Supplementary Information
Optimal indolence: a normative microscopic approach to work and leisure

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

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

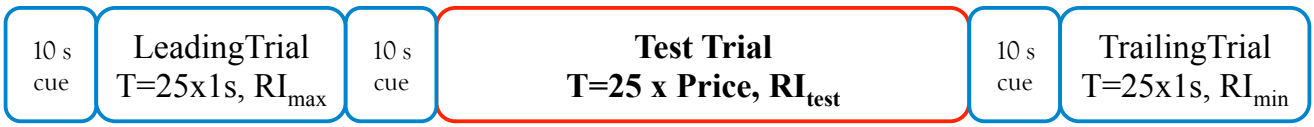

Figure S 1. Experimental procedure: triads of trials Subjects face triads of trials: 'leading', then 'test', then 'trailing'. Throughout a trial, the reward intensity and price are all held fixed; each trial lasts $T = 25$ times the price, plus a fixed, extra time (2s) on each occasion that the price is attained, during which the lever is retracted so that subjects cannot work. This enables the subject to harvest 25 rewards if it works for the entire trial duration. The leading trial involves maximal reward intensity and the shortest (1s) price; the trailing trial involves minimal reward intensity and the shortest (1s) price. Each trial is separated by a 10s cue during which house-lights are switched on, clearly indicating that a trial has ended and a new trial shall begin. The leading and trailing trials were provided so that subjects could calibrate and adequately evaluate the reward and price on test trials. Engaging in leisure on trailing trials also ensured that the subjects would not be fatigued on test trials.

S-1 Supplemental Methods

We formulate our model as a infinite-horizon (unichain) Semi-Markov Decision Process (SMDP) [1]. A state $\vec{s}$ contains all the information necessary for making a decision. 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 of that action, but is independent of all other states, actions and durations in the past. We further assume subjects jointly choose both the actions and their durations, as in [2–4].

A choice rule or *policy* $\pi([a, \tau_a]|\vec{s})$ specifies the subject's probability of taking action a for time τ_a in state $\vec{s}$. Under a given policy, we can define the expected reward rate, or the average reward per unit time

$$\rho^{\pi} = \lim_{\mathrm{T}\to\infty} \frac{\mathbb{E}_{\pi}\left[\sum_{\bar{t}=0}^{\mathrm{T}-1} r_{\bar{t}}([a_{t'}, \tau_{a_{t'}}]) - c_{\bar{t}}([a_{t'}, \tau_{a_{t'}}])\right]}{\mathrm{T}} \tag{S-1}$$

where $r_{t'}$ and $c_{t'}$ denote the benefits and costs at time points t'. Note that the expected reward rate is independent of the starting state.

Normatively, a subject should try to (approximately) maximise its expected return. The expected return or Q-value of taking action a, for duration τ_a from state $\vec{s}$ is

$$
\begin{aligned}
Q^{\pi}(\vec{s},[a,\tau_a]) &= \mathbb{E}_{\pi}\left[\sum_{\bar{t}=0}^{\infty}\left(r_{\bar{t}}([a_{t'},\tau_{a_{t'}}]) - c_{\bar{t}}([a_{t'},\tau_{a_{t'}}]) - \rho^{\pi}\tau_{a_{t'}}\right)|s_t=s, a_t=a, \tau_{a_t}=\tau_a\right] \\
&= \hat{r}(\vec{s},[a,\tau_a]) - \hat{c}(\vec{s},[a,\tau_a]) - \rho^{\pi}\tau_a + V^{\pi}(\vec{s'}) \\
&= \hat{r}(\vec{s},[a,\tau_a]) - \hat{c}(\vec{s},[a,\tau_a]) - \rho^{\pi}\tau_a + \sum_{a'}\int_{\tau_{a'}}\pi([a',\tau_{a'}]|\vec{s'})\,Q^{\pi}(\vec{s'},[a',\tau_{a'}])\,d\tau_{a'} \quad \text{(S-2)}
\end{aligned}
$$

where $V^{\pi}(\vec{s}) = \sum_a \int_{\tau_a} \pi([a,\tau_a]|\vec{s})\; Q^{\pi}(\vec{s},[a,\tau_a])$ is the *value* of state $\vec{s}$, averaged across all actions and their times. The subject pays an automatic *opportunity cost of time* $\rho^{\pi}\tau_a$ for taking action a for time τ_a [2–4]. The Q values in this formulation are approximately equivalent to those obtained using shallow, explicit exponential discounting over an infinite horizon [1, 5].

While simultaneously solving Eqs. (S-1) and (S-2) for the reward rate and the Q-values, we have more unknowns than equations. As conventional, we therefore set the value of a state to 0, and solve for the Q values relative to this baseline. The Q values reported here are therefore *differential* and not the actual ones. We drop differential denotations and simply refer to them as Q-values.

We used a stochastic, approximately-optimal softmax policy over action-duration pairs $[a,\tau_a]$ (see Eq. (4.5)). Subjects will be more likely to choose the action-duration with a greatest Q-value, but have a non-zero probability of choosing a suboptimal action-duration. Since arbitrarily long durations should be less likely to be chosen, this was combined with a prior probability density $\mu_a(\tau_a)$ of choosing duration τ_a to yield the net policy π that generates choices. The reward rate ρ^{π} depends on the policy, and vice-versa (Eqs (4.1-4.5)). Excluding this prior would a priori permit infinitely long leisure durations τ_L to be chosen with the same probability as short ones; these long leisure durations would significantly reduce the reward rate. On the other hand, all work durations τ_W that attain the price ($\tau_W \geq P - w$) would have an identical effect. Since the policy is over all action-durations ($[a,\tau_a]$), irrespective of whether they are of work and leisure, arbitrarily long leisure durations would have a greater effect on the reward rate than work durations. Including a prior that makes longer leisure durations less likely to be chosen normalizes the contributions of durations of work and leisure to the reward rate, affording both an equal role. We therefore employed an exponential prior for leisure $\mu_L(\tau_L)$; the exponential prior for work durations $\mu_W(\tau_W)$ did not matter as long its mean was not so short that it made attaining of the price much unlikely.

Since the policies depend on Q-values, which themselves recursively depend on the policies, except in the case of the optimal policy, we cannot solve for them in closed form. We use policy iteration to find them [1, 6]. Starting from an initial guess, each iteration involves updating the policy while holding the Q-values fixed, and then updating the Q-values while holding the policy fixed, until they are self-consistent, i.e. policy iteration has converged. Since, to our knowledge, policy iteration for stochastic policies has not been proved to converge to a unique policy, we executed the algorithm from different starting points. All policies reported in the main text are the *only* dynamic equilibria of policy iteration (irrespective of the starting point, they converged to the same equilibrium). An alternative would be to compute optimal Q-values (for which policy iteration provably converges to a unique equilibrium

[7]) and then make stochastic choices based on them; however, this would result in policies that are inconstant with their Q-values.

For convenience, we considered a weighted sum of linear and sigmoid benefits of leisure, with the same maximal slope as our canonical microscopic benefit-of-leisure functions

$$C_L(\tau) = \alpha\, K_L\, \tau + (1-\alpha)\, \frac{C_{L_{max}}}{1 + \exp\left[\, -4\frac{K_L}{C_{L_{max}}}(\tau - C_{L_{shift}})\,\right]} \tag{S-3}$$

where $C_{L_{max}}$ and $C_{L_{shift}}$ are the maximum and shift of the sigmoidal component and $\alpha \in [0,1]$ is the weight on the linear component (see Figure 2B)

S-2 Linear benefit-of-leisure yields exponential instrumental leisure duration distributions

If $C_L(\tau_L + \tau_{\mathrm{Pav}})$ is linear in duration τ_L, then, according to Eq. (4.2), the total Q-value of engaging in instrumental leisure in the post-reward state is also linear, $Q^\pi(\text{post}, [L, \tau_L]) = (K_L - \rho^\pi)(\tau_L + \tau_{\mathrm{Pav}}) + V^\pi([\text{pre}, 0])$. Then, according to the softmax policy, the probability of choosing to engage in instrumental leisure for time τ_L in the post-reward state is proportional to the exponential of the Q-value (minus the $\lambda\tau_L$ contributed by the effective prior probability density, see Eq. (4.5)). This probability is $\pi([L, \tau_L]\ |\text{post}) \propto \exp\left[-\{\beta(\rho^\pi - K_L) + \lambda\}\tau_L\right]$, which is an exponential distribution with mean $\mathbb{E}[\tau_L|\text{post}] = \frac{1}{\beta(\rho^\pi - K_L)+\lambda}$. Thus, for linear $C_L(\cdot)$, instrumental leisure bout durations are always exponentially distributed with a mean which depends on the reward rate. The greater the reward rate, the shorter is the mean leisure bout.

When $C_L(\tau_L + \tau_{\mathrm{Pav}})$ is nonlinear, it is typically not possible to derive the optimal policy analytically. We therefore report numerical results.

S-3 Derivation of Equation 5.1

We derive the result in Eq. (5.1). We consider a linear $C_L(\tau_L + \tau_{\mathrm{Pav}}) = K_L(\tau_L + \tau_{\mathrm{Pav}})$, and make two further simplifications: (i) the subject does not engage in leisure in the pre-reward state (and so works for the whole price when it works); and (ii) *a priori*, arbitrarily long leisure durations are possible ($\lambda = 0$). Then the reward rate in Eq. (1) becomes

$$\rho^\pi = \frac{RI + K_L\{\,\mathbb{E}[\tau_L|\text{post}] + \tau_{\mathrm{Pav}}\}}{P + \mathbb{E}[\tau_L|\text{post}] + \tau_{\mathrm{Pav}}} \tag{S-4}$$

As discussed in the *Results* section, the probability of engaging in instrumental leisure in the post-reward state is $\pi([L, \tau_L]\ |\text{post}) = \exp\left[-\{\beta(\rho^\pi - K_L)\}\tau_L\right]$, which is an exponential distribution with mean

$$\mathbb{E}[\tau_L|\text{post}] = \frac{1}{\beta(\rho^\pi - K_L)} \tag{S-5}$$

Re-arranging terms of this equation,

$$\rho^{\pi} = \frac{1}{\beta\, \mathbb{E}[\tau_L|\text{post}]} + K_L \tag{S-6}$$

Equating Eqs. (S-4) and (S-6) and solving for the mean instrumental leisure duration $\mathbb{E}[\tau_L|\text{post}]$, we derive

$$\mathbb{E}[\tau_L|\text{post}] = \frac{P + \tau_{\text{Pav}}}{\beta(RI - K_L P) - 1} \tag{S-7}$$

which is the second line of Eq. (5.1). This is the mean instrumental leisure duration as long as $RI - K_L P > 1/\beta$, and $\mathbb{E}[\tau_L|\text{post}] \to \infty$ otherwise. When the former condition holds, we may substitute Eq. (S-7) into Eq. (S-4) and solve for ρ^{π}

$$\begin{aligned} \rho^{\pi} &= \frac{(RI - K_L P)\ [\beta(RI + K_L \tau_{\text{Pav}}) - 1]}{(RI - K_L P)\ \beta(P + \tau_{\text{Pav}})} \\ &= \frac{\beta(RI + K_L \tau_{\text{Pav}}) - 1}{\beta(P + \tau_{\text{Pav}})} \end{aligned} \tag{S-8}$$

which is the first line of Eq. (5.1).

References

1. Puterman ML (2005) Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics). Wiley-Blackwell, 684 pp.

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

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

4. Dayan P (2012) Instrumental vigour in punishment and reward. Eur J Neurosci 35: 1152–1168.

5. Daw ND, Touretzky DS (2002) Long-term reward prediction in TD models of the dopamine system. Neural Computation 14: 2567–83.

6. Sutton R, Barto A (1998) Reinforcement learning: An introduction, volume 28. Cambridge University Press.

7. Singh S (1993) Soft dynamic programming algorithms: Convergence proofs. Proceedings of Workshop on Computational Learning .