Homeostasis After Injury: How Intertwined Inference and Control Underpin Post-Injury Pain and Behaviour

Pranav Mahajan
University of Oxford

Peter Dayan
Max Planck Institute for Biological Cybernetics,
University of Tübingen

Ben Seymour
University of Oxford

Abstract

Injuries are an unfortunate but inevitable fact of life, leading to an evolutionary mandate for powerful homeostatic processes of recovery and recuperation. The physiological responses of the body and the immune system must be coordinated with behaviour to allow protected time for this to happen, and to prevent further damage to the affected bodily parts. Reacting appropriately requires an internal control system that represents the nature and state of the injury and specifies and withholds actions accordingly. We bring the formal uncertainties embodied in this system into the framework of a partially observable Markov decision process (POMDP). We discuss nociceptive phenomena in light of this analysis, noting particularly the paradoxical behaviours associated with injury investigation, and the propensity for transitions from normative, tonic, to pathological, chronic, pain states. Importantly, these simulation results provide a quantitative account and enable us to sketch a much needed roadmap for future theoretical and experimental studies on injury, tonic pain, and transition to chronic pain. Ultimately, we seek novel ways to target chronic pain.

Keywords: Reinforcement Learning; Bayesian Decision Theory; Partially Observable Markov Decision Process; Pain Neuroscience; Homeostasis.

Introduction

Injuries are common across the lifespan of many species, and often lead to a period of vulnerability and reduced functionality whilst healing occurs. In both ecological and laboratory studies, characteristic behavioural changes such as increased anxiety and reduced activity are also often observed. They are considered to reflect adaptive changes and altered motivational priorities appropriate for safe recovery (Wall, 1979; Bolles & Fanselow, 1980; Williams, 2019). These changes are notoriously accompanied, and indeed potentially mediated, by forms of pain.

Organising appropriate behaviour is, by definition, a problem of control (Sutton & Barto, 2018), as operationalized in the context of pain by Seymour et al. (2023). Since the brain suffers inevitable uncertainties about the nature and extent of the injury and the status of the recovery process, we characterize this control problem as an example of what is known as a partially observable Markov decision process (POMDP; Kaelbling et al. (1998)). This framework extends Bayesian models of pain perception (Seymour et al., 2013; Büchel et al., 2014; Wiech, 2016) to address not only inferential aspects (for instance expectancy effects) but also control under uncertainty. The POMDP situates injury, recovery, and pain within the explanatory framework of neural reinforcement learning (Sutton & Barto, 2018; Dayan & Daw, 2008), where the solution is a policy mapping observations to actions or inaction.

We pay particular attention to the belief about the state of the injury constructed by the brain, arguing in the first instance that this determines tonic pain and also modulates phasic pain. In reality, beliefs will be critically informed by tonic nociceptor firing and nociceptor hyperexcitability. However, such signals, and particularly those arising via small diameter, unmyelinated, C-fibres (Debanne, 2004), may not be sufficiently robust, and may fail to characterize and track the resolution of the injury accurately as healing progresses. Other sources of injury-relevant information, such as physiological signals, sensory input (e.g., vision) (Höfle et al., 2010), and prior expectations (Mancini et al., 2022; Seymour & Mancini, 2020; Seymour et al., 2023), must be integrated over space and time to generate a precise, control-relevant representation (Büchel et al., 2014; Wiech, 2016).

Although our simulations are inevitably limited, they achieve two aims. First, they show how tonic pain, as the belief about the state of the injury in the first instance, is a critical, internally constructed and interpreted signal which has broadcast properties, reorganizing macroscopic patterns of behaviour. Second, our simulations bring to light several open questions about the nature and computational characterisation of tonic pain and relevant effects such as endogenous analgesia(Basbaum & Fields, 1978; Bannister & Dickenson, 2017), allowing us to sketch a roadmap for future experiments and theories to solve some of these challenges.

In our simulation results, we will dwell on a particular facet of the injury POMDP, which is the potentially large cost of the conventional policy solution to uncertainty, namely information gathering to determine which options are better, or at least less worse. This cost could accrue if, for instance, the only

way to learn about the current state of the injury is to attempt to use the injured body part. We examine such paradoxical actions of injury investigation (e.g. rubbing injured areas conventionally interpreted in the context of Gate Control Theory) in the light of our theory, and also suggest how the costs might lead to the development of pathological states of chronic pain, even when the periphery is apparently no longer presenting nociceptive input (Fitzcharles et al., 2021).

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

Results

Theory sketch

We propose that the brain treats injury in terms of a partially observable Markov Decision Problem (POMDP; Fig. 1A) (Drake, 1962; Astrom et al., 1965; Sondik, 1971; Kaelbling et al., 1998). A POMDP comprises states $\mathcal{S}$, actions $\mathcal{A}$, transitions $\mathcal{T}$, observations $\mathcal{O}$ and utilities $\mathcal{R}$; we make substantial simplifications in our construction of all of these. The decision-theoretic task for the brain is then to determine a policy which maps the history of observations to the probability of taking an action to maximize the so-called future *return*, which is a long-run measure of cumulative utility. The general Bayesian Decision Theoretic framework is further described in the Methods section.

For the purpose of our simulations, we construct a simplified injury POMDP describing the following situation: A patient contemplating whether or not to commit to a demanding activity (e.g. getting groceries), whilst being uncertain about their injury. The true state characterizes the current circumstance of the injury. We consider just two possible states $s_t \in \{0, 1\}$ for healthy or injured at time t. However, crucially, the brain lacks full information about the state – interoception is incomplete, as well as being noisy. The agent thus only has a so-called *belief* state $b_t \in [0, 1]$, which is its current distribution over the true state. The actions include everything that could be done; however, we make the radical simplification of considering just three to four actions: a_{act}, which involves anything physically demanding but can collect resources; $a_{\text{r\&r}}$, which allows time for recovery and recuperation; and a_{que}, which involves assessing the injury, for instance trying to walk on a recently broken leg (Fig. 1B). We also allow for a null action a_{nul} that does nothing (we omit this from a few experiments for simplicity). Transitions specify how states change, probabilistically, based on the actions – in general, with recovery and worsening of the injury, although here, we restrict ourselves to the case that the state actually remains constant (and so drop the time index for s). Observations include all the exterocepive and interoceptive information available to the brain about the injury, requiring multisensory integration, but we reduce it to a single observation channel for simplicity in our simulations. Investigatory actions such as probing the injured area might provide both types of information.

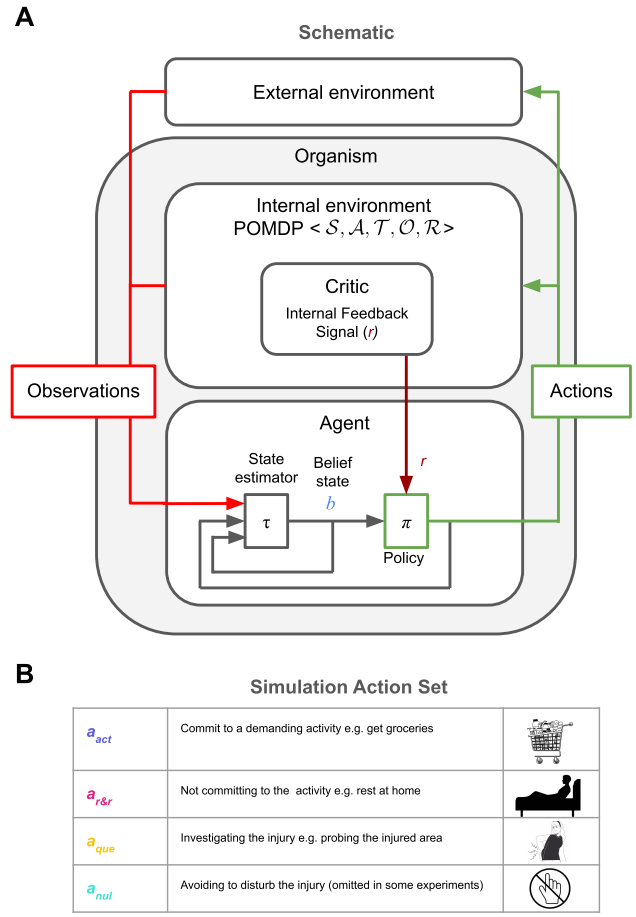

Figure 1: (A) Schematic of the injury POMDP, with an internal environment generating observations and conferring utilities and an internal agent inferring a belief state (Kaelbling et al., 1998) to choose optimal actions. (B) The POMDP action set used for simulations.

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

tential harm); and negative for resting while uninjured, reflecting unmet resource needs (again, as a proxy for the very long run effects of homeostatic deterioration). Since the agent only knows its belief b_t about s, the expected utility is internally constructed.

We illustrate the injury POMDP in a simplified setting where the agent chooses a single, final, choice of a_{act} or $a_{\text{r\&r}}$, but can choose to a_{que} along the way if it is sufficiently uncertain about the injury state. The true state remains fixed throughout the decision-making episode, and the return is the cumulative utility of the trajectory of choices. The one-step expected utilities in the belief space of the injury POMDP used in our simulations are shown in Fig. 2A. Note, that these expected utilities differ from long-run return for actions a_{que} and a_{nul}. Choosing a_{que} and a_{nul} also results in sampling observations from distributions shown in Fig. 2B. We see a_{que} provides more discriminating or discerning observations about the internal state and is therefore more informative than a_{nul}. In our results, we often report state-action values $Q(b_t, a)$, which represent the expected long-run return for belief state b_t and action a. The optimal action in any b_t is the one with the highest Q value.

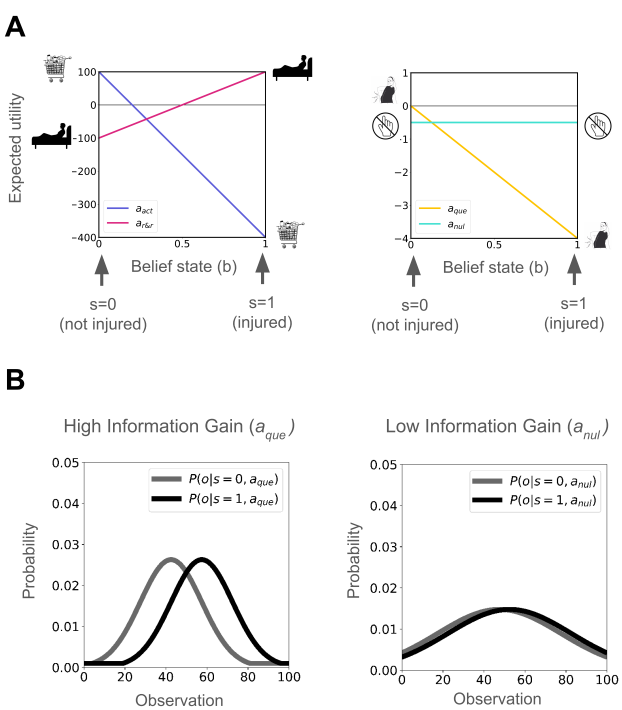

Figure 2: Injury POMDP: (A) Expected utilties of actions a_{act}, $a_{\text{r\&r}}$, a_{que} and a_{nul} in belief space. Utilities: when injured: $r(s=1, a_{\text{act}}) = -400, r(s=1, a_{\text{r\&r}}) = +100, r(s=1, a_{\text{que}}) = -4$; when uninjured: $r(s=0, a_{\text{act}}) = +100, r(s=0, a_{\text{r\&r}}) = -100, r(s=0, a_{\text{que}}) = 0$. $r(a_{\text{nul}}) = -0.5$ regardless of the internal state as a minor opportunity cost. (B) Observation processes for actions a_{que} and a_{nul}. (Further details in Methods - model simulation)

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

Normative consequences

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

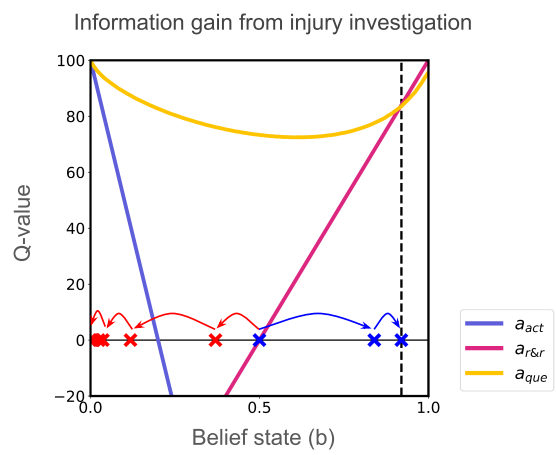

Figure 3: Action values after value iteration. Red arrows show belief updates under true state $s = 0$ (not injured), whereas blue arrows show belief updates under true state $s = 0$ (injured). The agent takes multiple a_{que} actions, reaching either belief thresholds, where either choosing a_{act} or $a_{\text{r\&r}}$ is more valuable than a_{que} and thus terminating the episode.

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

When starting from an uncertain belief state of $b_0 = 0.5$, the agent progressively samples observations by choosing a_{que} at the cost of some (self-constructed) phasic pain. This is because there is a value to the information that can be gained given this uncertainty (see Fig. 3). With every sample, the agent accumulates evidence about the internal state and updates its belief until it can commit to either action a_{act} or $a_{\text{r\&r}}$

(Fig. 3, red and blue arrows). The belief state acts as the context for driving optimal behaviour. When the true internal state is not injured $s = 0$, the agent updates its beliefs (red arrows) and chooses the optimal decision a_{act} in the end. When the true internal state is injured $s = 1$, the agent updates its beliefs (blue arrows) and chooses the optimal decision $a_{\text{r\&r}}$ in the end. In conclusion, this demonstrates agent choosing injury investigating/probing actions despite associated with costs so as to accrue enough evidence to take the optimal decision. Note, actions investigating an injury extend beyond "rubbing the injured area" (explained by Gate Control Theory as "touch inhibits pain") to include non-contact explorations like bending a painful back or moving a painful joint etc.

Trade-off between information gain and phasic pain: Consider the example of assessing your back after an injury. Mild discomfort from actions like stretching (e.g., "pins and needles") can provide valuable information about the injury's state, outweighing the discomfort. In contrast, persistent tonic pain from bending offers little clarity about healing, as the sensory feedback may not distinguish between injury or recovery, making such actions less valuable.

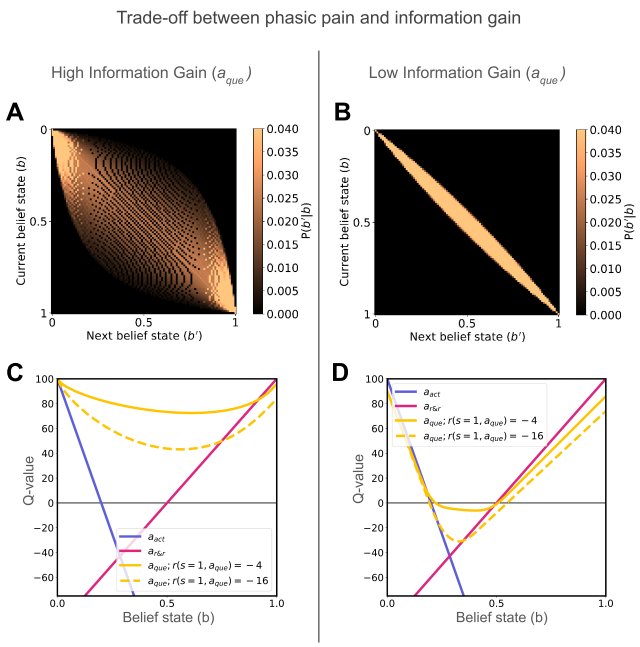

Figure 4: Didactic demonstration of a 2×2 factorial design. (A, B) Belief transitions reflect the influence of high vs. low informativeness of the observation process. (C, D) Action value plots show how phasic pain ($r(s = 1, a_{\text{que}}) = -4$ vs. $r(s = 1, a_{\text{que}}) = -16$) and high vs low information gain influence a_{que}.

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

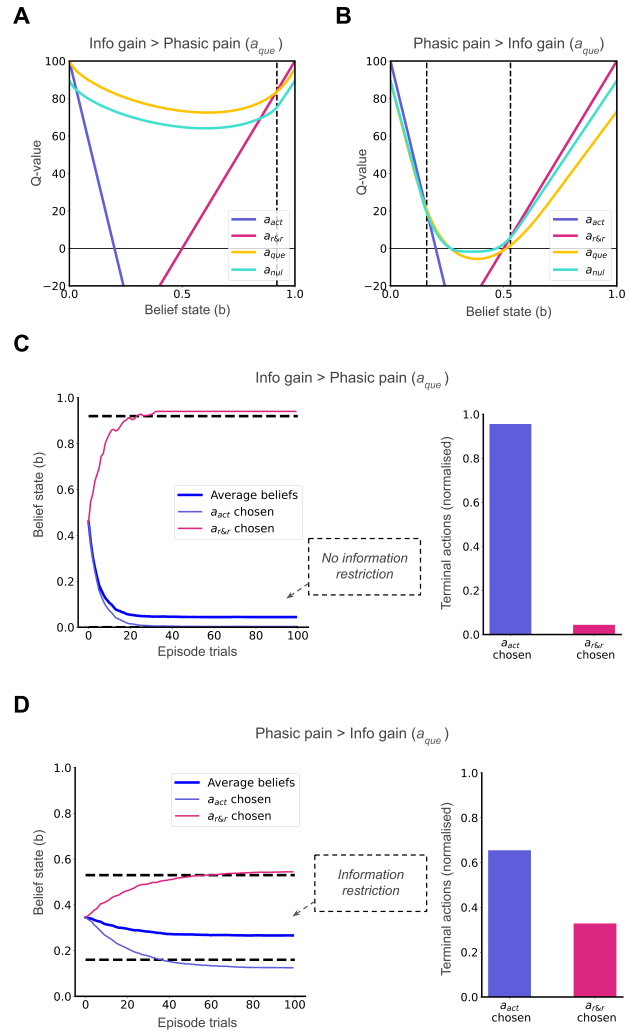

Figure 5: Two possible scenarios along the recovery where the internal state is $s = 0$: (A) Q-values when information gain from a_{que} outweighs phasic pain costs; (B) Q-values when phasic pain costs outweigh information gain. (C) Faster adaptive hidden state resolution is seen when more informative a_{que} is chosen over less informative a_{nul} (D) Maladaptive information restriction occurs when less informative a_{nul} is chosen over more informative a_{que}. Belief state trajectories are averaged over 1000 simulation episodes (dark blue) or only the episodes ending in a_{act} (violet) and $a_{\text{r\&r}}$ (pink). 'Episode trials' represent steps within a decision-making episode.

We find higher information gain of the observation process results in greater belief updates (Fig. 4A), and thus confers a higher action value of a_{que} across various uncertain belief states (Fig. 4C). This contrasts with less informative observation process which results in small belief updates (Fig. 4B) and lower value (Fig. 4D). Further, when observation process

is the same, higher cost of a_{que} decreases the value of choosing a_{que} (Fig. 4C,D). This dynamic illustrates the trade-off between exploratory behaviour and pain avoidance. This experiment also highlights a key facet of our framework: the informativeness of observations and the motivational (phasic) punishment used for credit assignment are disentangled/treated independently. In conclusion, the trade-off between information gain and phasic pain determines the value of an action in our model.

Dysfunctional consequences

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

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

To investigate this route to pain chronification, we simulated the injury POMDP in which the true internal state is no longer injured ($s = 0$), as in the late stages of recovery. In this scenario, the action to investigate the injury (a_{que}) is nevertheless slightly painful when $b_t > 0$, due to the uncertainty of the injury. Action a_{que} is more informative than a_{nul}, using the observation processes shown in Fig. 2B. The action a_{nul} has no cost associated and is less informative than a_{que}. The difference from $a_{\text{r\&r}}$ is that it is not a terminating action - the agent could subsequently change its mind and execute one of the three other action.

We varied the the cost of a_{que} from $r(s = 1, a_{\text{que}}) = -4$ to $r(s = 1, a_{\text{que}}) = -16$, while keeping the observation processes fixed to explore different action values where information gain exceeds phasic pain and vice versa respectively (Fig. 5A and B). We set the initial belief at the mid-point between the thresholds for choosing actions a_{act} and $a_{\text{r\&r}}$. When the information gained from a_{que} outweighs its associated pain, the individual's beliefs adaptively update, correctly resolving the hidden state to $b_t = 0$, leading ultimately to the optimal action ($a_{\text{r\&r}}$) over time (Fig. 5C, averaged belief state trajectories). However, when the phasic pain of a_{que} exceeds the information gain, it is less valued than a_{nul} and a_{nul} is preferred. Due to the less informative nature of a_{nul}, the beliefs do not change greatly. This leads to information restriction, as observed in belief state trajectories averaged over all simulation episodes (Fig. 5D). Thus, we can observe a prolonged incorrect belief that the injury persists, leading to slower recovery or persistent injury mis-inference, which in our model correlates with increased tonic pain (as per eqn. 11). Additionally, this behaviour results in more suboptimal avoidance behaviour - we see an increase in agent choosing $a_{\text{r\&r}}$ (Fig. 5D, right). This simulation illustrates how restricted information due to pain avoidance can contribute to the prolonged incorrect belief that the injury persists, providing a computational account for the chronic pain pathway described by (Seymour et al., 2023).

Aberrant priors: We next study the adverse effects of having a strongly held incorrect belief about the severity of an injury e.g. an aberrant prior. A prior (belief at the start of the decision-making episode) that underestimates the injury leads to further unnecessary harm, whereas a prior that overestimates the injury leads to over-protective behaviour. Both of our simulations use the same Q-values as Fig. 5A.

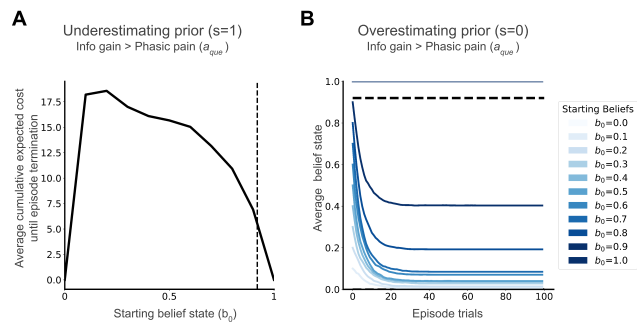

Figure 6: (A) Underestimation of injury when injured ($s = 1$), modelled as start belief states closer to 0, increases average phasic pain when injured. (B) Overestimation of injury when not injured ($s = 0$), modelled as start belief states closer to 1 but short of the decision threshold, can lead to prolonged incorrect beliefs, as seen in mean belief state trajectories. Belief state trajectories are averaged over 1000 simulation episodes.

In the first experiment, we use an underestimating prior when the true state is injured ($s = 1$). With aberrant starting belief states closer to zero (but greater than zero, as this is a decision threshold for a_{act}), we observe that the agent chooses a_{que} over a_{nul}. This results in greater average cumulative expected costs (interpreted as cumulative phasic pain) accrued during the episode in the interest of accurate state inference and subsequent optimal decision-making (Fig. 6A). These excess costs decrease as the starting belief becomes closer to $b_0 = 1$, i.e., being better aligned with the true state (Fig. 6A). This demonstrates a pathway for excess phasic pain through injury investigation, due to maladaptive priors underestimating an injury when the true state is injured.

In our second experiment, we use an overestimating prior when the true state is not injured ($s = 0$). If the starting belief state is aberrantly closer to 1 (itself a possible consequence of probability distortions in the context of risk aversion; Gagne & Dayan, 2022), then it contributes to prolonged incorrect belief that the injury persists, seen in belief trajectories in Fig. 6B. Such prolonged incorrect belief decreases as the starting belief becomes closer to $b_0 = 0$, i.e., again being better aligned with the true state (Fig. 6B). This demonstrates as pathway for persistent tonic and potentially chronic pain, due to mal-

adaptive priors overestimating the injury when the true state is recovered.

Discussion

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

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

Our results align with aspects of the Fear-Avoidance model of pain chronification (Vlaeyen & Linton, 2000; Vlaeyen et al., 2016), which posits that fear of movement leads to avoidance and physical deconditioning, thereby reducing opportunities to update beliefs about injury resolution. By formalising information restriction and avoidance dynamics, our framework provides a computational explanation for how avoidance behaviour may paradoxically sustain injury beliefs, consistent with theories of ex-consequentia reasoning (Arntz et al., 1995; van Vliet et al., 2018; Meulders, 2019). Similarly, our simulations with aberrant priors, particularly those reflecting overestimation, align with self-fulfilling prophecy and catastrophisation models (Jepma et al., 2018; Flink et al., 2013), which highlight how catastrophic thinking enhances attentional demand and hinders disengagement from pain (Van Damme et al., 2002, 2004).

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

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

Chronic pain likely arises from complex interactions between learning processes, internal and external environments, and behavioural feedback loops. By offering a computational framework, our account provides a foundation for making theories of this condition explicit. We map potential future research directions in the next section.

Roadmap for Future Work

In this section, we aim to lay out pertinent questions and possible foci for theoreticians and experimentalists. These are a direct product of our computational framework, highlighting its indispensable role in charting future studies.

What does the perceived tonic pain signal represent?

In the first instance, we associate the belief state with tonic pain, where a higher belief in being injured corresponds to greater tonic pain. This builds on inferential theories of pain (Yoshida et al., 2013; Büchel et al., 2014; Wiech, 2016), extending them to include injury and homeostatic regulation. In this framework, belief guides optimal actions via value functions constructed on belief states, $Q(b, a)$.

An alternative theoretical perspective is to associate tonic pain directly with the value $V(b)$ under a contemplated policy, treating phasic pain as a temporal difference error in $V(b)$. While this value-based view of tonic pain is appealing, it faces several challenges: (1) Fear is traditionally modelled as aversive value , making it difficult to computationally distinguish fear from tonic pain. (2) Committing to a disastrous policy elicits heightened fear in the current state, not tonic pain. (3) A value-like representation of tonic pain implies that injury-investigation actions producing phasic pain would paradoxically reduce tonic pain due to Bayesian decision theory assigning value to information gain. This hypothesis, suggesting endogenous analgesia, motivates future experiments (see next subsection). In any event, it will be worthwhile to mine the analogy from appetitive processing of the separation between liking (putatively the opposite of pain) and wanting (pu-

tatively the correlates of quantities such as $V(b)$) (Dayan, 2022; Berridge et al., 2009).

Another alternative view conceptualises tonic pain as a multi-attribute predictive state representation. Tonic pain encompasses various aversive sensations and emotions e.g., paraesthesia, stinging, stabbing, or burning pain. Predictive state representations, such as General Value Functions (GVFs) (Sutton et al., 2011) in belief space, could model these distinct attributes, creating features for value construction without directly signalling value itself. This representation could underpin the cognitive aspects of tonic pain. Additionally, GVFs replace traditional discount factors with time-varying continuation functions (Sherstan, 2020), potentially capturing temporal dynamics—for instance, how committing to a disastrous policy results in tonic pain only at the future moment of disaster, not immediately. Future work should experimentally tease apart these possibilities, with endogenous analgesia offering a promising avenue for designing such experiments.

How does endogenous control affect phasic and tonic pain?

The computational representation of tonic pain can only make complete sense in the light of the mechanisms of endogenous control (Basbaum & Fields, 1978). We propose designing experiments to disentangle the effects of (1) information gain and (2) uncontrollability on phasic and tonic pain.

For information gain, if actions that yield information (often accompanied by phasic pain) reduce tonic pain, this would support an information-based mechanism of endogenous control, where phasic and tonic pain are inversely modulated. It would also suggest value-like elements in tonic pain. Conversely, if information gain does not reduce tonic pain, the evidence would challenge a value-based interpretation. Similarly, one can test whether tonic pain correlates with the belief state over injury.

A further key factor is uncontrollability. While it is known to increase the propensity for Pavlovian withdrawal bias (Mahajan et al., 2024), uncontrollability's effects on phasic and tonic pain remain unclear. One hypothesis is that it makes phasic pain increase, consistent with its role as driving attention in some learning models (Zhang et al., 2018). However, prior work (Zhang et al., 2018) suggests the opposite effect for tonic pain, which may be suppressed to facilitate learning. Since pain is strongly influenced by attention and distraction, appropriate controls would be essential in such experiments. Testing the role of uncontrollability is particularly relevant, as gaining control has been hypothesised as a pathway out of chronic pain (Seymour, 2019; Meulders, 2020).

Where do utilities come from?

Utilities are the most contentious aspects of an MDP or a POMDP. In the first instance, here, we provide reasonable utilities that could be constructed by an internal critic. However, their origins remain an open question. One approach is to derive utilities from a convex drive function (Keramati & Gutkin, 2014). Preliminary explorations (Appendix Fig 7) show this approach predicts a normative increase in punishment sensitivity with greater deviation from homeostatic balance. It also predicts adaptive behaviours such as risk-aversion (e.g., spreading out an injury) and risk-seeking (through competition among multiple drives). However, the functional form of the drive function may be overly restrictive. An alternative is to initialise values directly (Dayan, 2022), akin to potential-based reward shaping (Wiewiora, 2003). Yet this approach still begs the question: where do the parameters for the drive function or the initial values come from?

A potential solution is to construct "intrinsic" rewards via bi-level optimisation (Chentanez et al., 2004; Barto & Simsek, 2005; Singh et al., 2009). Sparse feedback (e.g. a fitness function) in the outer loop would guide the construction of intrinsic rewards or value initialisations in the inner loop. This framework could extend Dayan (2022) by deriving intrinsic value initialisations from sparse feedback in the outer loop. Notably, such an approach reimagines pain as an intrinsically shaped feedback signal slowly adapting to environmental changes over a lifetime. This perspective might explain nociplasticity in chronic pain (Fitzcharles et al., 2021) as the agent solving the correct problem in an incorrectly perceived or shifting environment (Huys et al., 2015).

Methods

Bayesian decision theoretic approach to homeostasis

We build upon the Bayesian Decision Theoretic (BDT) framework (Dayan & Daw, 2008; Huys et al., 2015) utilising POMDPs (Drake, 1962; Astrom et al., 1965; Sondik, 1971; Kaelbling et al., 1998; Ross et al., 2008) and extend it to the problem of homeostasis.

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

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

$$b_t(s) = P(s_t = s|h_t, b_0) \tag{1}$$

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

$$\begin{aligned} b_t(s') &= \tau(b_{t-1}, a_{t-1}, o_t)(s') \\ &= \frac{O(s', a_{t-1}, o_t) \sum_{s \in \mathcal{S}} \mathcal{T}(s, a_{t-1}, s') b_{t-1}(s)}{P(o_t | b_{t-1}, a_{t-1})} \end{aligned} \tag{2}$$

where the $P(o_t|b_{t-1}, a_{t-1})$ is the marginal distribution over observations o_t (independent of s', acting as a normalising factor for b_t).

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

$$V^{\pi}(b) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{a_t \in \mathcal{A}} \rho(b_t, a_t) \pi(b_t, a_t) | b_0 = b\right] \tag{3}$$

where γ is a temporal discount function and

$$\rho(b_t, a_t) = \sum_{s \in \mathcal{S}} b_t(s) r(s, a_t) \tag{4}$$

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

$$V^{\pi}(b) = \sum_{a \in \mathcal{A}} \pi(b,a) \left[\rho(b,a) + \gamma \sum_{o \in O} P(o|b,a) V^{\pi}(\tau(b,a,o))\right] \tag{5}$$

where the sum over O is interpreted as the expected future return over the infinite horizon of executing action a, assuming the policy π is followed afterwards. It is known that there is a deterministic policy $\pi^*(b,a)$ that optimizes the value of all belief states (Ross et al., 2007)

$$\pi^*(b) = \arg\max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{a \in \mathcal{A}} \rho(b_t, a) \pi(b_t, a) | b_0 = b\right] \tag{6}$$

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

$$V^*(b) = \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \sum_{o \in \mathcal{Z}} P(o|b,a) V^*(\tau(b,a,o))\right] \tag{7}$$

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

$$Q^*(b,a) = \rho(b,a) + \gamma \sum_{o \in O} P(o|b,a) V^*(\tau(b,a,o)) \tag{8}$$

Model implementation and simulation details

We do not aim to model the entire lifetime of an agent but rather an episode in the agent's lifetime. Here we maintain a binary hidden state $s = \{0, 1\}$ and we do not model transitions. Since we do not model transitions in s_t within the POMDP episode, the state estimator equation 2 is reduced to the following,

$$b_t(s) = \tau(b_{t-1}, a_{t-1}, o_t)(s) = \frac{O(s, a_{t-1}, o_t) b_{t-1}(s)}{P(o_t | b_{t-1}, a_{t-1})} \tag{9}$$

This is similar to the Tiger POMDP environment by (Kaelbling et al., 1998), except that the a_{que} is also belief-dependent. Further we refer to the belief state over s as $P(s=1) = b_t(s=1) = b$ and $P(s=0) = b_t(s=0) = 1 - b$.

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

From the action set $\mathcal{A} = \{a_{\text{act}}, a_{\text{r\&r}}, a_{\text{que}}, a_{\text{nul}}\}$ described in Fig. 1, a_{act} and $a_{\text{r\&r}}$ result in episode termination and a_{que} and a_{nul} do not. We then used the Q values of each action in each belief state to determine the decision thresholds for choosing actions a_{act} and $a_{\text{r\&r}}$ which terminate the episode, defined as the most uncertain belief state b where the optimal action is no longer a non-terminal action.

The maximum trials within a POMDP episode were set to 100. Results were averaged over 1000 independent POMDP episodes for a fixed utility function. The discount factor was set to $\gamma = 1$. Further details on parameters and configurations for each simulation are provided in Appendix 1.

Definition of phasic and tonic pain in our model simulations

In the first instance, we define phasic pain based on equation 4, as belief-dependent negative feedback (expected negative costs of an action), only when concerning the pain/injury systems.

$$\text{Phasic Pain} \propto \sum_{s \in \mathcal{S}} b(s) \zeta(s,a) \tag{10}$$

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

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

$$\text{Tonic Pain} \propto b(s=1) \tag{11}$$

Acknowledgments

PM would like to also thank Michael Browning, Rafal Bogacz, Suyi Zhang, Charlie Yan, Shuangyi Tong, Danielle Hewitt for their feedback on presentations or earlier drafts of the manuscript. 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). PD is funded by the Max Planck Society and the Humboldt Foundation. 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.

References

Antcliff, D., Keeley, P., Campbell, M., Woby, S., & McGowan, L. (2016). Exploring patients' opinions of activity pacing and a new activity pacing questionnaire for chronic pain and/or fatigue: a qualitative study. *Physiotherapy*, *102*(3), 300–307.

Arntz, A., Rauner, M., & Van den Hout, M. (1995). "if i feel anxious, there must be danger": Ex-consequentia reasoning in inferring danger in anxiety disorders. *Behaviour research and therapy*, *33*(8), 917–925.

Astrom, K. J., et al. (1965). Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, *10*(1), 174–205.

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

Bannister, K., & Dickenson, A. (2017). The plasticity of descending controls in pain: translational probing. *The Journal of physiology*, *595*(13), 4159–4166.

Barto, A. G., & Simsek, O. (2005). Intrinsic motivation for reinforcement learning systems. In *Proceedings of the thirteenth yale workshop on adaptive and learning systems* (pp. 113–118).

Basbaum, A. I., & Fields, H. L. (1978). Endogenous pain control mechanisms: review and hypothesis. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *4*(5), 451–462.

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, *60*(6), 503–515.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward:'liking','wanting', and learning. *Current opinion in pharmacology*, *9*(1), 65–73.

Bolles, R. C., & Fanselow, M. S. (1980). A perceptual-defensive-recuperative model of fear and pain. *Behavioral and Brain Sciences*, *3*(2), 291–301.

Büchel, C., Geuter, S., Sprenger, C., & Eippert, F. (2014). Placebo analgesia: a predictive coding perspective. *Neuron*, *81*(6), 1223–1239.

Cannon, A. (1929). Organization for physiological homeostasis physiological reviews.

Chentanez, N., Barto, A., & Singh, S. (2004). Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, *17*.

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

Dayan, P. (2022). "liking" as an early and editable draft of long-run affective value. *PLoS Biology*, *20*(1), e3001476.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453.

Dayan, P., Roiser, J., & Viding, E. (2020). The first steps on long marches: The costs of active observation. In *Psychiatry reborn: Biopsychosocial psychiatry in modern medicine*, (pp. 213–228). Oxford University Press.

Debanne, D. (2004). Information processing in the axon. *Nature Reviews Neuroscience*, *5*(4), 304–316.

Drake, A. W. (1962). *Observation of a markov process through a noisy channel*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Fitzcharles, M.-A., Cohen, S. P., Clauw, D. J., Littlejohn, G., Usui, C., & Häuser, W. (2021). Nociplastic pain: towards an understanding of prevalent pain conditions. *The Lancet*, *397*(10289), 2098–2110.

Flink, I. L., Boersma, K., & Linton, S. J. (2013). Pain catastrophizing as repetitive negative thinking: a development of the conceptualization. *Cognitive behaviour therapy*, *42*(3), 215–223.

Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. *Computational Psychiatry (Cambridge, Mass.)*, *1*, 2.

Gagne, C., & Dayan, P. (2022). Peril, prudence and planning as risk, avoidance and worry. *Journal of Mathematical Psychology*, *106*, 102617.

Gordon, A., & Ziv, A. (2021). *The way out: A revolutionary, scientifically proven approach to healing chronic pain*. Penguin.

Hauskrecht, M. (2000). Value-function approximations for partially observable markov decision processes. *Journal of artificial intelligence research*, *13*, 33–94.

Höfle, M., Hauck, M., Engel, A. K., & Senkowski, D. (2010). Pain processing in multisensory environments. *e-Neuroforum*, *1*, 23–28.

Hull, C. L. (1943). Principles of behavior: an introduction to behavior theory.

Huys, Q. J., Guitart-Masip, M., Dolan, R. J., & Dayan, P. (2015). Decision-theoretic psychiatry. *Clinical Psychological Science*, *3*(3), 400–421.

Jepma, M., Koban, L., van Doorn, J., Jones, M., & Wager, T. D. (2018). Behavioural and neural evidence for self-reinforcing expectancy effects on pain. *Nature human behaviour*, *2*(11), 838–855.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1-2), 99–134.

Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *Elife*, *3*, e04811.

Kim, J., Esteves, J. E., Cerritelli, F., & Friston, K. (2022). An active inference account of touch and verbal communication in therapy. *Frontiers in Psychology*, *13*.

Klein, C. (2015). *What the body commands: The imperative theory of pain*. MIT Press.

Kochenderfer, M. J., Wheeler, T. A., & Wray, K. H. (2022). *Algorithms for decision making*. MIT press.

Konstantinou, K., Dunn, K. M., Ogollah, R., Lewis, M., van der Windt, D., Hay, E. M., ... others (2018). Prognosis of sciatica and back-related leg pain in primary care: the atlas cohort. *The Spine Journal*, *18*(6), 1030–1040.

Lovejoy, W. S. (1991). Computationally feasible bounds for partially observed markov decision processes. *Operations research*, *39*(1), 162–175.

Mahajan, P., Tong, S., Lee, S. W., & Seymour, B. (2024). Balancing safety and efficiency in human decision making. *bioRxiv*, 2024–01.

Mancini, F., Zhang, S., & Seymour, B. (2022). Computational and neural mechanisms of statistical pain learning. *Nature Communications*, *13*(1), 6613.

Mathys, C., et al. (2016). How could we get nosology from computation. *Comput Psychiatry New Perspect Ment Illn*, *20*, 121–38.

McParlin, Z., Cerritelli, F., Rossettini, G., Friston, K. J., & Esteves, J. E. (2022). Therapeutic alliance as active inference: The role of therapeutic touch and biobehavioural synchrony in musculoskeletal care. *Frontiers in Behavioral Neuroscience*, 224.

Meulders, A. (2019). From fear of movement-related pain and avoidance to chronic pain disability: a state-of-the-art review. *Current Opinion in Behavioral Sciences*, *26*, 130–136.

Meulders, A. (2020). Fear in the context of pain: Lessons learned from 100 years of fear conditioning research. *Behaviour Research and Therapy*, *131*, 103635.

Moseley, G. L. (2003). A pain neuromatrix approach to patients with chronic pain. *Manual therapy*, *8*(3), 130–140.

Padilla, S. L., Qiu, J., Soden, M. E., Sanz, E., Nestor, C. C., Barker, F. D., ... others (2016). Agouti-related peptide neural circuits mediate adaptive behaviors in the starved state. *Nature neuroscience*, *19*(5), 734–741.

Ross, S., Chaib-draa, B., & Pineau, J. (2007). Bayes-adaptive POMDPs. *Advances in neural information processing systems*, *20*.

Ross, S., Pineau, J., Paquet, S., & Chaib-Draa, B. (2008). Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, *32*, 663–704.

Seymour, B. (2019). Pain: a precision signal for reinforcement learning and control. *Neuron*, *101*(6), 1029–1041.

Seymour, B., Crook, R. J., & Chen, Z. S. (2023). Post-injury pain and behaviour: a control theory perspective. *Nature Reviews Neuroscience*, 1–15.

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

Seymour, B., Dolan, R., et al. (2013). Emotion, motivation, and pain. *Textbook of Pain*, 248–255.

Seymour, B., & Mancini, F. (2020). Hierarchical models of pain: Inference, information-seeking, and adaptive control. *NeuroImage*, *222*, 117212.

Sherstan, C. (2020). Representation and general value functions.

Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society* (pp. 2601–2606).

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations research*, *21*(5), 1071–1088.

Sondik, E. J. (1971). *The optimal control of partially observable markov processes*. Stanford University.

Sondik, E. J. (1978). The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, *26*(2), 282–304.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., & Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th international conference on autonomous agents and multiagent systems-volume 2* (pp. 761–768).

Tolman, E. C. (1932). *Purposive behavior in animals and men*. Univ of California Press.

Van Damme, S., Crombez, G., & Eccleston, C. (2002). Retarded disengagement from pain cues: the effects of pain catastrophizing and pain expectancy. *Pain*, *100*(1-2), 111–118.

Van Damme, S., Crombez, G., & Eccleston, C. (2004). Disengagement from pain: the role of catastrophic thinking about pain. *Pain*, *107*(1), 70–76.

van Vliet, C. M., Meulders, A., Vancleef, L. M., & Vlaeyen, J. W. (2018). The opportunity to avoid pain may paradoxically increase fear. *The Journal of Pain*, *19*(10), 1222–1230.

Vlaeyen, J. W., Crombez, G., & Linton, S. J. (2016). The fear-avoidance model of pain. *Pain*, *157*(8), 1588–1589.

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

Vroomen, P. C., De Krom, M., & Knottnerus, J. (2002). Predicting the outcome of sciatica at short-term follow-up. *British Journal of General Practice*, *52*(475), 119–123.

Wall, P. D. (1979). Three phases of evil: the relation of injury to pain. *Brain and Mind*, *69*, 293.

Wiech, K. (2016). Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. *Science*, *354*(6312), 584–587.

Wiewiora, E. (2003). Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, *19*, 205–208.

Williams, A. C. d. C. (2019). Persistence of pain in humans and other mammals. *Philosophical Transactions of the Royal Society B*, *374*(1785), 20190276.

Yoshida, W., Seymour, B., Koltzenburg, M., & Dolan, R. J. (2013). Uncertainty increases pain: evidence for a novel mechanism of pain modulation involving the periaqueductal gray. *Journal of Neuroscience*, *33*(13), 5638–5646.

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

Appendix

Appendix 1: Simulation parameters and details

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

Summary: The utilities mentioned in Fig. 2 were fixed throughout, with the exception of experiments where phasic pain exceeds information gain for a_{que}, therefore we vary $r(s=1,a_{\text{que}})=-4$ to $r(s=1,a_{\text{que}})=-16$. The observation process for a_{que} and a_{nul} were always high information gain (HI) and low information gain (LI) respectively, as shown in Fig. 2, with the exception of the Fig. 4, where we vary the information gain of a_{que}. For the HI process, the means of observation distributions for $s=0$ and $s=1$, were apart by 15 units and the observation distributions had the std. dev. of 15 units. For the LI process, the means of observation distributions for $s=0$ and $s=1$, were apart by 5 units and the observation distributions had the std. dev. of 30 units. Action a_{nul} was omitted from Normative Consequences results for simplicity but included in Dysfunctional Consequences results.

Appendix 2: HRRL results

In this appendix, we provide some results directly provided by the HRRL framework (Keramati & Gutkin, 2014). In traditional reinforcement learning models, punishment sensitivity is represented as a scalar multiplier applied to the negative reward (i.e., punishments) for learning value functions (Seymour et al., 2012). In contrast, within the HRRL framework, rewards and punishments are intrinsic to the organism and depend on the drive function determined by the internal environment (a potential internal critic). This framework intrinsically scales feedback signals according to punishment sensitivity or non-linear utilities imposed by a convex drive function. As illustrated in Fig. 7A, a simple uni-dimensional drive function is given by $D(s)=(s-s^*)^{n/m}$, where $(n/m)>1$ ensures convexity. Action-dependent deviations Δ from the setpoint result in feedback $r(s_t,a_t)=D(s_t)-D(s_{t+1})=D(s_t)-D(s_t+\Delta)$. For an injury-induced offset $d=s-s^*$, the feedback in the injured state becomes $-((d+\Delta)^{n/m}-d^{n/m})$, which exceeds the feedback in the healthy state ($d=0$), $-\Delta^{n/m}$. These results follow directly from the HRRL framework proposed by Keramati & Gutkin (2014).

We further present analytic results on risk regulation, derived from HRRL, in the context of injury. It has been observed that animals perform flexible, goal-directed behaviours to satisfy physiological needs (Cannon, 1929; Tolman, 1932; Hull, 1943). Here, we test the meta-control mechanism by which homeostatic priorities regulate risk (e.g., from risk-averse to risk-seeking behaviour).

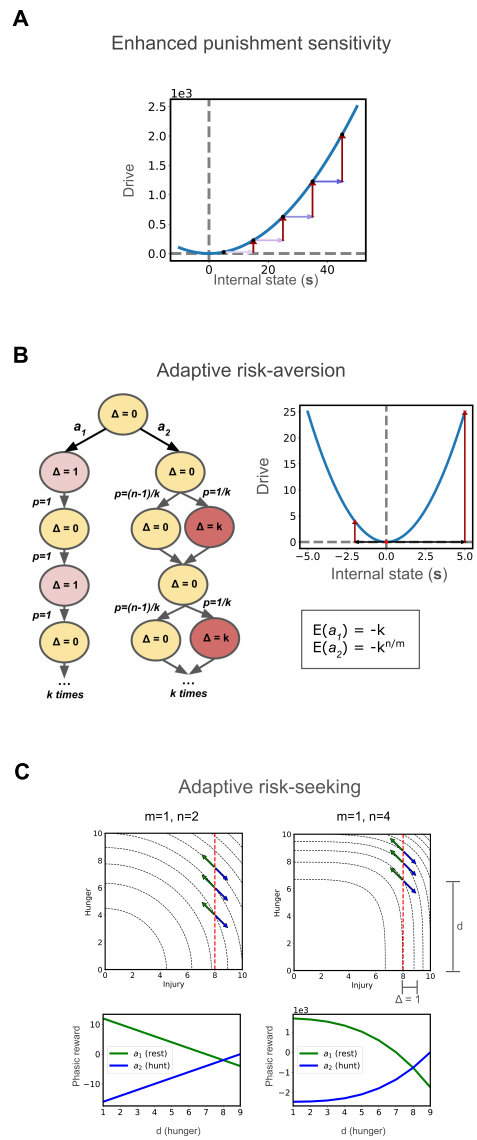

Figure 7: Predictions using HRRL (Keramati & Gutkin, 2014) (A) Normative increase in punishment sensitivity with increase in deviation from homeostatic setpoint (B) Risk-aversion due to convexity of the drive function. The agent prefers to take on equivalent certain damage rather than uncertain larger damage. (C) Risk-seeking via competition of drives. Here showing different phasic rewards (feedback signals) resulting from competition between hunger and injury state drives, for different drive function shapes.

The HRRL framework predicts risk-averse behaviour when the internal state deviates from the setpoint (e.g., during injury). To demonstrate this, we consider a thought experiment using an MDP with two actions, a_1 and a_2 (Fig. 7B). Action a_1 is risk-averse, resulting in deviations $\Delta=1$ repeated k times,

returning to the setpoint each time. Action a_2 is risk-seeking, causing a single large deviation $\Delta = k$ with probability $1/k$, and no deviation otherwise. Although the average cumulative deviation is k for both actions, the convexity of the drive function assigns cumulative returns of $-k$ for a_1 and $-k^{n/m}$ for a_2. For $n > m$, the agent selects the risk-averse action a_1.

The HRRL framework also predicts inhibitory effects on irrelevant drives. For example, in starvation, animals engage in high-risk behaviours that are typically maladaptive. Starving rodents forage in predator-prone areas and modulate aggressive behaviours within territories of depleted nutrients (Padilla et al., 2016). We model this using an MDP analogous to Buridan's ass but where the agent is hungry and injured. With a constant injury offset $d = 8$, we vary the hunger offset $d(\text{hunger})$. The agent has two actions: a_1 (rest), which improves the injury state by $\Delta = 1$ but worsens hunger by an equivalent amount, and a_2 (hunt), which does the reverse. Simulations of two different drive functions reveal that as hunger offset increases, the agent values hunting (despite pain) over resting (Fig. 7C). The effect is non-linear, depending on the drive function parameters. Thus, heightened hunger can inhibit the injury drive, demonstrating a form of endogenous analgesia. This computational approach complements the motivational view of pain (Klein, 2015), grounded in HRRL.

Appendix 3: Our mathematical model complements the Fear-Avoidance model of pain chronification

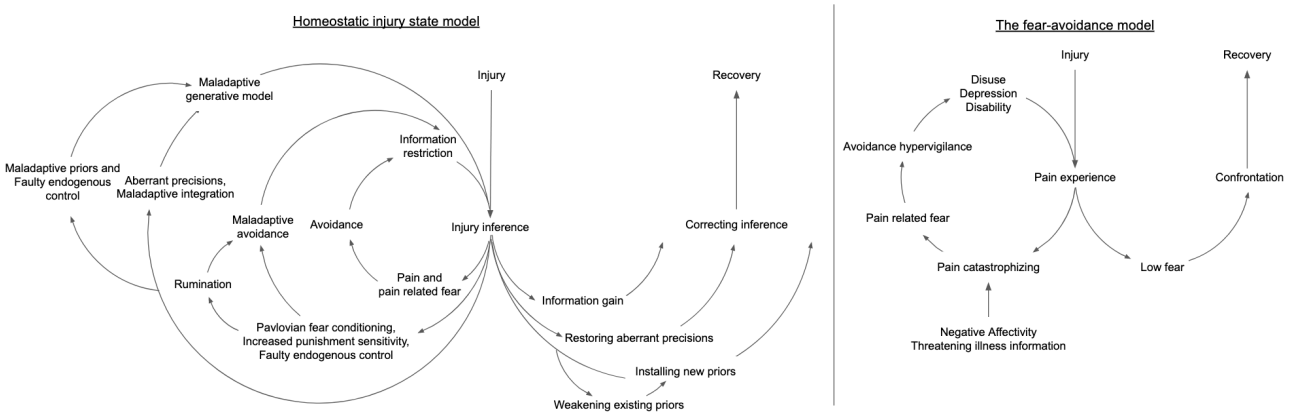

Figure 8: Comparison of predictions of our homeostatic injury state model to that of the Fear Avoidance model and how they play a complementary role in educating patients about their pain.