Explaining Animal Learning through Reinforcement Learning, Reward Parameterization, and Evolving World Models

Stanford | Department of Statistics

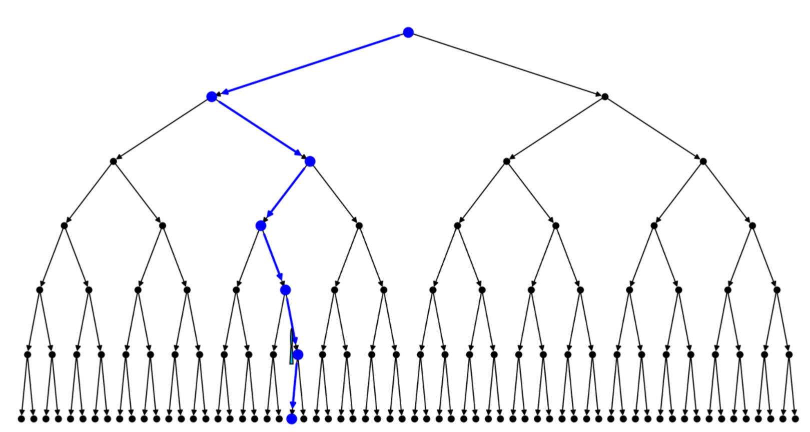
Camila Blank Aditi Jha Scott W. Linderman

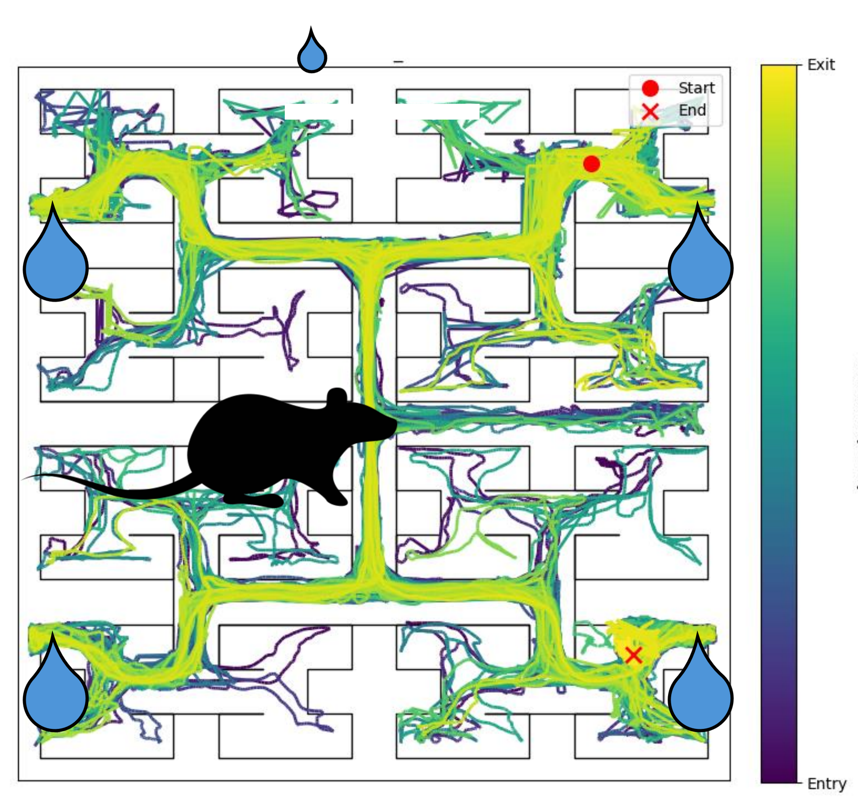
Motivation

- Why? Gain insight on the neural processes underlying a mouse's decision-making process in curiosity-driven navigation
- How? Combine reinforcement learning with multiple frameworks for intrinsic rewards
- Result? Quantify contributions of extrinsic and intrinsic rewards, track an evolving world model, and observe effects on cohorts with stimulated neural circuits
- What's different? We focus on modeling the learning process itself rather than just learned behavior

Mouse Maze Dataset

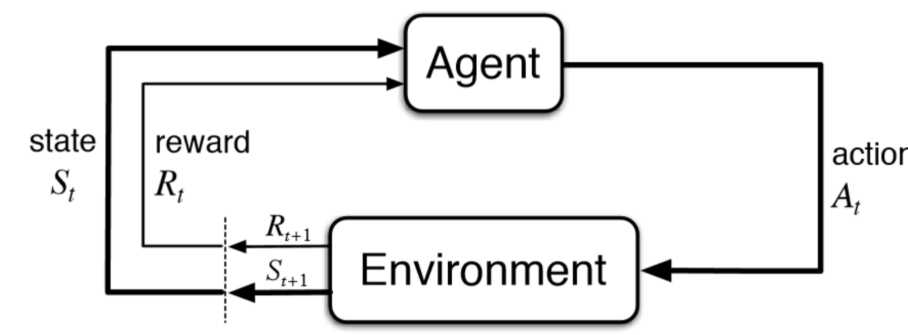
- Water-starved mice
  - Excitatory: C21
  - Control: saline
- Maze structure:
  - 127-node binary tree
  - Four randomly alternating water ports
- Task structure:
  - 10 sessions (1/day)
  - 45 min each





Markov Decision Processes

- Next state is solely a function of the current state (Markov Property)



Algorithms

1. Q-learning (control):

- $Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right)$ (for each goal)

2. Uncertainty reward:

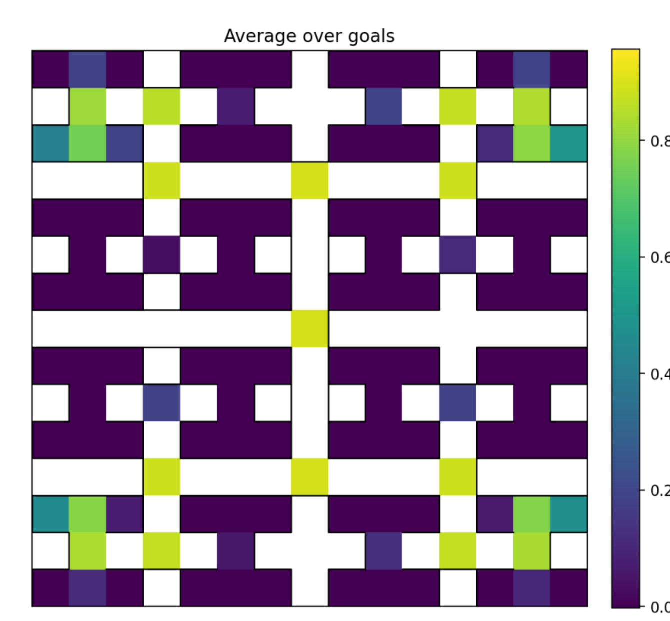
- Bayesian dynamics as world model
- Prior: $P(s'|s,a) \sim Dir(\alpha_1^{s,a}, \alpha_2^{s,a}, \ldots, \alpha_{|S|}^{s,a})$
- Mean given by posterior: $\hat{P}(s'|s,a) = \frac{\alpha_{s'}^{s,a}}{\sum_{i=1}^{|S|} \alpha_i^{s,a}}$
- $r_U^{t,k}(s,a,s') = \eta_U \cdot KL(P_{t,k}(s'|s,a) \parallel P_{t-1,k}(s'|s,a)$

3. Novelty reward:

- $r_N^{t,k}(s,a,s') = \eta_N \cdot \frac{1}{\sqrt{N(s')}}$

4. Epsilon decay

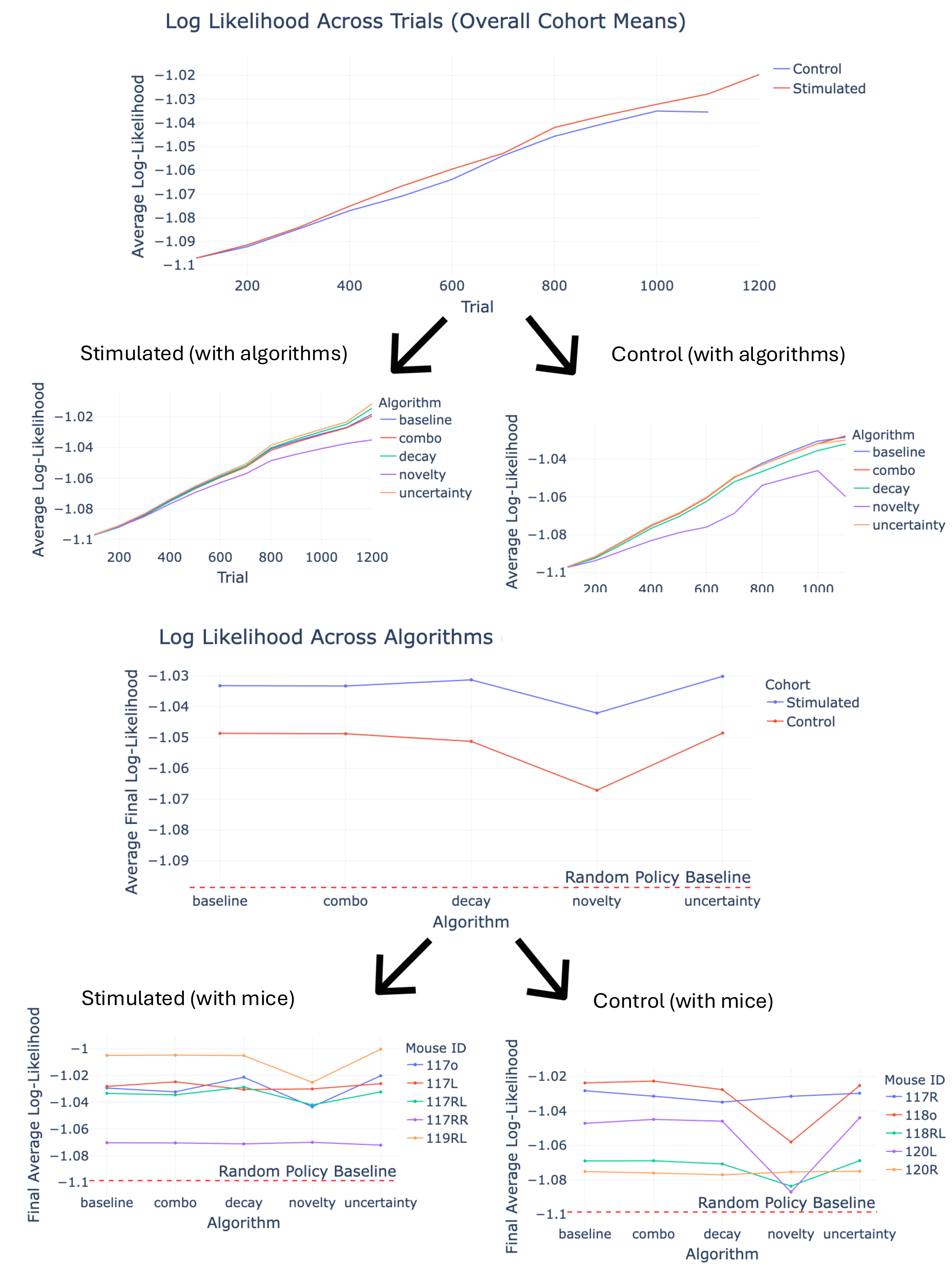
5. Combined (all of the above)



Tuning hyperparameters via log-likelihood optimization

- Minimize: $loss = -\frac{\sum_{j=1}^{N} \sum_{i=1}^{T_j} log\pi_j(a_{ij}|s_{ij})}{\# \, total \, timesteps}$
- π_j = softmax policy for $Q\_list[j]$ for trial j with $\beta = 1.0$

Uncertainty succeeds marginally



Discussion

- Results suggest that reducing uncertainty may be a source of intrinsic reward in mice
- Generally, Q-learning algorithms more effectively predict stimulated mouse behavior
- Next step is inverse reinforcement learning → derive the reward parameterization from the ground truth data