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 potentially catastrophic dangers. Conventional exploration-exploitation solutions collapse punishment and reward into a single signal, whereby early losses can be overcome by later gains. But 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 here this promotes safe but efficient learning, and is optimised by arbitrating Pavlovian avoidance on instrumental decision-making according to uncertainty. In a human approach-withdrawal experiment, we show that this flexible avoidance model captures both choice and reaction times. These results show that the Pavlovian fear system has a more sophisticated role in decision-making that 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 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 a point that cannot be overcome, or worse still, leads 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]. 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]), which effectively creates a multi-attribute exploration-exploitation architecture. 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. This may be important for models of chronic pain, as aberrant exploration has been proposed as a mechanism underlying a class of pathogenic model called fear-avoidance models, in which excessive fear is proposed in inhibit exploration to find relief [Vlaeyen and Crombez, 2020].

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 humans can be shown to use this sort of flexible meta-control strategy. We first we 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 - which can be thought of as general uncontrollability. We demonstrate the safety-efficiency trade-off in a range of simulation environments, and in a human three-armed VR maze task motivated by previous multi-arm avoidance tasks. Next, we provide direct experimental evidence of this model, in terms of choice and reaction times, in a novel virtual reality based approach-withdrawal task involving pain. And finally, we show how the model leads to predictions for clinical studies of chronic pain.

RESULTS

Pavlovian avoidance learning model

Our model consist of a Pavlovian punishment (fear) learning system and an integrated instrumental learning system (Fig. 1). The Pavlovian system acquires punishment expectations for each stimulus/state, and converts these expectations into action propensities via Pavlovian responses. Here we consider simple withdrawal responses, but note most animals have a broad set of species specific defensive responses [Bolles, 1970]. For simplicity, we don't include a Pavlovian reward system here. The instrumental system acquires a combined reward and punishment expectations for each stimulus-action or state-action pair, and 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, to test whether a flexible ω confers an advantage. For simulations, we use grid-world like environments, since they serve as a didactic-tool for understanding of Pavlovian-Instrumental interactions [Dayan et al., 2006]. Since Pavlovian biases influence not only choices but also reaction times, we also extend our model to reinforcement learning diffusion decision making (RLDDM) models [Pedersen et al., 2017, Fontanesi et al., 2019, Fengler et al., 2022].

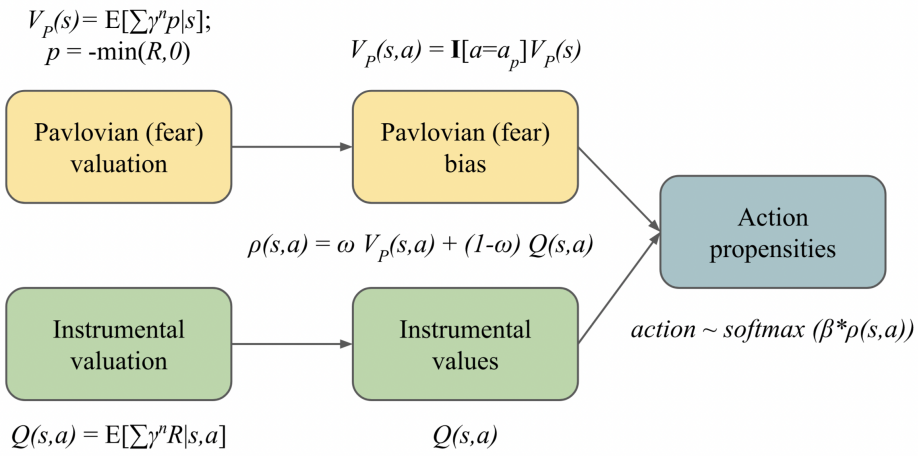

Figure 1. Pavlovian and Instrumental valuations are combined to arrive at action propensities used for (softmax) action selection. The Pavlovian bias influences protective behaviours through safer (Boltzmann) exploration and the arbitration between the Pavlovian and Instrumental system are performed using the parameter ω. See Methods for technical details; notations for the illustration follow Dorfman and Gershman [2019].

Experiment 1: A simulated flexible fear-commissioning model effectively balances safety and efficiency, and provides a basis for punishment sensitivity

We first aimed to (1) provide an empirical demonstration of the safety-efficiency trade-off arising in multi-attribute architecture with instrumental actions and Pavlovian bias and (2) test our hypothesis that an agent with a flexible fear-commissioning scheme can explore safely with little inefficiency in reward acquisition. We further aimed to use this demonstration to (3) tease apart how (constant vs flexible) Pavlovian bias affects sampling asymmetry and instrumental value propagation from the goal and (4) to test whether our model reproduces punishment-sensitive behaviour. We consider a simple fully-observable grid-world environment with stochastic state transitions and fixed starting state and fixed rewarding goal state. Fig. 2A inset (near the rewarding goal) provides an explanation of how misalignment of Pavlovian bias and the instrumental action can lead to a safety-efficiency dilemma. The gridworld action with the Pavlovian bias is assumed to be hard-wired (i.e. the withdrawal action is pretrained by evolution), however the Pavlovian state value is not and is learned during the episode. The simulation results (Fig. 2B, 2C) 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. But 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 looses the 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). It encourages cautious exploration early on when uncertainty is higher and reduces the Pavlovian biases as the uncertainty is resolved (Fig. 2E). The differences in learned instrumental value functions is visible in (Fig. 2D), which prompted us to further tease apart the effect of constant Pavlovian bias on value propagation.

Therefore, we next relocated the (previously fixed) goal from the bottom left corner to the bottom right halfway through i.e. at trial 500 of 1000 trials. Such a reward relocation would required the agent to relearn the value propagation from the goal to efficiently acquire it in subsequent trials. These simulations illustrate that persistent Pavlovian influence leads to persistent rigidity while the flexible fear commissioning scheme is able to efficiently locate the goal (Fig. 2G, 2H). This can be seen via diminished (instrumental) reward value

propagation through the states even in moderate constant Pavlovian influence ($\omega = 0.5$) (Fig 2J) as opposed to better value propagation with flexibly modulated ω (Fig. 2I). Pavlovian withdrawal/avoidance bias in action selection causes a sampling asymmetry observed in state visit plots (Fig. 2G, 2H) leading to observed effects in impaired instrumental value learning (Fig. 2I, Fig. 2J). In conclusion, having a flexibly modulated Pavlovian withdrawal bias plays an adaptive role in safe exploration with little expense to sample efficiency, in contrast to a fixed Pavlovian withdrawal bias which may provide safety at the cost of sample efficiency, sampling asymmetry in state visits and reduced instrumental value propagation.

We next aimed to test whether the model reproduces punishment-sensitive behaviour, given that its mechanism differs from a simple instrumental agent with varying punishment sensitivity [Seymour et al., 2012]. We therefore designed a T-maze environment(Fig. 2K) with 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. 2M). This can also be seen in Fig. 2L and Fig. 2N with state visits (showcasing sampling asymmetry) and diminishing positive value propagation leading to the goal on the right. However, after introducing a Pavlovian influence, the 'right' goal visits immediately decrease with increase in ω. This is also reflected in state visits (2L), through restricted exploration towards the goal on the right and improved exploration towards goal on the left. And the instrumental value propagation from the goal on the right is reduced as observed in Fig. 2N. The safety efficiency trade-off can also be observed through Fig. 2M.

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 sensitive to punishments. This is extremely helpful in softly ensuring an upper bound on losses, by (conservatively) foregoing decisions resulting in immediate loss, but followed by a much larger rewards. This is where the safety-efficiency trade-off marks clear distinction from the exploration-exploitation trade-off, in which earlier losses can be overcome by gains later on.

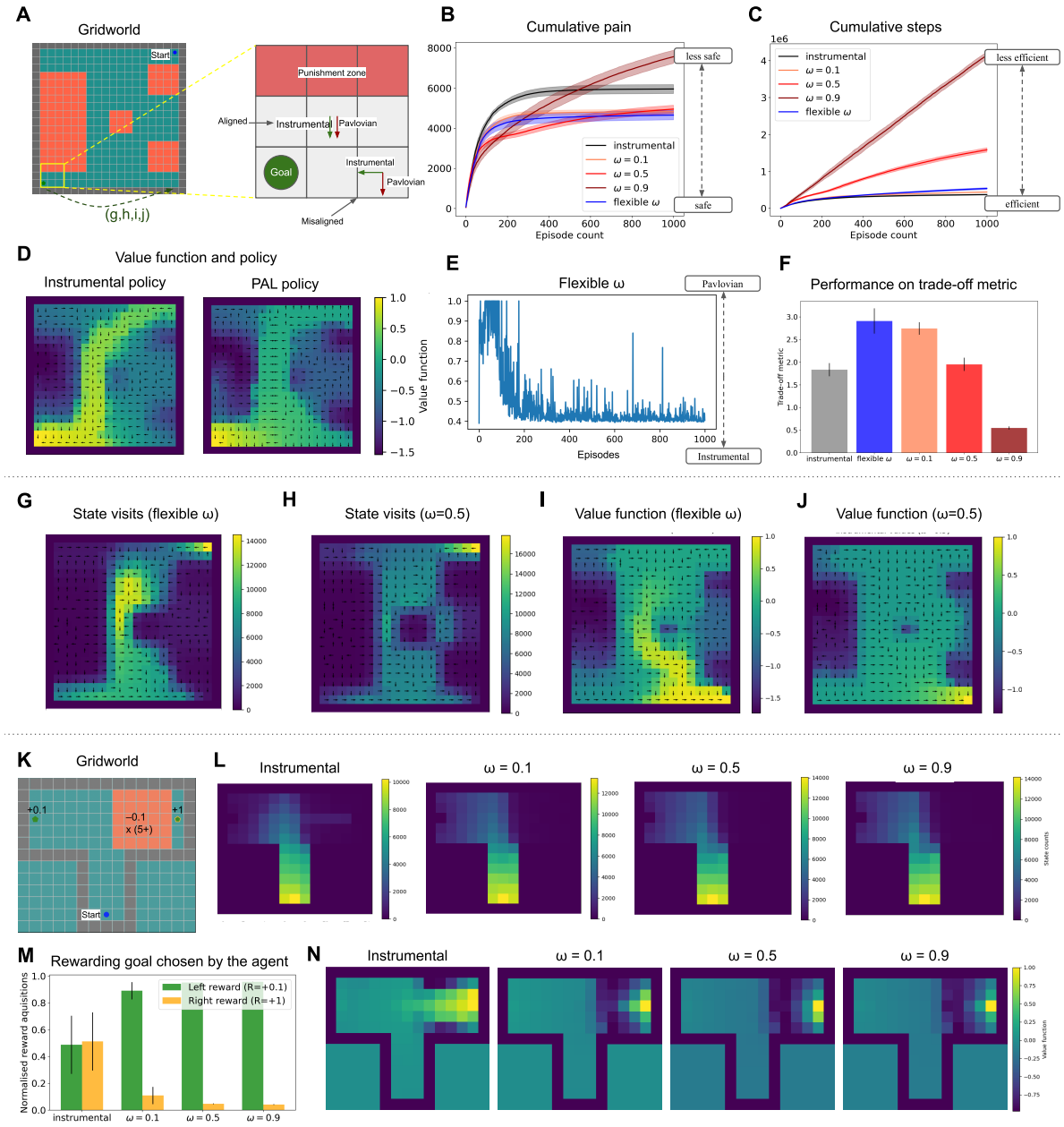

Figure 2. (A) Gridworld environment with starting state in top-right corner and rewarding goal state ($R = +1$) in bottom-left corner and the red states are painful ($R = -0.1$). The gridworld layout follows Gehring and Precup [2013]. The goal is fixed for results A-F but is relocated to bottom-right corner half-way through at episode 500 for results G-J. Inset provides didactic example of misalignment between Pavlovian bias and Instrumental action. (B) Cumulative pain accrued by fixed and flexible ω agents whilst learning over 1000 episodes as a measure of safety. (C) 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. (D) The instrumental values (heatmap) and policy (arrows) learned by the flexible ω agent at the end of the learning duration (Pavlovian bias not included in the plot and shows only the learned Instrumental value function). (E) Plot of flexibly modulated ω arbitration parameter over the learning duration showing the transition from a higher Pavlovian bias to a more instrumental agent over episodes as learning about the environment reduces uncertainty (F) Comparison of different agents using a trade-off metric and to be used only for didactic purposes (more details in Methods). (G) and (H) present the cumulative state visits over states over the complete learning duration (1000 episodes) for flexible and fixed ω agents. (I) and (J) present the instrumental values learned by the end of the learning duration by flexible and fixed ω agents. (K) T-maze gridworld environment with annotated rewards and punishments. (L) State visit count plots for $\omega = 0, 0.1, 0.5, 0.9$. Note that $\omega = 0$ is instrumental agent. (M) Proportion of the rewarding goal chosen by the agent (N) Value function plots for $\omega = 0, 0.1, 0.5, 0.9$.

Experiment 2: Humans show safety-efficiency trade-off in a three-armed maze task

Having provided a basic empirical in-silico demonstration of the safety-efficiency trade-off, we next aimed to provide a simple demonstration of the trade-off in humans. We designed a VR maze task, inspired by a previous manipulandum-based task that was designed to show punishment sensitivity in decision making, in which people were shown to choose a longer route to goal to avoid the possibility of receiving pain [Meulders et al., 2016, van Vliet et al., 2020, 2021, Glogan et al., 2021]. These tasks are inspired by chronic pain scenarios such as lower back pain, where lifting an object off the ground by bending in the knees might be safer but less efficient compared to bending the back. Accordingly, our task involved three possible routes to reach a goal, with the distance goal inversely proportional to the amount of pain. 16 healthy subjects (11 female, average age 30.25 yrs) performed the task, as shown in (Fig. 3A), in which they did 20 sessions, with a brief electrical pain stimulus to the non-dominant hand when in each punishment state. The resulting trade-off can be observed in Fig. 3B. This shows clearly that people choose a longer route to avoid more pain.

This punishment sensitive behaviour in humans can be shown in a simulated agent (Fig. 3C) using a Pavlovian bias up until $\omega = 0.7$ as observed in our simulations (Fig. 3D), but in theory could also be due to an instrumental bias in actions or due to other factors. Beyond $\omega = 0.7$, a high fixed Pavlovian bias may incur unnecessarily high cumulative pain and steps as it's reward value propagation is diminished (Fig. 3F) and attempts to restrict itself to pain-free states (Fig. 3E) whilst searching for reward (despite stochastic transitions which may lead to slightly more painful encounters). Our simulations with this environment help us show that the Pavlovian fear system assists in avoidance acquisition. The agent with higher Pavlovian impetus is more likely to navigate via a longer path but involving no pain. However it is unclear if the behaviour in this VR maze task is solely due to Pavlovian withdrawal biases. Furthermore, as shown in 2A, a misalignment of the Pavlovian bias and the instrumental action can cause an implicit safety-efficiency trade-off even without an apparent explicit trade-off present in the environment. In conclusion, this experiment shows an explicit safety-efficiency trade-off in a 3-armed VR maze task (inspired by previous manipulandum-tasks), which can be captured by our Pavlovian-avoidance learning (PAL) model in simulations.

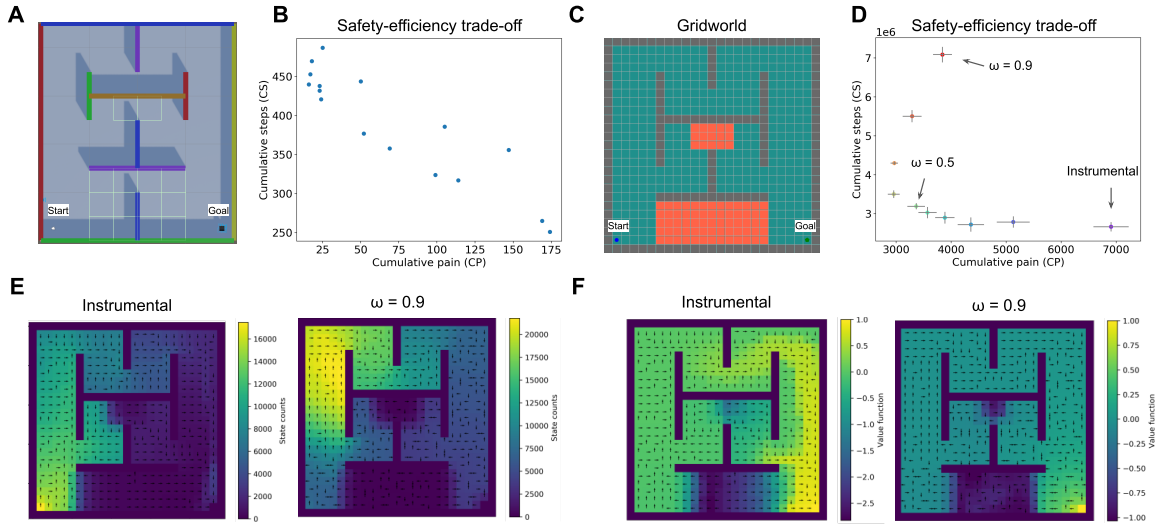

Figure 3. (A) Top-view of VR maze with painful regions annotated by highlighted borders, start in bottom-left corner and goal in bottom-right corner imitating the gridworld in G (B) Cumulative steps required to reach the goal vs cumulative pain acquired by participants in the VR maze task (C) Gridworld environment with short troute vs long route with varying pain (D) Cumulative steps required to reach the goal vs cumulative pain accrued by fixed ω agents ranging from $\omega = 0$ to $\omega = 0.9$ (E) State visit count plots for $\omega = 0, 0.9$ i.e. instrumental and highly Pavlovian agent. (F)Value function plots for $\omega = 0, 0.9$.

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

Next, we aimed to eperimentally test whether we can identify if Pavlovian bias in choices and reaction times is modulated by outcome uncertainty in human decision making, in accordance with our model. 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, according to the model simulation experiments presented earlier. 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. We used a hierarchical Bayesian estimation of model parameters, to increase the reliability across tasks (including the Go-No Go tasks [Zorowitz et al., 2023]).

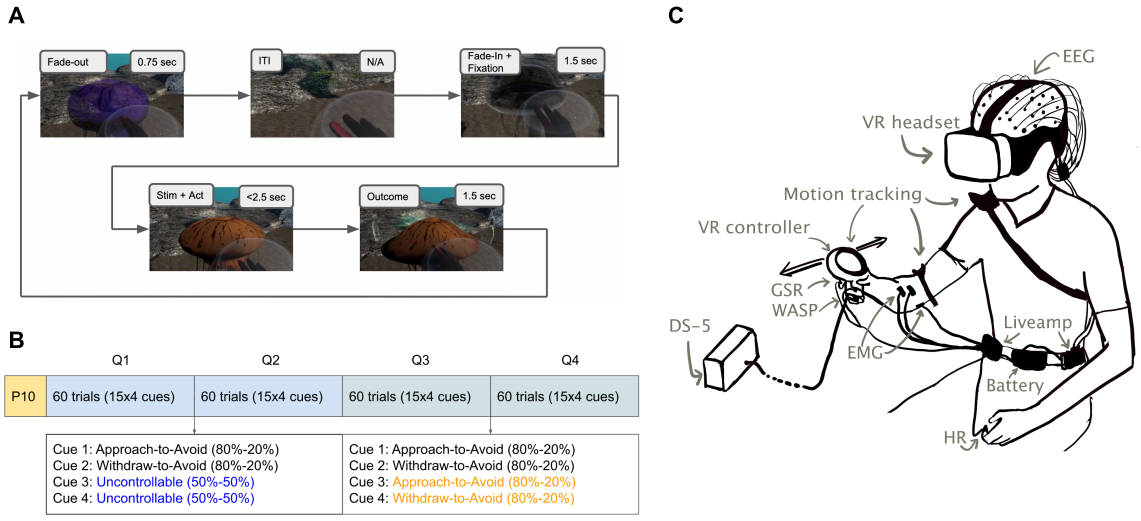

Figure 4. (A) Trial protocol: The participant is expected to take either an approach action (touch the jellyfish) or withdrawal action (withdraw the hand towards oneself) within the next 2.5 seconds once the jellyfish changes colour. The participant was requested to bring their hand at the centre of 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 found that the action bias b, instrumental learning and the Pavlovian responses competed for behavioural control, as observed in previous studies [Guitart-Masip et al., 2012, Cavanagh et al., 2013]. However, unlike previous studies that have treated Pavlovian bias as fixed, we found that the flexible Pavlovian bias better explained the behavioural data (Fig.5 (B) and (E). As in 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. Note that here, the baseline bias is not learned as it is with a Pavlovian bias. RW+bias+Pavlovian(fixed) model further included 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 $(1-\omega)$ We found that the flexible ω model fits significantly better than the fixed ω model (Fig. 5B). It can be seen that ω tracks associability, which decreases over the trials (Fig. 5C) (also resembling Fig. 2E).

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. The Pavlovian bias and the baseline bias as a result are now included in the drift rate. This model is able to capture how a Pavlovian withdrawal bias can speed up withdrawals and slow down approaches. We found that the RLDDM replicates the results for model-fitting (Fig. 5E) and flexible ω (Fig. 5F).

Hierarchically fitted model parameter distributions, subject-wise parameter distributions and tables of model parameters with LOOIC and WAIC values for all RL models and RLDDM models is available in the supplementary materials.

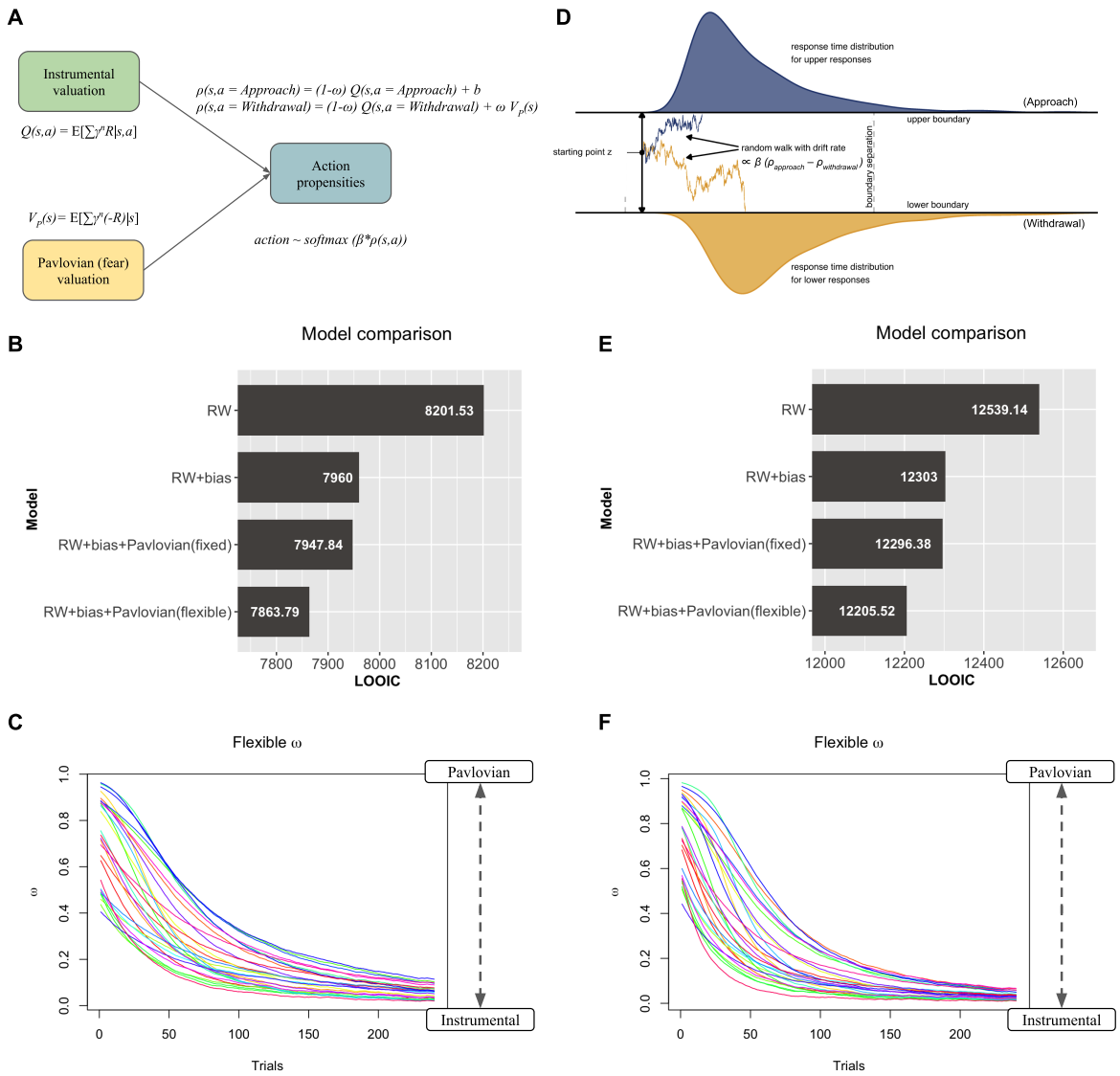

Figure 5. (A) Simplified model from Fig. 1 for the Approach-Withdrawal task (B) Model comparison shows that the model with flexible Pavlovian bias fits best in terms of LOOIC (C) Flexible ω from the RL model over 240 trials for 28 participants. (D) Simplified illustration of RLDDM for the Approach-Withdrawal 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) (E) Model comparison shows that the model with flexible Pavlovian bias fits best in terms of LOOIC (F) Flexible ω from the RLDDM over 240 trials for 28 participants

Model predictions: Adapting fear responses in a chronic pain gridworld

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]. To operationalise this in a gridworld, (Fig. 6A), 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. With a constant Pavlovian fear influence, the agent struggles to complete episodes (Fig. 6B), resembling effects seen in rodent models of anxiety [Laughlin et al., 2020] (Fig. 6D). 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.

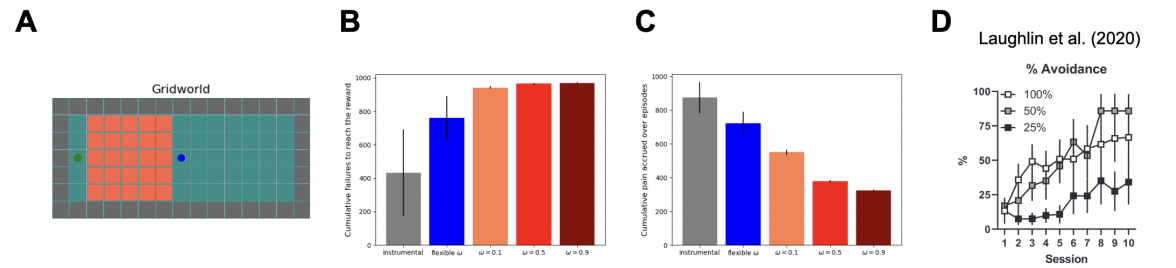

Figure 6. (a) Gridworld with start at the centre (blue) and goal at the left-end (green). (b) Cumulative failures to reach the goal as a measure of efficiency (c) Cumulative pain accrued as a measure of safety (d) Results from Laughlin et al. [2020] showing 25% of the (anxious) rats fail signalled active 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.

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 uncertainty 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. The latter makes novel predictions in clinical conditions, including maladaptive persistent avoidance in chronic pain in which it may be difficult to 'unlearn' an injury.

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 behavior, 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] 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. 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.

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, uncertainty computation could be possibly performed within the basal ganglia using scaled prediction errors [Mikhael and Bogacz, 2016, Moeller et al., 2022]. 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 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. 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 12.

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

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

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

$$Q(s,a) := Q(s,a) + \alpha(\delta) \tag{5}$$

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

$$\delta = R + \gamma Q(s', \mathrm{argmax}_{a'}(Q(s',a'))) - Q(s,a) \tag{6}$$

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:

$$\rho(s,a) = Q(s,a) \tag{7}$$

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:

$$\pi(a|s) = \frac{e^{(\rho(s,a)/\tau)}}{\sum_{a' \in A} e^{(\rho(s,a')/\tau)}} \tag{8}$$

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:

$$\tau(i) = \frac{\tau_0}{1 + \tau_k i} \tag{9}$$

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:

$$\rho(s,a_n) = (1-\omega)Q(s,a_n); \text{ where } a_n \in A_n = A \setminus A_p. \tag{10}$$

$$\rho(s,a_p) = (1-\omega)Q(s,a_p) + \omega(V_p(s)); \text{ where } a_p \in A_p \tag{11}$$

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

$$\rho(s,a) = (1-\omega)Q(s,a) + \omega(V_p(s,a)); \text{ where } V_p(s,a) = \mathbb{I}[a = a_p]V_p(s) \tag{12}$$

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 8.

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]. We maintain a running average of absolute TD-errors (TDE) at each state using the following update rule:

$$\Delta\Omega = \alpha_\Omega * \alpha(\mid \text{TDE} \mid - \Omega) \tag{13}$$

where Ω is the absolute TDE 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 TDE. To obtain parameter $\omega \in [0,1]$ from this absolute TDE estimator $\Omega \in [0,\infty)$, we scale it linearly using scalar κ and clip it between $[0,1]$ as follows:

$$\omega = \min(\kappa\Omega, 1) \tag{14}$$

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 TDE, we have the Rescorla-Wagner rule equivalent - punishment prediction error (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 it's 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 this 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. Ideally we would want this metric to give a higher value when cumulative pain

and cumulative steps, both quantities are as less as possible. One approach could be to calculate the metric inversely proportional to the product of the two, but such a metric could achieve a high value in extreme cases such as when high Pavlovian impetus policy achieves a very low cumulative pain but significantly high cumulative steps, and so this doesn't help us capture the nature of the trade-off improvement. Hence a better metric would be something that rewards improvements in safety and efficiency independently, thus we define it as follows;

$$\text{Trade-off metric} = \frac{1}{CP^2 + CS^2} \tag{15}$$

where CP and CS 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 and the metric can be 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 8 with free parameter $\beta = 1/\tau$ and $\beta > 0$. The learning rule for RW models was:

$$Q(s,a) := Q(s,a) + \alpha(R - Q(s,a)) \tag{16}$$

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:

$$V_p(s) := V_p(s) + \alpha(p - V_p(s)) \tag{17}$$

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

For RW+bias model,

$$\rho(s,a) = Q(s,a) + b; \text{ if } a = \text{Approach} \tag{18}$$

$$\rho(s,a) = Q(s,a); \text{ else} \tag{19}$$

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,

$$\rho(s,a) = (1-\omega)Q(s,a) + b; \text{ if } a = \text{Approach} \tag{20}$$

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

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 13 and 14 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.

$$v = \rho(s, a = \text{Approach}) - \rho(s, a = \text{Withdrawal}) \tag{22}$$

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].

VR maze experiment

Experiment and recruitment

A VR maze experiment was constructed following the simulation environment in Fig. 3C, 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. Different walls were coloured with different colours to help in navigation. 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 3A) 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 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 similar to the Approach-Withdrawal task.

Analysis

All participant trajectories inside the maze were discretised into 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 were plotted against each other to observe the trade-off.

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 COSYNE 2022 reviewers and CCN 2023 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 part supported by the NIHR Oxford Health Biomedical Research Centre. 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.

DATA AND CODE AVAILABILITY

All code and data will be available open source, released upon acceptance of the paper.

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.

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*.

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.

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.

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.

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.

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.

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.

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 Crombez, G. (2020). Behavioral conceptualization and treatment of chronic pain. *Annual review of clinical psychology*, 16:187–212.

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.