach empirically often yields a higher average weighted sum of rewards than alternatives in our simulations (supplementary Fig. B.1). This advantage is not always guaranteed in all kinds of tasks and depends on τ and MDP utilities. However, these differences could be quantified behaviourally by examining aggregate choices/consumption (adjusted for subjective utilities, (Lak et al., 2014)), especially under stable reward priorities.

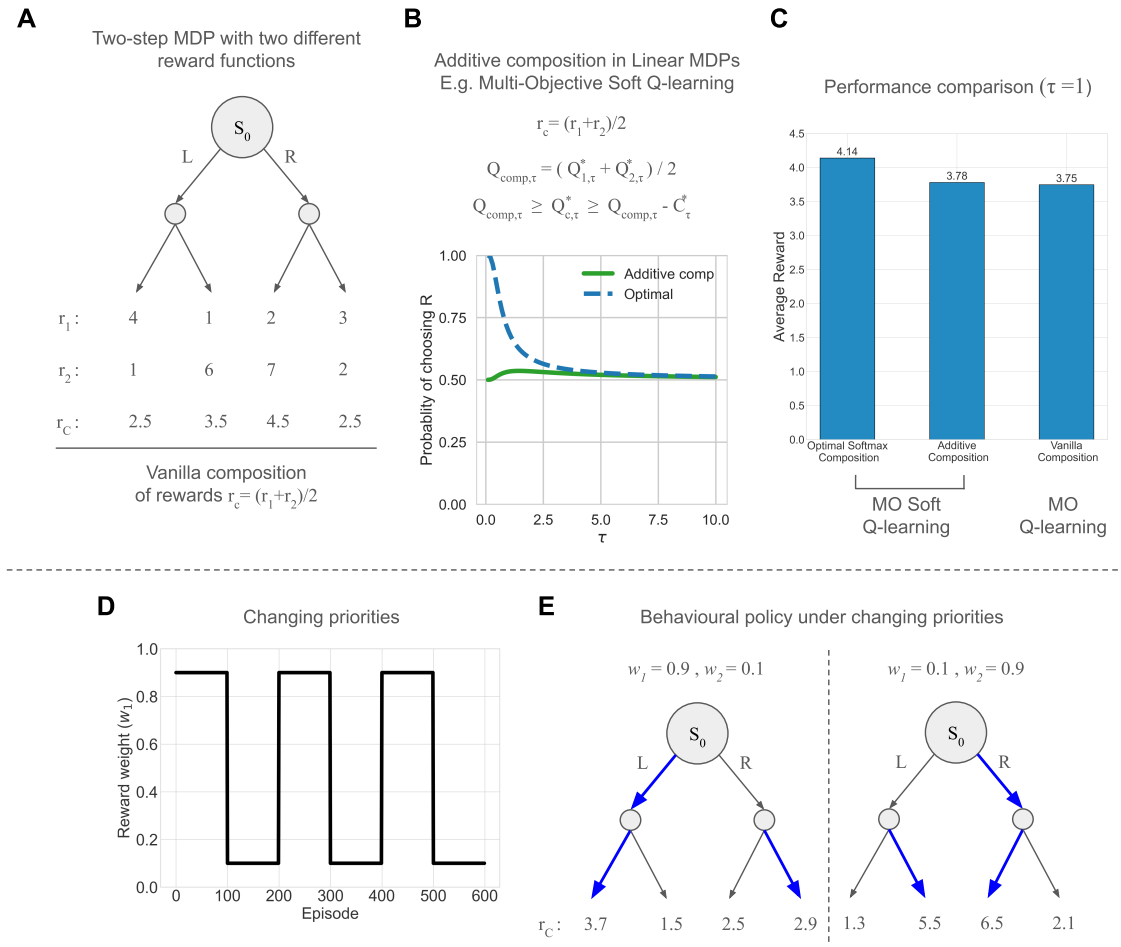


Figure B.1: MO Q-learning comparison: (A) Two-step MDP with two (diverging) reward functions from the main text. (B) Action selection probabilities under additive composition in Linear MDPs, which perform better than additive composition in MDPs (standard RL) for $\tau > 0$. (C) Multi-objective (MO) soft Q-learning leads to better performance than MO Q-learning in this task. MO SARSA comparison: (D) Reward weight w_1, denoting change in priorities and $w_2 = 1 - w_1$. (E) Distinct behavioural policies under different priorities, which then affect the valuation of future states. Bold blue lines indicate the preferred action, and r_c in this figure is calculated under vanilla/additive composition of rewards.

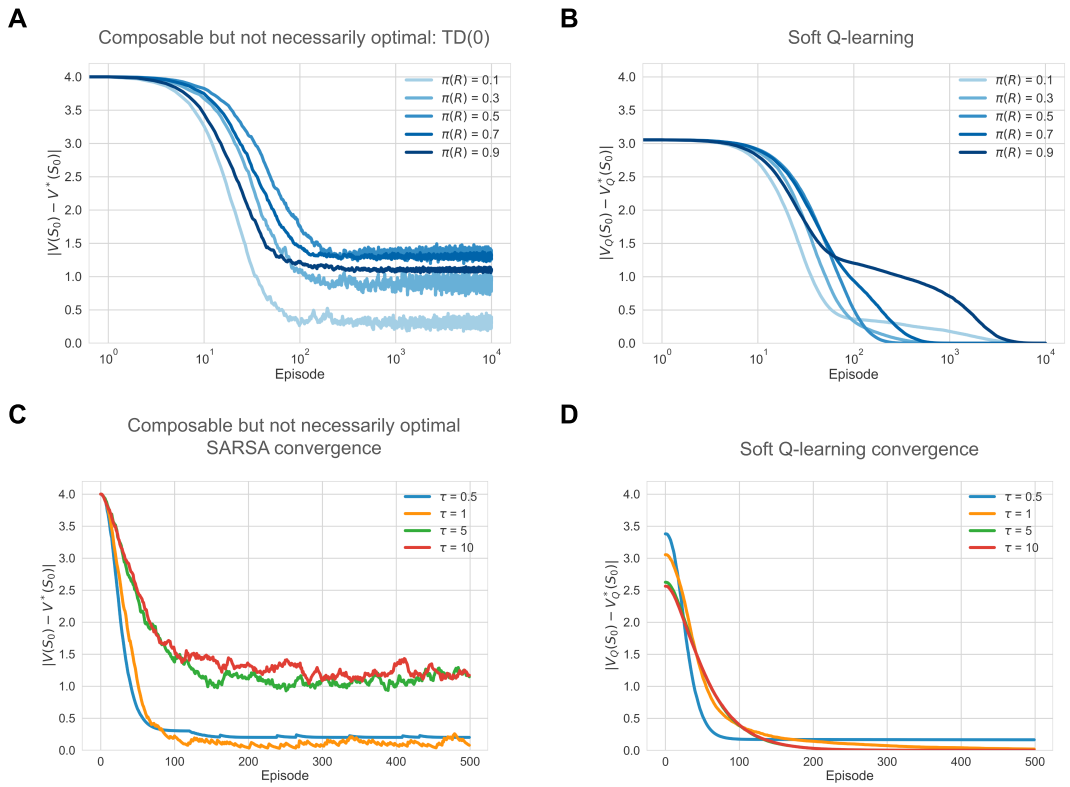


Figure B.2: Demonstration of the off-policy learning of optimal values in linear MDP under sub-optimal trajectories (A & B) and under optimal control at different τ (C & D). (A) TD(0) learns the value of the policy used for data collection but fails to learn the optimal policy from the collected data; (B) Soft Q-learning learns the optimal value under any policy. We use random policies with static probabilities, $\pi(a = R) \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\pi(a = L) = 1 - \pi(a = R)$. Other metaparameters are $\tau = 1$, learning rate $= 0.1$ and π^d as a uniform distribution, but the results are not affected by changing any of these. (C) SARSA (aka TD-control) learns the value of the exploring policy used for data collection, but fails to learn the optimal policy from the collected data. (D) Soft Q-learning learns the optimal value under any policy. Note, the plot has different values at episode 0, because the $V_Q^*(S_0)$ is dependent on τ. All plots were averaged over 30 runs, values were initialised to 0, and the learning rate was set to 0.1.

Figure B.3: MO SARSA with state-dependent TD-errors, uses $\tilde{\delta}_i = w_i \delta_i$, same as the second model from Millidge et al. (2024a). Thus the update is $\Delta Q_i = \alpha w_i \tilde{\delta}_i = \alpha w_i^2 \delta_i$ instead of simply $\Delta Q_i = \alpha \delta_i$. We find this does increase stability/reliability, but may not converge to the optimal values. MO SARSA with state-dependent updates, simply include it as $\Delta Q_i = \alpha w_i \delta_i$, similar to how it shows up in feature-based accounts. We find that this does not quite improve the stability. Neither can completely avoid the interference and unintended unlearning effect caused by the on-policy nature of MO SARSA.

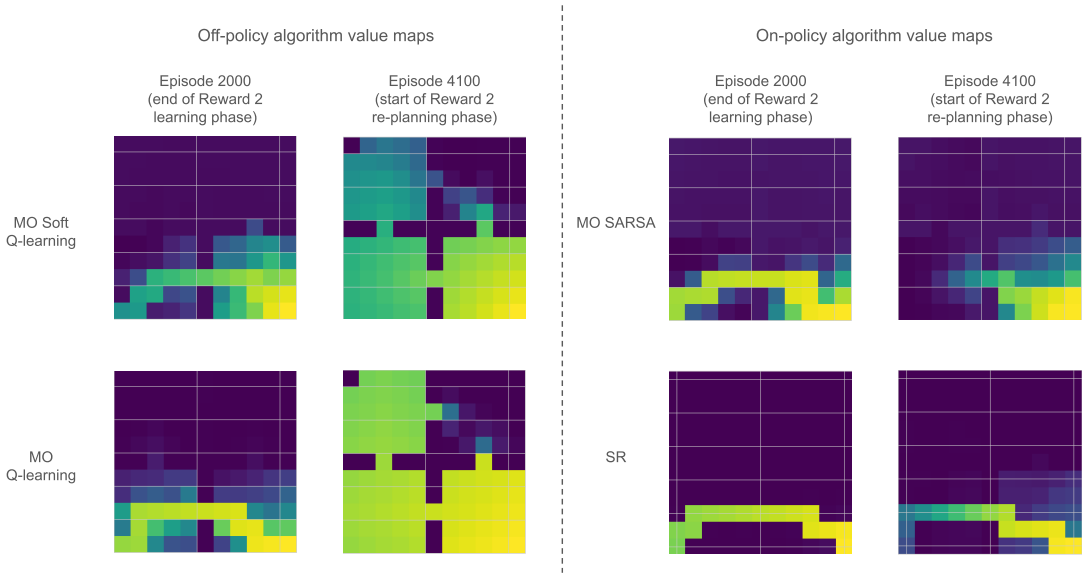


Figure B.4: Differences in value propagation between multi-objective (MO) off-policy and on-policy algorithms, while exploring under different policies. Off-policy algorithms propagate optimal values throughout the environment, whereas on-policy algorithms do not. This further explains the efficiency of off-policy MO algorithms in re-planning to a previously experienced reward function.

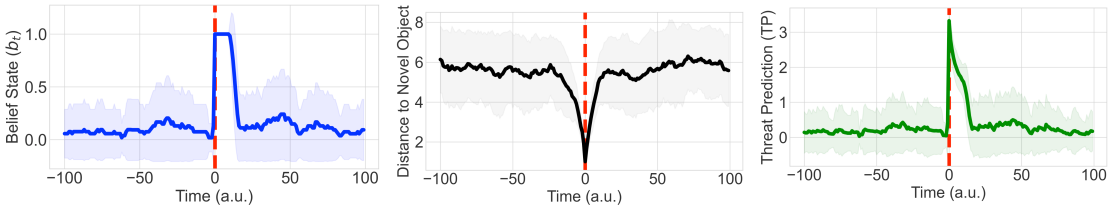


Figure B.5: An additional experiment showing that gating the aversive initialisation with a belief state produces the desired temporal asymmetry in TP responses. (A) Belief state that acts like a switch turned on in the vicinity of the novel object and turned off randomly 10-20 steps after avoiding the object (B). Resultant distance to novel object showing the approach-retreat bout, and (C) The Threat-Prediction response on the retreat start, which decays along with the gradient of the value initialisation.

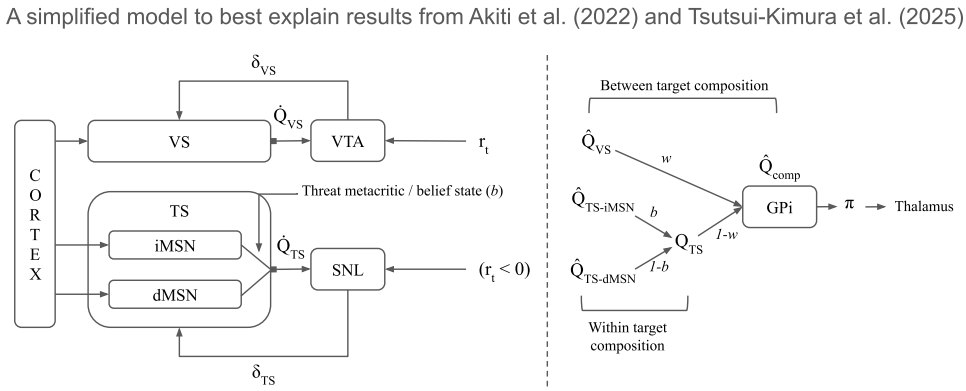


Figure B.6: A simplified model to best explain results from Akiti et al. (2022) and Tsutsui-Kimura et al. (2025).

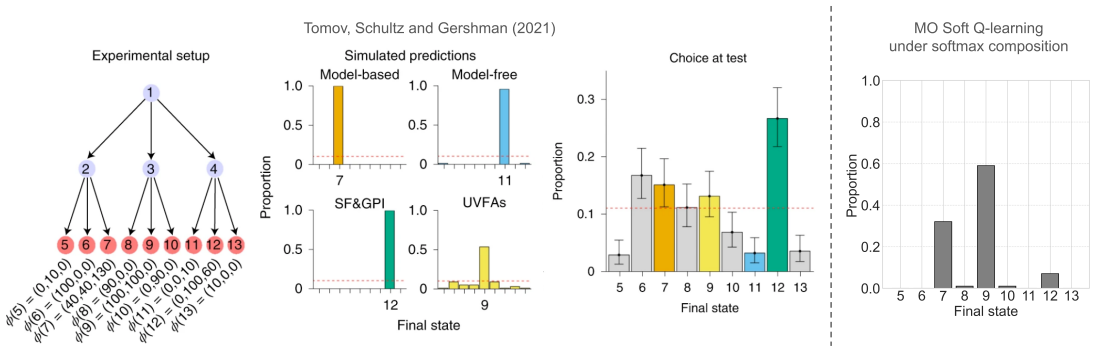


Figure B.7: Soft maximum composition may not match human behaviour when they are explicitly asked to optimise the weighted sum of rewards across different modalities (Tomov et al., 2021). There are two possible explanations. First, like all multi-objective (MO) model-free RL, our model generalises in values, not task structure, whereas human behaviour in these experiments shows generalisation in task structure. Second, the policy under weighted softmax composition can deviate significantly from the policy optimising the weighted summation of rewards. The left side of the figure shows the two-stage MDP by Tomov et al. (2021), where the participants were playing a medieval trading game. Different terminal states lead to different quantities of 3 resources (denoted by ϕ), and at the start of each episode, they get to know the price/cost of each resource, setting the reward weights, requiring them to maximise the weighted sum of these multiple rewards. First 100 episodes had different weights: $w_{train} = \{[1, -1, 0], [-1, 1, 0], [1, -2, 0], [-2, 1, 0]\}$ and then tested their responses on 101st episode on a novel weight combination $w_{test} = [1, 1, 1]$. They refined the experiment 3 times, and here we show the final iteration, which gathers support for the SF&GPI strategy, which predicts the final state 12 to be chosen on the 101st episode. However, it is important to note that only 60% of the participants managed to learn their tasks and produce average rewards greater than 0, and the rest were excluded. We find that the soft maximum composition of soft Q-learning results in favouring the final state 9 (same as UVFA, which generalises in values), and partly also 7 (same as MB) for $\tau < 1$ across all of their experiment iterations (here showing for the final one). Since weights cannot be negative in MO Soft Q-learning, the loss was included in the resource quantities (ϕ), composed with the absolute value of the weights for composition. Figures adapted from Tomov et al. (2021) with permission from Springer Nature.

C

Appendix for Chapter 6

Contents

C.1 Our mathematical model complements the Fear-Avoidance model of pain chronification

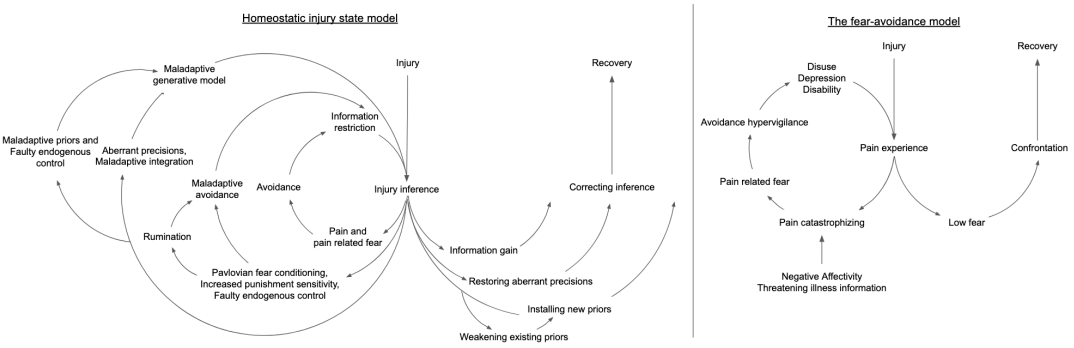


Figure C.1: Comparison of predictions of our homeostatic injury state model to that of the Fear Avoidance model and how they play a complementary role in educating patients about their pain.

D

Appendix for Chapter 7

Contents

D.1 Unifying view on dynamical systems vs optimal control

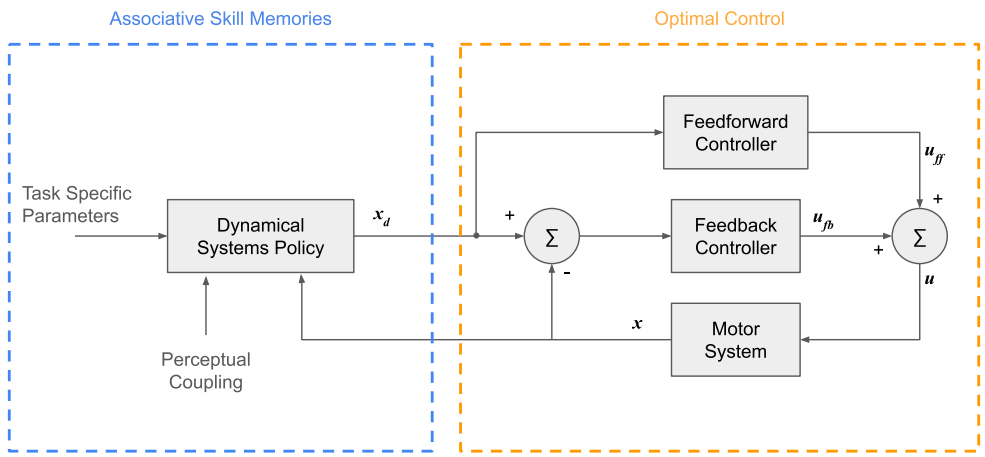


Figure D.1: Unifying view proposed by (Schaal et al., 2007), where the dynamic systems policy from Associative Skill Memories can employ an optimal control-based low-level controller.

D.2 Demonstration data used in skill memory expression task

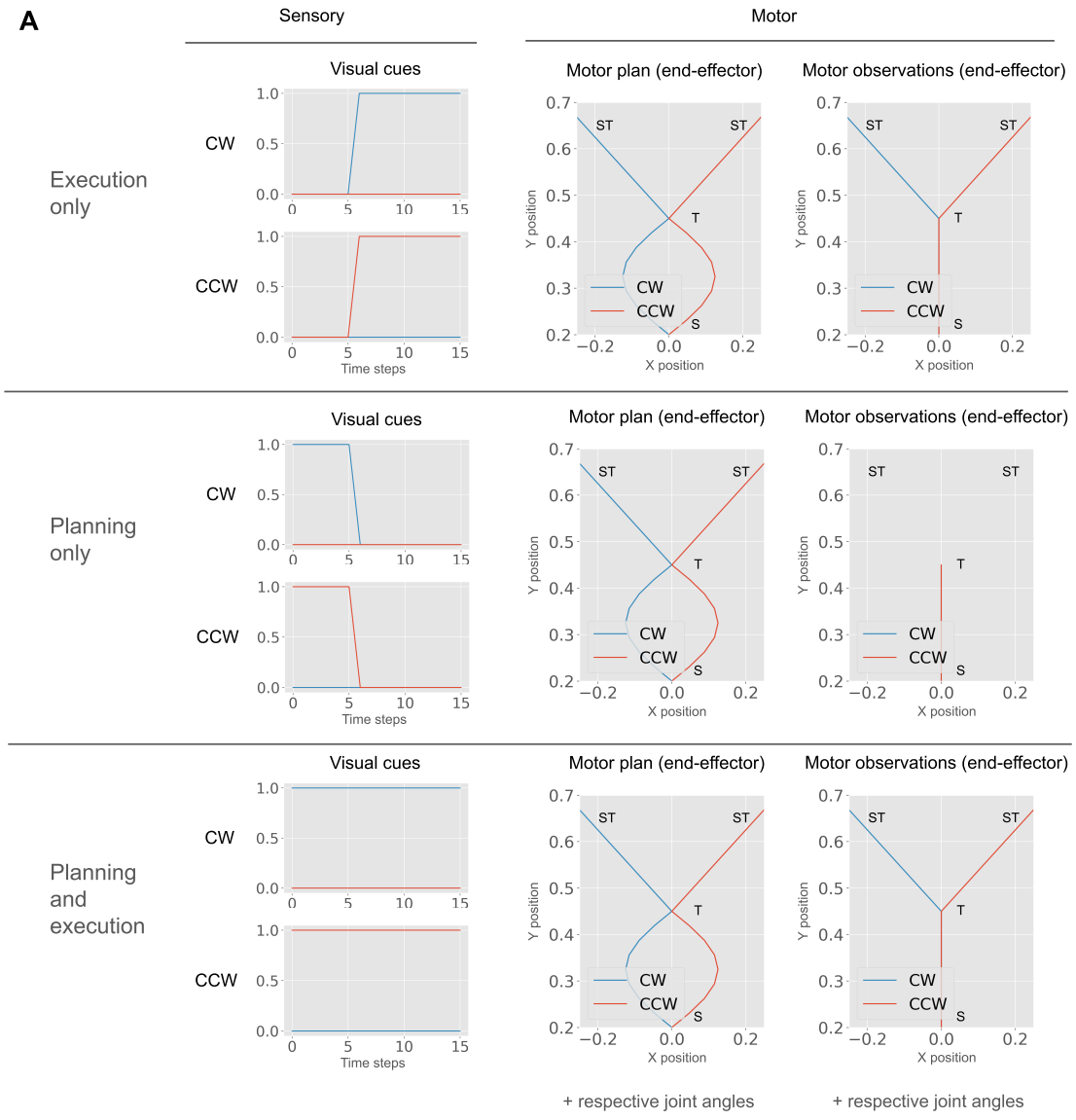


Figure D.2: Plots of the sensory and (hypothetically optimal) motor sequences used for demonstrations in robot experiments on skill memory expression inspired by Sheahan et al. (2016).

D.3 Additional results in the skill memory expression task

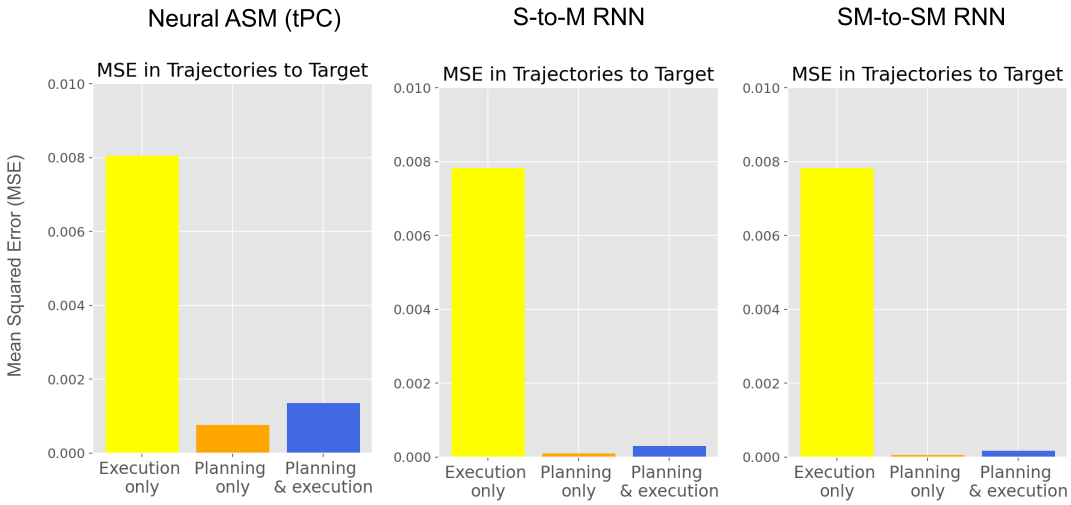


Figure D.3: Mean Squared Errors (MSE) in trajectories to the target as an alternate metric for quantifying the results in the skill memory expression task, inspired by Sheahan et al. (2016). MSE will show an inverse relationship to the DCD metric that we constructed or seen in the trajectory plots of Sheahan et al. (2016), i.e. failure to separate the skill memories results in low DCD and high MSE and vice versa.

D.4 Learning rate varies with number of skills

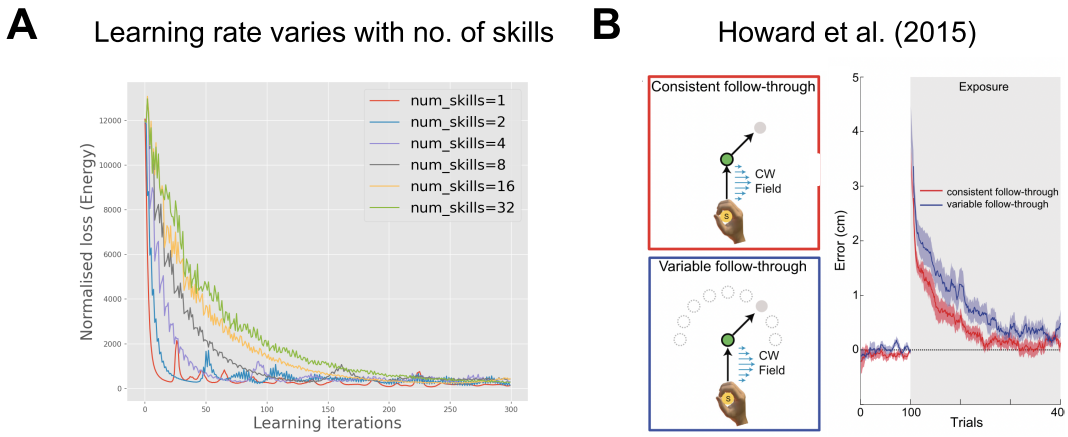


Figure D.4: Lesser the number of distinct skills to memorise, faster is the learning rate of our model, as seen in energy (normalised loss) over epochs. This potentially explains why consistent follow-throughs improve learning rates as variable follow-throughs can split the learning into different skill memories rather than a single memory, as observed by Howard et al. (2015). (Adapted from Howard et al. (2015), under CC BY 4.0 license. Figure was cropped and a title line was added.)

Bibliography

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

Ahmadreza Ahmadi and Jun Tani. How can a recurrent neurodynamic predictive coding model cope with fluctuation in temporal patterns? robotic experiments on imitative interaction. *Neural Networks*, 92:3–16, 2017.

Woo-Young Ahn, Nathaniel Haines, and Lei Zhang. Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hbayesdm package. *Computational Psychiatry (Cambridge, Mass.)*, 1:24, 2017.

Korleki Akiti, Iku Tsutsui-Kimura, Yudi Xie, Alexander Mathis, Jeffrey E Markowitz, Rockwell Anyoha, Sandeep Robert Datta, Mackenzie Weygandt Mathis, Naoshige Uchida, and Mitsuko Watabe-Uchida. Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron*, 110(22): 3789–3804, 2022.

Micah Allen, Andrew Levy, Thomas Parr, and Karl J Friston. In the body's eye: the computational anatomy of interoceptive inference. *PLoS Computational Biology*, 18(9): e1010490, 2022.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Ryunosuke Amo. Prediction error in dopamine neurons during associative learning. *Neuroscience Research*, 199:12–20, 2024.

Ryunosuke Amo, Sara Matias, Akihiro Yamanaka, Kenji F Tanaka, Naoshige Uchida, and Mitsuko Watabe-Uchida. A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nature neuroscience*, 25(8):1082–1092, 2022.

Deborah Antcliff, Philip Keeley, Malcolm Campbell, Steve Woby, and Linda McGowan. Exploring patients' opinions of activity pacing and a new activity pacing questionnaire for chronic pain and/or fatigue: a qualitative study. *Physiotherapy*, 102(3):300–307, 2016.

Arnoud Arntz, Michael Rauner, and Marcel Van den Hout. "if i feel anxious, there must be danger": Ex-consequentia reasoning in inferring danger in anxiety disorders. *Behaviour research and therapy*, 33(8):917–925, 1995.

Karl J Astrom et al. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

Maite Azcorra, Zachary Gaertner, Connor Davidson, Qianzi He, Hailey Kim, Shivathmihai Nagappan, Cooper K Hayes, Charu Ramakrishnan, Lief Fenno, Yoon Seok Kim, et al. Unique functional responses differentially map onto genetic subtypes of dopamine neurons. *Nature neuroscience*, 26(10):1762–1774, 2023.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29, 2016.

Benedicte M Babayan, Naoshige Uchida, and Samuel J Gershman. Belief state representation in the dopamine system. *Nature communications*, 9(1):1891, 2018.

Dominik R Bach and Peter Dayan. Algorithms for survival: a comparative perspective on emotions. *Nature Reviews Neuroscience*, 18(5):311–319, 2017.

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

Tali M Ball and Lisa A Gunaydin. Measuring maladaptive avoidance: from animal models to clinical anxiety. *Neuropsychopharmacology*, pages 1–9, 2022.

Kirsty Bannister and AH Dickenson. The plasticity of descending controls in pain: translational probing. *The Journal of physiology*, 595(13):4159–4166, 2017.

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1): 1–23, 2017.

Andrew G Barto and Ozgür Simsek. Intrinsic motivation for reinforcement learning systems. In *Proceedings of the thirteenth yale workshop on adaptive and learning systems*, pages 113–118. Yale University Press, New Haven, CO., 2005.

Allan I Basbaum and Howard L Fields. Endogenous pain control mechanisms: review and hypothesis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 4(5):451–462, 1978.

Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):3625, 2020.

Richard Bellman. On the theory of dynamic programming. *Proceedings of the national Academy of Sciences*, 38(8):716–719, 1952.

Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, June 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553380. URL `https://doi.org/10.1145/1553374.1553380`.

Benjamin A Berg, Geoffrey Schoenbaum, and Michael A McDannald. The dorsal raphe nucleus is integral to negative prediction errors in pavlovian fear. *European Journal of Neuroscience*, 40(7):3096–3101, 2014.

Rafal Bogacz. A tutorial on the free-energy framework for modelling perception and learning. *Journal of mathematical psychology*, 76:198–211, 2017.

Rafal Bogacz. Dopamine role in learning and action inference. *Elife*, 9, 2020.

Robert C Bolles. Species-specific defense reactions and avoidance learning. *Psychological review*, 77(1):32, 1970.

Robert C Bolles. The avoidance learning problem. In *Psychology of learning and motivation*, volume 6, pages 97–145. Elsevier, 1972.

Robert C Bolles and Michael S Fanselow. A perceptual-defensive-recuperative model of fear and pain. *Behavioral and Brain Sciences*, 3(2):291–301, 1980.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2):294–311, 2002.

Bart van den Broek, Wim Wiegerinck, and Hilbert Kappen. Risk Sensitive Path Integral Control, March 2012. URL `http://arxiv.org/abs/1203.3523`. arXiv:1203.3523 [cs].

Paul L Brown and Herbert M Jenkins. Auto-shaping of the pigeon's key-peck 1. *Journal of the experimental analysis of behavior*, 11(1):1–8, 1968.

Christian Büchel, Stephan Geuter, Christian Sprenger, and Falk Eippert. Placebo analgesia: a predictive coding perspective. *Neuron*, 81(6):1223–1239, 2014.

Christopher J Burke, Philippe N Tobler, Michelle Baddeley, and Wolfram Schultz. Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431–14436, 2010.

Regina M Carelli, Stephanie G Ijames, and Alison J Crumling. Evidence that separate neural circuits in the nucleus accumbens encode cocaine versus "natural"(water and food) reward. *Journal of Neuroscience*, 20(11):4255–4266, 2000.

James F Cavanagh, Ian Eisenberg, Marc Guitart-Masip, Quentin Huys, and Michael J Frank. Frontal theta overrides pavlovian learning biases. *Journal of Neuroscience*, 33 (19):8541–8548, 2013.

Guillermo A. Cecchi, Lejian Huang, Javeria Ali Hashmi, Marwan Baliki, María V. Centeno, Irina Rish, and A. Vania Apkarian. Predictive Dynamics of Human Pain Perception. *PLOS Computational Biology*, 8(10):e1002719, October 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002719. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002719`. Publisher: Public Library of Science.

Jing-Yu Chang, Patricia H Janak, and Donald J Woodward. Comparison of mesocorticolimbic neuronal responses during cocaine and heroin self-administration in freely moving rats. *Journal of Neuroscience*, 18(8):3098–3115, 1998.

Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 17, 2004.

George I Christopoulos, Philippe N Tobler, Peter Bossaerts, Raymond J Dolan, and Wolfram Schultz. Neural correlates of value, risk, and risk aversion contributing to decision making under risk. *Journal of Neuroscience*, 29(40):12574–12583, 2009.

Mark M Churchland, M Yu Byron, Stephen I Ryu, Gopal Santhanam, and Krishna V Shenoy. Neural variability in premotor cortex provides a signature of motor preparation. *Journal of Neuroscience*, 26(14):3697–3712, 2006a.

Mark M Churchland, Gopal Santhanam, and Krishna V Shenoy. Preparatory activity in premotor and motor cortex reflects the speed of the upcoming reach. *Journal of neurophysiology*, 96(6):3130–3146, 2006b.

Mark M Churchland, John P Cunningham, Matthew T Kaufman, Stephen I Ryu, and Krishna V Shenoy. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68(3):387–400, 2010.

Paul Cisek. Preparing for speed. focus on "preparatory activity in premotor and motor cortex reflects the speed of the upcoming reach". *Journal of neurophysiology*, 96(6): 2842–2843, 2006.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press USA, New York, 2015.

Michel-Pierre Coll, Zoey Walden, Pierre-Alexandre Bourgoin, Veronique Taylor, Pierre Rainville, Manon Robert, Dang Khoa Nguyen, Pierre Jolicoeur, and Mathieu Roy. Pain reflects the informational value of nociceptive inputs. *Pain*, 165(10):e115–e125, October 2024. ISSN 1872-6623. doi: 10.1097/j.pain.0000000000003254.

Sven Collette, Wolfgang M Pauli, Peter Bossaerts, and John O'Doherty. Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, 6: e29718, 2017.

Anne GE Collins and Michael J Frank. Opponent actor learning (opal): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological review*, 121(3):337, 2014.

Steve Collins, Andy Ruina, Russ Tedrake, and Martijn Wisse. Efficient bipedal robots based on passive-dynamic walkers. *Science*, 307(5712):1082–1085, 2005.

Jackson J Cone, Samantha M Fortin, Jenna A McHenry, Garret D Stuber, James E McCutcheon, and Mitchell F Roitman. Physiological state gates acquisition and expression of mesolimbic reward prediction signals. *Proceedings of the National Academy of Sciences*, 113(7):1943–1948, 2016.

Geert Crombez, Christopher Eccleston, Stefaan Van Damme, Johan WS Vlaeyen, and Paul Karoly. Fear-avoidance model of chronic pain: the next generation. *The Clinical journal of pain*, 28(6):475–483, 2012.

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Kelvin JA Davies. Adaptive homeostasis. *Molecular aspects of medicine*, 49:1–7, 2016.

Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

P Dayan, JP Roiser, and E Viding. The first steps on long marches: The costs of active observation. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine*, pages 213–228. Oxford University Press, 2020.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

Peter Dayan. "liking" as an early and editable draft of long-run affective value. *PLoS Biology*, 20(1):e3001476, 2022.

Peter Dayan and Laurence F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational neuroscience. MIT Press, Cambridge, Mass., 2001. ISBN 978-0-262-04199-7 978-0-262-54185-5.

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

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. *Advances in neural information processing systems*, 5, 1992.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

Peter Dayan, Yael Niv, Ben Seymour, and Nathaniel D Daw. The misbehavior of value and the discipline of the will. *Neural networks*, 19(8):1153–1160, 2006.

Kristopher De Asis, J Hernandez-Garcia, G Holland, and Richard Sutton. Multi-step reinforcement learning: A unifying algorithm. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Dominique Debanne. Information processing in the axon. *Nature Reviews Neuroscience*, 5(4):304–316, 2004.

Tim Denison and Martha J. Morrell. Neuromodulation in 2035: The Neurology Future Forecasting Series. *Neurology*, 98(2):65–72, January 2022. ISSN 1526-632X. doi: 10.1212/WNL.0000000000013061.

Simon Desch, Petra Schweinhardt, Ben Seymour, Herta Flor, and Susanne Becker. Endogenous modulation of pain relief: evidence for dopaminergic but not opioidergic involvement. *bioRxiv*, 2022.

Amir Dezfouli and Bernard W Balleine. Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35(7):1036–1051, 2012.

A Dickinson and BW Balleine. The role of learning in motivation in gallistel cr (ed.), stevens' handbook of experimental psychology (vol. 3, pp. 497–533), 2002.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Hayley M Dorfman and Samuel J Gershman. Controllability governs the balance between pavlovian and instrumental action selection. *Nature communications*, 10(1):1–8, 2019.

Kenji Doya. Canonical cortical circuits and the duality of bayesian inference and optimal control. *Current Opinion in Behavioral Sciences*, 41:160–167, 2021.

Alvin W Drake. *Observation of a Markov process through a noisy channel*. PhD thesis, Massachusetts Institute of Technology, 1962.

Zack Dulberg and Jonathan D Cohen. On the duality of pain and pleasure processing: Why two dimensions of valence may be better than one. *bioRxiv*, pages 2025–01, 2025.

Zack Dulberg, Rachit Dubey, Isabel M Berwian, and Jonathan D Cohen. Having multiple selves helps learning agents explore and adapt in complex changing worlds. *Proceedings of the National Academy of Sciences*, 120(28):e2221180120, 2023.

J. Dum and A. Herz. Endorphinergic modulation of neural reward systems indicated by behavioral changes. *Pharmacology Biochemistry and Behavior*, 21(2):259–266, August 1984. ISSN 0091-3057. doi: 10.1016/0091-3057(84)90224-7. URL `https://www.sciencedirect.com/science/article/pii/0091305784902247`.

Krishnamurthy Dvijotham and Emanuel Todorov. A unifying framework for linearly solvable control. *arXiv preprint arXiv:1202.3715*, 2012.

Stefan Elfwing and Ben Seymour. Parallel reward and punishment control in humans and robots: Safe reinforcement learning using the maxpain algorithm. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 140–147. IEEE, 2017.

Jens Ellrich and Hanns Christian Hopf. The r3 component of the blink reflex: normative data and application in spinal lesions. *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control*, 101(4):349–354, 1996.

Noémi Éltető, Dezső Nemeth, Karolina Janacsek, and Peter Dayan. Tracking human skill learning with a hierarchical bayesian sequence model. *PLoS Computational Biology*, 18 (11):e1009866, 2022.

Kazuki Enomoto, Naoyuki Matsumoto, Sadamu Nakai, Takemasa Satoh, Tatsuo K Sato, Yasumasa Ueda, Hitoshi Inokawa, Masahiko Haruno, and Minoru Kimura. Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proceedings of the National Academy of Sciences*, 108(37):15462–15467, 2011.

Neir Eshel, Michael Bukwich, Vinod Rao, Vivian Hemmelder, Ju Tian, and Naoshige Uchida. Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, 525(7568):243–246, 2015.

Dominic A Evans, A Vanessa Stempel, Ruben Vale, and Tiago Branco. Cognitive control of escape behaviour. *Trends in cognitive sciences*, 23(4):334–348, 2019.

Michael S Fanselow and Laurie S Lester. A functional behavioristic approach to aversively motivated behavior:: Predatory imminence as a determinant of the topography of defensive behavior. In *Evolution and learning*, pages 185–212. Psychology Press, 2013.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395, 2020.

Anatol G Feldman and Mindy F Levin. The origin and use of positional frames of reference in motor control. *Behavioral and brain sciences*, 18(4):723–744, 1995.

Alexander Fengler, Michael Joshua Frank, Krishn Bera, and Mads Lund Pedersen. Beyond drift diffusion models: Fitting a broad class of decision and rl models with hddm. *bioRxiv*, 2022.

EE Fetz. Are movement parameters recognizably coded in the activity of single neurons? *Behav Brain Sci*, 15:679–690, 1992.

Howard L Fields. A motivation-decision model of pain: the role of opioids. In *Proceedings of the 11th world congress on pain*, pages 449–459. IASP press Seattle, 2006.

Howard L Fields. How expectations influence pain. *Pain*, 159:S3–S10, 2018.

Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614): 1898–1902, 2003.

Mary-Ann Fitzcharles, Steven P Cohen, Daniel J Clauw, Geoffrey Littlejohn, Chie Usui, and Winfried Häuser. Nociplastic pain: towards an understanding of prevalent pain conditions. *The Lancet*, 397(10289):2098–2110, 2021.

Ida Landström Flink, Katja Boersma, and Steven J Linton. Pain catastrophizing as repetitive negative thinking: a development of the conceptualization. *Cognitive behaviour therapy*, 42(3):215–223, 2013.

Laura Fontanesi, Sebastian Gluth, Mikhail S Spektor, and Jörg Rieskamp. A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review*, 26(4):1099–1121, 2019.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

Karl J Friston, Jean Daunizeau, James Kilner, and Stefan J Kiebel. Action and behavior: a free-energy formulation. *Biological cybernetics*, 102(3):227–260, 2010.

Karl J Friston, A David Redish, and Joshua A Gordon. Computational nosology and precision psychiatry. *Computational Psychiatry (Cambridge, Mass.)*, 1:2, 2017.

MA Fullana, JE Dunsmoor, KRJ Schruers, HS Savage, DR Bach, and BJ Harrison. Human fear conditioning: From neuroscience to the clinic. *Behaviour research and therapy*, 124:103528, 2020.

Chris Gagne and Peter Dayan. Peril, prudence and planning as risk, avoidance and worry. *Journal of Mathematical Psychology*, 106:102617, 2022.

Christopher Gagne and Peter Dayan. Two steps to risk sensitivity. *Advances in Neural Information Processing Systems*, 34:22209–22220, 2021.

Francesca Gandolfo, Ferdinando A Mussa-Ivaldi, and Emilio Bizzi. Motor learning by field approximation. *Proceedings of the National Academy of Sciences*, 93(9): 3843–3846, 1996.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Javier García, Fern, and O Fernández. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. ISSN 1533-7928. URL `http://jmlr.org/papers/v16/garcia15a.html`.

Matthew PH Gardner, Geoffrey Schoenbaum, and Samuel J Gershman. Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, 285 (1891):20181645, 2018.

Chris Gaskett. Reinforcement learning under circumstances beyond its control. 2003.

Clement Gehring and Doina Precup. Smart exploration in reinforcement learning using absolute temporal difference errors. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1037–1044, 2013.

Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Samuel J Gershman. The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38(33):7193–7200, 2018.

Samuel J Gershman. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394, 2020.

Samuel J. Gershman. Just looking: The innocent eye in neuroscience. *Neuron*, 109(14): 2220–2223, July 2021. ISSN 08966273. doi: 10.1016/j.neuron.2021.05.022. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627321003755`.

Samuel J Gershman and Naoshige Uchida. Believing in dopamine. *Nature Reviews Neuroscience*, 20(11):703–714, 2019.

Samuel J Gershman, Marie-H Monfils, Kenneth A Norman, and Yael Niv. The computational nature of memory modification. *Elife*, 6:e23763, 2017.

Samuel J Gershman, Marc Guitart-Masip, and James F Cavanagh. Neural signatures of arbitration between pavlovian and instrumental action selection. *PLoS computational biology*, 17(2):e1008553, 2021.

Paul W Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(supplement_3):15647–15654, 2011.

Eveliina Glogan, Kristof Vandael, Rena Gatzounis, and Ann Meulders. When do we not face our fears? investigating the boundary conditions of costly pain-related avoidance generalization. *The Journal of Pain*, 2021.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. Compositional preference models for aligning lms. *arXiv preprint arXiv:2310.13011*, 2023.

Alan Gordon and Alon Ziv. *The way out: A revolutionary, scientifically proven approach to healing chronic pain*. Penguin, 2021.

Isobel Green, Ryunosuke Amo, and Mitsuko Watabe-Uchida. Shifting attention to orient or avoid: a unifying account of the tail of the striatum and its dopaminergic inputs. *Current Opinion in Behavioral Sciences*, 59:101441, 2024.

Francesca Greenstreet, Hernando Martinez Vergara, Yvonne Johansson, Sthitapranjya Pati, Laura Schwarz, Stephen C Lenzi, Jesse P Geerts, Matthew Wisdom, Alina Gubanova, Lars B Rollik, et al. Dopaminergic action prediction errors serve as a value-free teaching signal. *Nature*, pages 1–10, 2025.

Marc Guitart-Masip, Quentin JM Huys, Lluis Fuentemilla, Peter Dayan, Emrah Duzel, and Raymond J Dolan. Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, 62(1):154–166, 2012.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

Tuomas Haarnoja, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. Composable deep reinforcement learning for robotic manipulation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6244–6251. IEEE, 2018.

Robert M Hardwick, Alexander D Forrence, John W Krakauer, and Adrian M Haith. Time-dependent competition between goal-directed and habitual response preparation. *Nature human behaviour*, 3(12):1252–1262, 2019.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

Milos Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of artificial intelligence research*, 13:33–94, 2000.

Ioannis Havoutis and Sylvain Calinon. Learning from demonstration for semi-autonomous teleoperation. *Autonomous Robots*, 43:713–726, 2019.

Matthias Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pages 105–111. Elsevier, 1994.

Jay A Hennig, Sandra A Romero Pinto, Takahiro Yamaguchi, Scott W Linderman, Naoshige Uchida, and Samuel J Gershman. Emergence of belief-like representations through reinforcement learning. *PLOS Computational Biology*, 19(9):e1011067, 2023.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Marion Höfle, Michael Hauck, Andreas K Engel, and Daniel Senkowski. Pain processing in multisensory environments. *e-Neuroforum*, 1:23–28, 2010.

Jon C Horvitz, Tripp Stewart, and Barry L Jacobs. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain research*, 759 (2):251–258, 1997.

R. Hoskin, C. Berzuini, D. Acosta-Kane, W. El-Deredy, H. Guo, and D. Talmi. Sensitivity to pain expectations: A Bayesian model of individual differences. *Cognition*, 182:127–139, January 2019. ISSN 1873-7838. doi: 10.1016/j.cognition.2018.08.022.

Ian S Howard, James N Ingram, David W Franklin, and Daniel M Wolpert. Gone in 0.6 seconds: the encoding of motor memories depends on recent sensorimotor states. *Journal of Neuroscience*, 32(37):12756–12768, 2012.

Ian S Howard, Daniel M Wolpert, and David W Franklin. The effect of contextual cues on the encoding of motor memories. *Journal of neurophysiology*, 109(10):2632–2644, 2013.

Ian S Howard, Daniel M Wolpert, and David W Franklin. The value of the follow-through derives from motor learning depending on future actions. *Current Biology*, 25(3):397–401, 2015.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

Mark W Howe, Patrick L Tierney, Stefan G Sandberg, Paul EM Phillips, and Ann M Graybiel. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *nature*, 500(7464):575–579, 2013.

Oliver J Hulme, Tobias Morville, and Boris Gutkin. Neurocomputational theories of homeostatic control. *Physics of life reviews*, 31:214–232, 2019.

Quentin JM Huys, Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P Roiser. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3): e1002410, 2012.

Quentin JM Huys, Marc Guitart-Masip, Raymond J Dolan, and Peter Dayan. Decision-theoretic psychiatry. *Clinical Psychological Science*, 3(3):400–421, 2015.

Quentin JM Huys, Martin Gölzer, Eva Friedel, Andreas Heinz, Roshan Cools, Peter Dayan, and Raymond J Dolan. The specificity of pavlovian regulation is associated with recovery from depression. *Psychological medicine*, 46(5):1027–1035, 2016.

Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2):328–373, 2013.

Ryota Ishikawa, Genta Ono, and Jun Izawa. Bayesian surprise intensifies pain in a novel visual-noxious association. *Cognition*, 257:106064, April 2025. ISSN 0010-0277. doi: 10.1016/j.cognition.2025.106064. URL `https://www.sciencedirect.com/science/article/pii/S0010027725000046`.

Marc Jeannerod. The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain sciences*, 17(2):187–202, 1994.

Marieke Jepma, Leonie Koban, Johnny van Doorn, Matt Jones, and Tor D Wager. Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature human behaviour*, 2(11):838–855, 2018.

Joshua P. Johansen, Jason W. Tarpley, Joseph E. LeDoux, and Hugh T. Blair. Neural substrates for expectation-modulated fear learning in the amygdala and periaqueductal gray. *Nature Neuroscience*, 13(8):979–986, August 2010. ISSN 1546-1726. doi: 10.1038/nn.2594. URL `https://www.nature.com/articles/nn.2594`. Publisher: Nature Publishing Group.

Eric Jonas and Konrad Paul Kording. Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, 13(1):e1005268, January 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005268. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005268`. Publisher: Public Library of Science.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134, 1998.

Daniel Kahneman. Prospect theory: An analysis of decision under risk. *Econometrica*, 47 (2):263–291, 1979.

Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

Sham Kakade and Peter Dayan. Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6):549–559, 2002.

LJ Kamin, CJ Brimer, and AH Black. Conditioned suppression as a monitor of fear of the cs in the course of avoidance training. *Journal of comparative and physiological psychology*, 56(3):497, 1963.

Hilbert J Kappen. Linear theory for control of nonlinear stochastic systems. *Physical review letters*, 95(20):200201, 2005.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *Elife*, 3:e04811, 2014.

Eun Joo Kim, Omer Horovitz, Blake A Pellman, Lancy Mimi Tan, Qiuling Li, Gal Richter-Levin, and Jeansok J Kim. Dorsal periaqueductal gray-amygdala pathway conveys both innate and learned fear responses in rats. *Proceedings of the National Academy of Sciences*, 110(36):14795–14800, 2013.

Hyeonjin Kim, Jihyun K Hur, Mina Kwon, Soyeon Kim, Yoonseo Zoh, and Woo-Young Ahn. Causal role of the dorsolateral prefrontal cortex in modulating the balance between pavlovian and instrumental systems in the punishment domain. *Plos one*, 18 (6):e0286632, 2023.

HyungGoo R Kim, Athar N Malik, John G Mikhael, Pol Bech, Iku Tsutsui-Kimura, Fangmiao Sun, Yajun Zhang, Yulong Li, Mitsuko Watabe-Uchida, Samuel J Gershman, et al. A unified framework for dopamine signals across timescales. *Cell*, 183(6): 1600–1616, 2020.

Joohan Kim, Jorge E Esteves, Francesco Cerritelli, and Karl Friston. An active inference account of touch and verbal communication in therapy. *Frontiers in Psychology*, 13, 2022.

Jungwoo Kim, Suhwan Gim, Seng Bum Michael Yoo, and Choong-Wan Woo. A computational mechanism of cue-stimulus integration for pain in the brain. *Science Advances*, 10(37):eado8230, September 2024. doi: 10.1126/sciadv.ado8230. URL `https://www.science.org/doi/full/10.1126/sciadv.ado8230`. Publisher: American Association for the Advancement of Science.

Mykel J Kochenderfer, Tim A Wheeler, and Kyle H Wray. *Algorithms for decision making*. MIT press, 2022.

Kika Konstantinou, Kate M Dunn, Reuben Ogollah, Martyn Lewis, Danielle van der Windt, Elaine M Hay, ATLAS Study Team, et al. Prognosis of sciatica and back-related leg pain in primary care: the atlas cohort. *The Spine Journal*, 18(6): 1030–1040, 2018.

Wouter Kool, Samuel J Gershman, and Fiery A Cushman. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, 28(9): 1321–1333, 2017.

Timothy A Krausz, Alison E Comrie, Ari E Kahn, Loren M Frank, Nathaniel D Daw, and Joshua D Berke. Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron*, 111(21):3465–3478, 2023.

Kai A. Krueger and Peter Dayan. Flexible shaping: how learning in small steps helps. *Cognition*, 110(3):380–394, March 2009. ISSN 1873-7838. doi: 10.1016/j.cognition.2008.11.014.

Lea K Krugel, Guido Biele, Peter NC Mohr, Shu-Chen Li, and Hauke R Heekeren. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences*, 106(42): 17951–17956, 2009.

Talira Kucina, Lindsay Wells, Ian Lewis, Kristy de Salas, Amelia Kohl, Matthew A Palmer, James D Sauer, Dora Matzke, Eugene Aidman, and Andrew Heathcote. Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. *Nature Communications*, 14(1):2234, 2023.

Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E Borm, Simon RW Stott, Enrique M Toledo, J Carlos Villaescusa, et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell*, 167(2):566–580, 2016.

Kevin S LaBar, J Christopher Gatenby, John C Gore, Joseph E LeDoux, and Elizabeth A Phelps. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fmri study. *Neuron*, 20(5):937–945, 1998.

Armin Lak, William R Stauffer, and Wolfram Schultz. Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences*, 111(6):2343–2348, 2014.

Pablo Lanillos and Gordon Cheng. Adaptive robot body learning and estimation through predictive coding. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4083–4090. IEEE, 2018.

Lindsay C Laughlin, Danielle M Moloney, Shanna B Samels, Robert M Sears, and Christopher K Cain. Reducing shock imminence eliminates poor avoidance in rats. *Learning & Memory*, 27(7):270–274, 2020.

Jee Hang Lee, Su Yeon Heo, and Sang Wan Lee. Controlling human causal inference through *in silico* task design. *Cell Reports*, 43(2):113702, February 2024a. ISSN 2211-1247. doi: 10.1016/j.celrep.2024.113702. URL `https://www.sciencedirect.com/science/article/pii/S2211124724000305`.

Rachel S Lee, Yotam Sagiv, Ben Engelhard, Ilana B Witten, and Nathaniel D Daw. A feature-specific prediction error model explains dopaminergic heterogeneity. *Nature Neuroscience*, pages 1–13, 2024b.

Sang Wan Lee and Ben Seymour. Decision-making in brains and robots—the case for an interdisciplinary approach. *Current Opinion in Behavioral Sciences*, 26:137–145, 2019.

Sang Wan Lee, Shinsuke Shimojo, and John P O'Doherty. Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3): 687–699, 2014.

Sungwoo Lee, Younghyun Oh, Hyunhoe An, Hyebhin Yoon, Karl J Friston, Seok Jun Hong, and Choong-Wan Woo. Life-inspired interoceptive artificial intelligence for autonomous and adaptive agents. *arXiv preprint arXiv:2309.05999*, 2023.

Lucas Lehnert, Stefanie Tellex, and Michael L Littman. Advantages and limitations of using successor features for transfer in reinforcement learning. *arXiv preprint arXiv:1708.00102*, 2017.

Carl W Lejuez, Jennifer P Read, Christopher W Kahler, Jerry B Richards, Susan E Ramsey, Gregory L Stuart, David R Strong, and Richard A Brown. Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, 8(2):75, 2002.

Máté Lengyel and Peter Dayan. Hippocampal contributions to control: the third way. *Advances in neural information processing systems*, 20, 2007.

Daniel Levenstein, Veronica A. Alvarez, Asohan Amarasingham, Habiba Azab, Zhe S. Chen, Richard C. Gerkin, Andrea Hasenstaub, Ramakrishnan Iyer, Renaud B. Jolivet, Sarah Marzen, Joseph D. Monaco, Astrid A. Prinz, Salma Quraishi, Fidel Santamaria, Sabyasachi Shivkumar, Matthew F. Singh, Roger Traub, Farzan Nadim, Horacio G. Rotstein, and A. David Redish. On the Role of Theory and Modeling in Neuroscience. *Journal of Neuroscience*, 43(7):1074–1088, February 2023. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1179-22.2022. URL `https://www.jneurosci.org/content/43/7/1074`. Publisher: Society for Neuroscience Section: Viewpoints.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Jian Li, Daniela Schiller, Geoffrey Schoenbaum, Elizabeth A Phelps, and Nathaniel D Daw. Differential roles of human striatum and amygdala in associative learning. *Nature neuroscience*, 14(10):1250–1252, 2011.

T Ed Li, Mufeng Tang, and Rafal Bogacz. Predictive coding model detects novelty on different levels of representation hierarchy. *Neural computation*, 37(8):1373–1408, 2025.

Romain Ligneul, Zachary F Mainen, Verena Ly, and Roshan Cools. Stress-sensitive inference of task controllability. *Nature Human Behaviour*, 6(6):812–822, 2022.

Jack Lindsey and Ashok Litwin-Kumar. Action-modulated midbrain dopamine activity arises from distributed control policies. *arXiv preprint arXiv:2207.00636*, 2022.

Jack Lindsey, Jeffrey E Markowitz, Winthrop F Gillis, Sandeep Robert Datta, and Ashok Litwin-Kumar. Dynamics of striatal action selection and reinforcement learning. *bioRxiv*, 2024.

Yana Litovsky, George Loewenstein, Samantha Horn, and Christopher Y Olivola. Loss aversion, the endowment effect, and gain-loss framing shape preferences for noninstrumental information. *Proceedings of the National Academy of Sciences*, 119 (34):e2202700119, 2022.

Tomas Ljungberg, Paul Apicella, and Wolfram Schultz. Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of neurophysiology*, 67(1): 145–163, 1992.

William S Lovejoy. Computationally feasible bounds for partially observed markov decision processes. *Operations research*, 39(1):162–175, 1991.

Nicholas John Mackintosh. *Conditioning and associative learning*. Clarendon Press Oxford, 1983.

Pranav Mahajan, Veeky Baths, and Boris Gutkin. Doing what's not wanted: Conflict in incentives and misallocation of behavioural control can lead to drug-seeking despite adverse outcomes. *Addiction Neuroscience*, 8:100115, 2023.

Pranav Mahajan, Shuangyi Tong, Sang Wan Lee, and Ben Seymour. Balancing safety and efficiency in human decision making. *bioRxiv*, pages 2024–01, 2024.

Pranav Mahajan, Peter Dayan, and Ben Seymour. Homeostasis after injury: How intertwined inference and control underpin post-injury pain and behaviour. *bioRxiv*, pages 2025–02, 2025a.

Pranav Mahajan, Mufeng Tang, T Li, Ioannis Havoutis, and Ben Seymour. Neural associative skill memories for safer robotics and modelling human sensorimotor repertoires. *arXiv preprint arXiv:2505.09760*, 2025b.

Tiago V Maia. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & behavior*, 38(1):50–67, 2010.

Frederic Maire and Vadim Bulitko. Apprenticeship learning for initial value functions in reinforcement learning. *Planning and Learning in A Priori Unknown or Dynamic Domains*, page 23, 2005.

Flavia Mancini, Suyi Zhang, and Ben Seymour. Computational and neural mechanisms of statistical pain learning. *Nature Communications*, 13(1):6613, 2022.

Flavia Mancini, Pranav Mahajan, Anna á V Guttesen, Jakub Onysk, Ingrid Scholtes, Nicholas Shenker, Michael Lee, and Ben Seymour. Enhanced behavioural and neural sensitivity to punishments in chronic pain and fatigue. *Brain*, 148(6):2151–2162, 2025.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

Paul Masset, Pablo Tano, HyungGoo R Kim, Athar N Malik, Alexandre Pouget, and Naoshige Uchida. Multi-timescale reinforcement learning in the brain. *Nature*, pages 1–9, 2025.

Christoph Mathys et al. How could we get nosology from computation. *Comput Psychiatry New Perspect Ment Illn*, 20:121–38, 2016.

Masayuki Matsumoto and Okihide Hikosaka. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248):837–841, 2009.

Takazumi Matsumoto, Wataru Ohata, Fabien CY Benureau, and Jun Tani. Goal-directed planning and goal understanding by extended active inference: Evaluation through simulated and physical robot experiments. *Entropy*, 24(4):469, 2022.

Marcelo G Mattar and Nathaniel D Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018.

Zoe McParlin, Francesco Cerritelli, Giacomo Rossettini, Karl J Friston, and Jorge E Esteves. Therapeutic alliance as active inference: The role of therapeutic touch and biobehavioural synchrony in musculoskeletal care. *Frontiers in Behavioral Neuroscience*, page 224, 2022.

William Menegas, Benedicte M Babayan, Naoshige Uchida, and Mitsuko Watabe-Uchida. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *elife*, 6:e21886, 2017.

William Menegas, Korleki Akiti, Ryunosuke Amo, Naoshige Uchida, and Mitsuko Watabe-Uchida. Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature neuroscience*, 21(10):1421–1430, 2018.

Ann Meulders. From fear of movement-related pain and avoidance to chronic pain disability: a state-of-the-art review. *Current Opinion in Behavioral Sciences*, 26: 130–136, 2019.

Ann Meulders, Debora Vansteenwegen, and Johan WS Vlaeyen. The acquisition of fear of movement-related pain and associative learning: a novel pain-relevant human fear conditioning paradigm. *Pain*, 152(11):2460–2469, 2011.

Ann Meulders, Mathijs Franssen, Riet Fonteyne, and Johan WS Vlaeyen. Acquisition and extinction of operant pain-related avoidance behavior using a 3 degrees-of-freedom robotic arm. *Pain*, 157(5):1094–1104, 2016.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49(2-3):267–290, 2002.

John G Mikhael and Rafal Bogacz. Learning reward uncertainty in the basal ganglia. *PLoS computational biology*, 12(9):e1005062, 2016.

John G Mikhael, HyungGoo R Kim, Naoshige Uchida, and Samuel J Gershman. The role of state uncertainty in the dynamics of dopamine. *Current Biology*, 32(5):1077–1087, 2022.

Kevin J Miller, Amitai Shenhav, and Elliot A Ludvig. Habits without values. *Psychological review*, 126(2):292, 2019.

Beren Millidge, Alexander Tschantz, Anil K Seth, and Christopher L Buckley. On the relationship between active inference and control as inference. In *International workshop on active inference*, pages 3–11. Springer, 2020.

Beren Millidge, Yuhang Song, Armin Lak, Mark E Walton, and Rafal Bogacz. Reward bases: A simple mechanism for adaptive acquisition of multiple reward types. *PLOS Computational Biology*, 20(11):e1012580, 2024a.

Beren Millidge, Mufeng Tang, Mahyar Osanlouy, Nicol S. Harper, and Rafal Bogacz. Predictive coding networks for temporal prediction. *PLOS Computational Biology*, 20 (4):1–31, April 2024b. doi: 10.1371/journal.pcbi.1011183.

Kevin GC Mizes, Jack Lindsey, G Sean Escola, and Bence P Ölveczky. Dissociating the contributions of sensorimotor striatum to automatic and visually guided motor sequences. *Nature Neuroscience*, 26(10):1791–1804, 2023.

Kevin GC Mizes, Jack Lindsey, G Sean Escola, and Bence P Ölveczky. The role of motor cortex in motor sequence execution depends on demands for flexibility. *Nature Neuroscience*, pages 1–10, 2024.

Anahit Mkrtchian, Jessica Aylward, Peter Dayan, Jonathan P Roiser, and Oliver J Robinson. Modeling avoidance in mood and anxiety disorders using reinforcement learning. *Biological psychiatry*, 82(7):532–539, 2017a.

Anahit Mkrtchian, Jonathan P Roiser, and Oliver J Robinson. Threat of shock and aversive inhibition: Induced anxiety modulates pavlovian-instrumental interactions. *Journal of Experimental Psychology: General*, 146(12):1694, 2017b.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518 (7540):529–533, 2015.

Dean Mobbs, Drew B Headley, Weilun Ding, and Peter Dayan. Space, time, and fear: survival computations along defensive circuits. *Trends in cognitive sciences*, 24(3): 228–241, 2020.

Moritz Moeller, Sanjay Manohar, and Rafal Bogacz. Uncertainty–guided learning with scaled prediction errors in the basal ganglia. *PLoS computational biology*, 18(5): e1009816, 2022.

Vishwanathan Mohan, Ajaz Bhat, and Pietro Morasso. Muscleless motor synergies and actions without movements: From motor neuroscience to cognitive robotics. *Physics of Life Reviews*, 30:89–111, 2019.

Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, Nathaniel Douglass Daw, and Samuel J Gershman. The successor representation in human reinforcement learning. *Nature human behaviour*, 1(9):680–692, 2017.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

Genela Morris, Alon Nevet, David Arkadir, Eilon Vaadia, and Hagai Bergman. Midbrain dopamine neurons encode decisions for future action. *Nature neuroscience*, 9(8): 1057–1063, 2006.

G Lorimer Moseley. A pain neuromatrix approach to patients with chronic pain. *Manual therapy*, 8(3):130–140, 2003.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. In *International Conference on Machine Learning*, pages 25303–25336. PMLR, 2023.

Michael Moutoussis, Richard P Bentall, Jonathan Williams, and Peter Dayan. A temporal difference account of avoidance learning. *Network: Computation in Neural Systems*, 19(2):137–160, 2008.

O Hobart Mowrer. Two-factor learning theory: summary and comment. *Psychological review*, 58(5):350, 1951.

Orval Mowrer. Learning theory and behavior. 1960.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287, 1999.

Ryu Nishimoto and Jun Tani. Learning to generate combinatorial action sequences utilizing the initial sensitivity of deterministic dynamical systems. *Neural Networks*, 17 (7):925–933, 2004.

Ryunosuke Nishimoto, Jun Namikawa, and Jun Tani. Learning multiple goal-directed actions through self-organization of a dynamic neural network model: A humanoid robot experiment. *Adaptive Behavior*, 16(2-3):166–181, 2008.

Yael Niv and Geoffrey Schoenbaum. Dialogues on prediction errors. *Trends in cognitive sciences*, 12(7):265–272, 2008.

Yael Niv, Nathaniel Daw, and Peter Dayan. How fast to work: Response vigor, motivation and tonic dopamine. *Advances in neural information processing systems*, 18, 2005.

Yael Niv, Nathaniel D Daw, and Peter Dayan. Choice values. *Nature neuroscience*, 9(8): 987–988, 2006.

Yael Niv, Nathaniel D Daw, Daphna Joel, and Peter Dayan. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3): 507–520, 2007.

Yael Niv, Jeffrey A Edlund, Peter Dayan, and John P O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562, 2012.

CL Nord, RP Lawson, Quentin JM Huys, S Pilling, and Jonathan P Roiser. Depression is associated with enhanced aversive pavlovian control over instrumental behaviour. *Scientific reports*, 8(1):1–10, 2018.

Peter J Norton and Daniel J Paulus. Transdiagnostic models of anxiety disorder: Theoretical and empirical underpinnings. *Clinical Psychology Review*, 56:122–137, 2017.

Gaspard Oliviers, Rafal Bogacz, and Alexander Meulemans. Learning probability distributions of sensory inputs with monte carlo predictive coding. *PLOS Computational Biology*, 20(10):e1012532, 2024.

Martin O'Neill and Wolfram Schultz. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron*, 68(4):789–800, 2010.

JB Overmier. Theories of instrumental learning. In *Animal learning: Survey and analysis*, pages 349–384. Springer, 1979.

Sindhu Padakandla, Prabuchandran KJ, and Shalabh Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.

Ronald Parr and Stuart Russell. Reinforcement learning with hierarchies of machines. *Advances in neural information processing systems*, 10, 1997.

Thomas D Parsons. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in human neuroscience*, 9:660, 2015.

Peter Pastor, Mrinal Kalakrishnan, Ludovic Righetti, and Stefan Schaal. Towards associative skill memories. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pages 309–315. IEEE, 2012.

Peter Pastor, Mrinal Kalakrishnan, Franziska Meier, Freek Stulp, Jonas Buchli, Evangelos Theodorou, and Stefan Schaal. From dynamic movement primitives to associative skill memories. *Robotics and Autonomous Systems*, 61(4):351–361, 2013.

Edward H Patzelt, Catherine A Hartley, and Samuel J Gershman. Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience*, 1:e18, 2018.

Mads Lund Pedersen, Michael J Frank, and Guido Biele. The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic bulletin & review*, 24(4):1234–1251, 2017.

Jennifer N Perusini and Michael S Fanselow. Neurobehavioral perspectives on the distinction between fear and anxiety. *Learning & Memory*, 22(9):417–425, 2015.

Mathias Pessiglione, Ben Seymour, Guillaume Flandin, Raymond J Dolan, and Chris D Frith. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045, 2006.

Corrado Pezzato, Mohamed Baioumy, Carlos Hernandez Corbato, Nick Hawes, Martijn Wisse, and Riccardo Ferrari. Active inference for fault tolerant control of robot manipulators with sensory faults. In *International Workshop on Active Inference*, pages 20–27. Springer, 2020.

Payam Piray and Nathaniel D Daw. Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature communications*, 12(1):4942, 2021.

Jean-Francois Poulin, Jian Zou, Janelle Drouin-Ouellet, Kwang-Youn A Kim, Francesca Cicchetti, and Rajeshwar B Awatramani. Defining midbrain dopaminergic neuron diversity by single-cell gene expression profiling. *Cell reports*, 9(3):930–943, 2014.

Jean-Francois Poulin, Giuliana Caronia, Caitlyn Hofer, Qiaoling Cui, Brandon Helm, Charu Ramakrishnan, C Savio Chan, Daniel A Dombeck, Karl Deisseroth, and Rajeshwar Awatramani. Mapping projections of molecularly defined dopamine neuron subtypes using intersectional genetic approaches. *Nature neuroscience*, 21(9): 1260–1271, 2018.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Charlotte Prévost, Mimi Liljeholm, Julian M Tyszka, and John P O'Doherty. Neural correlates of specific and general pavlovian-to-instrumental transfer within human amygdalar subregions: a high-resolution fmri study. *Journal of Neuroscience*, 32(24): 8383–8390, 2012.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1): 79–87, 1999.

Roger Ratcliff, Cynthia Huang-Pollock, and Gail McKoon. Modeling individual differences in the go/no-go task with a diffusion model. *Decision*, 5(1):42, 2018.

Ethan B Richman, Nicole Ticea, William E Allen, Karl Deisseroth, and Liqun Luo. Neural landscape diffusion resolves conflicts between needs across time. *Nature*, pages 1–9, 2023.

Mike JF Robinson and Kent C Berridge. Instant transformation of learned repulsion into motivational "wanting". *Current Biology*, 23(4):282–289, 2013.

Amit Rogel, Richard Savery, Ning Yang, and Gil Weinberg. Robogroove: Creating fluid motion for dancing robotic arms. In *Proceedings of the 8th International Conference on Movement and Computing*, pages 1–9, 2022.

Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32: 663–704, 2008.

Mathieu Roy, Daphna Shohamy, Nathaniel Daw, Marieke Jepma, G Elliott Wimmer, and Tor D Wager. Representation of aversive prediction errors in the human periaqueductal gray. *Nature neuroscience*, 17(11):1607–1612, 2014.

David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.

Evan M Russek, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, 13(9):e1005768, 2017.

Stuart J Russell and Andrew Zimdars. Q-decomposition for reinforcement learning agents. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 656–663, 2003.

Brian F Sadacca, Joshua L Jones, and Geoffrey Schoenbaum. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *elife*, 5:e13665, 2016.

Michael Sailer, Jan Ulrich Hense, Sarah Katharina Mayr, and Heinz Mandl. How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in human behavior*, 69:371–380, 2017.

Noor Sajid, Philip J Ball, Thomas Parr, and Karl J Friston. Active inference: demystified and compared. *Neural computation*, 33(3):674–712, 2021.

Tommaso Salvatori, Yuhang Song, Yujian Hong, Simon Frieder, Lei Sha, Zhenghua Xu, Rafal Bogacz, and Thomas Lukasiewicz. Associative memories via predictive coding, 2021.

Stefan Schaal, Peyman Mohajerian, and Auke Ijspeert. Dynamics systems vs. optimal control—a unifying view. *Progress in brain research*, 165:425–445, 2007.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

Wolfram Schultz. Multiple reward signals in the brain. *Nature reviews neuroscience*, 1(3): 199–207, 2000.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

Eric Schulz and Peter Dayan. Computational psychiatry for computers. *Iscience*, 23(12), 2020.

Anil K Seth. Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11):565–573, 2013.

Ben Seymour. Pain: a precision signal for reinforcement learning and control. *Neuron*, 101(6):1029–1041, 2019.

Ben Seymour and Flavia Mancini. Hierarchical models of pain: Inference, information-seeking, and adaptive control. *NeuroImage*, 222:117212, 2020.

Ben Seymour, John P. O'Doherty, Peter Dayan, Martin Koltzenburg, Anthony K. Jones, Raymond J. Dolan, Karl J. Friston, and Richard S. Frackowiak. Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667, June 2004. ISSN 1476-4687. doi: 10.1038/nature02581.

Ben Seymour, John P O'doherty, Martin Koltzenburg, Katja Wiech, Richard Frackowiak, Karl Friston, and Raymond Dolan. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature neuroscience*, 8(9):1234–1240, 2005.

Ben Seymour, Nathaniel Daw, Peter Dayan, Tania Singer, and Ray Dolan. Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, 27(18): 4826–4831, 2007.

Ben Seymour, Nathaniel D Daw, Jonathan P Roiser, Peter Dayan, and Ray Dolan. Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience*, 32(17):5833–5842, 2012.

Ben Seymour, RJ Dolan, et al. Emotion, motivation, and pain. *Textbook of Pain*, pages 248–255, 2013.

Ben Seymour, Robyn J Crook, and Zhe Sage Chen. Post-injury pain and behaviour: a control theory perspective. *Nature Reviews Neuroscience*, 24(6):378–392, 2023a.

Ben Seymour, Robyn J. Crook, and Zhe Sage Chen. Post-injury pain and behaviour: a control theory perspective. *Nature Reviews Neuroscience*, 24(6):378–392, June 2023b. ISSN 1471-0048. doi: 10.1038/s41583-023-00699-5. URL `https://www.nature.com/articles/s41583-023-00699-5`. Publisher: Nature Publishing Group.

Reza Shadmehr, Maurice A Smith, and John W Krakauer. Error correction, sensory prediction, and adaptation in motor control. *Annual review of neuroscience*, 33(1): 89–108, 2010.

Tali Sharot. The optimism bias. *Current biology*, 21(23):R941–R945, 2011.

Hannah R Sheahan, David W Franklin, and Daniel M Wolpert. Motor planning, not execution, separates motor memories. *Neuron*, 92(4):773–779, 2016.

Tingke Shen and Peter Dayan. Risking your tail: Modeling individual differences in risk-sensitive exploration using bayes adaptive markov decision processes. *eLife*, 13, 2024.

Herbert A Simon et al. Theories of bounded rationality. *Decision and organization*, 1(1): 161–176, 1972.

Peter Simor, Zsofia Zavecz, Kata Horváth, Noémi Éltető, Csenge Török, Orsolya Pesthy, Ferenc Gombos, Karolina Janacsek, and Dezso Nemeth. Deconstructing procedural memory: Different learning trajectories and consolidation of sequence and statistical learning. *Frontiers in Psychology*, 9:2708, 2019.

Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, pages 2601–2606. Cognitive Science Society, 2009.

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Edward J Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978.

Edward Jay Sondik. *The optimal control of partially observable Markov processes*. Stanford University, 1971.

Yuhang Song, Beren Millidge, Tommaso Salvatori, Thomas Lukasiewicz, Zhenghua Xu, and Rafal Bogacz. Inferring neural activity before plasticity as a foundation for learning beyond backpropagation. *Nature neuroscience*, 27(2):348–358, 2024.

Margarida Sousa, Pawel Bujalski, Bruno F Cruz, Kenway Louie, Daniel C McNamee, and Joseph J Paton. A multidimensional distributional map of future reward in dopamine neurons. *Nature*, pages 1–9, 2025.

Juliana K Sporrer, Jack Brookes, Samson Hall, Sajjad Zabbah, Ulises Daniel Serratos Hernandez, and Dominik R Bach. Functional sophistication in human escape. *Iscience*, 26(11), 2023.

Klaas Enno Stephan, Will D. Penny, Jean Daunizeau, Rosalyn J. Moran, and Karl J. Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, July 2009. ISSN 10538119. doi: 10.1016/j.neuroimage.2009.03.025. URL `https://linkinghub.elsevier.com/retrieve/pii/S1053811909002638`.

Peter Sterling. Allostasis: a model of predictive regulation. *Physiology & behavior*, 106 (1):5–15, 2012.

Xin Sui, Peter Dayan, and Kevin Lloyd. Exploring optimal risk-sensitive behavior in the balloon analogue risk task (bart). In *Computational Psychiatry Conference 2023*, 2023.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768, 2011.

Richard S Sutton, Michael Bowling, and Patrick M Pilarski. The alberta plan for ai research. *arXiv preprint arXiv:2208.11173*, 2022.

Abby Tabor, Michael A Thacker, G Lorimer Moseley, and Konrad P Körding. Pain: a statistical account. *PLoS computational biology*, 13(1):e1005142, 2017.

Yuji K Takahashi, Hannah M Batchelor, Bing Liu, Akash Khanna, Marisela Morales, and Geoffrey Schoenbaum. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, 95(6):1395–1405, 2017.

Yuji K Takahashi, Thomas A Stalnaker, Lauren E Mueller, Sevan K Harootonian, Angela J Langdon, and Geoffrey Schoenbaum. Dopaminergic prediction errors in the ventral tegmental area reflect a multithreaded predictive model. *Nature Neuroscience*, 26(5):830–839, 2023.

Deborah Talmi, Ben Seymour, Peter Dayan, and Raymond J Dolan. Human pavlovian–instrumental transfer. *Journal of Neuroscience*, 28(2):360–368, 2008.

Aviv Tamar, Huan Xu, and Shie Mannor. Scaling up robust mdps by reinforcement learning. *arXiv preprint arXiv:1306.6189*, 2013.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Mufeng Tang, Helen Barron, and Rafal Bogacz. Learning grid cells by predictive coding. *arXiv preprint arXiv:2410.01022*, 2024a.

Mufeng Tang, Helen Barron, and Rafal Bogacz. Sequential memory with temporal predictive coding. *Advances in Neural Information Processing Systems*, 36, 2024b.

Jun Tani. Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(3):421–436, 1996.

Pablo Tano, Peter Dayan, and Alexandre Pouget. A local temporal difference code for distributional reinforcement learning. *Advances in neural information processing systems*, 33:13662–13673, 2020.

Beverly E Thorn and Kim E Dixon. Coping with chronic pain: A stress-appraisal coping model. In *Coping with chronic illness and disability: Theoretical, empirical, and clinical aspects*, pages 313–335. Springer, 2007.

E Thorndike. Biological lectures from the marine laboratory at woods' holl, usa, for 1899. *Nature*, 62:411, 1900.

Emanuel Todorov. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

Emanuel Todorov. General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pages 4286–4292. IEEE, 2008.

Emanuel Todorov. Compositionality of optimal control laws. *Advances in neural information processing systems*, 22, 2009a.

Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009b.

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

Momchil S Tomov, Eric Schulz, and Samuel J Gershman. Multi-task reinforcement learning in humans. *Nature Human Behaviour*, 5(6):764–773, 2021.

Alexander Tschantz, Laura Barca, Domenico Maisto, Christopher L Buckley, Anil K Seth, and Giovanni Pezzulo. Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference. *Biological Psychology*, 169:108266, 2022.

Iku Tsutsui-Kimura, Zhiyu Melissa Tian, Ryunosuke Amo, Yizhou Zhuo, Yulong Li, Malcolm G Campbell, Naoshige Uchida, and Mitsuko Watabe-Uchida. Dopamine in the tail of the striatum facilitates avoidance in threat–reward conflicts. *Nature Neuroscience*, pages 1–16, 2025.

Dennis C Turk and Thomas E Rudy. Cognitive factors and persistent pain: A glimpse into pandora's box. *Cognitive therapy and research*, 16(2):99–122, 1992.

Athina Tzovara, Christoph W Korn, and Dominik R Bach. Human pavlovian fear conditioning conforms to probabilistic learning. *PLoS computational biology*, 14(8): e1006243, 2018.

Gonzalo P Urcelay. A psychological mechanism for the growth of anxiety. 2024.

Stefaan Van Damme, Geert Crombez, and Chris Eccleston. Retarded disengagement from pain cues: the effects of pain catastrophizing and pain expectancy. *Pain*, 100 (1-2):111–118, 2002.

Stefaan Van Damme, Geert Crombez, and Christopher Eccleston. Disengagement from pain: the role of catastrophic thinking about pain. *Pain*, 107(1):70–76, 2004.

Benjamin Van Niekerk, Steven James, Adam Earle, and Benjamin Rosman. Composing value functions in reinforcement learning. In *International conference on machine learning*, pages 6401–6409. PMLR, 2019.

Maaike MH van Swieten and Rafal Bogacz. Modeling the effects of motivation on choice and learning in the basal ganglia. *PLoS Computational Biology*, 16(5):e1007465, 2020.

Christine M van Vliet, Ann Meulders, Linda MG Vancleef, and Johan WS Vlaeyen. The opportunity to avoid pain may paradoxically increase fear. *The Journal of Pain*, 19 (10):1222–1230, 2018.

Christine M van Vliet, Ann Meulders, Linda MG Vancleef, Elke Meyers, and Johan WS Vlaeyen. Changes in pain-related fear and pain when avoidance behavior is no longer effective. *The Journal of Pain*, 21(3-4):494–505, 2020.

Christine M van Vliet, Ann Meulders, Linda MG Vancleef, and Johan WS Vlaeyen. Avoidance behaviour performed in the context of a novel, ambiguous movement increases threat and pain-related fear. *Pain*, 162(3):875–885, 2021.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017.

Johan WS Vlaeyen. Learning to predict and control harmful events: chronic pain and conditioning. *Pain*, 156:S86–S93, 2015.

Johan WS Vlaeyen and Geert Crombez. Behavioral conceptualization and treatment of chronic pain. *Annual review of clinical psychology*, 16:187–212, 2020.

Johan WS Vlaeyen and Steven J Linton. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. *Pain*, 85(3):317–332, 2000.

Johan WS Vlaeyen, Geert Crombez, and Steven J Linton. The fear-avoidance model of pain. *Pain*, 157(8):1588–1589, 2016.

Patrick CAJ Vroomen, MCTFM De Krom, and JA Knottnerus. Predicting the outcome of sciatica at short-term follow-up. *British Journal of General Practice*, 52(475): 119–123, 2002.

Patrick D Wall. Three phases of evil: the relation of injury to pain. *Brain and Mind*, 69: 293, 1979.

Edgar T Walters, Robyn J Crook, G Gregory Neely, Theodore J Price, and Ewan St John Smith. Persistent nociceptor hyperactivity as a painful evolutionary adaptation. *Trends in neurosciences*, 46(3):211–227, 2023.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018a.

Jiexin Wang, Stefan Elfwing, and Eiji Uchibe. Deep reinforcement learning by parallelizing reward and punishment using the maxpain architecture. In *2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 175–180. IEEE, 2018b.

Mitsuko Watabe-Uchida and Naoshige Uchida. Multiple dopamine systems: weal and woe of dopamine. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 83, pages 83–95. Cold Spring Harbor Laboratory Press, 2018.

Mitsuko Watabe-Uchida, Neir Eshel, and Naoshige Uchida. Neural circuitry of reward prediction error. *Annual review of neuroscience*, 40(1):373–394, 2017.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 29(5):1229–1262, 2017.

James CR Whittington and Rafal Bogacz. Theories of error back-propagation in the brain. *Trends in cognitive sciences*, 23(3):235–250, 2019.

Katja Wiech. Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. *Science*, 354(6312):584–587, 2016.

Eric Wiewiora. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.

Amanda C de C Williams. Persistence of pain in humans and other mammals. *Philosophical Transactions of the Royal Society B*, 374(1785):20190276, 2019.

David R Williams and Harriet Williams. Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement 2. *Journal of the experimental analysis of behavior*, 12(4):511–520, 1969.

Yumeya Yamamori, Oliver J Robinson, and Jonathan P Roiser. Approach-avoidance reinforcement learning as a translational and computational model of anxiety-related avoidance. *bioRxiv*, pages 2023–04, 2023.

Yuichi Yamashita and Jun Tani. Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS computational biology*, 4(11):e1000220, 2008.

Jonas Zaman, Katja Wiech, and Johan WS Vlaeyen. Perceptual decision parameters and their relation to self-reported pain: a drift diffusion account. *The Journal of Pain*, 21 (3-4):324–333, 2020.

Peter R Zambetti, Bryan P Schuessler, Bryce E Lecamp, Andrew Shin, Eun Joo Kim, and Jeansok J Kim. Ecological analysis of pavlovian fear conditioning in rats. *Communications Biology*, 5(1):1–11, 2022.

Ruixun Zhang, Thomas J Brennan, and Andrew W Lo. The origin of risk aversion. *Proceedings of the National Academy of Sciences*, 111(50):17777–17782, 2014.

Suyi Zhang, Hiroaki Mano, Gowrishankar Ganesh, Trevor Robbins, and Ben Seymour. Dissociable learning processes underlie human pain conditioning. *Current Biology*, 26 (1):52–58, 2016.

Suyi Zhang, Hiroaki Mano, Michael Lee, Wako Yoshida, Mitsuo Kawato, Trevor W Robbins, and Ben Seymour. The control of tonic pain by active relief learning. *Elife*, 7: e31949, 2018.

Zhewei Zhang, Kauê M Costa, Angela J Langdon, and Geoffrey Schoenbaum. The devilish details affecting tdrl models in dopamine research. *Trends in Cognitive Sciences*, 2025.

Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

Christopher A Zimmerman, Scott S Bolkan, Alejandro Pan-Vazquez, Bichan Wu, Emma F Keppler, Jordan B Meares-Garcia, Eartha Mae Guthman, Robert N Fetcho, Brenna McMannon, Junuk Lee, et al. A neural mechanism for learning from delayed postingestive feedback. *Nature*, pages 1–10, 2025.

Samuel Zorowitz, Ida Momennejad, and Nathaniel D Daw. Anxiety, avoidance, and sequential evaluation. *Computational Psychiatry*, 4:1–17, 2020.

Samuel Zorowitz, Gili Karni, Natalie Paredes, Nathaniel Daw, and Yael Niv. Improving the reliability of the pavlovian go/no-go task. 2023.