The Moveable Feast of Predictive Reward Discounting in Humans

Luke Dickens
Brain & Behaviour Lab
Dept. of Computing
Imperial College London

Bernardo Caldas
Brain & Behaviour Lab
Dept. of Computing
Imperial College London

Benedikt Schoenhense
Brain & Behaviour Lab
Dept. of Bioengineering
Imperial College London

Guy-Bart Stan
Control Engineering Synthetic Biology Lab
Centre for Synthetic Biology and Innovation
Dept. of Bioengineering
Imperial College London

A. Aldo Faisal
Brain & Behaviour Lab
Dept. of Bioengineering
& Dept. of Computing
Imperial College London
& MRC Clinical Sciences Centre, London

Abstract

This work investigates the implicit discounting that humans use to compare rewards that may occur at different points in the future. We show that the way discounting is applied is not constant, but changes depending on context and in particular can be influenced by the apparent complexity of the environment. To investigate this, we conduct a series of neurophysics experiments, in which participants perform discrete-time, sequential, 2AC tasks with non-episodic characteristics and varying reward structure. The varying rewards in our games cause participants behaviour to change giving a characteristic signal of their future reward discounting. Model-free, model-based and hybrid reinforcement learning models are fit to participant data, as well as a lighter weight model which does not assume a learning mechanism. Results show that the complexity of the task affects the geometric discount factor, relating to the length of time that participants may wait for reward. This in turn indicates that participants may be optimising some hidden objective function that is not dependent on the discount factor.

Keywords: reinforcement learning, systems neuroscience, geometric-discount, planning horizon, human learning, human discount factor, psychophysics, task complexity

Acknowledgements

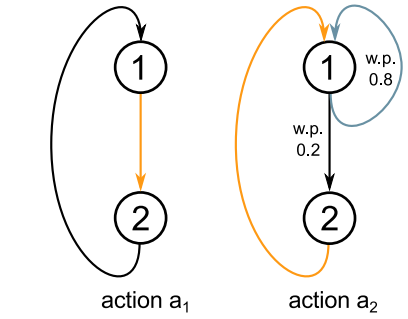
We are deeply indebted to EPSRC for their generous support of this research. This includes the Doctoral Prize Fellowship (Digital Economy) 2011/12, and project numbers $EP/M002187/1$ and $EP/G036004/1$. LDs present address: Dept. of Information Studies, University College London. Address for correspondence: aldo.faisal@imperial.ac.uk.

1 Introduction

Our brains adapt our behaviour in order to improve some measure of success in the world. Because our life is a continuous, ongoing experience, the mechanism by which we learn cannot always rely on having previously experienced present conditions, nor is it always possible to partition previous experience into well defined parts. In the terminology of reinforcement learning, we learn mostly *on-line* in a *non-episodic* environment, and we are (hopefully) *perpetual learners*, i.e. the learning is never completed.

In sequential decision making tasks, rewards may occur distributed over time, yet need to be related to the present time for decision making. To address this, a reward, r, that arises Δ units of time in the future, can be viewed as having the same value as an immediate reward $f(\Delta)r$, where $f(\Delta) \leq 1$ is a monotonically non-increasing function (i.e. f is a discount function). This is often viewed as a *trick* to allow learning (behavioural adaptation) to occur as soon as new experience is available. We call $f(\Delta)r$ the *present value* of future reward r. This discounting of future rewards may be more than just a mathematical convenience *trick*. For instance, it is often also used to explain how humans defer immediate gratification in favour of more substantial long-term outcomes, e.g. in neuroscience and economics. However, two agents that model the world in the same way, but use different discount functions f, will not have exactly the same preferences, nor will their decisions always coincide with the optimal choice evaluated over their lifetimes. As a consequence, a sophisticated on-line learning agent may sometimes wish to change their discount function f depending on the context, e.g. the complexity of the task, and we show evidence that humans do precisely that.

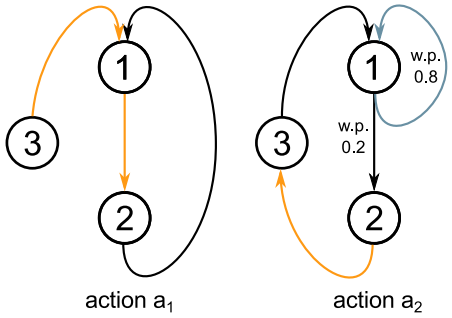
Game 1A	
$r(1, a_1, 2)$	\$0
$r(2, a_1, 1)$	\$25
$r(1, a_2, 1)$	\$25
$r(1, a_2, 2)$	\$0
$r(2, a_2, 1)$	\$x



Games type A

Game 2A		
$r(1, a_1, 2)$	\$0	
$r(2, a_1, 1)$	\$50	w.p. x
	\$0	w.p. (1-x)
$r(1, a_2, 1)$	\$10	
$r(1, a_2, 2)$	\$0	
$r(2, a_2, 1)$	\$10	

Game 1B	
$r(1, a_1, 2)$	\$0
$r(2, a_1, 1)$	\$5
$r(3, a_1, 1)$	\$x
$r(1, a_2, 1)$	\$20
$r(1, a_2, 2)$	\$0
$r(2, a_2, 3)$	\$0
$r(3, a_2, 1)$	\$25



Games type B

Game 2B		
$r(1, a_1, 2)$	\$0	
$r(2, a_1, 1)$	\$10	
$r(3, a_1, 1)$	\$80	w.p. x
	\$0	w.p. (1-x)
$r(1, a_2, 1)$	\$10	
$r(1, a_2, 2)$	\$0	
$r(2, a_2, 3)$	\$0	
$r(3, a_2, 1)$	\$10	

Figure 1: Transition dynamics (central) games played by participants in experiments, step-wise effects of the two actions, a_1 and a_2, shown separately for clarity. Reward structures (left/right) describe two games for each game type – $r(s, a, s')$ is reward for transitioning from state s to s' under action a. Non-deterministic binary outcomes are labelled w.p. (with probability). All games have non-deterministic transitions in state 1 for action a_2. Games 2A & 2B have some non-deterministic rewards. The value of x changes slowly during play.

Related work Reinforcement learning models offer a biologically plausible framework in which to study human behaviour in sequential learning tasks [3–5]. In particular, evidence for reward prediction errors in the brain, have a close analog to the temporal differences in the widely used temporal difference (TD) learning algorithm [4].

In discrete time tasks, some RL methods, including TD, use a geometrically discounted reward, often called $\gamma \in [0, 1]$. Here, a guaranteed reward, r, lying k time-steps in the future is assigned present value $\gamma^k r$. Values of $\gamma < 1$ allow on-line algorithms (those that continuously incorporate new experience) to *approximately* optimise the average reward per time-step [6]. However, for machine learning practitioner γ is typically a free parameter, which is chosen a priori to

suit the learning conditions. The choice of γ can have a profound effect on the trade-off between accuracy and speed of convergence [7], with larger γ leading to more accurate but slower learning.

The work from [1] tests human participants on episodic state-based 2AC choice tasks, and fits multiple RL models to the data, including model-based[1], model-free, and hybrid (combined model-based/model-free, see [2]) variants. They find that a participants' behaviour is best described by hybrid models, and estimate γ values close to 1. However, they use tasks with episodic structure, where the length of episodes is fixed. In these tasks, it is sufficient for participants to optimise the average reward per episode, and update learning after each episode, therefore avoiding the need for on-line learning .

Tasks with non-uniform length to their episodes (and unknown overall time constraints), require participants to optimise over a longer, uncertain period of time, and it is no longer feasible to use $\gamma = 1$. Work described in [5] investigates such a task, where they find that participants with chemically elevated seratonin levels appear to use lower γ values, than when under control conditions. The authors link lower γ values to more impulsive behaviours, i.e. geared towards more immediate gratification. This in turn, may help to explain some of the characteristic behavioural changes that are associated with recreational drug use. However, the work from [5] only fit one RL model to the data, without the same comparative evaluation across multiple models seen here and in [1]. More importantly, we investigate how different degrees of task complexity affect the induced reward discounting.

Contributions The proposed poster will demonstrate the following contributions:

- We develop a suite of discrete-time sequential state-based 2AC (two action choice) tasks, that induce behaviours designed to elicit human reward discounting characteristics.
- We consider a suite of model-free, model-based, and hybrid reinforcment learning (RL) methods, and infer the maximum-likelihood parameters for each model on each participant's data, including the geometric discount γ. We call these *bottom-up* analysis methods.
- We develop a (*top-down*) method for inferring γ without assuming a specific learning mechanism.
- We compare both approaches on synthetic data, and show that bottom-up methods (unsurprisingly) perform more accurately on data generated by a matching model. Conversely, our top method accurately recovers γs from data generated by a wide range of RL methods and parameters.
- Some model-free RL methods consistently outperform other RL models on experimental data. The top-down method consistently gives predictions of γ, which are more stable than the best bottom-up methods and have greater evidential weight (using Akaike's information criterion (AIC)).
- More complex tasks are associated with higher values of γ (using both bottom-up and top-down analysis), and γ has a strong relationship with the characteristic time-scale of the task. However, noise characteristics have at best a weak effect on the induced value of γ.

This last finding indicates that an individual's discounting function can change depending on context, and appears to be influenced by the characteristic time-scale of the task. This may be surprising to some RL researchers, as the vast majority of RL algorithms that use a discounting function keep that function fixed.

2 General Approach

We design sequential state-based 2AC tasks, whose reward structure change slowly, but unpredictably, during play at a medium time-scale. Participants are told to optimise the total reward over the duration of a game, and the changing reward structure forces them to continuously explore and adapt behaviour at the short time-scale to achieve this.

We define a participant's *planning timescale* as the effective number of time-steps in the future, k, that a reward must be for its *present value* to half that of its *immediate value*, i.e. $\frac{r}{2} = f(k)r$. The task is designed such that a user is repeatedly presented with a choice between an immediate small reward versus a longer term, larger reward. We refer to this as a choice between shorter and longer paths. The changes in the reward structure are such that there are periods of the experiment where even participants with very short planning timescale would still prefer to wait for larger rewards. Conversely, there are also periods where even participants with very long planning timescales prefer the short term rewards.

These changes in reward structure will therefore induce behavioural changes in the participants, causing them to switch between preferring shorter paths to preferring the longer paths, and vice versa. These switches can help us to identify each individual's planning timescale, l, in terms of a step-wise geometric discount γ (where $l = \frac{\ln 0.5}{\ln \gamma}$). A variety of games

[1] As with [1], we differentiate between model-free RL methods (pure value based learning), and model-based methods (e.g. including an environmental model to accelerate learning and improve predictions).

are used with different levels of complexity. Some games present a choice between waiting 1 or waiting 2 time-steps for a reward. Other games present a choice between waiting 1 or waiting 3 time-steps for a rewards. We also explore differences between games with deterministic but changing game rewards (Games 1A & 1B, see Section 4), versus games with probabilistic rewards with changing distributions (Games 2A & 2B).

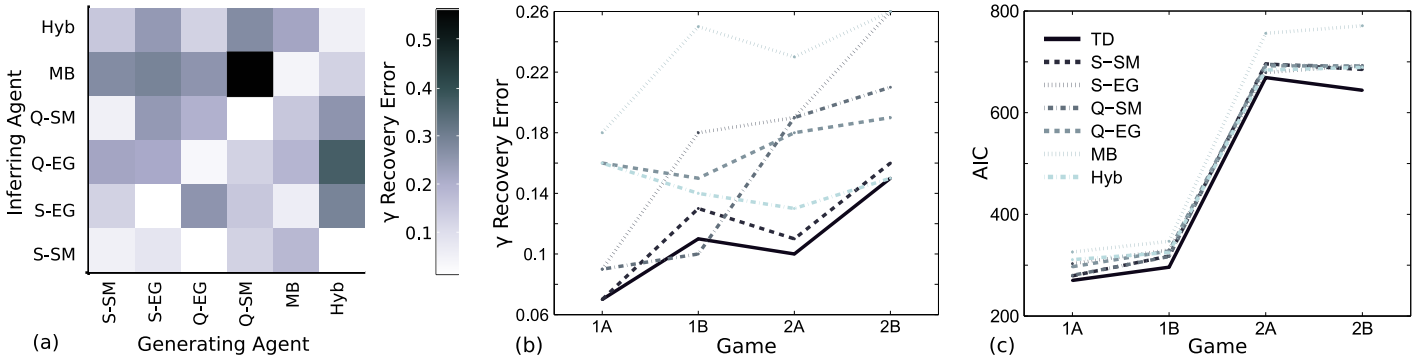


Figure 2: Performance on synthetic data (a) Average γ-recovery accuracy, $||\hat{\gamma} - \gamma||_2$. Data generated on game 1A using *generating agent* and a selection of parameters, then γ recovered by inferring agent. (b) Average γ-recovery accuracy on synthetic data averaged over all generating agents for a variety of generating parameters. (c) Corresponding averaged Akaike's information criterion (AIC). The best performance for each game is highlighted in bold.

3 Results

It is unclear a priori what learning approach participants will employ, whether or not it is equivalent to an RL method, and if so whether it is model-based, model-free or hybrid. We use a gradient based method to find the maximum-likelihood parameters for each method from observed traces of states, actions and rewards. A top-down method is also developed, which assumes that a geometric factor is used to discount future rewards, but without assuming a specific learning mechanism. For brevity the following shorthand is used for inference methods and RL algorithms: Top-down (TD), SARSA Softmax (S-SM), SARSA ε-greedy (S-EG), Q-Learning Softmax (Q-SM), Q-Learning ε-greedy (Q-EG), Model-based (MB) & Hybrid method combining Q-EG + MB (Hyb).

Synthetic Data Figure 2 (a) shows the root mean squared (RMS) γ-recovery accuracy between the recovered discount factor, $\hat{\gamma}$, and the generating value, γ, on synthetic data for a variety of methods. Each RL methods in the suite is used both to generate synthetic data, and to recover γ. As expected, the recovery performance is best when the inferring agent is the same as the generating agent. Figures 2 (b) and (c) respectively show the average RMS recovery accuracy and the average AIC of each bottom-up and the top-down method on synthetic data. Data is averaged over a selection of all agents, each with a variety of the parameters. The top-down method performs the best overall, and competes with the best recovering model in all cases. Also, the top-down method has the best average AIC[2] of all methods.

Experimental Data We next apply these methods to the experimental participant data. We fit each bottom-up and the top-down models to each participant's behaviour, and evaluate the strength of evidence for the given model (using AIC). On the majority of individual participants, and averaged overall, the top-down model has the lowest AIC of all models. Figure 3 (a) shows the mean recovered discount factors $\hat{\gamma}$ for each game using both the top-down (TD) method and the two best performing RL methods. We find $\hat{\gamma}$ in reasonable agreement between these algorithms, both at the population level and individually. To our knowledge, this is the first direct evidence that humans change their discounting function depending on the complexity of the associated game, where complexity is measured in terms of characteristic timescale between rewards. This shows an average $\hat{\gamma} \in [0.55, 0.6]$ for games with a long-path of length 2 (1A & 2A) and an average $\hat{\gamma} \in [0.75, 0.9]$ for games with a long path of length 3 (1B & 2B). There is no discernable difference between games with deterministic and non-deterministic rewards, e.g. Game 1A versus Game 2A.

To explore these results further, we plot in Figure 3 (b) the individually inferred discount factors $\hat{\gamma}$ against the average experienced path length, $\bar{l}$ (time between visits to state 1). The dotted line represents the orthogonal least squares fit to the data after individual variances have been normalised, and the positive slope is in agreement with our hypothesis that longer average pathlengths lead to longer horizons. The pair of variables $\hat{\gamma}$ and $\bar{l}$ have a Pearson's product-moment correlation coefficient of $r = 0.1833$, and the corresponding 1-tailed test for a significant positive linear relationship

[2]AIC is chosen over BIC for simplicity. However, BIC penalises complexity, in terms of number of parameters, more heavily. RL methods have more parameters than the top-down method, so the top-down approach would be preferred by either measure.

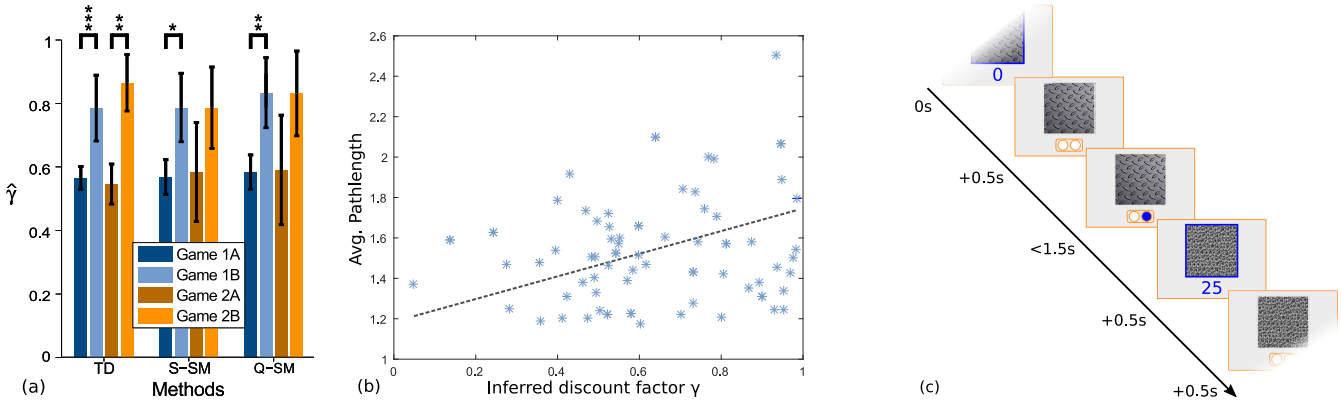


Figure 3: (a) Average inferred discount factors $\hat{\gamma}$ on per game basis, with top-down and two best performing RL methods shown. (b) Individually inferred discount factors $\hat{\gamma}$ versus the average pathlength, $\bar{l}$, experienced by each participant in each game. The dotted line shows the orthogonal least squares fit after normalising the variance. (c) Timeline of events in a single step of a task. The current state is presented (phase 0), a choice is requested (phase 1), the choice is then made – or assigned randomly on time-out (phase 2), the next state and reward is displayed (phase 3/0), and the game continues.

between the two variables gives a p-value of 0.055. The AIC values on participant data give greater support for model-free reinforcement learning models than the model-based or hybrid models (not shown). This is in contrast to the findings of [1] on fixed length episodic tasks. However, further investigations are required to determine if this is significant.

4 Methodology

Psychophysics experiments (N=24 participants) involved basic computer interaction and were conducted in accordance with local ethics committee guidelines. Participants were presented with sequential decision making tasks (see Figure 2 (c)). At each time-step, the participant is presented with an image representing the current state. After a pause of 0.5 seconds, the participant chooses one of two actions within a further 1.5 seconds (or a random choice is made and displayed). After another 0.5 seconds, the resulting state is presented along with a numeric reward, indicating the value of the most recent transition. The reward structure varies slowly so participants must continually explore and adapt. Participants are not told how long the task will last, and therefore must view the task as one with an unknown horizon. There were 4 fundamental tasks, shown in Figure 1, and each participant was presented all 4 tasks in a random order with breaks. Each task (game) repeatedly presents the participant with (limited) control over whether she takes a longer path leading to a larger reward, or a shorter path that leads to a smaller reward. The expected value of the larger reward is varied throughout play. The shorter path in each game is of length 1 time-step. Games of type A (Game 1A & 2A) have a long path length of 2 steps. Games of type B (Game 1B & 2B) have a long path length of 3 steps. Participants are instructed to try to achieve as large a total reward as possible over the lifetime of each game.

Top Down Method To predict γ with the top-down method, we (a) estimate the underlying policy at each time-step, (b) identify the time points in the trace when the dominant action in the policy changes, and (c) use gradient ascent to determine the maximum likelihood value for the gaussian/exponential filter width over recent rewards and a corresponding γ which best explains the switching points.

References

[1] N. Daw, S.J. Gershman, B. Seymour, P. Dayan, and R.J. Dolan. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215, 2011.

[2] J. Gläscher, N. Daw, P. Dayan, and J.P. O'Doherty. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron.*, 66(4):585–95, 2010.

[3] A. R. Otto, A. Skatova, S. Madlon-Kay, and N. D Daw. Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, 27(2):319–333, 2015.

[4] W. Schultz, P. Dayan, and PR. Montague. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–9, 1997.

[5] N. Schweighofer, M. Bertin, K. Shishida, Y. Okamoto, S.C. Tanaka, S. Yamawaki, and K. Doya. Low-serotonin levels increase delayed reward discounting in humans. *The Journal of Neuroscience*, 28(17):4528–4532, 2008.

[6] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[7] Huizhen Yu and D.P. Bertsekas. New error bounds for approximations from projected linear equations. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 1116–1123, Sept 2008.