Balancing safety and efficiency in human decision making

Pranav Mahajan[1], Shuangyi Tong [2], Sang Wan Lee[3], and Ben Seymour[1,2]

[1]**Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford**
[2]**Institute of Biomedical Engineering, University of Oxford**
[3]**Department of Brain and Cognitive Sciences, Korea Advanced Institute of Science and Technology (KAIST)**
[3]**Kim Jaechul Graduate School of AI, KAIST**
[3]**KAIST Center for Neuroscience-inspired Artificial Intelligence**

Corresponding author:
Pranav Mahajan, Ben Seymour

Email address: pranav.mahajan@ndcn.ox.ac.uk, ben.seymour@ndcn.ox.ac.uk

ABSTRACT

The safety-efficiency dilemma describes the problem of maintaining safety during efficient exploration and is a special case of the exploration-exploitation dilemma in the face of potential dangers. Conventional exploration-exploitation solutions collapse punishment and reward into a single feedback signal, whereby early losses can be overcome by later gains. However, the brain has a separate system for Pavlovian fear learning, suggesting a possible computational advantage to maintaining a specific fear memory during exploratory decision-making. In a series of simulations, we show this promotes safe but efficient learning and is optimised by arbitrating Pavlovian avoidance of instrumental decision-making according to uncertainty. We provide a basic test of this model in a simple human approach-withdrawal experiment, and show that this flexible avoidance model captures choice and reaction times. These results show that the Pavlovian fear system has a more sophisticated role in decision-making than previously thought, by shaping flexible exploratory behaviour in a computationally precise manner.

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, lead to system failure 'death' before you ever get the chance to benefit through exploitation, emphasizing 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 paper 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.

RESULTS

Pavlovian avoidance learning (PAL) model

Our model consists of a Pavlovian punishment (fear) learning system and an integrated instrumental learning system (Fig. 1). The Pavlovian system learns punishment expectations for each stimulus/state, with the corresponding Pavlovian responses manifest as action propensities. 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 Rescorla-Wagner updating 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. The Pearce-Hall associability maintains a running average of absolute temporal difference errors (δ) as per equation 14 and the flexible ω is implemented using associability (see equation 15 in Methods). 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].

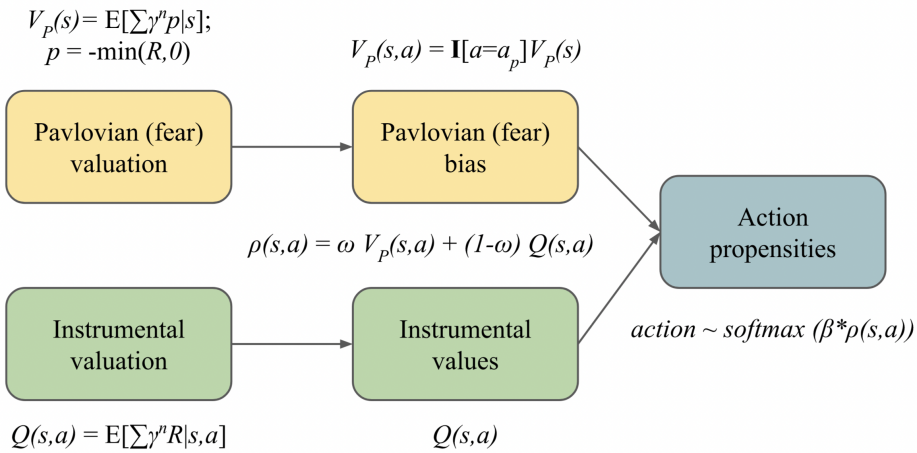

Figure 1. Pavlovian and Instrumental valuations are combined to arrive at action propensities used for (softmax) action selection. The Pavlovian bias influences protective behaviours through safer (Boltzmann) exploration and the arbitration between the Pavlovian and Instrumental systems is performed 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].

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. 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. 2B, and the Pavlovian state value is learned during the episode and shapes instrumental policy and value acquisition (Fig. 2C). Figure 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. 2C utilizes the flexible ω scheme utilized below.

Fig. 2D plots cumulative pain accrued over multiple episodes during learning and is our measure of safety. Fig. 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. 2D, 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 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. 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.

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 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. 2C) showing how the Pavlovian bias sculpts instrumental value acquisition.

In the supplementary methods, 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).

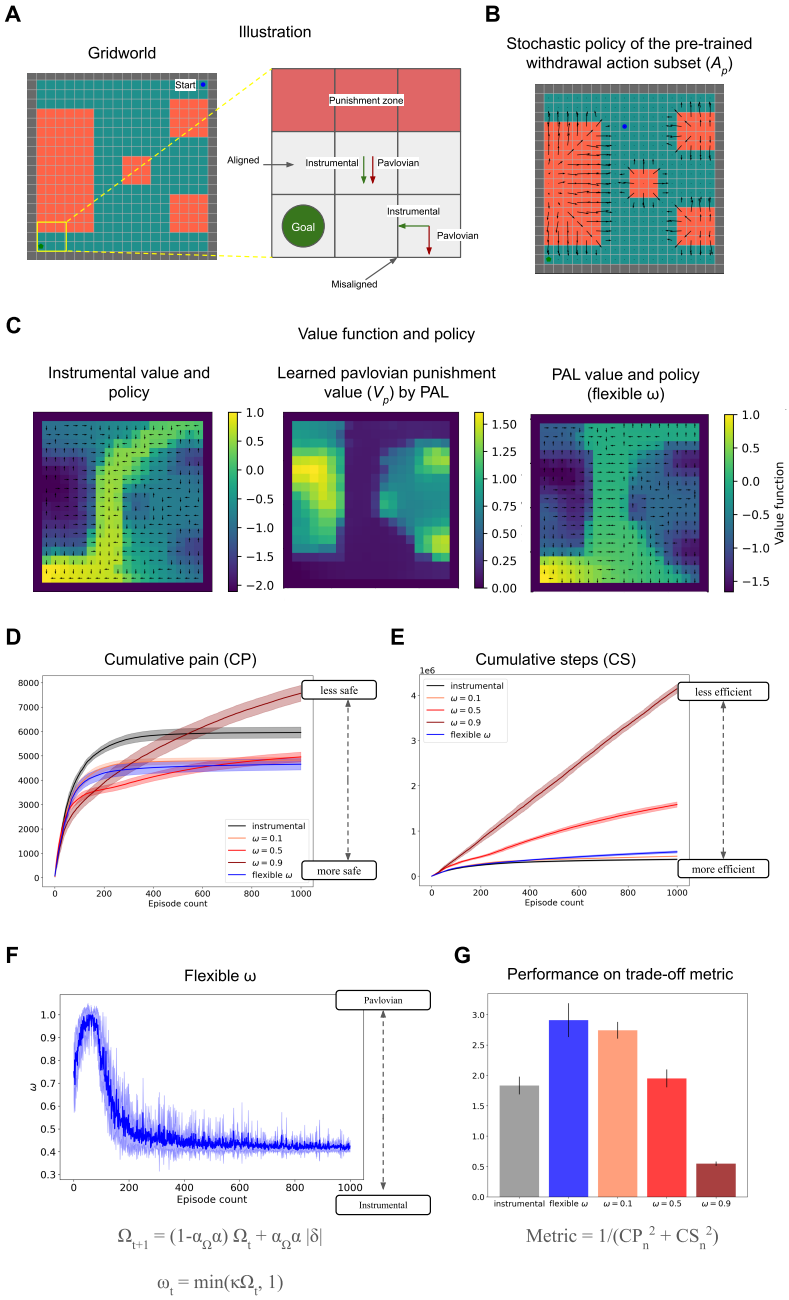

Figure 2. (A) Grid world environment with starting state in the top-right corner and rewarding goal state ($R = +1$) in the bottom-left corner and the red states are painful ($R = -0.1$). The grid world layout follows Gehring and Precup [2013]. Inset provides a didactic example of misalignment between 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. (C) The learned instrumental values and Pavlovian fear bias V_p(heatmap) and policy (arrows) 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. (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 16 and more details in Methods).

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

Observing the differences in the on-policy value functions with an without the Pavlovian influence (Fig. 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. 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. 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. 3B). Comparing the instrumental agent with other agents in Fig. 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 through Fig. 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. 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. 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. 3G) and attempts to restrict itself to pain-free states (Fig. 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. 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. 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.

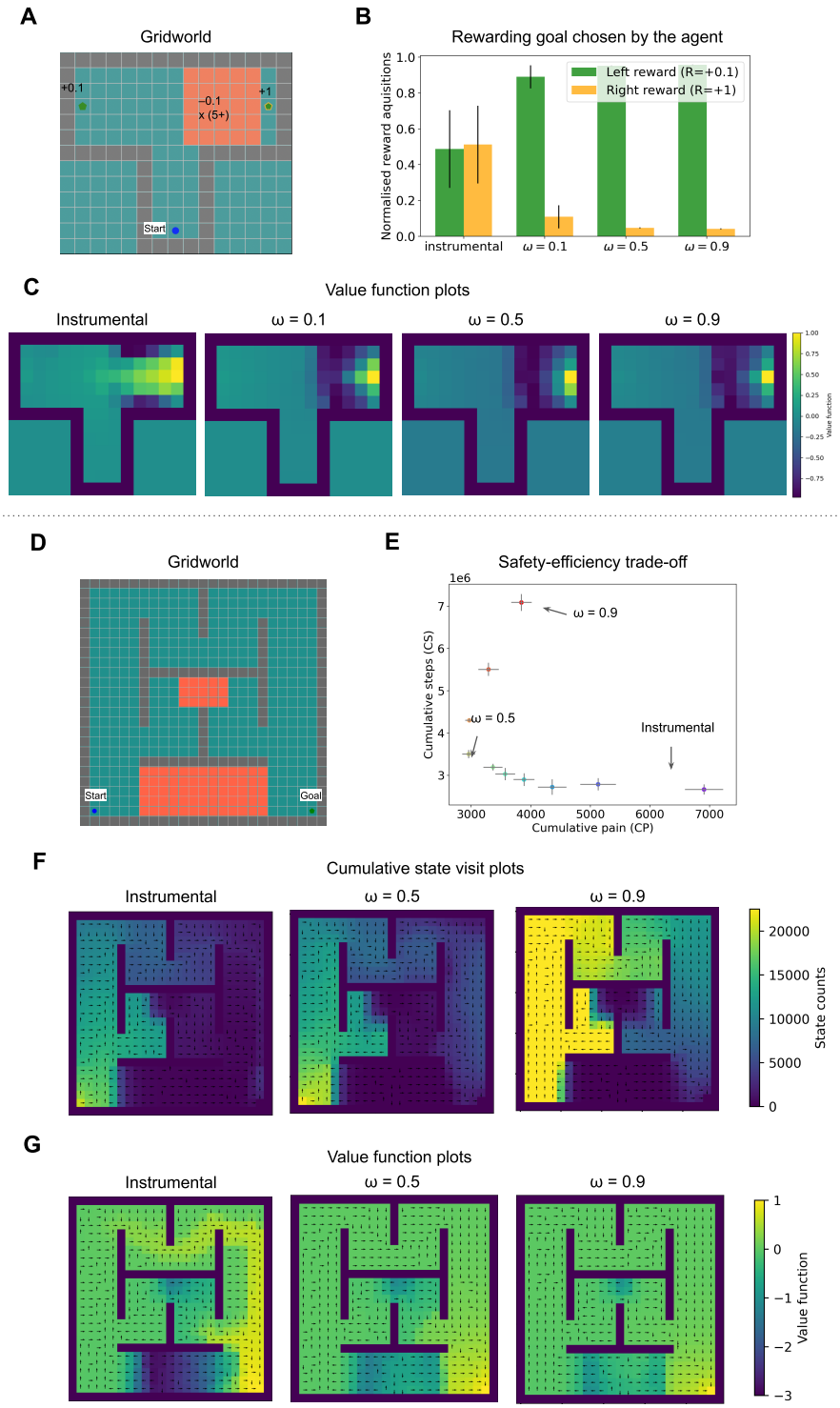

Figure 3. (A) T-maze grid world environment with annotated rewards and punishments. (B) Proportion of the rewarding goal chosen by the agent (C) Value function plots for $\omega = 0, 0.1, 0.5, 0.9$ shows diminished value propagation from the reward on the right (D) Grid world environment with three routes 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 $\omega = 0, 0.5, 0.9$.

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 the a Pavlovian bias in reaction times.

Fig. 4 describes the trial protocol (Fig.4A), block protocol (Fig.4B) and experimental setup (Fig. 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.

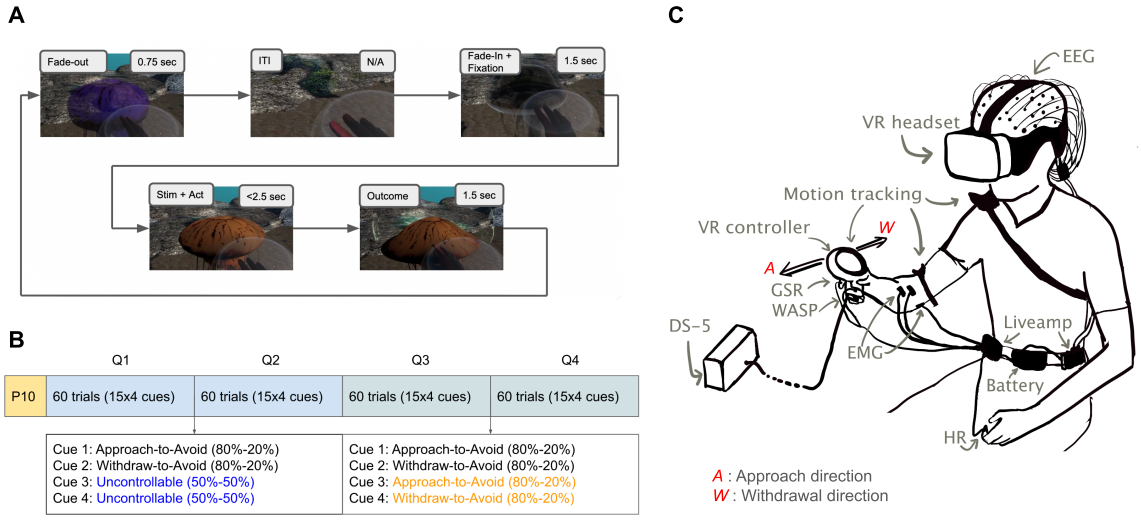

Figure 4. (A) Trial protocol: The participant is expected to take either an approach action (touch the jellyfish) or withdrawal action (withdraw the hand towards oneself) within the next 2.5 seconds once the jellyfish changes colour. The participant was requested to bring their hand at the centre of 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.5A) and RLDDM models Fig.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.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 number of approaches and withdrawal actions across all subjects and trials (Appendix A.5, Fig. 10A). 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. 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 includes both LOOIC and WAIC scores, showing the same result. Further, it can be seen that ω tracks associability, which decreases over the trials (Fig. 5C) (which also resembles Fig. 2E). Fig. 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. 5E) and flexible ω (Fig. 5F). Fig. 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. 5H) and that the best fitting model captures the withdrawal RT distribution well, but can be improved in the future to capture approach RT distribution better.

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 includes tables of model parameters with LOOIC and WAIC values for all RL models and RLDDM models.

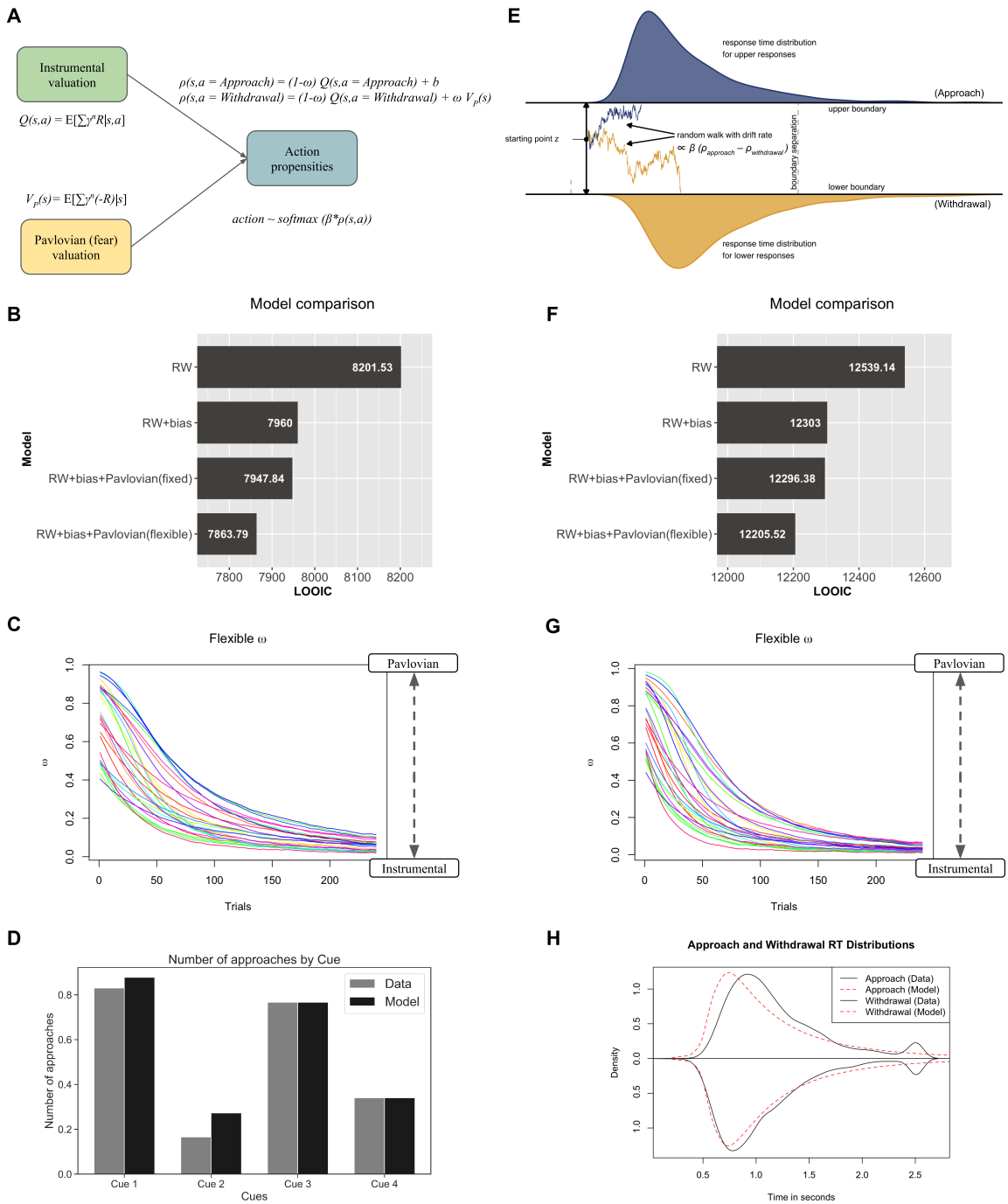

Figure 5. (A) Simplified model from Fig. 1 for the Approach-Withdrawal task (B) Model comparison shows that the model with flexible Pavlovian bias fits best in terms of LOOIC (C) Flexible ω from the RL model over 240 trials for 28 participants. (D) Number of approaches aggregated over all subjects and all trials in data and model predictions by the RL model with flexible ω. (normalized to 1) (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 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)

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 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] intrinsially 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.

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. 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 Pavlvoian 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.

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 and selects an action a according to its stochastic policy $\pi_t(s,a)$ (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 to the next state s' and the agent receives a scalar reward $R \in (-\infty, +\infty)$.

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:

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

And the optimal action-value function is defined as $Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a)$.

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 into pain (or punishment) $p \geq 0$:

$$p = -\min(R, 0), \tag{2}$$

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 :

$$V_p^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{\mathrm{k}=0}^{\infty} \gamma^{\mathrm{k}} p_{t+k|s_t=s}\right], \tag{3}$$

The subset of actions with the Pavlovian bias A_p are arrived at using a pretraining in the same environemnt with only punishments and random starting points. $V_p(s)$ then bias this pretrained subset of actions A_p according to equation 13.

The Pavlovian fear state-value functions are updated as follows:

$$V_p(s) := V_p(s) + \alpha(p + \gamma V_p(s') - V_p(s)) \tag{4}$$

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):

$$V(s) := V(s) + \alpha(R + \gamma V(s') - V(s)) \tag{5}$$

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

$$Q(s,a) := Q(s,a) + \alpha(\delta) \tag{6}$$

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

$$\delta = R + \gamma Q(s', \mathrm{argmax}_{a'}(Q(s',a'))) - Q(s,a) \tag{7}$$

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:

$$\rho(s,a) = Q(s,a) \tag{8}$$

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:

$$\pi(a|s) = \frac{e^{(\rho(s,a)/\tau)}}{\sum_{a' \in A} e^{(\rho(s,a')/\tau)}} \tag{9}$$

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:

$$\tau(i) = \frac{\tau_0}{1 + \tau_k i} \tag{10}$$

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., 2018]. 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:

$$\rho(s,a_n) = (1-\omega)Q(s,a_n); \text{ where } a_n \in A_n = A \setminus A_p. \tag{11}$$

$$\rho(s,a_p) = (1-\omega)Q(s,a_p) + \omega(V_p(s)); \text{ where } a_p \in A_p \tag{12}$$

The same can be compactly written as mentioned in the illustration (Fig. 1):

$$\rho(s,a) = (1-\omega)Q(s,a) + \omega(V_p(s,a)); \text{ where } V_p(s,a) = \mathbb{I}[a = a_p]V_p(s) \tag{13}$$

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 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 7) at each state using the following update rule:

$$\Omega_{t+1} = (1 - \alpha_\Omega * \alpha)\Omega_t + \alpha_\Omega * \alpha \mid \delta \mid \tag{14}$$

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:

$$\omega_t = \min(\kappa\Omega_t, 1) \tag{15}$$

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 supplementary materials 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:

$$\text{Trade-off metric} = \frac{1}{CP_n^2 + CS_n^2} \tag{16}$$

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). Participants from ages 18-60 were allowed to participate in the study. All subjects provided written informed consent for the experiment, which was approved by the local ethics board (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. 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 9 with free parameter $\beta = 1/\tau$ and $\beta > 0$. The learning rule for RW models was:

$$Q(s,a) := Q(s,a) + \alpha(R - Q(s,a)) \tag{17}$$

where $R = -1$ in case of electric shocks or $R = 0$ in case of neutral outcome. Punishment p can be defined as per equation 2 and thus $p = 1$ in case of electric shocks or $p = 0$ and the Pavlovian punishment value is calculated as

per:

$$V_p(s) := V_p(s) + \alpha(p - V_p(s)) \tag{18}$$

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

For RW+bias model,

$$\rho(s,a) = Q(s,a) + b; \text{ if } a = \text{Approach} \tag{19}$$

$$\rho(s,a) = Q(s,a); \text{ else} \tag{20}$$

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,

$$\rho(s,a) = (1-\omega)Q(s,a) + b; \text{ if } a = \text{Approach} \tag{21}$$

$$\rho(s,a) = (1-\omega)Q(s,a) + \omega(V_p(s)); \text{ if } a = \text{Withdrawal} \tag{22}$$

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 14 and 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.

$$v = \rho(s, a = \text{Approach}) - \rho(s, a = \text{Withdrawal}) \tag{23}$$

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 SCP 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.

ACKNOWLEDGEMENTS

PM would like to thank Michael Browning, Rafal Bogacz, Suyi Zhang, Charlie Yan, Maryna Alves Rosa Reges, Danielle Hewitt, Katja Wiech, the anonymous COSYNE 2022 reviewers, CCN 2023 reviewers and Science Advances reviewers for their feedback on earlier draft(s) of the subsections/extended abstracts of the manuscript. PM would like to thank Simon Desch for feedback on RLDDM fitting and Danielle Hewitt for suggestions and guidance on EEG data analysis. The work was funded by Wellcome Trust (214251/Z/18/Z, 203139/Z/16/Z and 203139/A/16/Z), IITP (MSIT 2019-0-01371) and JSPS (22H04998). This research was also partly supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

CODE AND DATA AVAILABILITY

Code and data: https://github.com/PranavMahajan25/Safety-Efficiency-Trade-off.

AUTHOR CONTRIBUTIONS

PM and BS designed the theory, simulation experiments and human VR experiments. PM programmed the grid world simulations, both VR tasks and analysed the data. PM collected the data for the VR Jellyfish task and ST and PM collected the data for the VR three-armed maze task. ST built the VR setup and provided support in troubleshooting any bugs. PM wrote the first draft which BS and PM heavily edited. SL provided feedback and theoretical insight on multiple iterations of the study. PM, ST, SL and BS edited and approved the manuscript. SL and BS acquired funding for the study.

REFERENCES

Ahn, W.-Y., Haines, N., and Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hbayesdm package. *Computational Psychiatry (Cambridge, Mass.)*, 1:24.

Altman, E. (1999). *Constrained Markov decision processes*, volume 7. CRC Press.

Bach, D. R. and Dayan, P. (2017). Algorithms for survival: a comparative perspective on emotions. *Nature Reviews Neuroscience*, 18(5):311–319.

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

Ball, T. M. and Gunaydin, L. A. (2022). Measuring maladaptive avoidance: from animal models to clinical anxiety. *Neuropsychopharmacology*, pages 1–9.

Berg, B. A., Schoenbaum, G., and McDannald, M. A. (2014). The dorsal raphe nucleus is integral to negative prediction errors in pavlovian fear. *European Journal of Neuroscience*, 40(7):3096–3101.

Bolles, R. C. (1970). Species-specific defense reactions and avoidance learning. *Psychological review*, 77(1):32.

Brown, P. L. and Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck 1. *Journal of the experimental analysis of behavior*, 11(1):1–8.

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., and Frank, M. J. (2013). Frontal theta overrides pavlovian learning biases. *Journal of Neuroscience*, 33(19):8541–8548.

Crombez, G., Eccleston, C., Van Damme, S., Vlaeyen, J. W., and Karoly, P. (2012). Fear-avoidance model of chronic pain: the next generation. *The Clinical journal of pain*, 28(6):475–483.

Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural networks*, 19(8):1153–1160.

Desch, S., Schweinhardt, P., Seymour, B., Flor, H., and Becker, S. (2022). Endogenous modulation of pain relief: evidence for dopaminergic but not opioidergic involvement. *bioRxiv*.

Dorfman, H. M. and Gershman, S. J. (2019). Controllability governs the balance between pavlovian and instrumental action selection. *Nature communications*, 10(1):1–8.

Elfwing, S. and Seymour, B. (2017). 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.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. (2020). Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. *Advances in Neural Information Processing Systems*, 33:22384–22395.

Fengler, A., Frank, M. J., Bera, K., and Pedersen, M. L. (2022). Beyond drift diffusion models: Fitting a broad class of decision and rl models with hddm. *bioRxiv*.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902.

Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review*, 26(4):1099–1121.

Fullana, M., Dunsmoor, J., Schruers, K., Savage, H., Bach, D., and Harrison, B. (2020). Human fear conditioning: From neuroscience to the clinic. *Behaviour research and therapy*, 124:103528.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Gaskett, C. (2003). Reinforcement learning under circumstances beyond its control.

Gehring, C. and Precup, D. (2013). 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.

Gershman, S. J., Guitart-Masip, M., and Cavanagh, J. F. (2021). Neural signatures of arbitration between pavlovian and instrumental action selection. *PLoS computational biology*, 17(2):e1008553.

Glogan, E., Vandael, K., Gatzounis, R., and Meulders, A. (2021). When do we not face our fears? investigating the boundary conditions of costly pain-related avoidance generalization. *The Journal of Pain*.

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., and Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, 62(1):154–166.

Heger, M. (1994). Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pages 105–111. Elsevier.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3):e1002410.

Huys, Q. J., Gölzer, M., Friedel, E., Heinz, A., Cools, R., Dayan, P., and Dolan, R. J. (2016). The specificity of pavlovian regulation is associated with recovery from depression. *Psychological medicine*, 46(5):1027–1035.

Kamin, L., Brimer, C., and Black, A. (1963). 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.

Kim, H., Hur, J. K., Kwon, M., Kim, S., Zoh, Y., and Ahn, W.-Y. (2023). 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.

Krugel, L. K., Biele, G., Mohr, P. N., Li, S.-C., and Heekeren, H. R. (2009). 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.

Kucina, T., Wells, L., Lewis, I., de Salas, K., Kohl, A., Palmer, M. A., Sauer, J. D., Matzke, D., Aidman, E., and Heathcote, A. (2023). Calibration of cognitive tests to address the reliability paradox for decision-conflict tasks. *Nature Communications*, 14(1):2234.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., and Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fmri study. *Neuron*, 20(5):937–945.

Laughlin, L. C., Moloney, D. M., Samels, S. B., Sears, R. M., and Cain, C. K. (2020). Reducing shock imminence eliminates poor avoidance in rats. *Learning & Memory*, 27(7):270–274.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., and Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature neuroscience*, 14(10):1250–1252.

Ligneul, R., Mainen, Z. F., Ly, V., and Cools, R. (2022). Stress-sensitive inference of task controllability. *Nature Human Behaviour*, 6(6):812–822.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Clarendon Press Oxford.

Maia, T. V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & behavior*, 38(1):50–67.

Meulders, A., Franssen, M., Fonteyne, R., and Vlaeyen, J. W. (2016). Acquisition and extinction of operant pain-related avoidance behavior using a 3 degrees-of-freedom robotic arm. *Pain*, 157(5):1094–1104.

Meulders, A., Vansteenwegen, D., and Vlaeyen, J. W. (2011). The acquisition of fear of movement-related pain and associative learning: a novel pain-relevant human fear conditioning paradigm. *Pain*, 152(11):2460–2469.

Mikhael, J. G. and Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS computational biology*, 12(9):e1005062.

Mkrtchian, A., Aylward, J., Dayan, P., Roiser, J. P., and Robinson, O. J. (2017a). Modeling avoidance in mood and anxiety disorders using reinforcement learning. *Biological psychiatry*, 82(7):532–539.

Mkrtchian, A., Roiser, J. P., and Robinson, O. J. (2017b). Threat of shock and aversive inhibition: Induced anxiety modulates pavlovian-instrumental interactions. *Journal of Experimental Psychology: General*, 146(12):1694.

Moeller, M., Manohar, S., and Bogacz, R. (2022). Uncertainty–guided learning with scaled prediction errors in the basal ganglia. *PLoS computational biology*, 18(5):e1009816.

Moskovitz, T., O'Donoghue, B., Veeriah, V., Flennerhag, S., Singh, S., and Zahavy, T. (2023). Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. In *International Conference on Machine Learning*, pages 25303–25336. PMLR.

Mowrer, O. (1960). Learning theory and behavior.

Mowrer, O. H. (1951). Two-factor learning theory: summary and comment. *Psychological review*, 58(5):350.

Nord, C., Lawson, R., Huys, Q. J., Pilling, S., and Roiser, J. P. (2018). Depression is associated with enhanced aversive pavlovian control over instrumental behaviour. *Scientific reports*, 8(1):1–10.

Norton, P. J. and Paulus, D. J. (2017). Transdiagnostic models of anxiety disorder: Theoretical and empirical underpinnings. *Clinical Psychology Review*, 56:122–137.

O'Neill, M. and Schultz, W. (2010). Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron*, 68(4):789–800.

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in human neuroscience*, 9:660.

Pedersen, M. L., Frank, M. J., and Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic bulletin & review*, 24(4):1234–1251.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042–1045.

Prévost, C., Liljeholm, M., Tyszka, J. M., and O'Doherty, J. P. (2012). 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.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., and Wager, T. D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. *Nature neuroscience*, 17(11):1607–1612.

Sailer, M., Hense, J. U., Mayr, S. K., and Mandl, H. (2017). 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.

Seymour, B., Daw, N., Dayan, P., Singer, T., and Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, 27(18):4826–4831.

Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., and Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience*, 32(17):5833–5842.

Sharot, T. (2011). The optimism bias. *Current biology*, 21(23):R941–R945.

Talmi, D., Seymour, B., Dayan, P., and Dolan, R. J. (2008). Human pavlovian–instrumental transfer. *Journal of Neuroscience*, 28(2):360–368.

Thorn, B. E. and Dixon, K. E. (2007). 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.

Turk, D. C. and Rudy, T. E. (1992). Cognitive factors and persistent pain: A glimpse into pandora's box. *Cognitive therapy and research*, 16(2):99–122.

Urcelay, G. P. (2024). A psychological mechanism for the growth of anxiety.

van Vliet, C. M., Meulders, A., Vancleef, L. M., Meyers, E., and Vlaeyen, J. W. (2020). Changes in pain-related

fear and pain when avoidance behavior is no longer effective. *The Journal of Pain*, 21(3-4):494–505.

van Vliet, C. M., Meulders, A., Vancleef, L. M., and Vlaeyen, J. W. (2021). Avoidance behaviour performed in the context of a novel, ambiguous movement increases threat and pain-related fear. *Pain*, 162(3):875–885.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432.

Vlaeyen, J. W. and Linton, S. J. (2000). Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art. *Pain*, 85(3):317–332.

Wang, J., Elfwing, S., and Uchibe, E. (2018). 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.

Watabe-Uchida, M. and Uchida, N. (2018). 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.

Yamamori, Y., Robinson, O. J., and Roiser, J. P. (2023). Approach-avoidance reinforcement learning as a translational and computational model of anxiety-related avoidance. *bioRxiv*, pages 2023–04.

Zhang, S., Mano, H., Ganesh, G., Robbins, T., and Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology*, 26(1):52–58.

Zhang, S., Mano, H., Lee, M., Yoshida, W., Kawato, M., Robbins, T. W., and Seymour, B. (2018). The control of tonic pain by active relief learning. *Elife*, 7:e31949.

Zorowitz, S., Karni, G., Paredes, N., Daw, N., and Niv, Y. (2023). Improving the reliability of the pavlovian go/no-go task.

Zorowitz, S., Momennejad, I., and Daw, N. D. (2020). Anxiety, avoidance, and sequential evaluation. *Computational Psychiatry*, 4:1–17.

APPENDIX

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

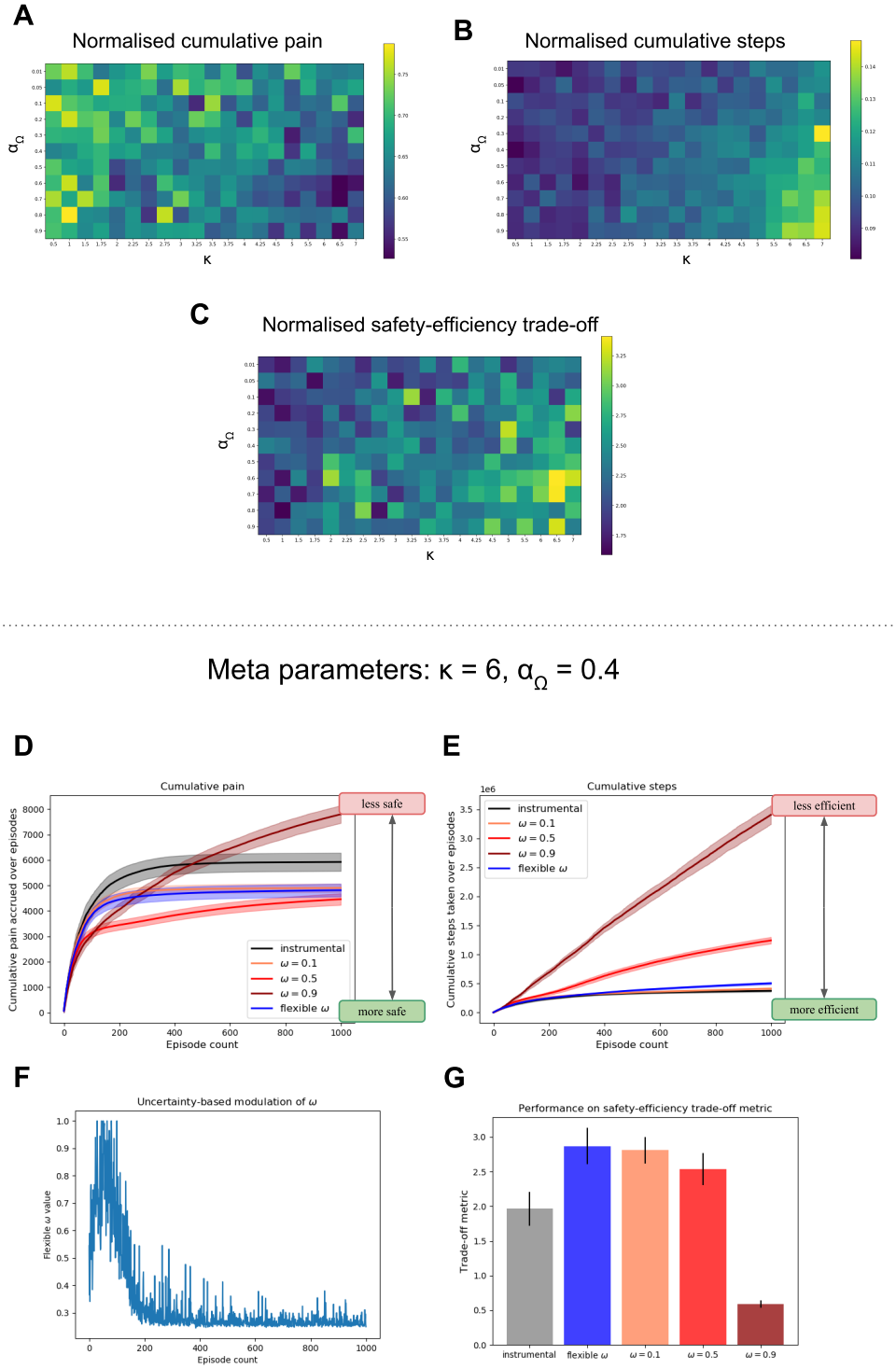

Figure 6. This figure shows the robustness of grid search for tuning the meta parameters for the associability-based ω in grid world simulations. We show that the results hold for a range of values close to the chosen meta-parameters. (A-C) Grid search results for the environment in Fig. 2 for varying 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.

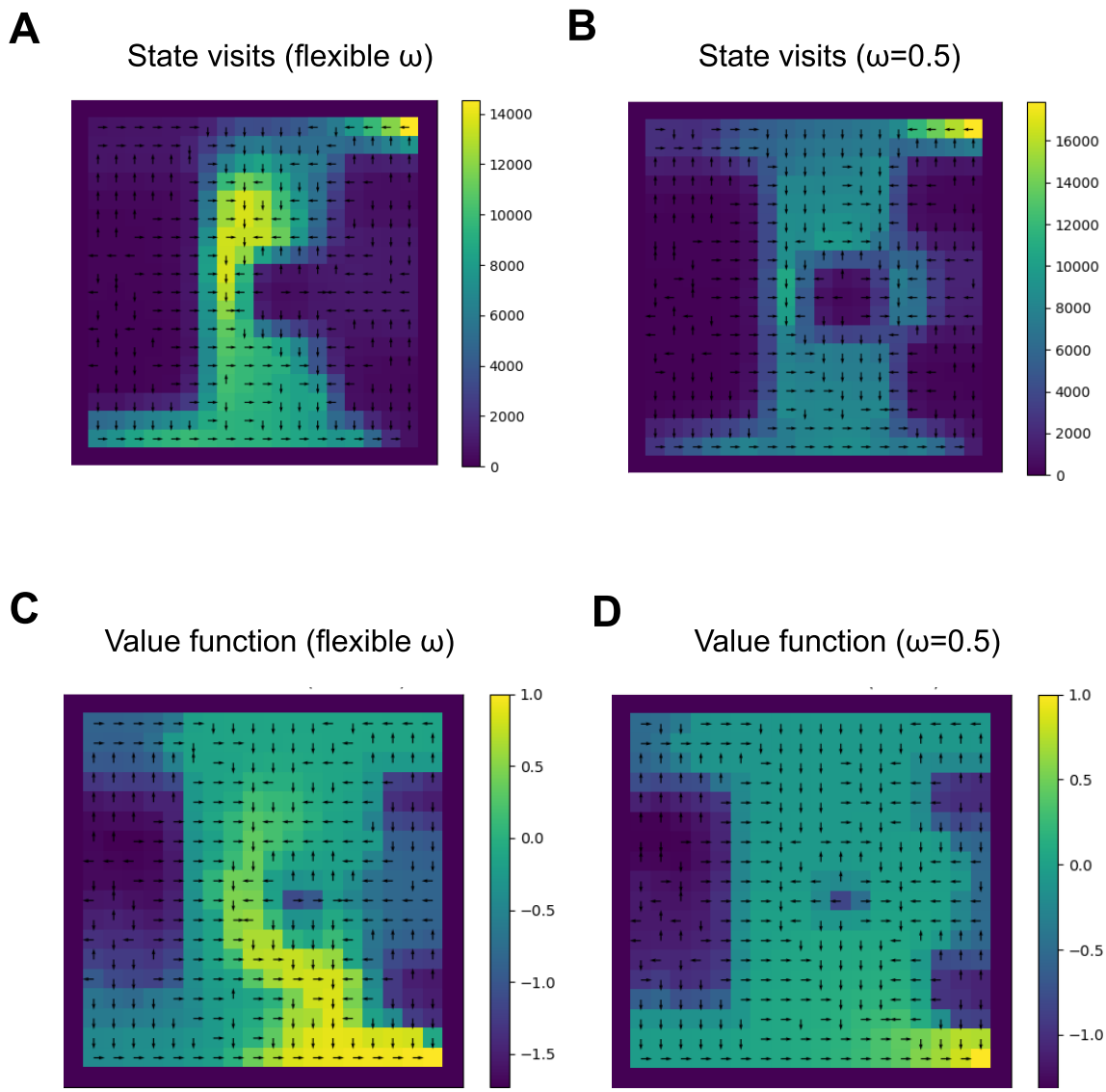

Figure 7. This figure shows cumulative state visit plots and value function plots of the flexible ω and fixed ω agents at the end of 1000 episodes when we relocate the reward goal from the bottom left corner (Fig. 2) to the bottom right corner on episode 500. Comparing state visit plots A & B and comparing 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

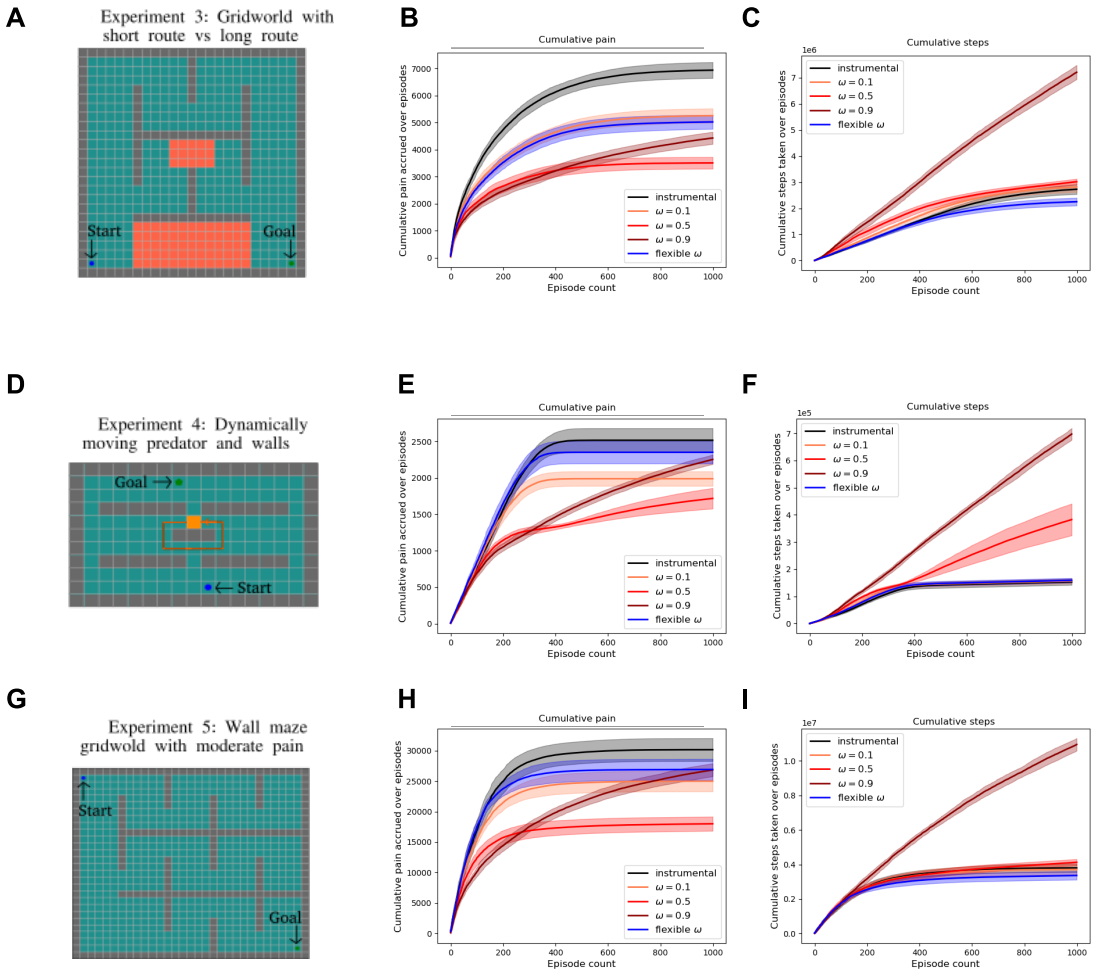

Figure 8. In this figure, we show the performance of fixed $\omega = 0.1, 0.5, 0.9$ and flexible ω agents on a range of grid world environments, namely (A) the three-route environment from Fig. 3, (D) an environment with a moving predator on routine path and (G) wall maze grid world from Elfwing and 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

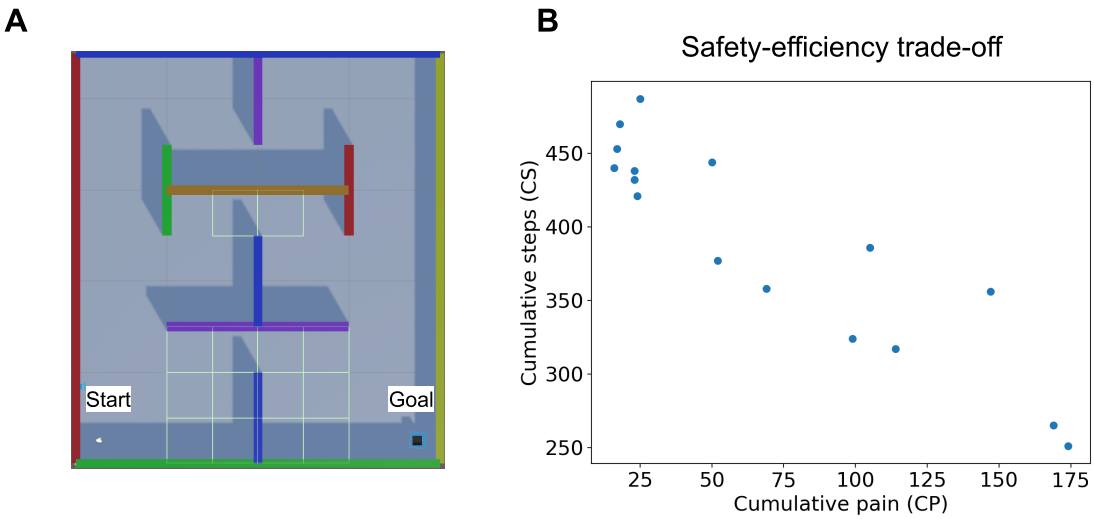

Figure 9. (A) Top-view of virtual reality (VR) maze with painful regions annotated by highlighted borders (B) Cumulative steps required to reach the goal vs cumulative pain acquired by participants over 20 episodes in the VR maze task. In this figure, we show the results of a VR maze replicating the 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. 10A) and we consider the few inactive approaches and inactive withdrawals due to timeout as approaches and withdrawals respectively (Fig. 10B).

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:

$$\text{Choice bias metric(cue X, cue Y)} = \%\text{ withdrawal choices on 'cue X'} - \%\text{ approach choices on 'cue Y'} \quad (24)$$

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. 10C). 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. 10E). 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. 10D & 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.

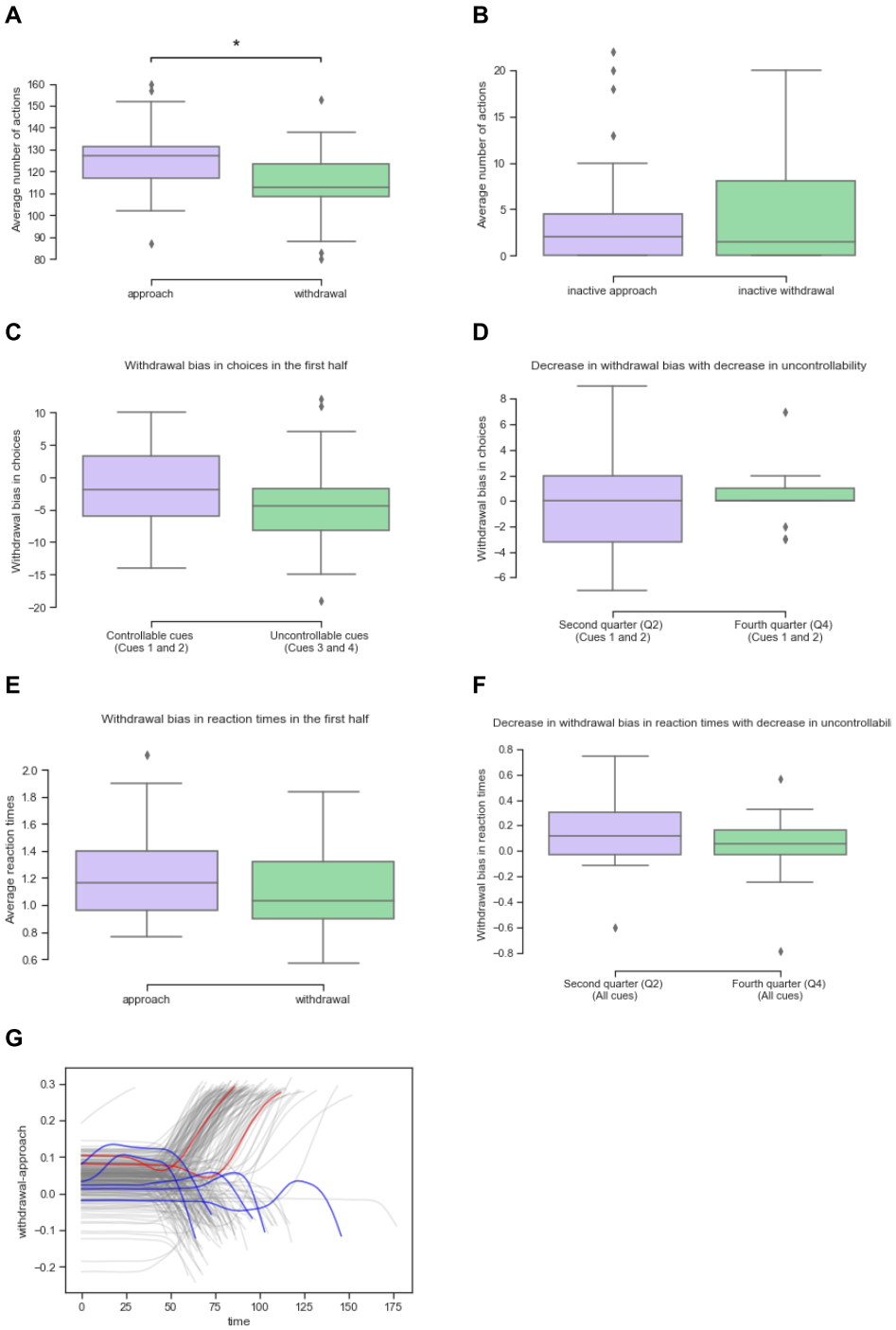

Figure 10. (A) Asymmetry in average approach and withdrawal responses over all subjects showing a baseline approach bias - MannwhitneyuResult(statistic=597.0, pvalue=0.0003981)(B) Incomplete approaches and incomplete withdrawals that were counted as approaches and withdrawals respectively. (C) Withdrawal bias in choices in the first half with uncontrollable cues - MannwhitneyuResult(statistic=492.5, pvalue=0.05037) (D) Decrease in withdrawal bias in choice with decrease in uncontrollability - MannwhitneyuResult(statistic=350.5, pvalue=0.76108) (E) Withdrawal bias in reaction times in the first half with uncontrollable cues - MannwhitneyuResult(statistic=475.0, pvalue=0.08820) (F) Decrease in withdrawal bias in reaction times with decrease in uncontrollability - MannwhitneyuResult(statistic=456.0, pvalue=0.14903) (G) Change-of-mind trials observed in motor data.

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

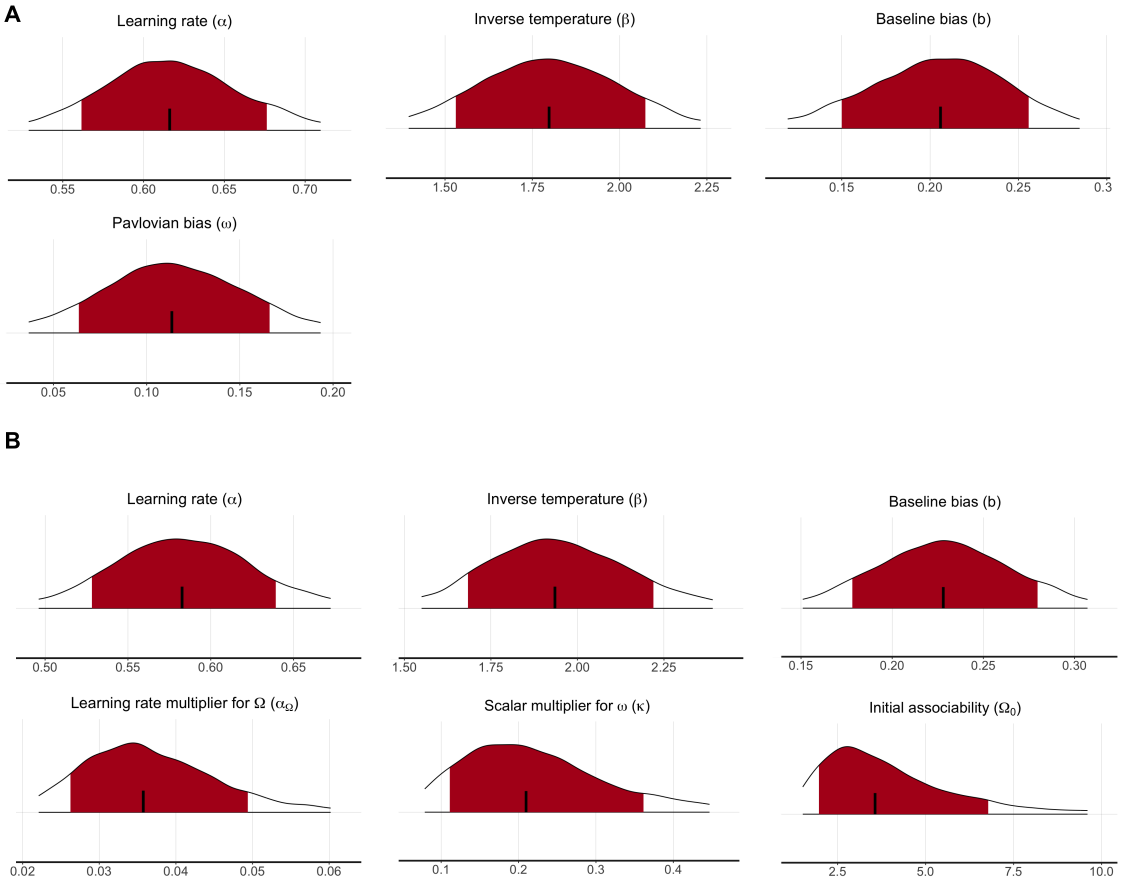

Figure 11. Group-level parameter distributions from (A) the RL model (M3) with fixed ω and (B) the RL model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

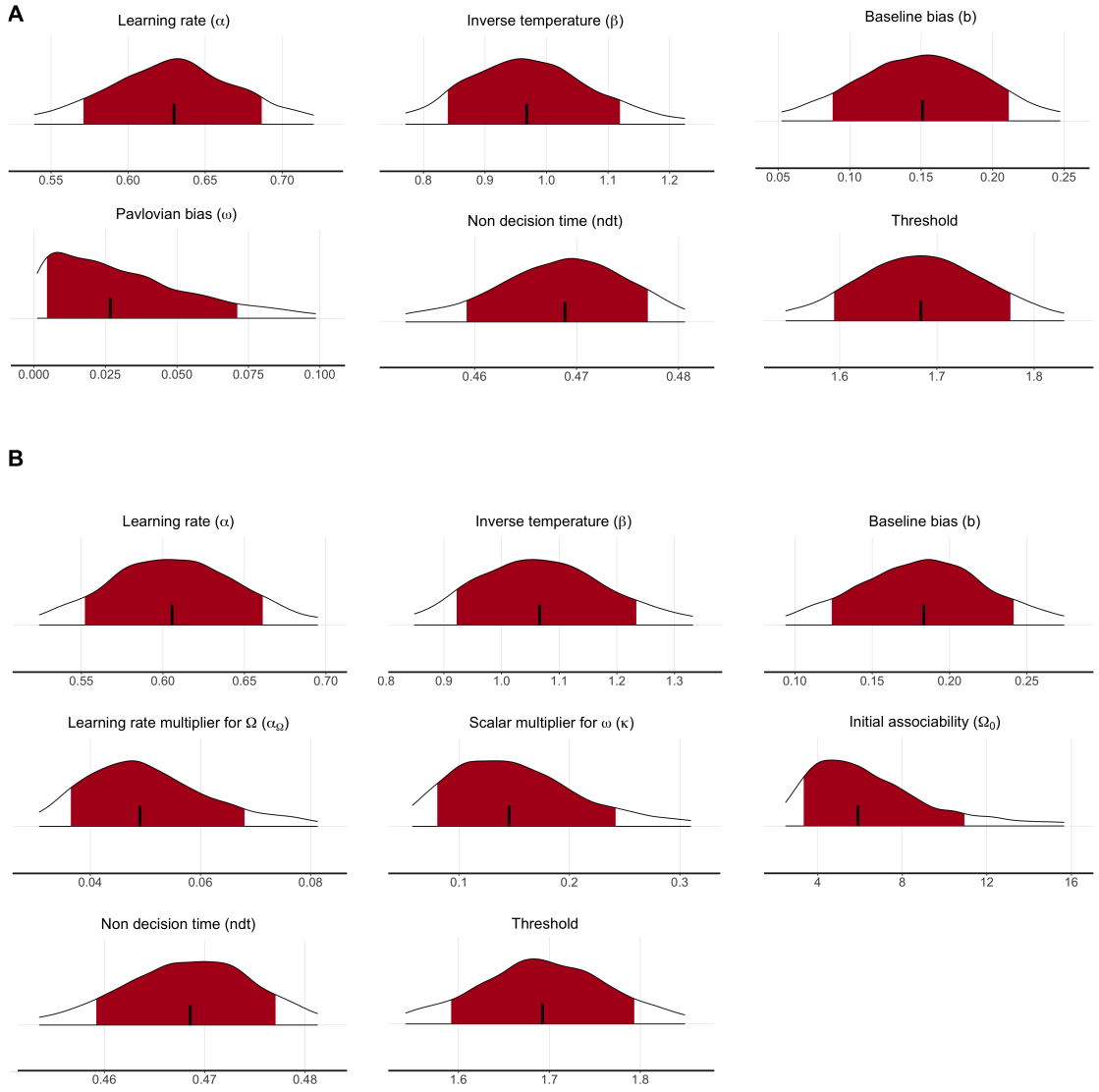

Figure 12. Group-level parameter distributions from (A) the RLDDM model (M3) with fixed ω and (B) the RLDDM model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

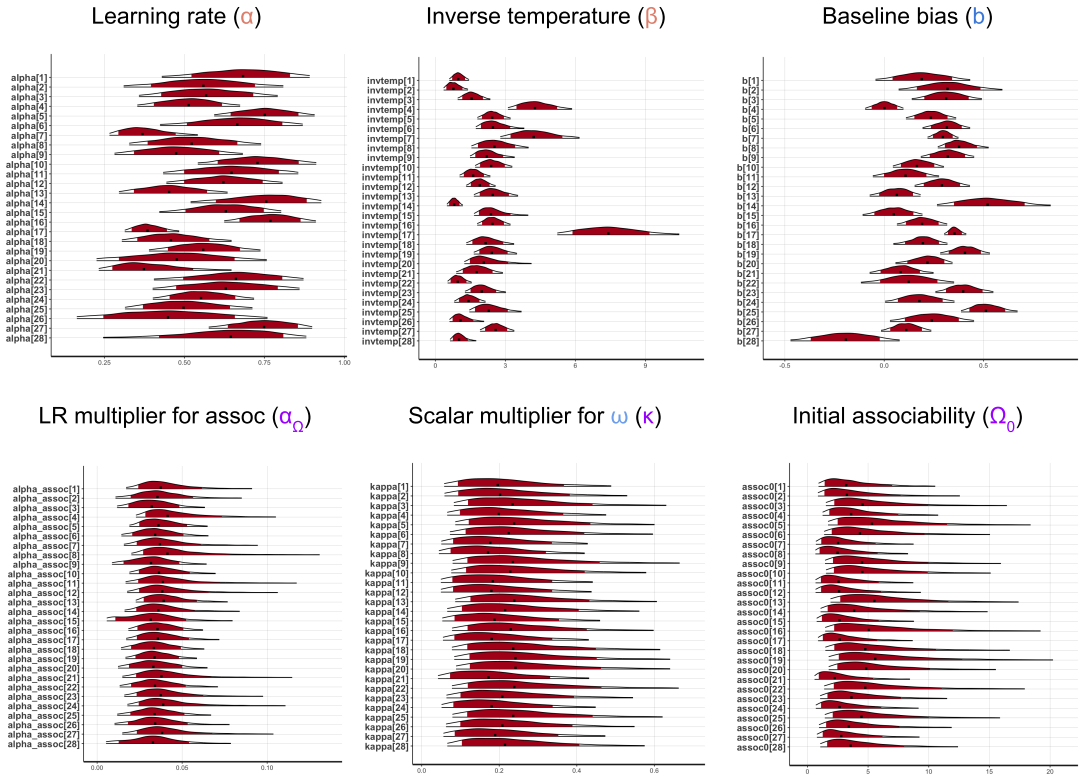

Figure 13. Subject-level parameter distributions from (A) the RL model (M3) with fixed ω and (B) the RL model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

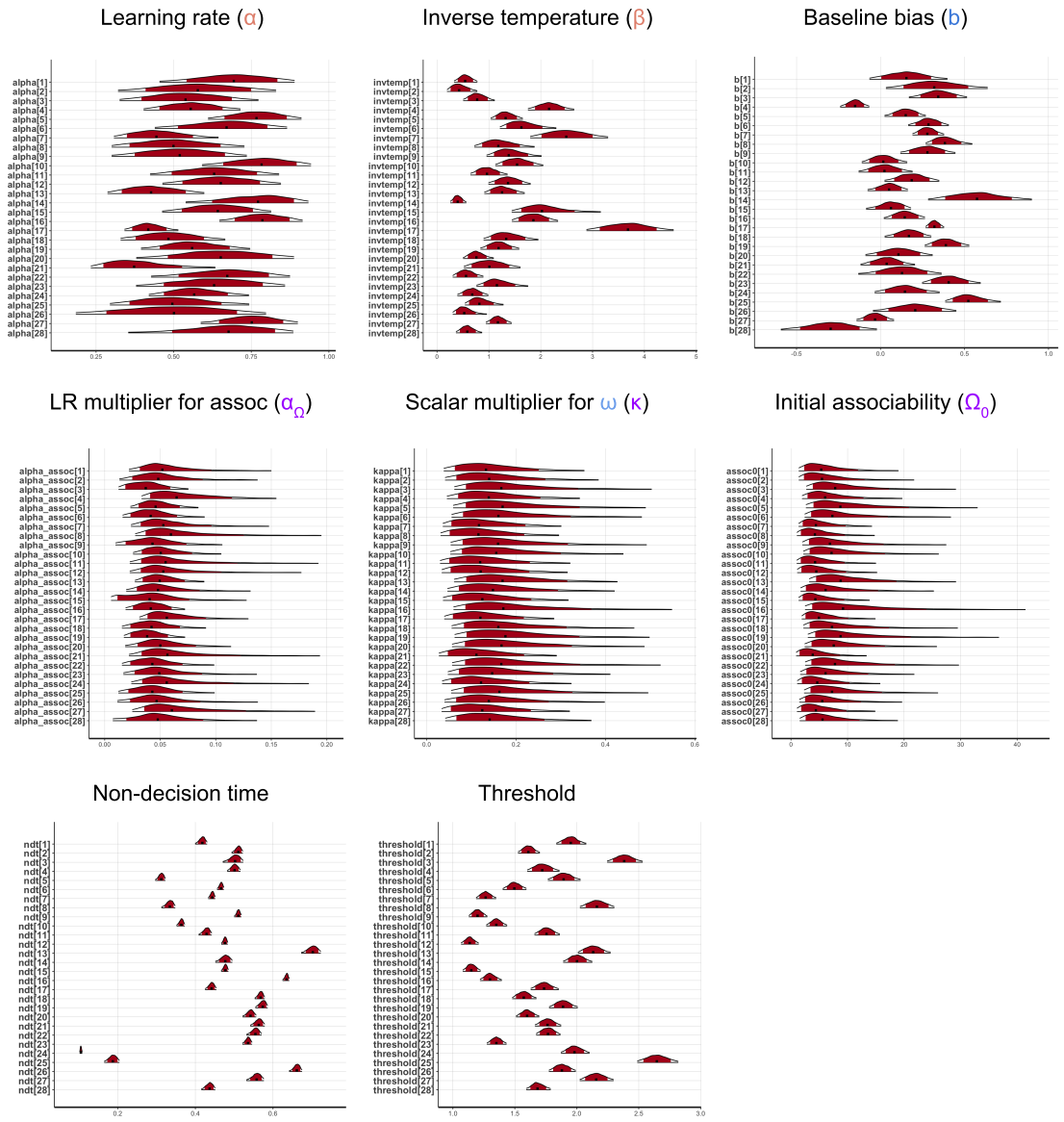

Figure 14. Subject-level parameter distributions from (A) the RLDDM model (M3) with fixed ω and (B) the RLDDM model (M4) with flexible ω. Shaded red regions denote 95% confidence intervals.

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

Please refer to Table 1 and Table 2.

Table 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 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

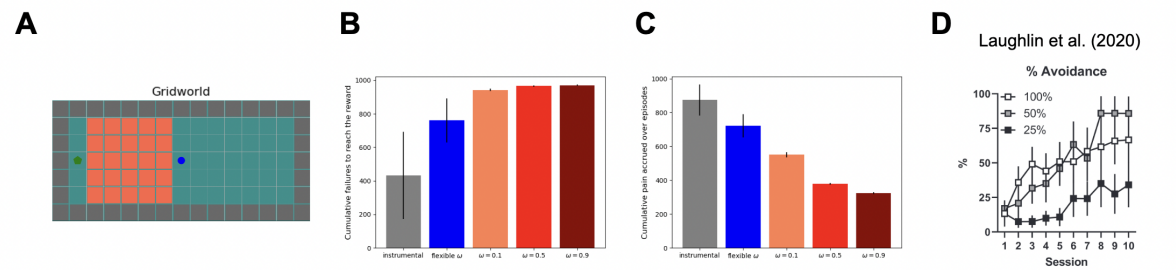

Figure 15. Pavlovian-instrumental interactions are invoked in a popular model of chronic pain, in which excessive Pavlovian fear of movement is self-punitive in a context in which active avoidance would reduce pain [Meulders et al., 2011, Crombez et al., 2012]. (A) Grid world with a start at the centre (blue) and goal at the left end (green, operationalises this. We augment the action set to include an additional "immobilize" 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] showing 25% of the (anxious) rats fail 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

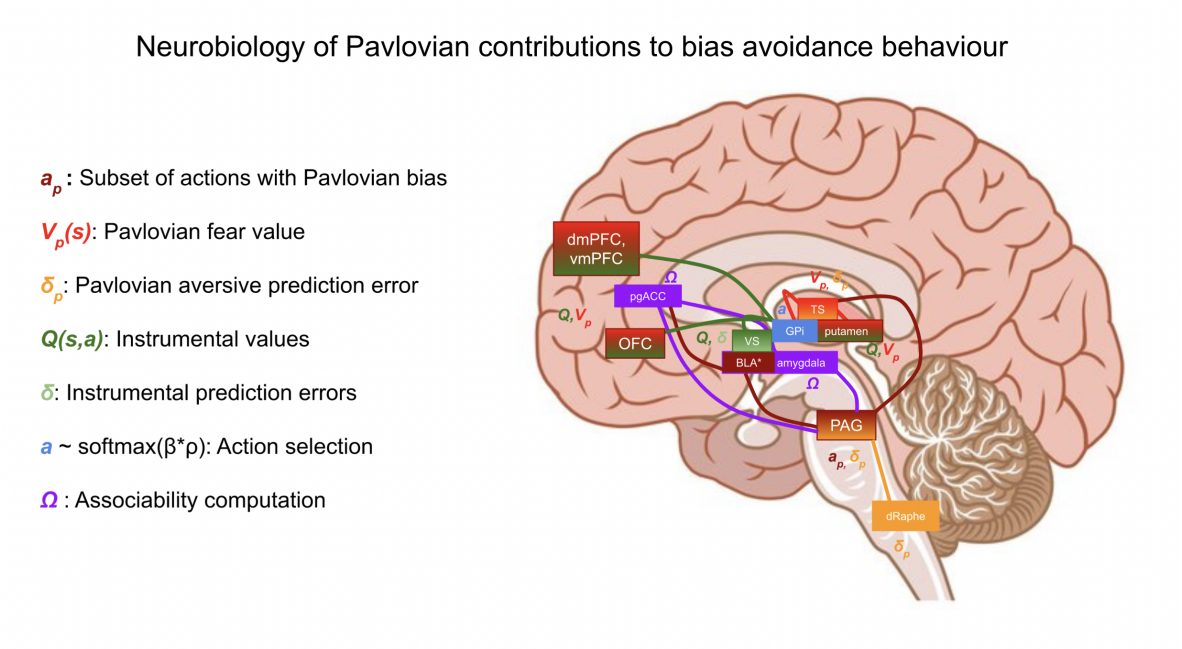

Figure 16. An overview of neurobiological substrates for the proposed PAL model based on relevant prior literature.