Computational Cognitive Science

Lecture 15: Temporal Discounting

Benjamin Peters

School of Informatics

University of Edinburgh

3 November, 2025

Recap

Last time we focused on reward-based learning and the fundamental trade-off between exploration and exploitation.

Goal selection

Today, we will focus on selecting actions in the presence of multiple goals.

- Multiple goals may be in conflict with each other.
- At any moment and agent may have to choose which goal to advance.
- Particularly salient when rewards have different time horizons.

Reading

- Chebolu, S., & Dayan, P. (2024). Optimal and sub-optimal temporal decisions can explain procrastination in a real-world task. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*(0). link

Goal selection conflict

- “I will have this one piece of cake.” (but also want to loose weight)
- “One more episode.” (need to get up early in the morning)
- “My flat really needs some cleaning.” (assignment is due on Wednesday)
- “It’s raining and cold, but I’ll go for a run.” (because I want to be healthier)

Reinforcement learning

Reinforcement learning describes how an agent learns to act in an environment to maximize total reward.

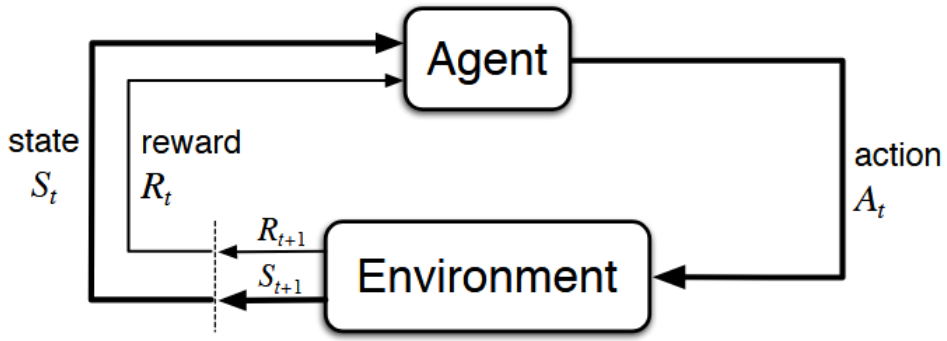


Figure 3.1: The agent–environment interaction in a Markov decision process.

At each time step t, the agent is in some state s_t, chooses an action a_t, receives a reward $R_t = R(s_t, a_t, s_{t-1})$, and transitions to a new state s_{t+1}.

Figure from Sutton & Barto book.

Reinforcement learning

The goal is to find a **policy** $\pi(a|s)$ (a mapping from state to probabilities for particular actions) that maximizes the **expected total reward** over time:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\inf} \gamma^k R_{t+k+1}$$

- The *discount rate* γ ($0 \leq \gamma \leq 1$) determines the present value of future rewards: a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately.
- $\gamma = 0$: myopic agent (only maximize R_{t+1}).
- $\gamma \to 1$: farsighted agent.

Reinforcement learning

How do we define a policy $\pi(a|s)$ that maximizes the expected total reward G_t?

Suppose we knew the *action-value function* $Q_\pi(s, a) = \mathbb{E}_\pi[G_t|s_t = s, a_t = a]$.

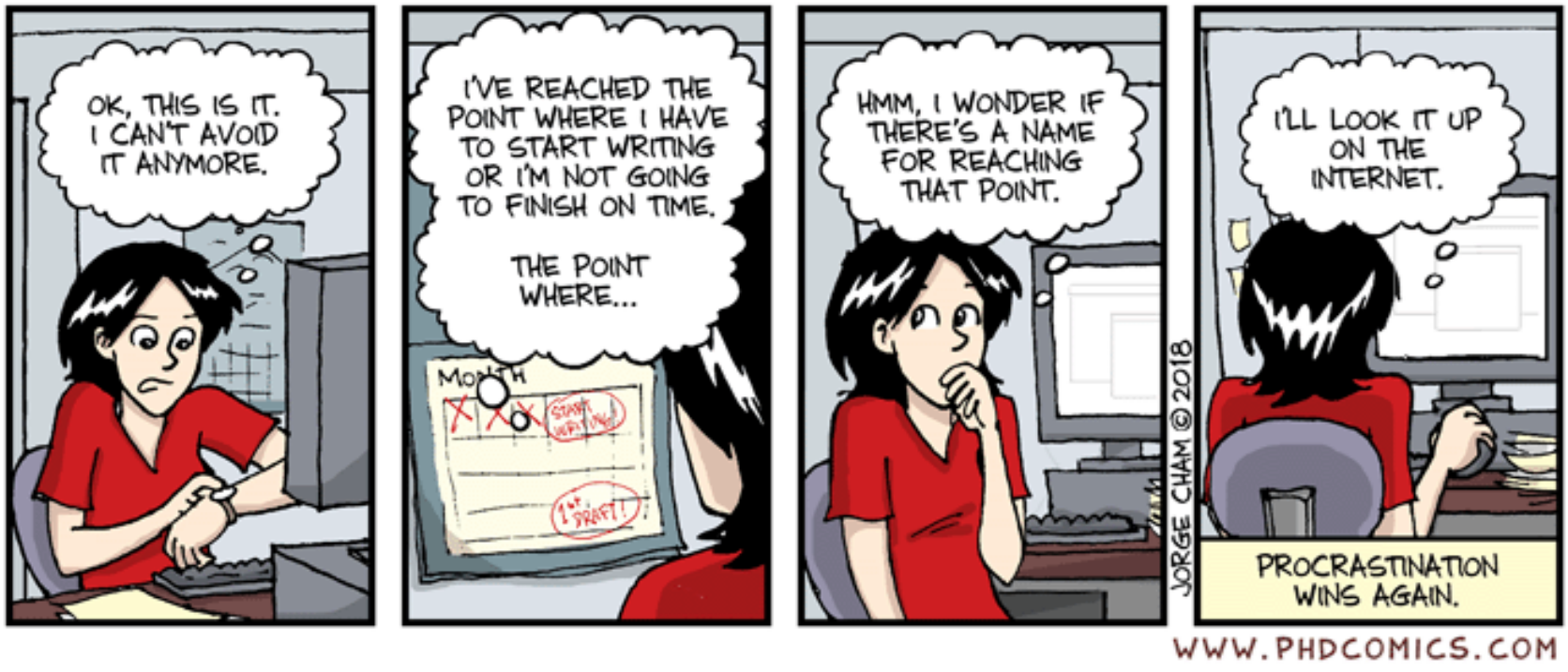
We can define the optimal policy $\pi^*(a|s)$ as always picking the action a^* that maximizes $Q_\pi(s, a)$.

If we have a model of the environment $P(s'|s, a)$, i.e., how the agent transitions from state s to state s' when taking action a, we can compute an optimal policy (via dynamic programming).

Reinforcement learning

For a given environment, $P(s'|s,a)$, and reward function $R(s',a,s)$, and discount rate γ, RL gives a way to determine what a rational agent should be doing in order to maximize the expected total reward.

Procrastination

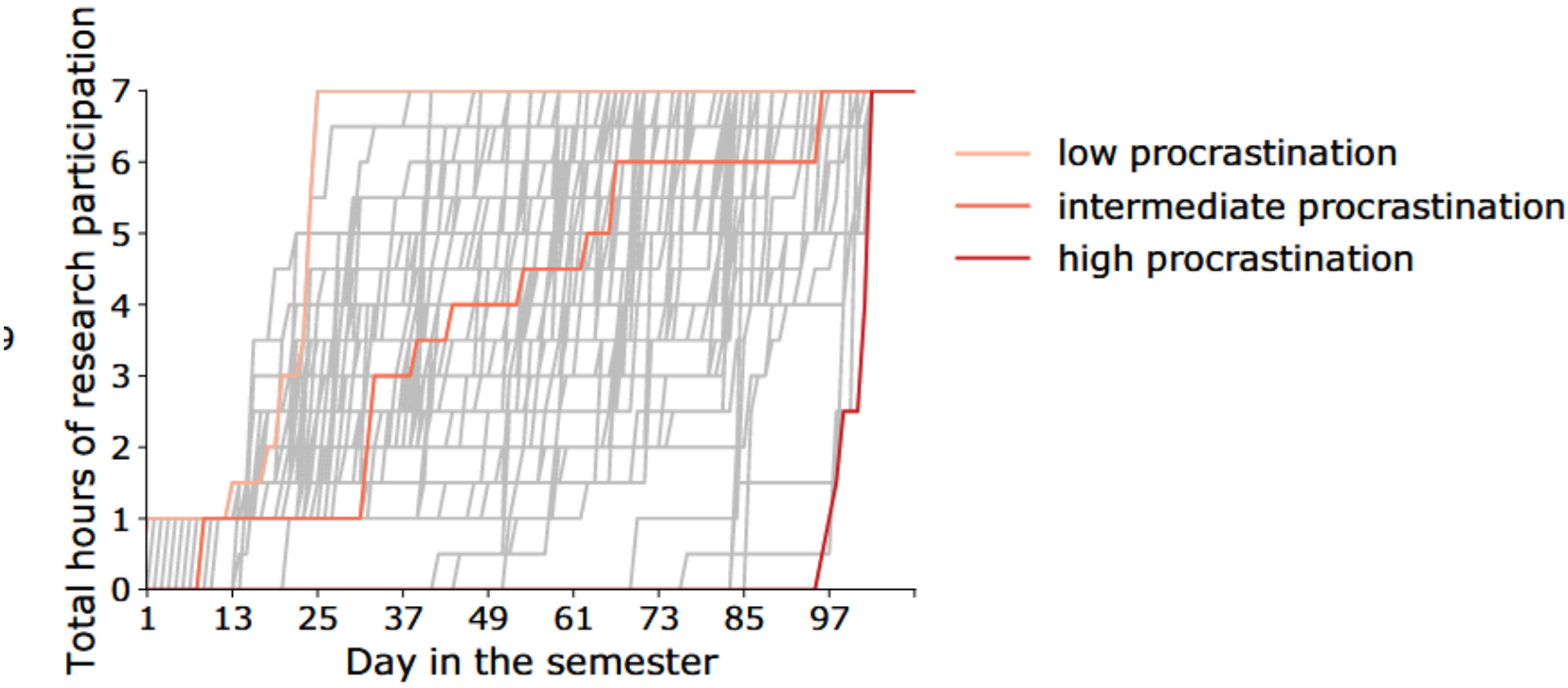


PhD Comics

Procrastination

- Procrastination is the voluntary delay of an intended course of action despite expecting to be worse off for the delay (Steel, 2007).
- Affects approximately 20% of all adults and $\sim$ 80% of college students.
- Aversive: 95% of procrastinators wish to reduce the behaviour.

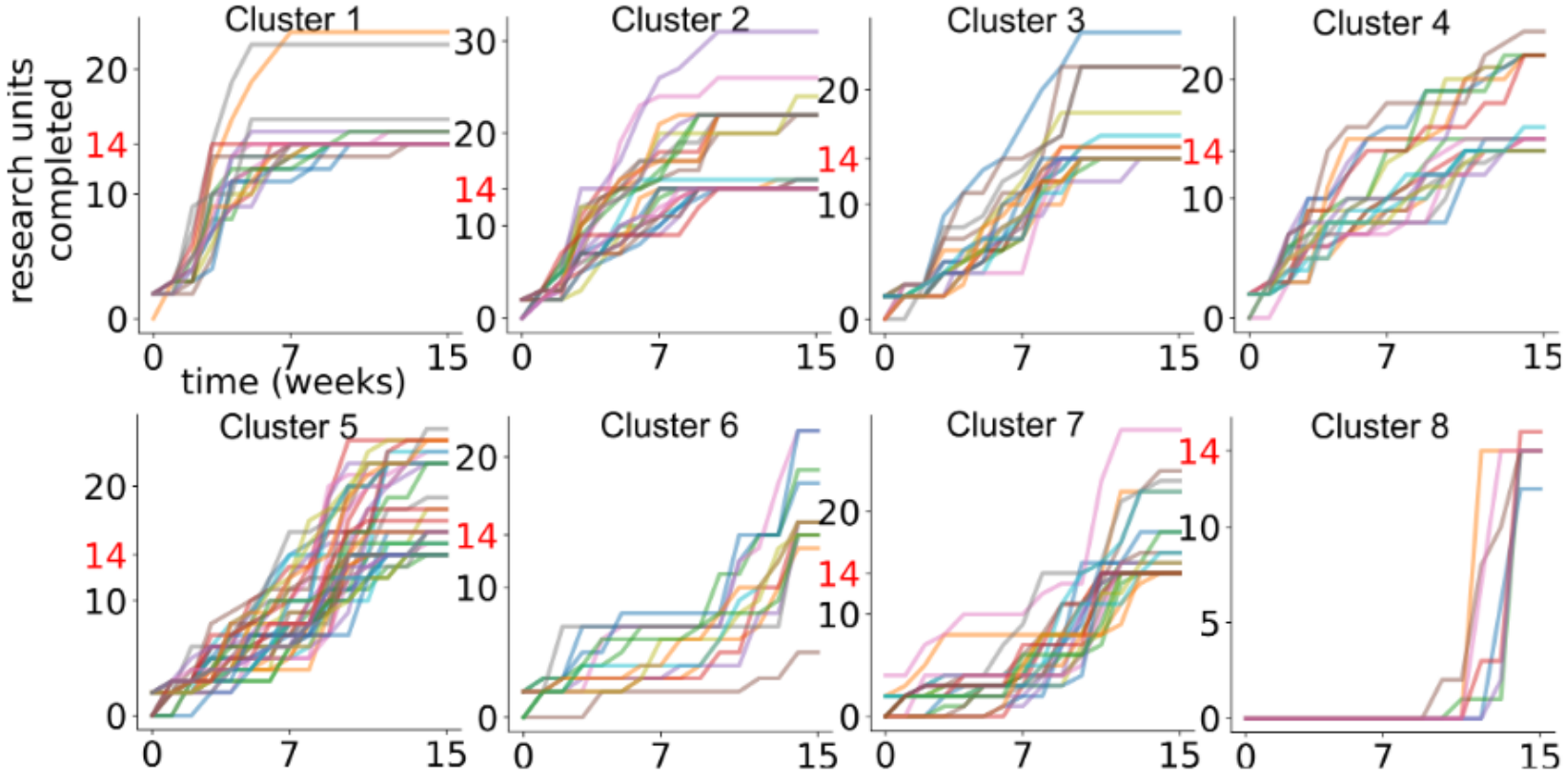
Measuring procrastination



- Zhang & Ma (2024) measured progress in a real-world task.
- Psychology students completed a self-paced 7-hour research participation requirement over the course of a semester. Students got grade-point incentive for an additional 4 hours.
- Result: positive correlation between when most of the work was completed and reward discounting rate.

Modeling procrastination

Chebolu & Dayan (2024) identify eight clusters of behaviour.



They use RL to explain procrastination behaviour as rational.

Environment

- The agent has 16 weeks ($0 \leq t \leq 15$) to complete at least 14 units and up to 22 units of work (one unit: $\frac{1}{2}$ hour).
- In week t, the agent can decide to complete some a_t number of units ($0 \leq a_t \leq 22 - s_t$).
- s_t: number of completed units in week t
- Binomial success probability η determines the efficacy to actually completing a work unit (e.g., bad time planing has lower efficiancy)

$$P(s' \mid s, a) = \binom{a}{s' - s} \eta^{s'-s} (1 - \eta)^{a-s'+s}$$

Reward structure

- Reward of completing required r_{unit} and additional units r_{extra}.
- Completion rewards are only given at the end of the 16 week period.
- Every work unit a_t comes with an effort cost r_{effort}.
- Vigour cost: actual effort increases with workload $r_{\text{effort}}(a) = r_{\text{effort}} a^k$ $(k \geq 1)$.
- Time not used for working $(22 - a_t$ units) is used to 'shirk' with reward r_{shirk} (alternative work, chores, relaxing)

Model

Obtain optimal action-value function $Q(s, a)$ by maximzing the expected total reward $\mathbb{E} \sum_{t=0}^{T} \gamma^t R(s, a, s')$.

Stochastic (softmax) policy:

$$\pi(a \mid s) = \frac{\exp(\beta Q(s, a))}{\sum_{a'} \exp(\beta Q(s, a'))}$$

where β is an inverse temperature parameter controlling how deterministically the agent selects the optimal action.

Reminder

The *action-value function*

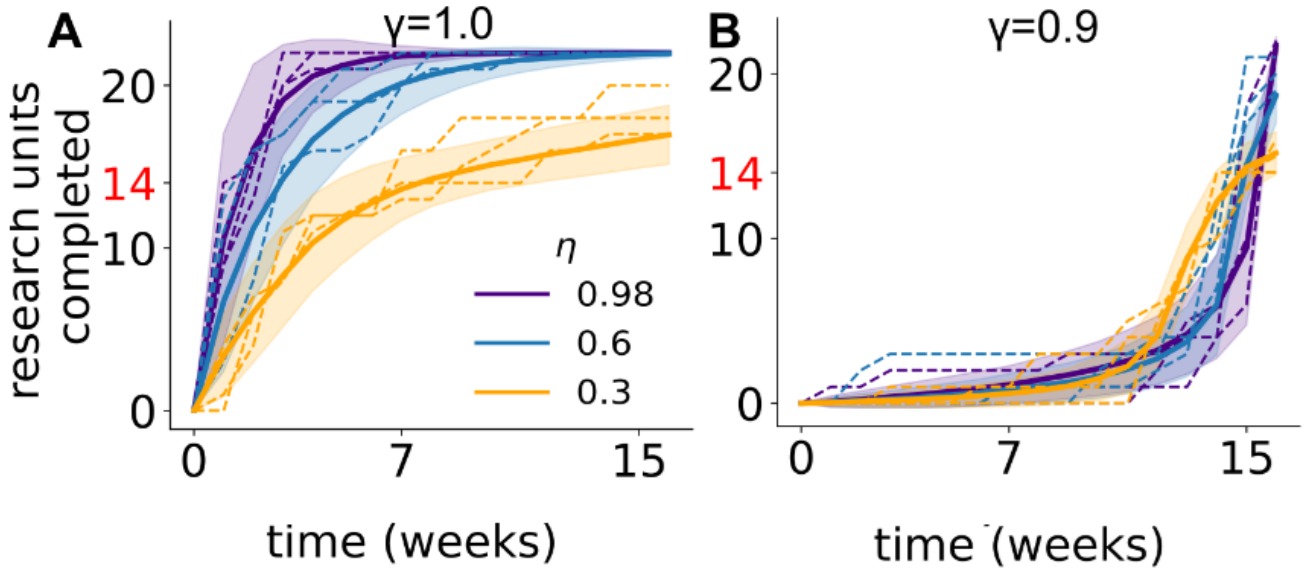
$$Q_\pi(s, a) = \mathbb{E}_\pi \left[G_t | s_t = s, a_t = a \right]$$

The **expected total reward** over time:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\text{inf}} \gamma^k R_{t+k+1}$$

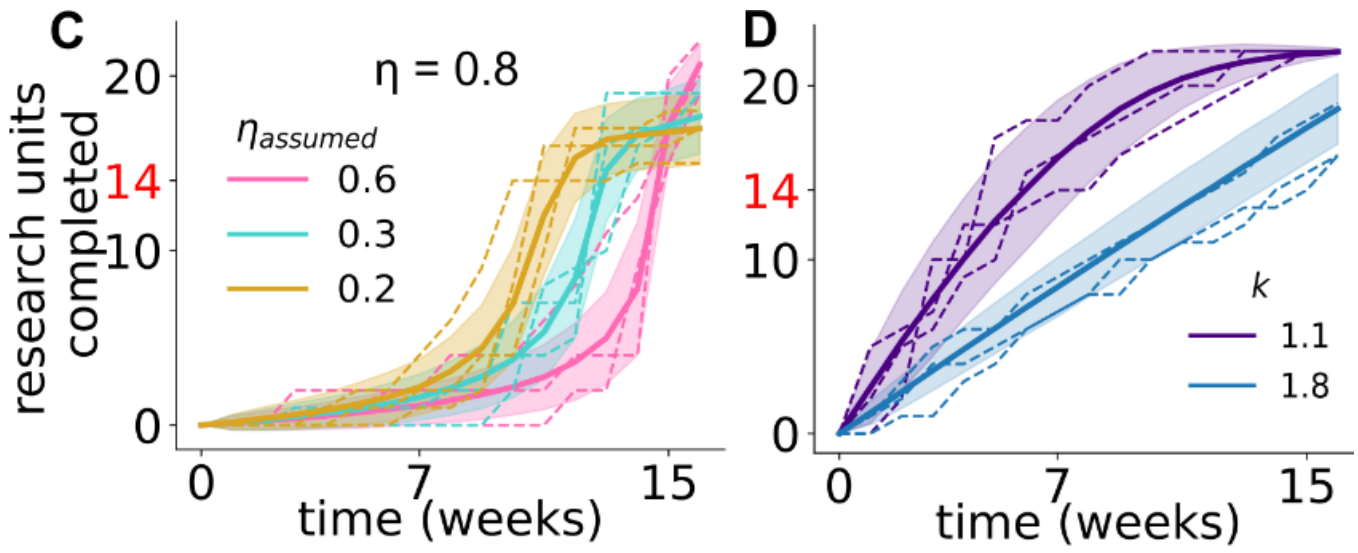
The reward at time t may be the sum of multiple rewards (r_{unit}, r_{extra}, $r_{\text{effort}}(a)$, r_{shirk}).

Results



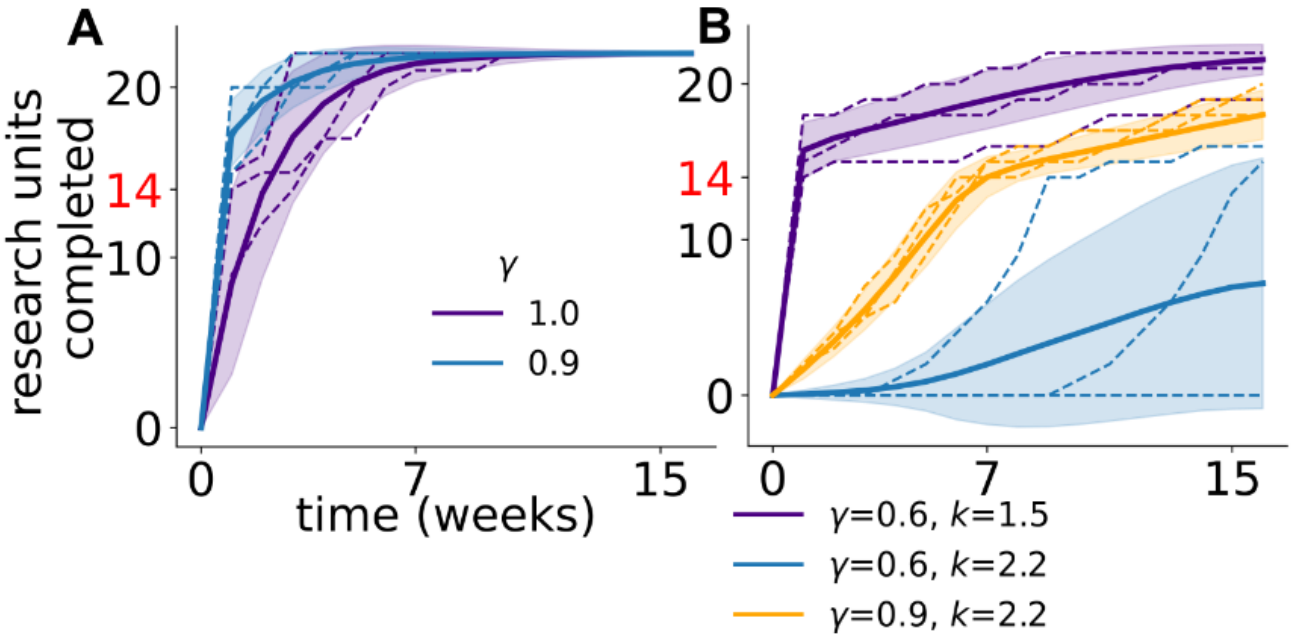
1. Temporal discounting induces a temporal preference for working later: it becomes optimal to put of work if $\gamma < 1$.
2. Efficacy affects the extent of delay: When $\gamma = .9$, efficacy η controls how late a subject can afford to delay working

Results



3. A gap between real (η) and assumed (η_{assumed}) efficacy leads to overestimation of delay. Early completions.
4. Convex effort costs could explain steady completion: If effort costs increase supra-linearly with the amount of work $\rightarrow$ work is more equally spread out across time.

Results



Some students might perceive rewards as arriving immediately upon completing the 7 hour-requirement, rather than at the end. Now, the reward is given as soon as all 14 units are completed.

5. This let's delay due to discounting disappear.
6. If convex effort costs ($k > 1$) are high, i.e., effort costs increase superlinearly with actions effor/actions a, it is more rational to spread out work again.

Result

Other interesting patterns (see paper)

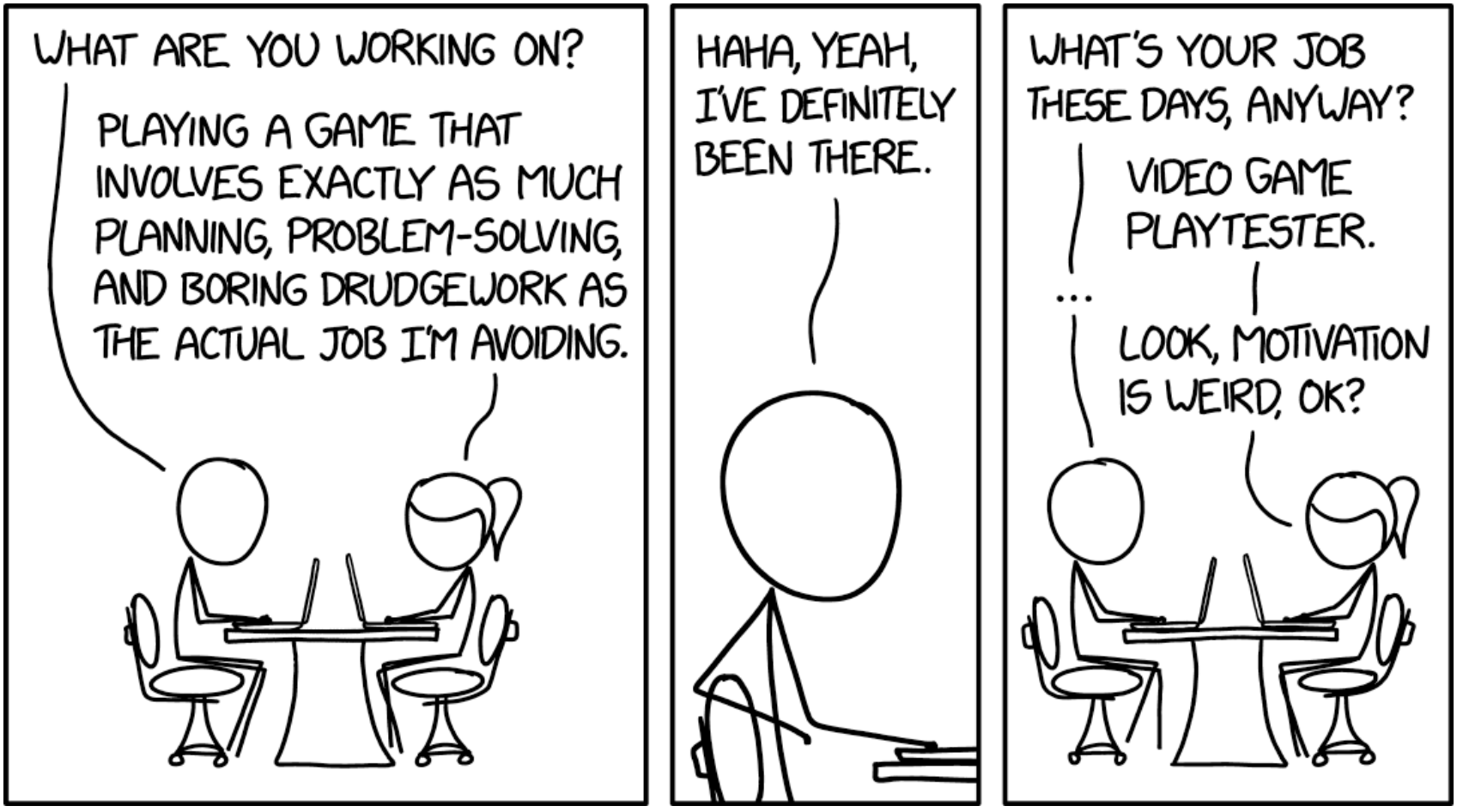
- If we use separate discounting factors for positive rewards, γ_r, and costs (negative rewards), γ_c, we can simulate that some subjects will delay most work until the end of the semester, even if they get immediate rewards upon completion.
- For $\gamma_c < \gamma_r$, future efforts are more discounted than future rewards.
- At any timepoint is seems more rational to do no work now, because we get positive rewards from shirking and the completion reward (only received after 14 units) is far away in terms of effort. Instead, it is better to do the work after now (e.g., at $t + 2$), because anticipated effort costs are heavily discounted.

Perfectionism

See model by Zhang & Ma, 2019.

- Perfectionism is known to be strongly associated with procrastination. Can be modeled by not giving proportional rewards, but make the final reward for completion a power-law function of the actually completed work $R(s_T) \propto s_T^{\alpha}$. Note, $0 \leq s_T \leq 1$ is a proportion here.
- *alpha* $= 1$: proportional reward ($X\%$ work completed gives $X\%$ of the reward).
- *alpha* $\rightarrow$ inf: all-or-none, only get reward if everything is completed. High *alpha* as model of perfectionism.
- If time is limited and work cannot fully be completed, rational strategy of high perfectionism subject is to not work at all.
- This may be aggravated if there is uncertainty if full completion can ever be achieved (even with unlimited time).

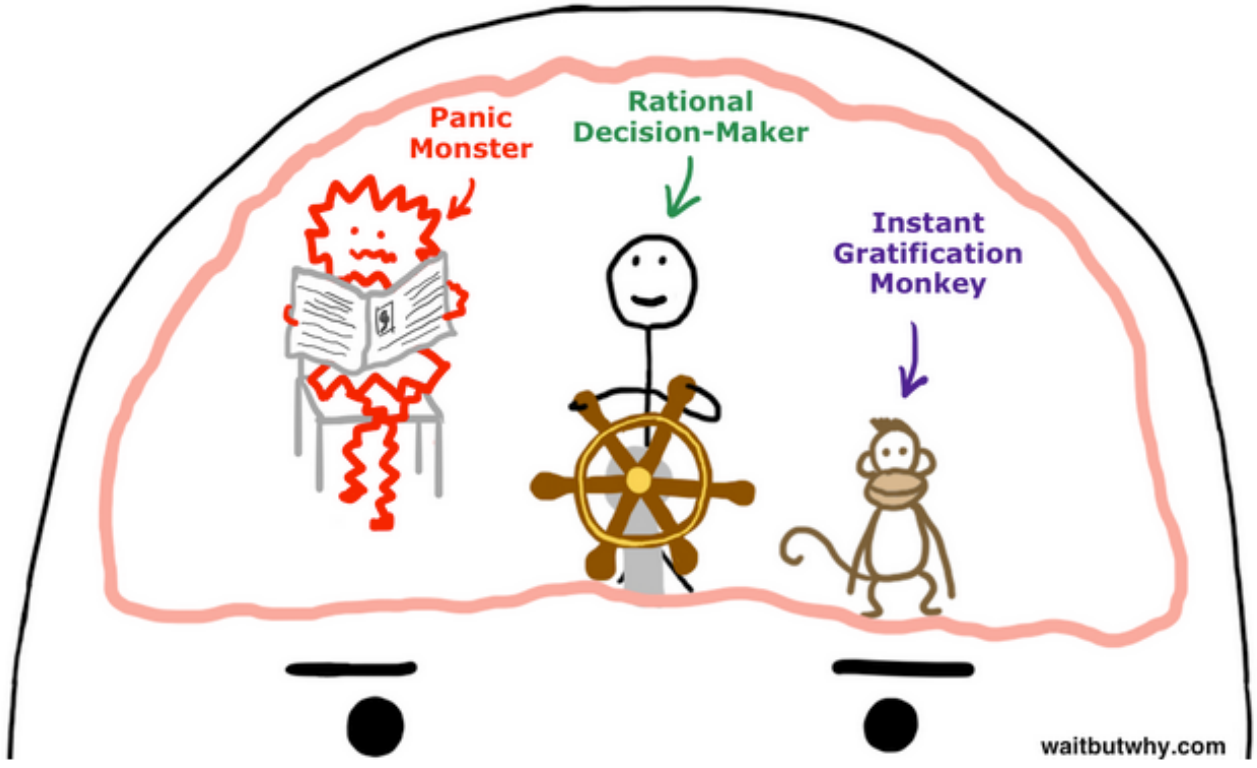
Procrastination



Questions

- Procastination is often associated with impulsivity. Can this be explained via discounting future rewards (and getting distracted)?
- Discpline and perserverance may stem from internal costs of unfinished tasks or devices to counteract preference reversals.
- Course completion is an extrinsic reward. Intrinsic motivation as reward for learning something new?

Other (non-scientific) explanatory frameworks



- Rational decision maker, panic monster, instant gratification monkey (waitbutwhy.com)

References

- Zhang, P. Y., & Ma, W. J. (2024). Temporal discounting predicts procrastination in the real world. *Scientific Reports, 14*(1), 14642. https://doi.org/10.1038/s41598-024-65110-4
- Chebolu, S., & Dayan, P. (2024). Optimal and sub-optimal temporal decisions can explain procrastination in a real-world task. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*(0). https://escholarship.org/uc/item/2mg517js
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction.* MIT Press.
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure.*Psychological Bulletin, 133*(1), 65–94.