Optimal indolence: How long to work and how long to play

BY RITWIK K. NIYOGI[1]† ,YANNICK-ANDRE BRETON[2], REBECCA B. SOLOMON[2], KENT CONOVER[2],PETER SHIZGAL[2], PETER DAYAN[1]

[1] *Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom,* [2] *Center for Studies in Behavioral Neurobiology, Concordia University, Montreal, Quebec, Canada*

Dividing limited time between work and leisure when both have their attractions is a common everyday decision. We provide a normative control-theoretic treatment of this decision that bridges economic and psychological accounts. We show how our framework applies to free-operant behavioural experiments in which subjects are required to work (depressing a lever) for sufficient total time (called the price) to receive a reward. When the microscopic benefit-of-leisure increases non-linearly with duration, the model generates behaviour that qualitatively matches various microfeatures of subjects' choices, including the distribution of leisure bout durations as a function of the payoff. We relate our model to traditional accounts by deriving macroscopic, molar, quantities from microscopic choices.

Keywords: work, leisure, normative, microscopic, reinforcement learning, economics

1. Introduction

What to do, when to do it and how long to do it for are fundamental questions for behaviour. Different options across these dimensions of choice yield different costs and benefits, making for a rich, complex, optimization problem.

One common decision is between working and not working (i.e., being at leisure). Working leads to rewards such as food and money; whereas leisure is supposed to be intrinsically beneficial. Since these activities are usually mutually exclusive, subjects must decide how to allocate time to each.

This decision has been studied by economists [1, 2, 3, 4, 5], behavioural psychologists [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16], ethologists [17] and neuroscientists [18, 19, 20, 21, 22, 23, 24]. Tasks involving free operant behaviour are particularly revealing, since subjects can choose what, when and how, minimally encumbered by direct experimenter intervention. We consider the cumulative handling time (CHT) schedule brain stimulation reward paradigm of Shizgal and colleagues [20, 21], in which animals have to invest quantifiable work to get rewards which are psychophysically stationary and repeatable.

Most previous investigations of time allocation have focused on *molar* or *macroscopic* characterisations of behaviour [1, 2, 4, 10, 18, 21, 22, 23, 25, 26, 27, 28, 29, 30, 31], capturing the average times allocated to work or leisure. Here, we characterize the detailed temporal topography of choice, i.e., the fine-scale *molecular* or *microscopic* structure of allocation [32, 33, 34, 35, 36, 37], that is lost in molar averages (see Figure 1C). We build an approximately normative, reinforcement-learning, account, in which microscopic choices approximately maximize net benefit. Our central intent is to understand the qualitative structure of the molecular behaviour of subjects, providing an account that can generalize to many experimental paradigms. Therefore, although we apply the model to a set of CHT experiments in rats it is the next stage of the programme to fit this behaviour quantitatively in detail.

† E-mail: ritwik.niyogi@gatsby.ucl.ac.uk

2. Task and Experiment

Consider a CHT task [20, 21] in which subjects choose between working–the facile task of holding down a light lever, and engaging in leisure, i.e., resting, grooming, exploring etc (Figure 1A). A brain stimulation reward (BSR; [38]) is given after the subject has accumulated work for an experimenter-defined total time-period called the *price* (P; see Table 1 for a description of all symbols). BSR does not suffer satiation and allows precise, psychophysically stable data to be collected over many months. We show data initially reported in [39] (and subsequently in Breton et al (in preparation), Solomon et al (in preparation)).

The objective strength of the BSR is the frequency of electrical stimulation pulses applied to the medial forebrain bundle. This is assumed to have a subjective worth, or *microscopic utility* (to distinguish it from the *macroscopic utility* described in [18, 19, 20, 21, 22, 23]) called the *reward intensity* (RI, in arbitrary units). The transformation from objective to subjective worth has been previously determined [40, 41, 42, 43, 44, 45], The ratio of the reward intensity to the price is called the *payoff*. Leisure is assumed to have an intrinsic subjective worth, although its utility remains to be quantified.

In the task, subjects face triads of trials: 'leading', 'test', then 'trailing' (Figure S1). Throughout a trial, the objective strength of the reward and price are all held fixed. The total time the subjects could work per trial is 25 times the price, enabling at most 25 rewards to be harvested. Leading and trailing trials involve maximal and minimal reward intensities respectively, and the shortest price (we use the qualifiers "short", "long", etc. to emphasise that the price is an experimenter determined *time-period*). We analyze the sandwiched test trials, which span a range of prices and reward intensities. Leading and trailing trials allow calibration, so subjects can stably assess RI and P on test trials. Subjects tend to be at leisure on trailing trials, limiting fatigue. Subjects repeatedly experience each test reward intensity and price over many months, and so can readily appreciate them after minimal experience on a given trial without uncertainty.

A behaviourally observed work or leisure bout is defined as a temporally continuous act of working or engaging in leisure, respectively. Of course, contiguous short working or leisure bouts are externally indistinguishable from one long bout. Subjects are free to distribute leisure bouts in between individual work bouts.

3. Molar and molecular analyses of data

The key molar statistic is the Time Allocation (TA), namely the proportion of the available time for working in a test trial that the subject spends pressing the lever. Figure 1B shows example TAs for a typical subject. *TA* increases with the reward intensity and decreases with the price. Conversely, a molecular analysis, shown in the *ethograms* in (Figure 1C, D), assesses the detailed temporal topography of choice, recording when, and for how long, each act of work or leisure occurred (after the first acquisition of the reward in the trial, i.e., after the 'pink' lever presses in Figure 1D). The *TA* can be derived from the molecular ethogram data, but not vice-versa, since many different molecular patterns (Figure 1C) share a single *TA*.

Qualitative characteristics of the molecular structure of the data (Figure 1D) include: (i) at high payoffs, subjects work almost continuously, engaging in little leisure inbetween work bouts; (ii) at low payoffs, they engage in leisure all at once, in long bouts after working, rather than distributing the same amount of leisure time into multiple short leisure bouts; (iii) subjects work continuously for the entire price duration, as long as the price is not very long (as shown by an analysis conducted by Y-AB, to be published separately); (iv) the duration of leisure bouts is variable.

4. Micro Semi Markov Decision Process Model

We consider whether key features of the data in Figure 1D might arise from the subject's making stochastic optimal control choices, i.e., ones that at least approximately maximise the expected return arising from all benefits and costs over entire trials. Following [24], we formulate this computational problem using the reinforcement learning framework of infinite horizon (Semi) Markov Decision Processes ((S)MDPs) [46, 47] (Figure 2A). Subjects not only choose which action a to take, i.e. to work (W) or engage in leisure (L), but also *the duration of the action* (τ_a). They pay an automatic *opportunity cost of time*: performing an action over a longer period denies the subject the opportunity to take other actions during that period, and thus extirpates any potential benefit from those actions.

Since trials are substantially extended, we assume the subjects do not worry about the time the trial ends, and instead make choices that would (approximately) maximize their average summed microscopic utility per unit time [24]. Nevertheless, for comparison with the data, we still terminate each trial at $25\times$ price, so actions can be *censored* by the end of the trial, preventing their completion.

Utility: The utility of the reward is RI. We assume that pressing the lever requires such minimal force that it does not incur any direct effort cost. We assume leisure to be intrinsically beneficial according to a function $C_L(\tau)$ of its duration (but formally independent of any other rewards or costs). The simplest such function is linear $C_L(\tau) = K_L\tau$ (Figure 2B, upper panel blue line), which would imply that the net utility of several short leisure bouts would be the same as a single bout of equal total length (Figure 2B, lower panel, blue line).

Alternatively, $C_L(\cdot)$ could be nonlinear (Figure 2B, upper panel, red curve). For this function, a single long leisure bout would be preferred to an equivalent time spent in several short bouts (Figure 2B, lower panel, red curve). If $C_L(\cdot)$ saturates, the rate of accrual of benefit-of-leisure $C_L(\tau)/\tau$ will peak at an optimal bout duration. We represent this class of functions with a sigmoid, although many other non-linearities are possible.

Finally, to encompass both extremes, we consider a weighted sum of linear and sigmoid $C_L(\cdot)$, with the same maximal slope (Figure 2B, green curve. Linear $C_L(\cdot)$ has weight $\alpha = 1$, Eq. (S-3))

Evidence from related tasks [48, 49] suggests that the leisure time will be subject to Pavlovian as well as instrumental influences [50, 51, 52]. Subjects exhibit high error rates and slow reaction times for trials with high net payoffs, even when this is only detrimental. We formalize this with a leisure time as a sum of a mandatory Pavlovian contribution τ_{Pav}, and an instrumental contribution τ_L, chosen, in the light of τ_{Pav}, to optimize the expected return. The Pavlovian component comprises a mandatory pause, which is curtailed by the subject's reengagement (conditioned-response) with the reward (unconditioned-stimulus)-predicting lever (conditioned-stimulus). We assume $\tau_{\mathrm{Pav}} = f_{\mathrm{Pav}}(RI, P)$ decreases with payoff – i.e., increases with price and decreases with reward intensity (Figure 2C). The net microscopic benefit-of-leisure is then $C_L(\tau_L + \tau_{\mathrm{Pav}})$ over a bout of total length $\tau_L + \tau_{\mathrm{Pav}}$.

State space: The state $\vec{s} \in \mathcal{S}$ in the model contains all the information required to make a decision. This comprises a binary component ('pre' or 'post'), reporting whether or not the subject has just received a reward; and a real-valued component, indicating if not, how much work $w \in [0, P)$ out of the price P has been performed. Alternatively, $P - w$ is how far the subject is from the price.

Transitions: At state $[\mathrm{pre}, w]$, the subject can choose to work (W) for a duration τ_W or engage in leisure (L) for a duration τ_L. If it chooses the latter, it enjoys a benefit-of-leisure $C_L(\tau_L)$ for time τ_L, after which it returns to the same state. If the subject chooses to work up to a time that is less than the price, (i.e. $w + \tau_W < P$), then its next state is $\vec{s'} = [\mathrm{pre}, w + \tau_W]$. However, if $w + \tau_W \geq P$, the subject gains the work reward RI and transitions to the post reward state $\vec{s'} = [\mathrm{post}]$, consuming time $P - w$. Although subjects can *choose* work durations τ_W that go beyond the price, they cannot physically work for longer than this time, since the lever is retracted as the reward is delivered.

In the post-reward state $\vec{s} = [\mathrm{post}]$, the subject can add *instrumental* leisure for time τ_L to the manda-

tory Pavlovian leisure τ_{Pav} discussed above. It receives utility $C_L(\tau_L + \tau_{\text{Pav}})$ over time $\tau_L + \tau_{\text{Pav}}$, and then transitions to state $\vec{s'} = [\text{pre}, 0]$. The cycle then repeats.

In all cases, the subject's next state in the future $\vec{s'}$ depends on its current state $\vec{s}$, the action a, and the duration τ_a, but is independent of all other states, actions and durations in the past, making the model an SMDP. The model is molecular, as it generates the topography of lever depressing and releasing. It is microscopic as it commits to particular durations of performing actions. We therefore refer to it as a micro SMDP.

Policy evaluation: A (stochastic) policy π determines the probability of each choice of action and duration. It is assumed to be evaluated according to the average reward rate (see Eq. (S-1)). In the SMDP, the state cycles between 'pre' and 'post' reward. The average reward rate is the ratio of the expected total microscopic utility accumulated during a cycle to the expected total time that a cycle takes. The former comprises RI from the reward and the expected microscopic utilities of leisure; the latter includes the price P and the expected duration engaged in leisure.

The total average reward rate is

$$\rho^\pi = \frac{RI + \mathbb{E}_{\pi([L,\tau_L]|\text{post})}\left[C_L(\tau_{\text{Pav}} + \tau_L)\right] + \int_0^P dw\, \mathbb{E}_{\pi_{w_L}}\left[\sum_{n_{L|[\text{pre},w]}} C_L(\tau_L)\right]}{P + \mathbb{E}_{\pi([L,\tau_L]|\text{post})}[\tau_L] + \tau_{\text{Pav}} + \int_0^P dw\, \mathbb{E}_{\pi_{w_L}}\left[\sum_{n_{L|[\text{pre},w]}} \tau_L\right]} \tag{4.1}$$

Here, $\pi([L,\tau_L]|\text{post})$ and π_{w_L} are the probabilities of engaging in instrumental leisure L for time τ_L in the post-reward and pre-reward state $[\text{pre}, w]$, respectively; $\mathbb{E}_\pi$ is the expectation over those probabilities. $n_{L|[\text{pre},w]}$ is the (random) number of times the subject engages in leisure in the pre-reward state $[\text{pre}, w]$.

For state $\vec{s} = \text{post}$, the action $a = [L, \tau_L]$ of engaging in leisure for time τ_L has differential value $Q^\pi(\text{post}, [L, \tau_L])$ (see Eq. (S-2))that includes three terms: (i) the microscopic utility of the leisure, $C_L(\tau_L + \tau_{\text{Pav}})$; (ii) opportunity cost $-\rho^\pi(\tau_L + \tau_{\text{Pav}})$ for the leisure time (the rate of which is determined by the overall average reward rate); and (iii) the long-run value $V^\pi([\text{pre}, 0])$ of the *next* state. The value of state $\vec{s}$ is defined as

$$V^\pi(\vec{s}) = \sum_a \int_{\tau_a} \pi([a,\tau_a]|\vec{s})\, Q^\pi(\vec{s}, [a,\tau_a])$$

averaging over the actions and durations that the policy π specifies at state $\vec{s}$. Thus

$$Q^\pi(\text{post}, [L,\tau_L]) = C_L(\tau_L + \tau_{\text{Pav}}) - \rho^\pi(\tau_L + \tau_{\text{Pav}}) \quad + V^\pi([\text{pre}, 0]) \tag{4.2}$$

Note the clear distinction between the immediate microscopic benefit-of-leisure $C_L(\tau_L + \tau_{\text{Pav}})$ and the net benefit of leisure, given by the overall Q value.

The value $Q^\pi([\text{pre}, w], [L, \tau_L])$ of engaging in leisure for τ_L in the pre-reward state has the same form, but without the contribution of τ_{Pav}, and with a different subsequent state

$$Q^\pi([\text{pre}, w], [L,\tau_L]) = C_L(\tau_L) - \rho^\pi \tau_L \qquad\qquad + V^\pi([\text{pre}, w]) \tag{4.3}$$

Finally, the value $Q^\pi([\text{pre}, w], [W, \tau_W])$ of working for time τ_W in the pre-reward state has two components, depending on whether or not the accumulated work time $w + \tau_W$ is still less than the price (defined using a delta function as $\delta(w + \tau_W < P)$).

$$\begin{aligned} Q^\pi([\text{pre}, w], [W,\tau_W]) = {} & \delta(w + \tau_W < P)[-\rho^\pi \tau_W \qquad\qquad + V^\pi([\text{pre}, w + \tau_W])] \\ & + \delta(w + \tau_W \geq P)[RI - \rho^\pi(P - w) + V^\pi(\text{post})] \end{aligned} \tag{4.4}$$

Policy: We assume the subject's policy π is stochastic, based on a *softmax* of the (differential) value of each choice, i.e., favouring actions and durations with greater expected returns. Random behavioural lapses make extremely long leisure or work bouts unlikely; we therefore consider a probability density $\mu_a(\tau_a)$ of choosing duration τ_a (potentially depending on the action a), which is combined with the softmax like prior and likelihood. For leisure bouts, we assume $\mu_L(\tau_L) = \lambda \exp(-\lambda \tau_L)$ is exponential with mean $1/\lambda = 10P$. The prior $\mu_W(\tau_W)$ for work bouts plays little role, provided its mean is not too short. This makes

$$\pi([a, \tau_a] \,|\vec{s}) = \frac{\exp\left[\beta\, Q^\pi(\vec{s}, [a, \tau_a])\right]\ \mu_a(\tau_a)}{\sum_{a'} \int_{\tau_{a'}} \exp\left[\beta\, Q^\pi(\vec{s}, [a', \tau_{a'}])\right]\ \mu_{a'}(\tau_{a'})\, d\tau_{a'}} \tag{4.5}$$

Subjects will be more likely to choose the action with a greatest Q-value, but have a non-zero probability of choosing a suboptimal action. The parameter $\beta \in [0, \infty)$ controls the degree of stochasticity in choices. Choices are completely random if $\beta = 0$, whereas $\beta \to \infty$ signifies optimal choices.

5. Micro SMDP policies

We first use the micro SMDP to study the issue of stochasticity, then consider the three main regimes of behaviour evident in the data in Figure 1D: when payoffs are high (subjects work almost all the time), low (subjects never work) and medium (when they divide their time). Finally, we discuss the molar consequences of the molecular choices made by the SMDP.

(*a*) *Stochasticity*

To illustrate the issues for the stochasticity of choice, we consider the case of a linear $C_L(\tau_L + \tau_{\mathrm{Pav}}) = K_L(\tau_L + \tau_{\mathrm{Pav}})$, and make two further simplifications: the subject does not engage in leisure in the pre-reward state (thus working for the whole price); and $\lambda = 0$, licensing arbitrarily long leisure durations. Then the Q-value of leisure is linear in τ_L, so the leisure duration distribution is exponential (see Text S2). The expected reward rate and mean leisure duration can be derived analytically (see Text S3).

As long as $RI - K_L P > \frac{1}{\beta}$

$$\begin{aligned} \rho^\pi &= \frac{\beta(RI + K_L \tau_{\mathrm{Pav}}) - 1}{\beta(P + \tau_{\mathrm{Pav}})} \\ \mathbb{E}[\tau_L|\mathrm{post}] &= \frac{P + \tau_{\mathrm{Pav}}}{\beta(RI - K_L P) - 1} \end{aligned} \tag{5.1}$$

Otherwise, if $RI - K_L P < \frac{1}{\beta}$, then $\rho^\pi \to K_L$ (middle panels in Figure 3) and the subject would choose to engage in leisure for the entire trial as $\mathbb{E}[\tau_L|\mathrm{post}] \to \infty$ (upper panels in Figure 3) .

Deterministically optimal behaviour requires $\beta \to \infty$. In that case, provided $RI > K_L P$, the subject would not engage in leisure at all ($\mathbb{E}[\tau_L|\mathrm{post}] = 0$), but would work the entire trial (interspersed by only Pavlovian leisure τ_{Pav}) with optimal reward rate $\rho^* = {}^{(RI + K_L \tau_{\mathrm{Pav}})}/_{(P + \tau_{\mathrm{Pav}})}$ (Figure 3, upper and middle panels, respectively, dashed black lines). However, if $RI < K_L P$, then it would engage in leisure for the entire trial. Thus time allocation functions would be step-functions of the reward intensity and price, as shown by the dashed black lines in the lower panels of Figure 3.

Of course, as is amply apparent in Figure 1D, actual behaviour shows substantial variability, motivating stochastic choices, with $\beta < \infty$. Since all the other quantities can be scaled, we set $\beta = 1$ without loss of generality. This leads to smoothly changing time allocation functions, expected leisure durations and reward rates, as shown by the solid lines in Figure 3.

(*b*) *High payoffs*

In the more general case, the payoff is high when the reward intensity is high, or the price is short, or both. Subjects work as much as possible, making the reward rate in Eq. (4.1) $\rho^\pi \approx {}^{(RI + C_L(\tau_{\mathrm{Pav}}))}/_{(P + \tau_{\mathrm{Pav}})}$ (dash-dotted line in Figure 4 A, upper panels). Since τ_{Pav} is small for high payoffs, $\rho^\pi \approx \frac{RI}{P}$ is just the payoff of the trial. The opportunity cost of leisure time $\rho^\pi(\tau_L + \tau_{\mathrm{Pav}})$ is then linear with a very steep slope, which dominates $C_L(\tau_L + \tau_{\mathrm{Pav}})$ (dashed line in Figure 4A upper panel), irrespective of which form it follows. The Q-value of engaging in leisure in the post-reward state then becomes the linear opportunity cost of leisure time, i.e. $Q^\pi(\mathrm{post}, [L, \tau_L]) \approx -\rho^\pi(\tau_L + \tau_{\mathrm{Pav}})$ (solid bold line in Figure 4A, upper panels).

From Eq.(4.5), the probability density of engaging in instrumental leisure for time τ_L is $\pi([L, \tau_L]\,|\mathrm{post}) \propto \exp\left[-(\beta\rho^\pi + \lambda)\tau_L\right]$. This is an exponential distribution with very short mean $\frac{1}{\beta\rho^\pi + \lambda}$ (Figure 4A, lower panels). The net post-reward leisure bout, consisting of both Pavlovian and instrumental components has the same distribution, only shifted by τ_{Pav}, i.e., a lagged exponential distribution with mean $\tau_{\mathrm{Pav}} + \frac{1}{\beta\rho^\pi + \lambda}$ (Figure 4F).

The probability of choosing to engage in leisure in a pre-reward state (i.e., after the potential resumption of working) is correspondingly also extremely small. Further, the steep opportunity cost of not working would make the distribution of any pre-reward leisure duration also be approximately a very short mean exponential (but not lagged by τ_{Pav}, Figure 4B,C). Therefore when choosing to work, the duration of the work bout chosen (τ_W) barely matters (as revealed by the identical Q-values and policies for different work bout durations in Figure 4D,E). That is, irrespective of whether the subject performs numerous short work bouts or pre-commits to working the whole price, it enjoys the same expected return. To the experimenter, the subject appears to work without interruption for the entire price. In sum, for high payoffs, the subject works almost continuously, with very short, lagged-exponentially distributed leisure bouts at the end of each work bout (Figure 5A, lowest panel). This accounts well for key feature (i) of the data.

(*c*) *Low payoffs*

At the other extreme, after discovering that the payoff is very low, subjects barely work (Figure 1D; top panel). Temporarily ignoring leisure consumed in the pre-reward state, the reward rate in Eq. (4.1) becomes

$$\rho^\pi \approx \frac{\mathbb{E}_{\pi([L,\tau_L]|\mathrm{post})}\left[C_L(\tau_{\mathrm{Pav}} + \tau_L)\right]}{P + \mathbb{E}_{\pi([L,\tau_L]|\mathrm{post})}[\tau_L] + \tau_{\mathrm{Pav}}}$$

shown by the dash-dotted line in Figure 6 A (upper panels), and is comparatively small. The opportunity cost of time grows so slowly that the Q-value of leisure is dominated by the microscopic benefit-of-leisure $C_L(\tau_L + \tau_{\mathrm{Pav}})$ (dashed curves in Figure 6A, upper panels).

We showed that for linear $C_L(\cdot)$, the Q-value is linear and the leisure duration distribution is exponential (shown again in Figure 6A, left panel). For initially supra-linear $C_L(\cdot)$, the Q-value becomes a bump (solid bold curve in Figure 6A upper panel, centre and right). The probability of choosing to engage in instrumental leisure for time τ_L is then the exponential of this bump, which yields a unimodal, gamma-like distribution (Figure 6A lower panel, centre and right). Thus for a low payoff, a subject would opt to consume leisure all at one go, if from the mode of this distribution. This accounts for key feature (ii) of the data.

The net duration of leisure in the post-reward state $\tau_L + \tau_{\mathrm{Pav}}$ is then almost the same unimodal gamma-like distribution (Figure 6F). If the Pavlovian component is increased, the instrumental component $\pi(\tau_L|\mathrm{post})$ will decrease leaving identical the distribution of their sum $Pr(\tau_L + \tau_{\mathrm{Pav}}|\mathrm{post})$ (compare Figure 6 A, lower right panel).

The location of the mode of the net leisure bout duration distribution (Figure 6F) is crucial. For shorter prices associated with low net payoffs, this mode lies much beyond the trial duration $T = 25P$.

Hence, a leisure bout drawn from this distribution would almost always exceed the trial duration, and so be *censored*, i.e. terminated by the end of the trial. Our model successfully predicts the molecular data in this condition (Figure 5A, upper panel). We discuss our model's predictions for long prices later.

The main effect of changing from partially linear to saturating $C_L(\cdot)$ is to decrease both the mean and the standard deviation of leisure bouts. The tail of the distribution (Figure 6A, left versus right panel) is shortened, since the Q-values of longer leisure bouts ultimately fail to grow.

Engaging in leisure in post- and pre-reward states are closely related. Thus, if the payoff is too low then the subject will also choose to engage in long leisure bouts in the pre-reward states (Figure 6 B,C). Correspondingly, the subject will be less likely to commit to longer work times and lose the benefits of leisure (Figure 6D,E). If behaviour is too deterministic, then the behavioural cycle from pre- to post-reward can fail to complete (leading to non-ergoditicty). This is not apparent in the behavioural data, so we do not consider it further.

(*d*) *Medium payoffs*

The opportunity costs of time for intermediate payoffs are also intermediate. Thus the Q-value of leisure (solid bold curves in Figure 7A, upper panels) depends delicately on the balance between the benefit-of-leisure and the opportunity cost (dashed and dashed-dotted lines in Figure 7A, upper panels, respectively). For the sigmoidal $C_L(\cdot)$, the combination of supra- and sub-linearity leads to a bimodal distribution for leisure bouts that is a weighted sum of exponential and a gamma-like distributions (Figure 7A, lower centre and right panels; F).

Bouts drawn from the exponential component will be short. However, the mode of the gamma-like distribution lies beyond the trial duration (Figure 7F), as in the low payoff case when the price is not long (Figure 6F). Bouts drawn from this will thus be censored. Altogether, this predicts a pattern of several work bouts interrupted by short leisure bouts, followed by a long, censored leisure bout (Figure 5A, middle panel). Occasionally, a long, but uncensored, duration can be drawn from the distribution in Figure 7F. The subject would then engage in a long, uncensored leisure bout before returning to work. Our model thus also accounts well for the details of the molecular data on medium payoffs, including variable leisure bouts (key feature (iv)).

(*e*) *Pre-commitment to working continuously for the entire price duration*

The micro SMDP model accounts for feature (iii) of the data, that subjects generally work continuously for the entire price duration. That is, subjects could choose to pre-commit by working for the entire price P, or divide P into multiple contiguous working bouts. In the latter case, even if Q-value of working is greater than that of engaging in leisure, the stochasticity of choice implies that subjects would have some chance of engaging in leisure instead, i.e., the pessimal choice (Figure 7B,C). Pre-committing to working continuously for the entire price avoids this corruption (Figure 7D,E). In Figure 7E, for any given state $[\text{pre}, w]$ the probability of choosing longer work bouts τ_W increases, until the price is reached. Corruption does not occur for a deterministic, optimal policy, so pre-commitment is unnecessary. This case is then similar to that for a high payoff (Figure 4D,E).

(*f*) *Molar behaviour from the micro SMDP*

If the micro SMDP model accounts for the molecular data, integrating its output should account for the molar characterizations of behaviour that were the target of most previous modelling. Consider first the case of a fixed short price $P = 4s$, across different reward intensities (Figure 8A). After an initial region in which different $C_L(\cdot)$ affect the outcome, the reward rate ρ^π in Eq. (4.1) increases linearly with the reward intensity (Figure 8A, upper left panel). Consequently, the opportunity cost of time increases linearly too. If $C_L(\cdot)$ is linear, the resultant linear Q-value of leisure in the post-reward state,

and hence, the mean of the exponential leisure bout duration distribution decreases (Figure 8A, upper and lower centre panels, respectively). If $C_L(\cdot)$ is sigmoidal, the bump corresponding to the Q-value of leisure shifts leftwards to smaller leisure durations (Figure 8A, upper right panel). Both the mode and the relative weight of the gamma-like distribution decrease as the reward intensity increases (Figure 8A, upper right panel). Thus, as the model smoothly transitions from low through medium to high reward intensities, time allocation increases smoothly from zero to one (Figure 8A, lower left panel).

The converse holds if the price is increased while holding the reward intensity fixed at a high value, making the time allocation decrease smoothly (Figure 8B). The reward rate ρ^{π} in Eq. (4.1) decreases hyperbolically, eventually reaching an asymptote (at a level depending on $C_L(\cdot)$, Figure 8B, upper left panel). For long prices, the mode of the unimodal distribution does not increase by much with further price increases. However, by design of the experiment, the trial duration increases with the price. When the trial is much shorter than this mode, most long leisure bouts are censored and time allocation is near zero (Figure 8B, lower right panel). As the trial duration approaches the mode, long leisure bouts are less likely to get censored (Figure 8C, upper left panel).

We therefore make the counterintuitive predict that as the price increases, subjects will eventually be observed to resume working after a long leisure bout. Thus with longer prices, proportionally more work bouts will be observed (Figure 8C, upper right panel). Consequently, time allocation would be observed to not decrease, and even increase with the price (see the foot of the red curve in Figure 8B lower left panel). Such behaviour would be observed for eventually sub-linear benefits-of-leisure. An increase in time allocation at long prices is not possible for linear $C_L(\cdot)$ (blue curve in Figure 8B lower left panel). As the price increases, so does the mean of the resultant exponential leisure bout duration distribution (Figure 8B centre panels) and long leisure bouts will still be censored.

In general, for the same reward intensity and price, less time is spent working for linear than saturating $C_L(\cdot)$ (compare the blue and red curves Figure 8A and B, lower left panels), since linear $C_L(\cdot)$ is associated with longer leisure bouts. Thus, larger payoffs are necessary to capture the entire range of time allocation. The effect of different $C_L(\cdot)$ on the reward rate at low payoffs is more subtle (compare blue and red curves in Figure 8A and B, upper left panels panels). This depends on the ratio of the expected microscopic benefit-of-leisure ($\mathbb{E}_{\pi([L,\tau_L]|\text{post})}\left[C_L(\tau_{Pav}+\tau_L)\right]$) and the expected leisure duration ($\mathbb{E}_{\pi([L,\tau_L]|\text{post})}[\tau_L]+\tau_{Pav}$) in the reward rate equation, Eq. (4.1). This is constant ($=K_L$) for a linear $C_L(\cdot)$. The latter term can be much greater for a saturating $C_L(\cdot)$, leading to a lower reward rate.

Figure 8 shows that the Pavlovian component of leisure τ_{Pav} will mainly be evident at shorter prices. At high reward intensities, instrumental leisure is negligible and leisure is mainly Pavlovian. That time allocation for real subjects saturates at 1, implies that τ_{Pav} decreases with payoff, as argued.

6. Discussion

Real time decision-making involves choices about when and for how long to execute actions as well as well as which to perform. We studied a simplified version of this problem, considering a paradigmatic case with economic, psychological, ethological and biological consequences, namely working for explicit external rewards versus engaging in leisure for its own implicit benefit. We offered a normative, microscopic framework accounting for subjects' temporal choices, showing the rich collection of effects associated with the way that the subjective benefit-of-leisure grows with its duration.

Our microscopic formulation involved an infinite horizon Semi-Markov Decision Process (SMDP) with three key characteristics: approximate optimization of the reward rate, stochastic choices as a function of the values of the options concerned, and an assumption that, a priori, temporal choices would never be infinitely extended (owing to either lapses or the greater uncertainty that accompanies the timing of longer intervals [53]). The metrics associated with this last assumption had little effect on the output of the model.

We exercised our model by examining a psychophysical paradigm called the cumulative handling

time (CHT) schedule involving brain stimulation reward. The CHT controls both the (average) minimum inter-reward interval and the amount of work required to earn a reward. More common schedules of reinforcement such as Fixed Ratio, or Variable Interval control one but not the other. This makes the CHT particularly useful for studying the choice of how long to either work or engage in leisure. Nevertheless, it would be straightforward to adapt our model to treat waiting schedules such as [54, 55, 56, 57, 58, 59, 60] or to add other facets. For instance, effort costs would lead to shorter work bouts rather than the pre-commitment to working for the duration of the price observed in the data. Costs of waiting through a delay would also lead subjects to quit waiting earlier than later. Other tasks with other work requirements could also be fitted into the model by changing the state and transition structure of the Markov chain. The main issue the CHT task poses for the model is that it is separated into episodic trials of different types making infinite horizon optimization an approximation. However, the approximation is likely benign, since the relevant trials are extended (each lasts 25 times the price), and the main effect is that work and leisure bouts can sometimes be censored at the ends of trials.

It is straightforward to account for subjects' behaviour in the CHT when payoffs are high (i.e., when the rewards are big and the price is short and the subjects work almost all the time) or low (vice-versa, when the subjects barely work at all). The medium payoff case involves a mixture of working and leisure, and is more challenging. Since the behaviour of the model is driven by relative utilities, the key quantity controlling the allocation of time is the microscopic benefit-of-leisure function. This qualitatively fits the medium payoff case when it is sigmoidal. Then, the predicted leisure duration distribution is a mixture of an exponential and a gamma-like component, with the weight on the longer, gamma-like component decreasing with payoff.

The microscopic benefit-of-leisure function reflects a subject's innate preference for the duration of leisure when only considering leisure. It is independent of the effects of all other rewards and costs. It is not the same as the Q-value of leisure, which is payoff dependent since it includes the opportunity cost of time (see Eq. (4.2)). For intuition about the consequences of different functions, consider the case of choosing between taking a long holiday all at one go, or taking multiple short holidays of the same net duration. Given a linear microscopic benefit-of-leisure function, these would be equally preferred; however, sigmoidal functions (or other functions with initially supra-linear forms) would prefer the former.

Stochasticity in choices had a further unexpected effect in tending to make subjects pre-commit to a single long work bout rather than dividing work up into multiple short bouts following on from each other. The more bouts the subject used for a single overall work duration, the more likely stochasticity would lead to a choice in favour of leisure, and thus the lower the overall reward rate. Pre-commitment to a single long duration avoids this. Our model therefore provides a novel reason for pre-commitment to executing a choice to completion: the avoidance of corruption due to stochasticity. If there was also a cost to making a decision – either from the effort expended, or from starting and stopping the action at the beginning and ends of bouts, then this effect would be further enhanced. Such switch costs would mainly influence pre-commitment during working rather than the duration of leisure, since there is exactly one behavioural switch in the latter no matter how long it lasts.

Even at very high payoffs, subjects are observed still to engage in short leisure bouts after receiving a reward – the so-called post-reinforcement pause (PRP). This is apparently not instrumentally appropriate, and so we consider PRPs to be Pavlovian. The PRP may consist of an obligatory initial component, which is curtailed by the subject's Pavlovian response to the lever. This obligatory component could be due to the enjoyment or "consumption" of the reward. The task was set up so that instrumental rather than Pavlovian components of leisure dominate, so for simplicity we assumed the latter to be a payoff-dependent constant (rather than being a random variable). We can only model PRPs rather crudely, given the paucity of independent data to fit – but our main conclusions are only very weakly sensitive to changes.

By integrating molecular choices we derived molar quantities. A standard molar psychological ac-

count assumes that subjects match their time allocation between work and leisure to the ratio of their payoffs as in a form of the generalized matching law [8, 9, 11, 14, 16]. This has been used to yield a 3-dimensional relationship known as a mountain, which directly relates time allocation to objective reward strength and price [19, 21]. However, the algorithmic mountain models depend on a rather simple assignment of utility to leisure that does not have the parametric flexibility to encompass the issues on which our molecular model has focused. Those issues can nevertheless have molar signatures – for instance, if the microscopic benefit-of-leisure is eventually sub-linear, then as the price becomes very long, extended leisure bouts are less likely to get censored and so, the subject would then be observed to resume working before the end of the trial. Integrating this, at long prices, time allocation would be observed not to decrease, and even increase with the price, a prediction not made by any existing macroscopic model. Experimentally testing whether this prediction holds true would shed light on the types of non-linear microscopic benefit-of-leisure functions and their parameters actually used by subjects.

Another standard molar (but computational) approach comes from the microeconomic theory of labour supply [1]. Subjects are assumed to maximize their *macroscopic* utility over combinations of work and leisure, [3, 5, 18]. If work and leisure were imperfect substitutes, so leisure is more valuable given that a certain amount of work has been performed, and/or vice-versa, then perfect maximizers would choose some of each. For our model, leisure and work are formally perfect substitutes; and we use stochasticity to capture the substantial variability evident at a molecular scale and thus also molar time allocation.

Behavioural economists have investigated real-life time allocation [2, 3, 5], including making predictions which seemingly contradict those made by labour supply theory accounts [4]. For instance, [4] found that New York City taxi drivers gave up working for the day once they attained a target income, even when customers were in abundance. Contrary to this finding, in the experimental data we model, subjects work nearly continuously when the payoff is high rather than giving up early.

One class of model that does make predictions at molecular as well as molar levels involves the continuous time Markov chains popular in ethology [17]. In these models, the entire stream of observed behaviour (work and leisure bouts) can be summarized by a small set of parametric distributions, and the effect of variables like payoff can be assessed with respect to how those parameters change. These models are descriptive, characterising what the animal does, rather than being normative.

Our micro-SMDP model has three revealing variants. One is a nano-scopic MDP, for which choices are made at the finest possible temporal granularity rather than having determinable durations (so a long work bout would turn into a long sequence of 'work-work-work…' choices). This model has a straightforward formal relationship to the micro-SMDP model [61]. The distinction between these formulations cannot be made behaviourally, but may be possible in terms of their neural implementations. The second, minor alteration, restricts transitions to those between work and leisure, precluding the above long sequences of choices. The third variant is to allow a wider choice of actions, notably a 'quit', which would force the subject to remain at leisure until the end of the trial. This is simpler, and can offer a normative account of behaviour for high and low payoffs. However, in various cases, subjects resume working after long leisure bouts, whereas this should formally not be possible following quitting.

Considered more generally, quitting can be seen as an extreme example of correlation between successive leisure durations – and it is certainly possible that quantitative analyses of the data will reveal subtler dependencies. One source of these could be fatigue (or varying levels of attention or engagement). The CHT procedure (with trailing trials enabling sufficient rest) was optimised to provide stable behavioural performance over long periods. However, fatigue together with the effect of payoff might explain aspects of the microstructure of the data, especially on medium payoff trials. Fatigue would lead to runs of work bouts interspersed with short leisure bouts, followed by a long leisure bout to reset or diminish the degree of fatigue. Note, however, that fatigue would make work and leisure imperfect substitutes.

We modelled epochs in a trial after the reward intensity and price are known for sure. The subjects

repeatedly experience the reward intensity and price conditions during training over many months, and so would be able to appreciate them after minimal experience on a given trial. However, before this minimal experience, subjects face partial observability, and have to decide whether to explore (by depressing the lever to find out about the benefits of working) or exploit the option of leisure (albeit in ignorance of the price). This leads to a form of optimal stopping problem. However, the experimental regime is chosen broadly so that subjects almost always explore to get at least one sample of the reward and the price (the pink shaded bouts in Figure 1D).

Finally, having raised computational and algorithmic issues, we should consider aspects of the neural implementation of the microscopic behaviour. The neuromodulator dopamine is of particular interest. Previous macroscopic analyses have revealed that an increase in the tonic release of the neuromodulator dopamine shifts the 3-dimensional relationships towards longer prices [21, 22, 23], as if, for instance, dopamine multiplies the intensity of the reward. Equally, models of instrumental vigour have posited that tonic dopamine signals the average reward rate, thus realizing the opportunity cost of time [24, 62, 63]. This would reduce the propensity to be at leisure. Finally, it has been suggested as being involved in overcoming the cost of effort [64], a factor that could readily be incorporated into the model. While the ability to discriminate between these various factors is lost in macroscopic analyses, we hope that a microscopic analysis will distinguish them.

7. Methods

Trials were simulated by generating work and leisure bouts from their respective distributions within the cycle in Figure 2A and arbitrarily ending trials at $25 \times price$. All computations and simulations were performed using MATLAB.

Acknowledgments

The authors thank Laurence Aitchison for fruitful discussions. RKN and PD received funding from the Gatsby Charitable Foundation. Y-AB, RBS, KC and PS received funding from Canadian Institutes of Health Research grant $MOP74577$, Fond de recherche Québec - Santé (Group grant to the Groupe de recherche en neurobiologie comportementale, Shimon Amir, P.I.), and Concordia University Research Chair (Tier I).

Author Contributions

Project was formulated by RKN, PD, PS, based on substantial data, analyses and experiments of Y-AB, KC, RS, PS. RKN, PD formalised the model, RKN implemented and ran the model; RKN analysed the molecular ethogram data; Y-AB formalised and implemented a CTMC model. All authors wrote the manuscript.

References

[1] Frank RH. Microeconomics and Behavior. McGraw-Hill Higher Education; 2005.

[2] Kagel JH, Battalio RC, Green L. Economic Choice Theory: An Experimental Analysis of Animal Behavior. Cambridge University Press; 1995.

[3] Battalio RC, Green L, Kagel JH. Income-Leisure Tradeoffs of Animal Workers. The American Economic Review. 1981;71(4):621–632.

[4] Camerer C, Babcock L, Loewenstein G, Thaler R. Labor Supply of New York City Cabdrivers: One Day at a Time. Q J Econ. 1997 May;112(2):407–441.

[5] Green L, Kagel JH, Battalio RC. Consumption-leisure tradeoffs in pigeons: Effects of changing marginal wage rates by varying amount of reinforcement. J Exp Anal Behav. 1987 Jan;47(1):17–28.

[6] Skinner BF. The behavior of organisms: an experimental analysis. New York: Appleton-Century-Crofts; 1938.

[7] Skinner BF. Selection by consequences. Science (New York, NY). 1981 Jul;213(4507):501–4.

[8] Herrnstein RJ. Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav. 1961 Jul;4:267–72.

[9] Herrnstein RJ. Formal properties of the matching law. J Exp Anal Behav. 1974 Jan;21(1):159–64.

[10] Baum WM, Rachlin HC. Choice as time allocation. J Exp Anal Behav. 1969 Nov;12(6):861–74.

[11] Baum WM. On two types of deviation from the matching law: bias and undermatching. J Exp Anal Behav. 1974 Jul;22(1):231–42.

[12] Baum WM. Optimization and the matching law as accounts of instrumental behavior. J Exp Anal Behav. 1981 Nov;36(3):387–403.

[13] Green L, Rachlin H. Economic substitutability of electrical brain stimulation, food, and water. J Exp Anal Behav. 1991 Mar;55(2):133–43.

[14] McDowell JJ. On the falsifiability of matching theory. J Exp Anal Behav. 1986 Jan;45(1):63–74.

[15] Dallery J, McDowell JJ, Lancaster JS. Falsification of matching theory's account of single-alternative responding: Herrnstein's k varies with sucrose concentration. J Exp Anal Behav. 2000 Jan;73(1):23–43.

[16] McDowell JJ. On the classic and modern theories of matching. J Exp Anal Behav. 2005 Jul;84(1):111–27.

[17] Haccou P, Meelis E. Statistical Analysis of Behavioural Data: An Approach Based on Time-structured Models. Oxford University Press, USA; 1992.

[18] Conover KL, Shizgal P. Employing labor-supply theory to measure the reward value of electrical brain stimulation. Games and Economic Behavior. 2005 Aug;52(2):283–304.

[19] Arvanitogiannis A, Shizgal P. The reinforcement mountain: allocation of behavior as a function of the rate and intensity of rewarding brain stimulation. Behav Neurosci. 2008 Oct;122(5):1126–38.

[20] Breton YA, Marcus JC, Shizgal P. Rattus Psychologicus: construction of preferences by self-stimulating rats. Behav Brain Res. 2009 Aug;202(1):77–91.

[21] Hernandez G, Breton YA, Conover K, Shizgal P. At what stage of neural processing does cocaine act to boost pursuit of rewards? PloS one. 2010 Jan;5(11):e15081.

[22] Trujillo-Pisanty I, Hernandez G, Moreau-Debord I, Cossette MP, Conover K, Cheer JF, et al. Cannabinoid receptor blockade reduces the opportunity cost at which rats maintain operant performance for rewarding brain stimulation. J Neurosci. 2011 Apr;31(14):5426–35.

[23] Hernandez G, Trujillo-Pisanty I, Cossette MP, Conover K, Shizgal P. Role of Dopamine Tone in the Pursuit of Brain Stimulation Reward. J Neurosci. 2012 Aug;32(32):11032–11041.

[24] Niv Y, Daw ND, Joel D, Dayan P. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology. 2007 Apr;191(3):507–20.

[25] Baum WM. From molecular to molar: a paradigm shift in behavior analysis. J Exp Anal Behav. 2002 Jul;78(1):95–116.

[26] Baum WM. Molar versus as a paradigm clash. J Exp Anal Behav. 2001 May;75(3):338–41; discussion 367–78.

[27] Baum WM. Molar and molecular views of choice. Behavioural Processes. 2004 Jun;66(3):349–59.

[28] Baum WM. Introduction to molar behavior analysis. Mexican Journal of Behavior Analysis. 1995;21:7–25.

[29] Baum WM. Time-based and count-based measurement of preference. J Exp Anal Behav. 1976 Jul;26(1):27–35.

[30] Rachlin H. A molar theory of reinforcement schedules. J Exp Anal Behav. 1978 Nov;30(3):345–60.

[31] Hineline PN. Beyond the molar-molecular distinction: we need multiscaled analyses. J Exp Anal Behav. 2001 May;75(3):342–7; discussion 367–78.

[32] Ferster C, Skinner BF. Schedules of reinforcement. New York: Appleton-Century-Crofts; 1957.

[33] Gilbert TF. Fundamental dimensional properties of the operant. Psych Rev. 1958 Sep;65(5):272–82.

[34] Shull RL, Gaynor ST, Grimes JA. Response rate viewed as engagement bouts: effects of relative reinforcement and schedule type. J Exp Anal Behav. 2001 May;75(3):247–74.

[35] Williams J, Sagvolden G, Taylor E, Sagvolden T. Dynamic behavioural changes in the Spontaneously Hyperactive Rat: 2. Control by novelty. Behav Brain Res. 2009 Mar;198(2):283–90.

[36] Williams J, Sagvolden G, Taylor E, Sagvolden T. Dynamic behavioural changes in the Spontaneously Hyperactive Rat: 1. Control by place, timing, and reinforcement rate. Behav Brain Res. 2009 Mar;198(2):273–82.

[37] Williams J, Sagvolden G, Taylor E, Sagvolden T. Dynamic behavioural changes in the Spontaneously Hyperactive Rat: 3. Control by reinforcer rate changes and predictability. Behav Brain Res. 2009 Mar;198(2):291–7.

[38] Olds J, Milner P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. J Comp and Phys Psych. 1954 Dec;47(6):419–27.

[39] Breton Y, Conover K, Shizgal P. Probability discounting of brain stimulation reward in the rat.; 2009. 39th Annual Meeting of the Society for Neuroscience (Neuroscience 2009).

[40] Gallistel CR, Leon M. Measuring the subjective magnitude of brain stimulation reward by titration with rate of reward. Behav Neurosci. 1991 Dec;105(6):913–25.

[41] Simmons JM, Gallistel CR. Saturation of subjective reward magnitude as a function of current and pulse frequency. Behav Neurosci. 1994 Feb;108(1):151–60.

[42] Hamilton AL, Stellar JR, Hart EB. Reward, performance, and the response strength method in self-stimulating rats: validation and neuroleptics. Phys & Behav. 1985 Dec;35(6):897–904.

[43] Mark TA, Gallistel CR. Subjective reward magnitude of medial forebrain stimulation as a function of train duration and pulse frequency. Behav Neurosci. 1993 Apr;107(2):389–401.

[44] Leon M, Gallistel CR. The function relating the subjective magnitude of brain stimulation reward to stimulation strength varies with site of stimulation. Behav Brain Res. 1992 Dec;52(2):183–93.

[45] Sonnenschein B, Conover K, Shizgal P. Growth of brain stimulation reward as a function of duration and stimulation strength. Behav Neurosci. 2003 Oct;117(5):978–94.

[46] Sutton RS, Barto AG. Reinforcement learning: An introduction. vol. 28. Cambridge University Press; 1998.

[47] Puterman ML. Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics). Wiley-Blackwell; 2005.

[48] Guitart-Masip M, Fuentemilla L, Bach DR, Huys QJM, Dayan P, Dolan RJ, et al. Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. J Neurosci. 2011 May;31(21):7867–75.

[49] Shidara M, Aigner TG, Richmond BJ. Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. J Neurosci. 1998 Apr;18(7):2613–25.

[50] Breland K, Breland M. The misbehavior of organisms. Am Psychol. 1961;16(11):681–684.

[51] Dayan P, Niv Y, Seymour B, Daw ND. The misbehavior of value and the discipline of the will. Neural Net. 2006 Oct;19(8):1153–60.

[52] Takikawa Y, Kawagoe R, Itoh H, Nakahara H, Hikosaka O. Modulation of saccadic eye movements by predicted reward outcome. Exp Brain Res. 2002 Jan;142(2):284–91.

[53] Gibbon J. Scalar expectancy theory and Weber's law in animal timing. Psych Rev. 1977;84(3):279–325.

[54] Miyazaki K, Miyazaki KW, Doya K. Activation of dorsal raphe serotonin neurons underlies waiting for delayed rewards. J Neurosci. 2011 Jan;31(2):469–79.

[55] Miyazaki KW, Miyazaki K, Doya K. Activation of dorsal raphe serotonin neurons is necessary for waiting for delayed rewards. J Neurosci. 2012 Aug;32(31):10451–7.

[56] Fletcher PJ. Effects of combined or separate 5,7-dihydroxytryptamine lesions of the dorsal and median raphe nuclei on responding maintained by a DRL 20s schedule of food reinforcement. Brain Res. 1995 Mar;675(1-2):45–54.

[57] Jolly DC, Richards JB, Seiden LS. Serotonergic mediation of DRL 72s behavior: receptor subtype involvement in a behavioral screen for antidepressant drugs. Biol Psychiatry. 1999 May;45(9):1151–1162.

[58] Ho MY, Al-Zahrani SS, Al-Ruwaitea AS, Bradshaw CM, Szabadi E. 5-hydroxytryptamine and impulse control: prospects for a behavioural analysis. J Psychopharmacol. 1998;12(1):68–78.

[59] Bizot JC, Thibot MH, Le Bihan C, Soubri P, Simon P. Effects of imipramine-like drugs and serotonin uptake blockers on delay of reward in rats. Possible implication in the behavioral mechanism of action of antidepressants. J Pharmacol Exp Ther. 1988 Sep;246(3):1144–1151.

[60] Bizot J, Le Bihan C, Puech AJ, Hamon M, Thibot M. Serotonin and tolerance to delay of reward in rats. Psychopharmacology (Berl). 1999 Oct;146(4):400–412.

[61] Sutton R, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence. 1999;112:181–211.

[62] Cools R, Nakamura K, Daw ND. Serotonin and dopamine: unifying affective, activational, and decision functions. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2011 Jan;36(1):98–113.

[63] Dayan P. Instrumental vigour in punishment and reward. Eur J Neurosci. 2012 Apr;35(7):1152–1168.

[64] Salamone JD, Correa M. Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. Behav Brain Res. 2002 Dec;137(1-2):3–25.

Figure Legends

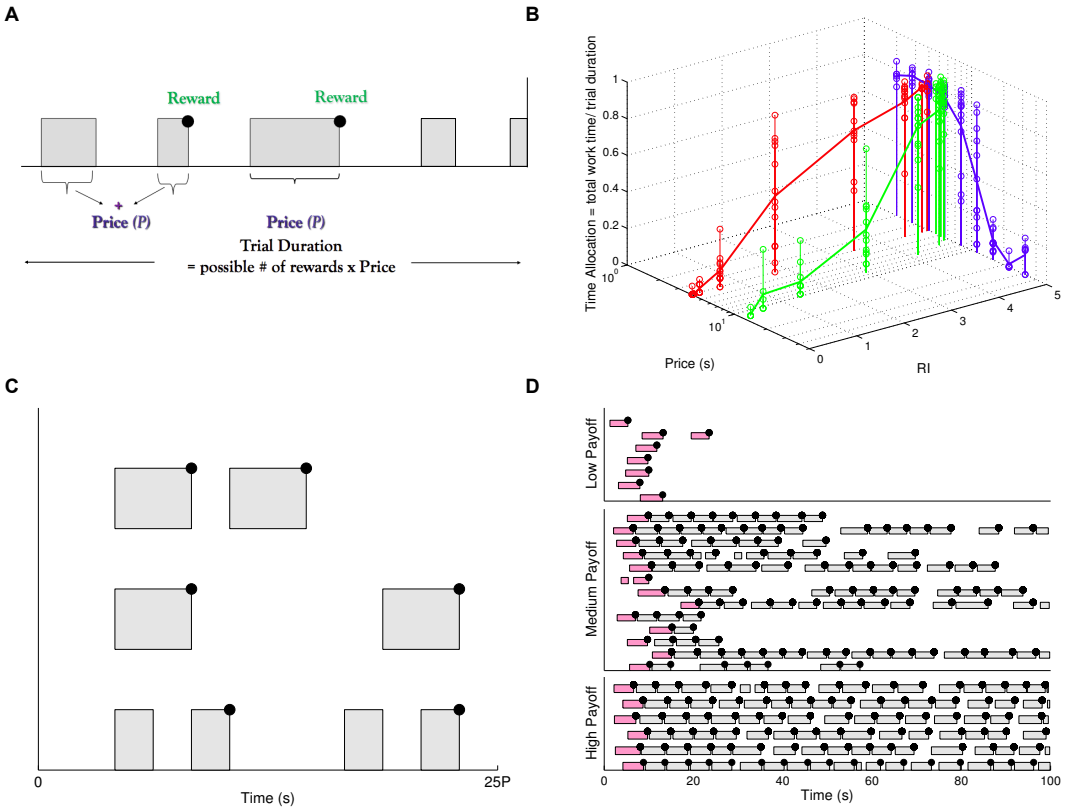

Figure 1. Task and key features of the data. A) Cumulative handling time (CHT) task. Grey bars denote work (depressing a lever), white gaps show leisure. The subject must accumulate work up to a total period of time called the *price* (P) in order to obtain a single reward (black dot) of subjective reward intensity RI. The trial duration is $25 \times$ price (plus 2s each time the price is attained, during which the lever is retracted so it cannot work; not shown). The reward intensity and price are held fixed within a trial. B) Molar time allocation (TA) functions of a typical subject as a function of reward intensity and price. Red curves: effect of reward intensity, for a fixed short price; blue curves: effect of price, for a fixed high reward intensity; green curves: joint effect of reward intensity and price. C) A molecular analysis may reveal different microstructures of working and engaging in leisure. The three rows show three different hypothetical trials. All three microstructures have the same molar TA, but are clearly distinguishable. D) Molecular *ethogram* showing the detailed temporal topography of working and engaging in leisure for the subject in B). Upper, middle and lower panels show high, medium and low payoffs, respectively, for a fixed, short price. Following previous reports using rat subjects, releases shorter than 1 second are considered part of the previous work bout (since subjects remain at the lever during this period). *Graphically*, this makes some work bouts *appear* longer than others. The subject mostly pre-commits to working continuously for the entire price duration. When the payoff is high, the subject works almost continuously for the entire trial, engaging in very short leisure bouts inbetween work bouts. When the payoff is low, the subject engages in a long leisure bout after receiving a reward. This leisure bout is potentially longer than the trial, whence it would be censored. The part of a trial before the reward, price and probability of reward delivery are certainly known is coloured pink and not considered further. Data collected by Y-AB and RS and initially reported in [39].

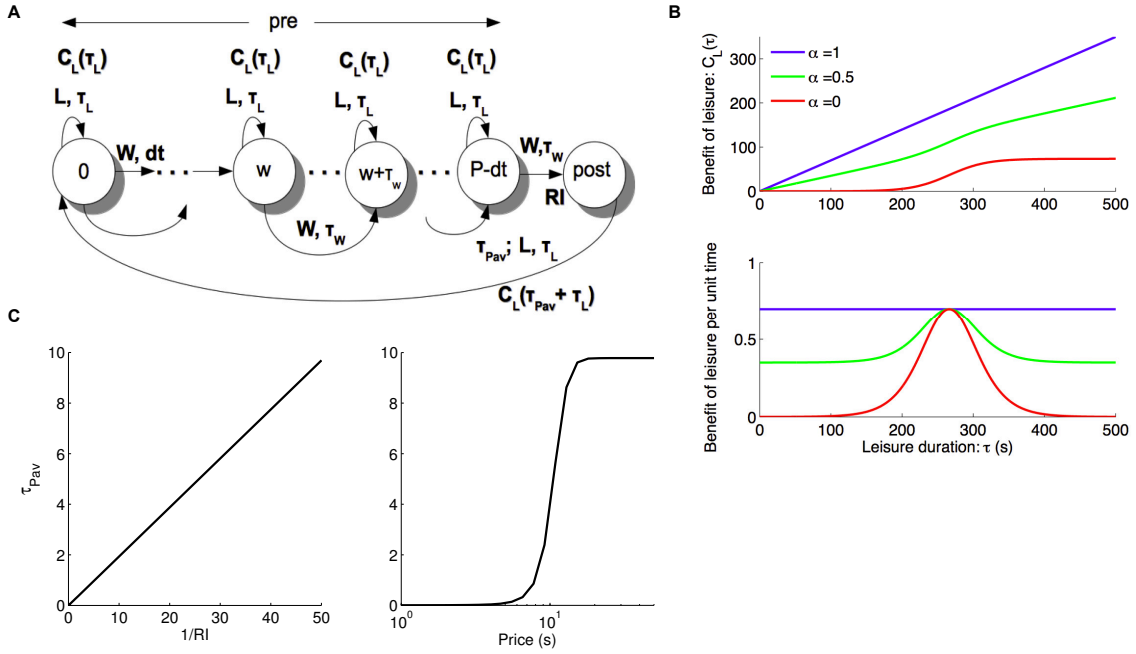

Figure 2. Model and leisure functions. A) The infinite horizon micro semi-Markov decision process (SMDP). States are characterised by whether they are pre- or post-reward. Subjects choose not only whether to work or to engage in leisure, but also for how long to do so. Pre-reward states are further defined by the amount of work time w that the subject has so far invested. At a pre-reward state state [pre,w], the subject can choose to work (W) for a duration τ_W or engage in leisure (L) for a duration τ_L. Working for τ_W transitions the subject to a subsequent pre-reward state [pre,$w+\tau_W$] if $w+\tau_W < P$, and to the post-reward state if $w+\tau_W \geq P$. Engaging in leisure for τ_L transitions the subject to the same state. For working, only transitions to the post-reward state are rewarded, with reward intensity RI. Engaging in leisure for τ_L has a benefit $C_L(\tau_L)$. In the post-reward state, the subject is assumed already to have been at leisure for a time τ_{Pav}, which reflects Pavlovian conditioning to the lever. By choosing to engage in instrumental leisure for a duration τ_L, it gains a microscopic benefit-of-leisure $C_L(\tau_{\mathrm{Pav}}+\tau_L)$ and then returns to state [pre,0] at the start of the cycle whence the process repeats. B) Upper panel: canonical microscopic benefit-of-leisure functions $C_L(\cdot)$; lower panel: the net microscopic benefit-of-leisure per unit time spent in leisure. For simplicity we considered linear $C_L(\cdot)$ (blue), whose net benefit per unit time is constant, sigmoidal $C_L(\cdot)$ (red), which is initially supra-linear but eventually saturates and so has a unimodal net benefit per unit time; and a weighted sum of these two (green). See Eq. (S-3) for details. C) Time τ_{Pav} is the Pavlovian component of leisure, reflecting conditioning to the lever. It is decreasing with reward intensity (here, inversely) and increasing with price (here sigmoidally), so that it decreases with payoff.

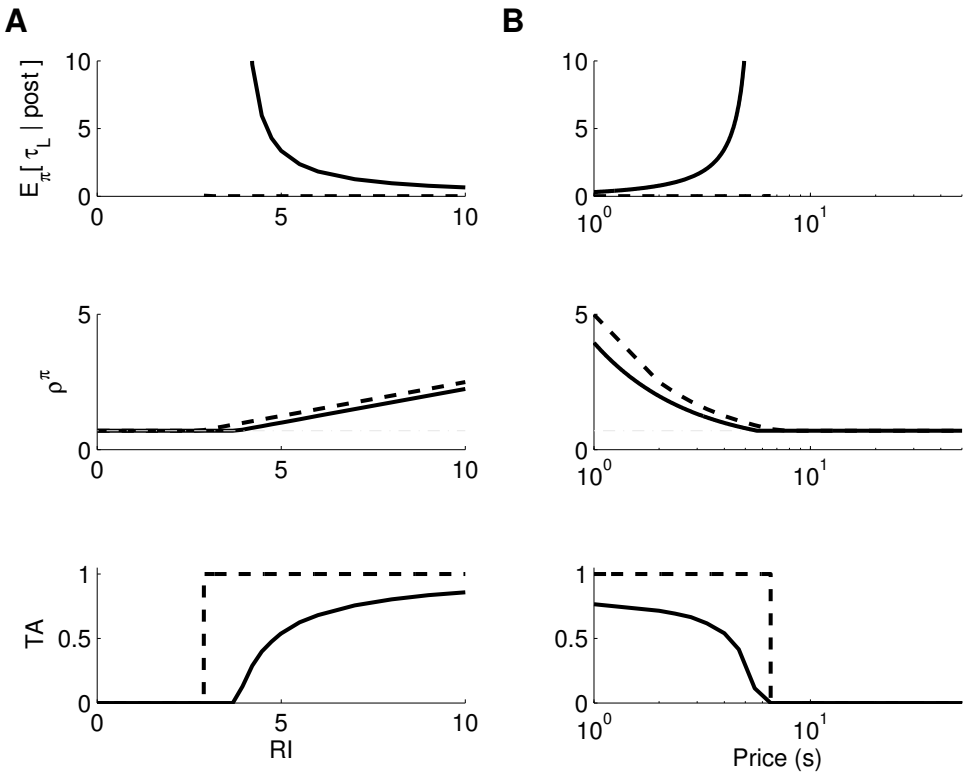

Figure 3. Effect of stochasticity. We use a linear microscopic benefit-of-leisure function ($\alpha = 1$) to demonstrate the effect of stochasticity on: upper panels, mean instrumental leisure, post-reward; middle panels, expected reward rate; lower panels, time allocation as a function of A) reward intensity and B) price. Solid and dashed black lines denote stochastic ($\beta = 1$) and deterministic, optimal ($\beta \to \infty$) choices, respectively. Grey dotted line in middle panels are $\rho^{\pi} = K_L$. Time allocations are step functions under a deterministic, optimal policy but smooth under a stochastic one. Price $P = 4s$ in A, while reward intensity $RI = 4.96$ in B).

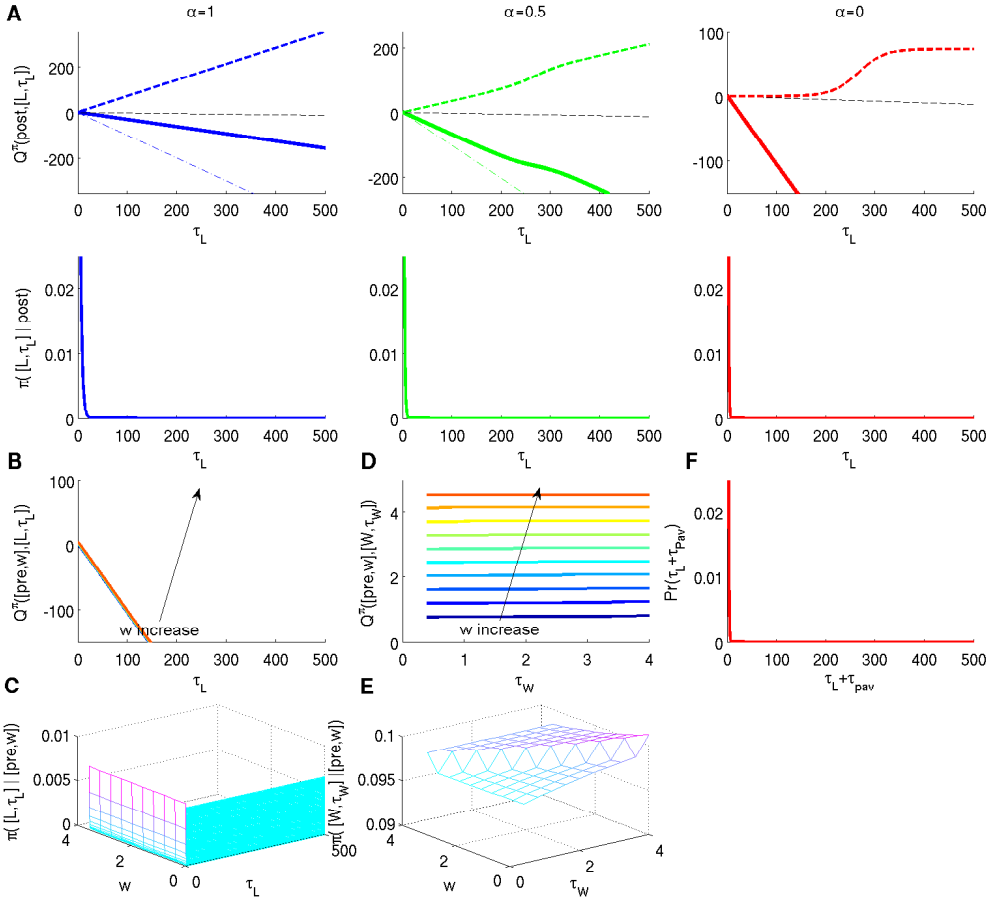

Figure 4. Q-values and policies for a high payoff. A) Upper and lower panels show Q-values and policies for engaging in instrumental leisure for time τ_L, respectively, in the post-reward state for three canonical $C_L(\cdot)$. In upper panels, solid bold curves show Q-values; coloured dashed and dash-dotted lines show $C_L(\cdot)$ and the opportunity cost of time, respectively. Black dashed line is the linear component from the effective prior probability density for leisure time $-\lambda\tau_L$. Note the different y-axis scales. B;D) Q-values and C;E) policies for (B;C) engaging in leisure for time τ_L and (D;E) working for time τ_W in a pre-reward state [pre,w]. Blue to red shows increasing w, i.e., subject is furthest away from the price for blue, and nearest to it for red. F) Probability of engaging in leisure for net time $\tau_L + \tau_{\mathrm{Pav}}$ in the post-reward state for sigmoid $C_L(\cdot)$ ($\alpha = 0$). This is the same as the lower right panel in A) but shifted by τ_{Pav}. Reward intensity, $RI = 4.96$, price $P = 4$s.

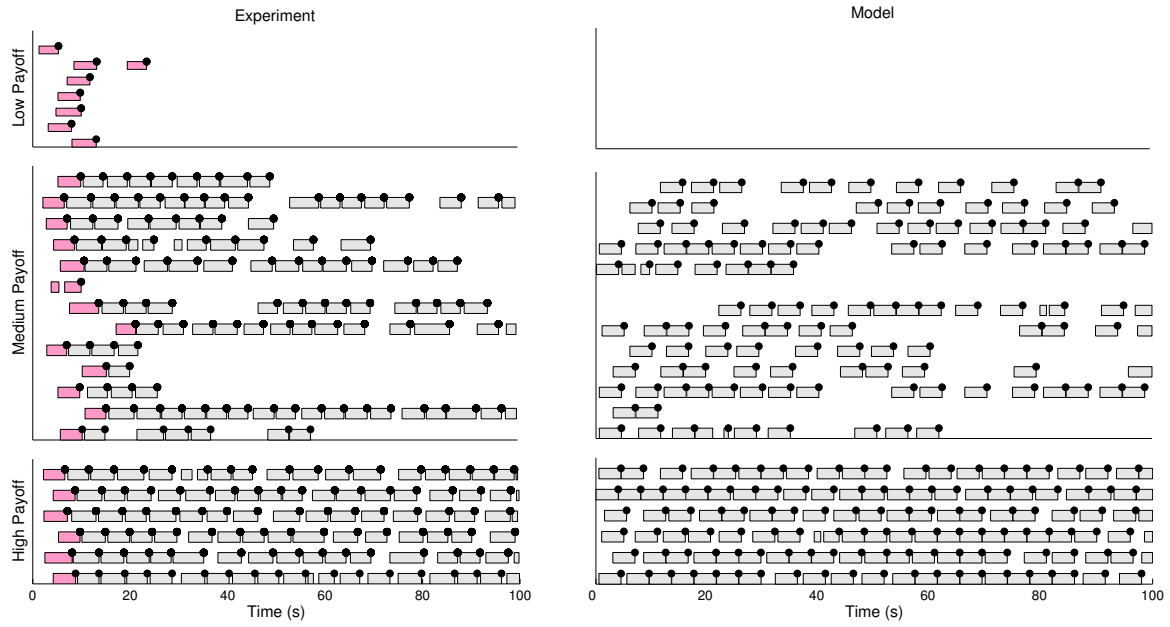

Figure 5. Micro SMDP model with stochastic, approximately optimal choices accounts for key features of the molecular data. Ethogram data from left: experiment and right: micro SMDP model. Upper, middle and lower panels show high, medium and low payoffs, respectively. Pink bars show work bouts before the subject knows what the reward and price are. These are excluded from all analyses, and so do not appear on the model plot.

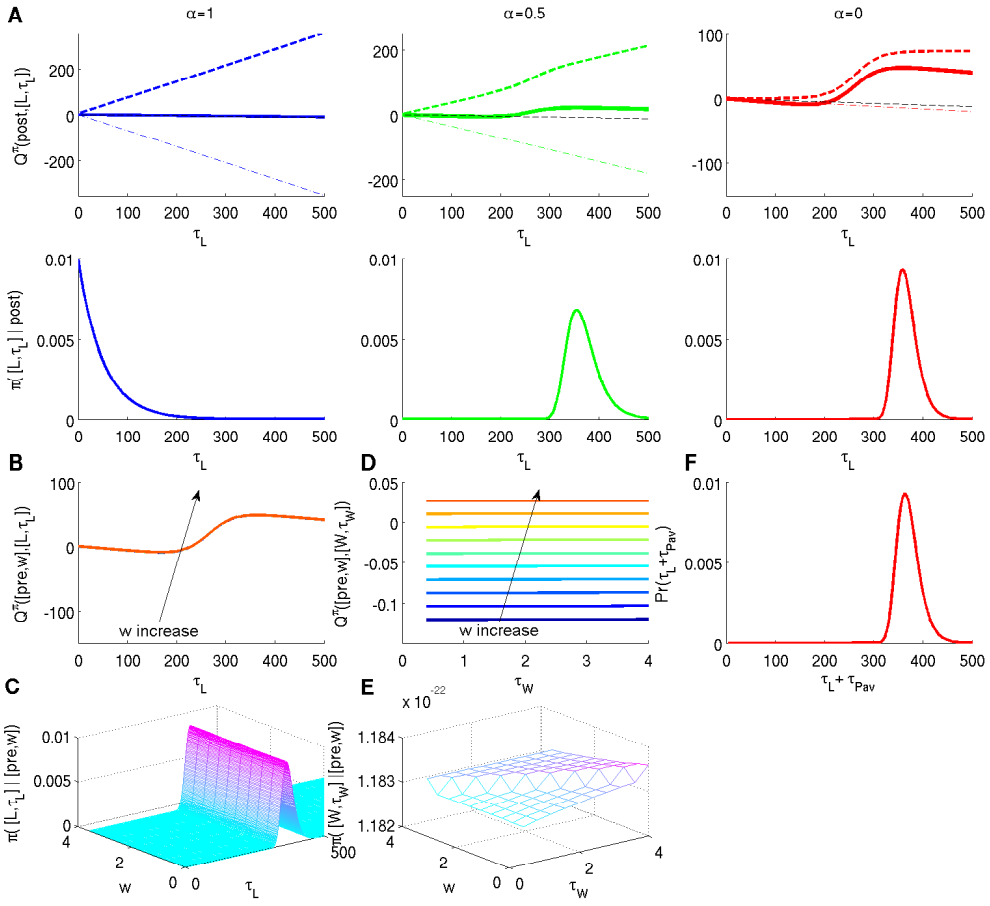

Figure 6. Q**-values and policies for a low payoff.** Panel positions as in Figure 4. Reward intensity, $RI = 0.04$, price $P = 4$s.

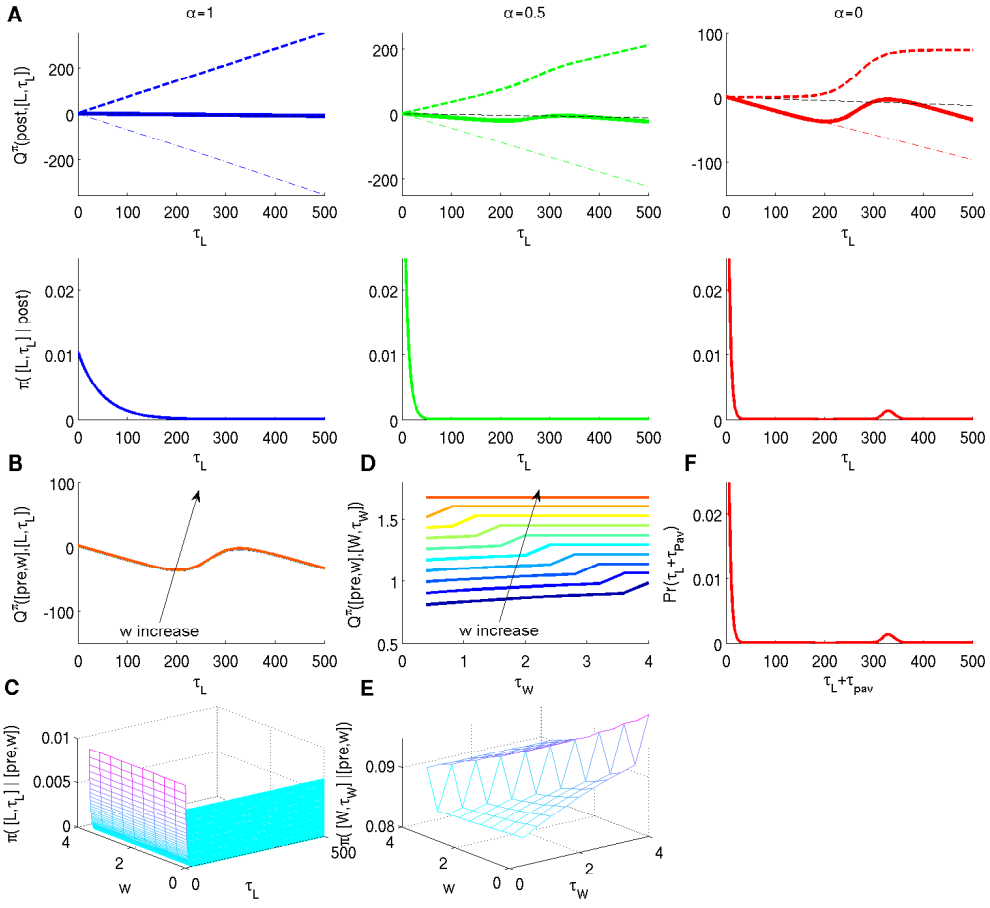

Figure 7. Q**-values and policies for a medium payoff.** Panel positions as in Figure 4. Reward intensity, $RI = 1.76$, price $P = 4$s.

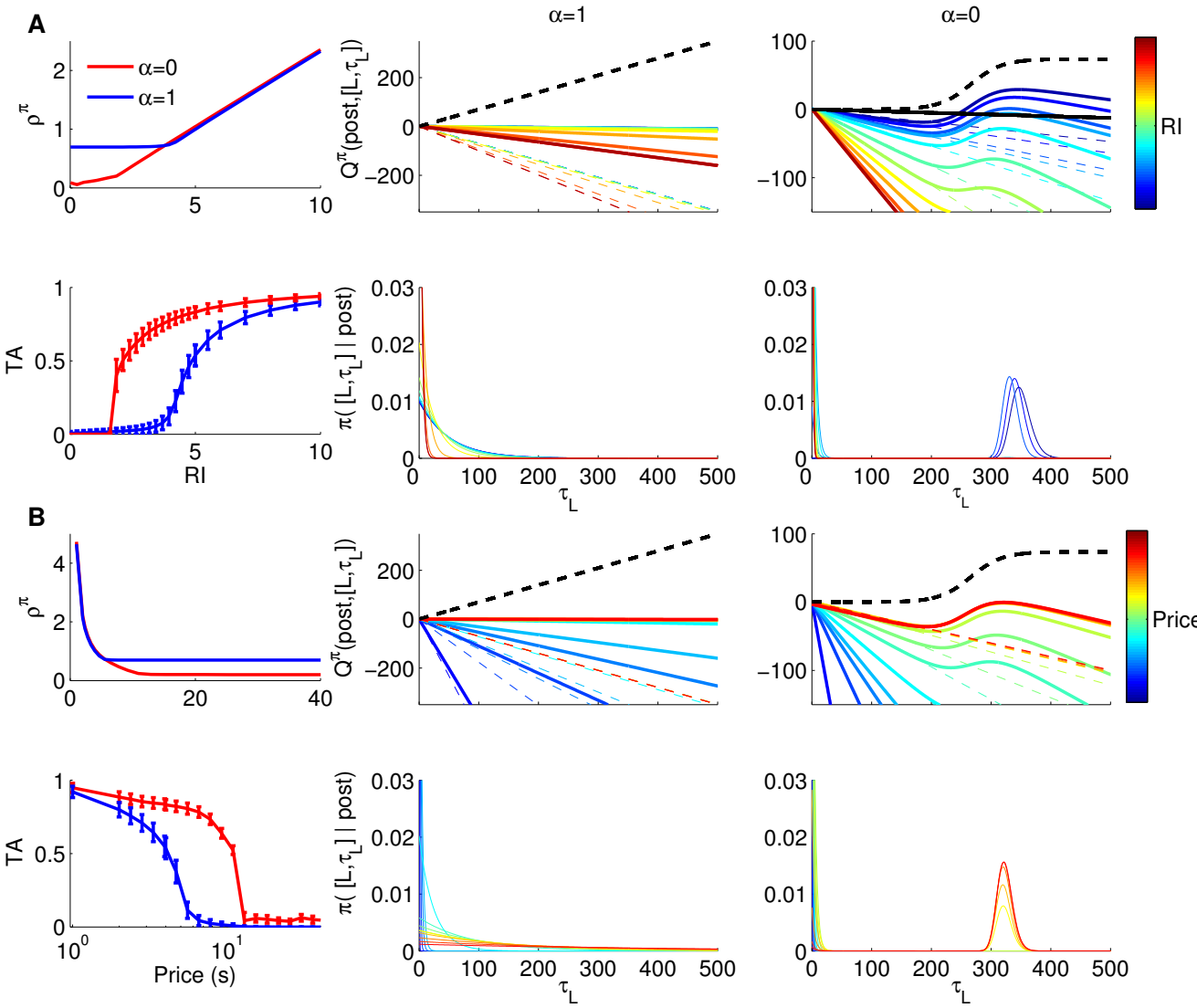

Figure 8. Macroscopic characterisations of behaviour. A) Effect of reward intensity for a short price ($P = 4$s). Upper and lower left panels: reward rate ρ^{π} and time allocation TA, respectively. Blue and red curves are for linear ($\alpha = 1$) and sigmoid ($\alpha = 0$) $C_L(\cdot)$ respectively; error bars are standard deviations. Centre and right panels: Q-values and policies for engaging in instrumental leisure for time τ_L in the post-reward state for linear (centre) and sigmoid (right) $C_L(\cdot)$.Black dashed line in upper panel shows $C_L(\cdot)$; dashed and solid bold coloured curves show the opportunity cost of time and Q-values, respectively. Blue to red denotes increasing reward intensity. B) Effect of price for a high reward intensity ($RI = 4.96$). Panel positions as in A). Note that the abscissa in the upper left panel is on a linear scale to demonstrate the hyperbolic relationship between reward rate and price. Blue to red in the centre and right panels denotes increasing price. C) Left: probability of engaging in leisure for net time $\tau_L + \tau_{\mathrm{Pav}}$ in the post-reward state, and right: ethograms for two long prices (red: $P = 30.1$s and pink: $P = 21.4s$). Reward intensity is fixed at $RI = 4.96$. As the price is increased, reward rate asymptotes (B, upper left panel) and hence the mode of this probability distribution does not increase by much. The trial duration, proportional to the price does increase. Therefore more of the probability mass (grey shaded area) is included in each trial. Samples drawn from this distribution for the lower price get censored more often. For a longer price, the subject is more often observed to resume working after a long leisure bout. The effect is an increase in observed time allocation.

Table 1: List of symbols

Symbol	Meaning
$1/\lambda$	mean of exponential effective prior probability density for leisure time
$\alpha \in [0, 1]$	weight on linear component of microscopic benefit-of-leisure
$\beta \in [0, \infty)$	inverse temperature or degree of stochasticity-determinism parameter
CHT	Cumulative Handling Time
$C_L(\cdot)$	microscopic benefit-of-leisure
$C_{L_{max}}$	maximum of sigmoidal microscopic benefit-of-leisure
$C_{L_{shift}}$	shift of sigmoidal microscopic benefit-of-leisure
$\delta(\cdot)$	delta function
$\mathbb{E}_\pi$	expected value with respect to policy π
K_L	slope of linear microscopic benefit-of-leisure
L	leisure
$\mu_a(\tau_a)$	effective prior probability density of choosing duration τ_a
P	Price
$\pi([a, \tau_a] \mid \vec{s})$	policy or choice rule: probability of choosing action a, for duration τ_a from state $\vec{s}$
post	post-reward
pre	pre-reward
$Q(\vec{s}, [a, \tau_a])$	expected return or (differential) Q-value of taking action a, for duration τ_a from state $\vec{s}$
ρ	reward rate
$\rho\ \tau_a$	opportunity cost of time for taking action a for duration τ_a
RI	(subjective) Reward Intensity
$\frac{RI}{P}$	payoff
$\vec{s}$	state
TA	Time Allocation
τ_L	duration of instrumental leisure
τ_{Pav}	Pavlovian component of post-reward leisure
τ_W	duration of work
W	work
$w \in [0, P)$	amount of work time so far executed out of the price
$V(\vec{s})$	expected return or value of state $\vec{s}$