COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH
BIOSTATISTICS

Characterizing Human Reward-based Decision-making Behavior with Reinforcement Learning Models

Xingche Guo

Department of Biostatistics, Columbia University

Joint work with Donglin Zeng (University of Michigan) and Yuanjia Wang (Columbia University)

At the 2024 ICSA Applied Statistics Symposium

06/18/2024

Mental health - Major Depressive Disorder (MDD)

Scientific finding:

An individual's learning ability and decision-making may be altered by MDD *(Pizzagalli, et al. 2005)*.

Try to Answer:

How does MDD affect the decision-making and reward learning?

- Learn slow?
- Not sensitive to reward?
- Easy to distract?
- etc…

Task:

Behavior cloning/ imitation learning
(Ross and Bagnell, 2010)

EMBARC Study:

A clinical trial for exploring how biomarkers affect the treatment outcome for MDD *(Trivedi et al., 2016)*.

Data Types

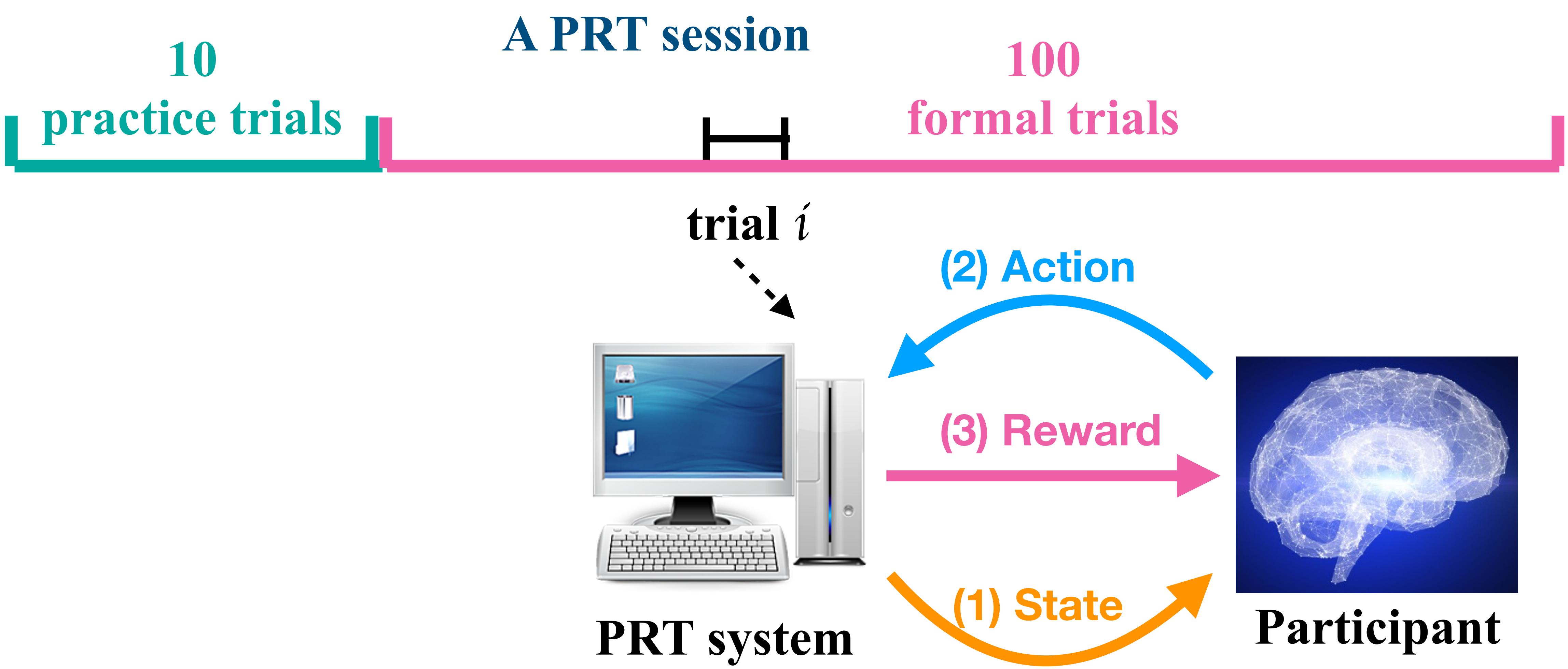
- **Demographical and clinical data**
- **Neuroimaging data:**
  - **Task EEG/fMRI**
  - **Resting-state EEG/fMRI**
  - **etc…**
- **Human behavioral data:**
  - **Probabilistic reward task *(Pizzagalli et al., 2005)***
  - **Emotion conflict task *(Etkin et al., 2006)***
  - **etc…**

Experimental Design

- **MDD group vs Health Control group (Today's focus)**
- **In MDD group: Treatment vs Placebo**

Probabilistic reward task (PRT):

A computer-based behavioral experiment that measures the subject's ability to modify behavior in response to rewards.
(Pizzagalli et al., 2005)



Participant's goal: **learn from the PRT system (to maximize rewards).**

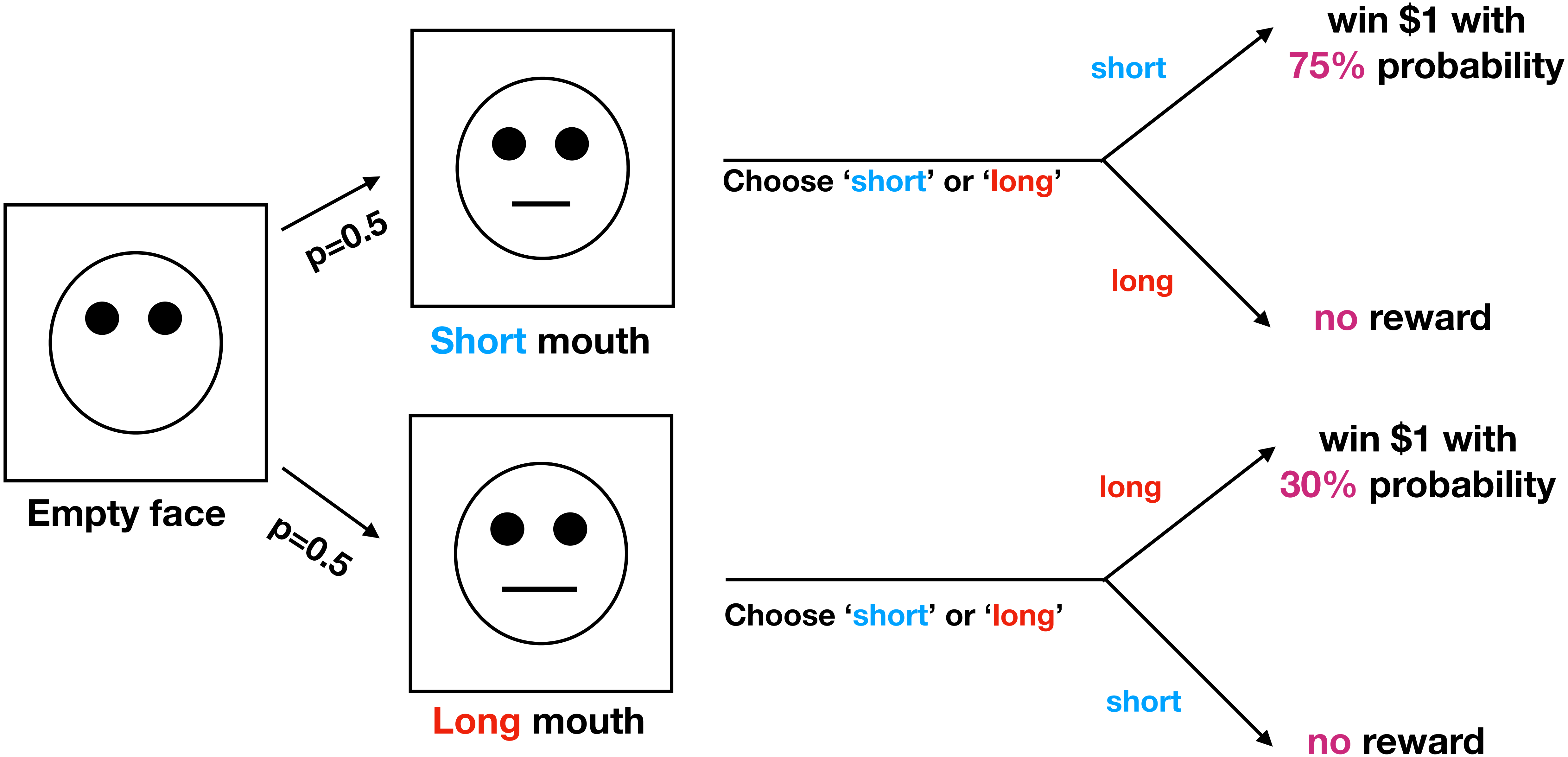
Our goal: **understand how the participant learns the PRT system (not interested in PRT system).**

Probabilistic reward task (PRT):

Demo (single trial)

You are told the task is to identify the correct mouth.

You don't know the reward generating mechanism.



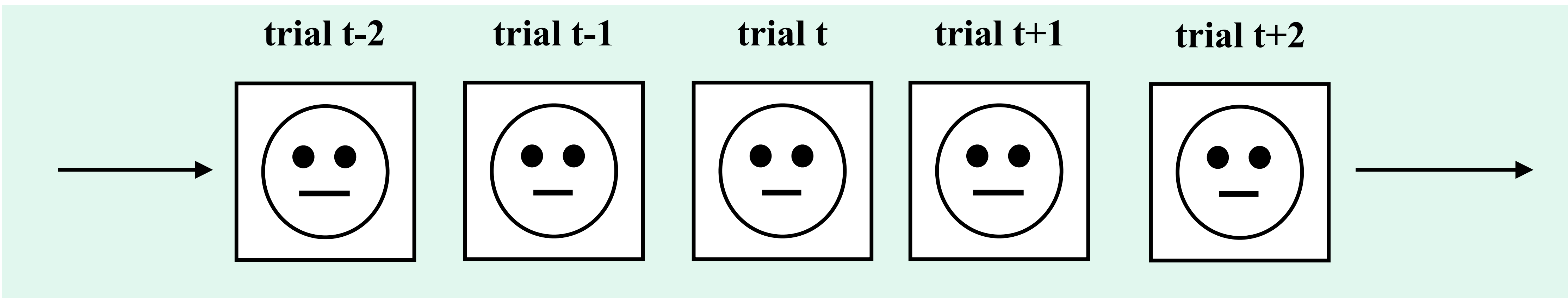
Small difference in mouth size

Rewards are imbalanced

Probabilistic reward task (PRT):

Demo (multiple trials)

A PRT session



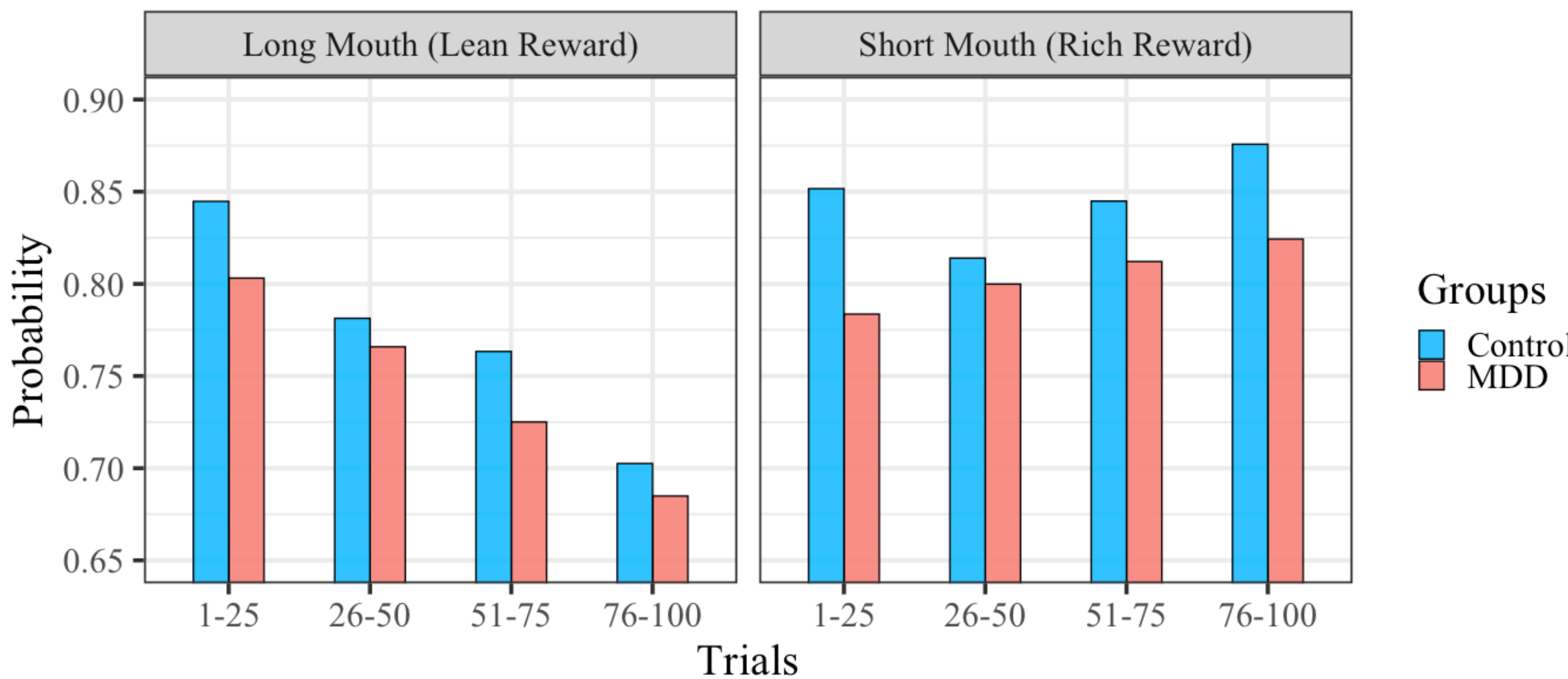
States: (S)	long S_{t-2}	short S_{t-1}	short S_t	long S_{t+1}	short S_{t+2}	......
Actions: (A)	long A_{t-2}	short A_{t-1}	short A_t	short A_{t+1}	short A_{t+2}	
Rewards: (R)	0 R_{t-2}	1 R_{t-1}	1 R_t	0 R_{t+1}	0 R_{t+2}	
Prob. of winning:	30%	75%	75%	0%	75%	

What is observed from PRT?

Conditional correct answer rate for MDD and Control groups (evenly divide 100 trials to 4 blocks).

$P\left(\textbf{Action} = \textbf{'long'} \mid \textbf{State} = \textbf{'long'}\right)$ **Lean Reward**

$P\left(\textbf{Action} = \textbf{'short'} \mid \textbf{State} = \textbf{'short'}\right)$ **Rich Reward**



- **Subjects tend to prioritize states with higher rewards as trial progresses.**
- **Subjects in MDD perform worse in PRT than subjects in Control.**

Classical RL models *(Huys et al. 2013)*

Problem setups for PRT

Problem size: subjects (i = 1, …, n) from a group, trials (t = 1, …, T) for each session.

State space (S): {0, 1}: 0 = 'long mouth' (lean); 1 = 'short mouth' (rich).

Action space (A): {0, 1}: 0 = 'long mouth'; 1 = 'short mouth'.

Reward space (R): {0, 1}: 0 = 'no reward'; 1 = 'win reward'.

Data for one group: $\{\ldots, S_{it}, A_{it}, R_{it}, \ldots\}$, $i = 1,\ldots,n;\ \ t = 1,\ldots,T$.

Classical RL models *(Huys et al. 2013)*

Q-learning model

Expected reward (own estimate):

$$Q_{it}(a,s) = \mathbb{E}^{(\text{est})}\left(R_{it} \mid A_{it}=a, S_{it}=s\right)$$

Minimize reward prediction error: $R_{it} - Q_{it}(a,s)$

Update expected reward (gradient descent):

$$Q_{i,t+1}(a,s) = Q_{it}(a,s) + \beta_i\left(R_{it} - Q_{it}(a,s)\right)$$

$$\left(a = A_{it},\ s = S_{it}\right)$$

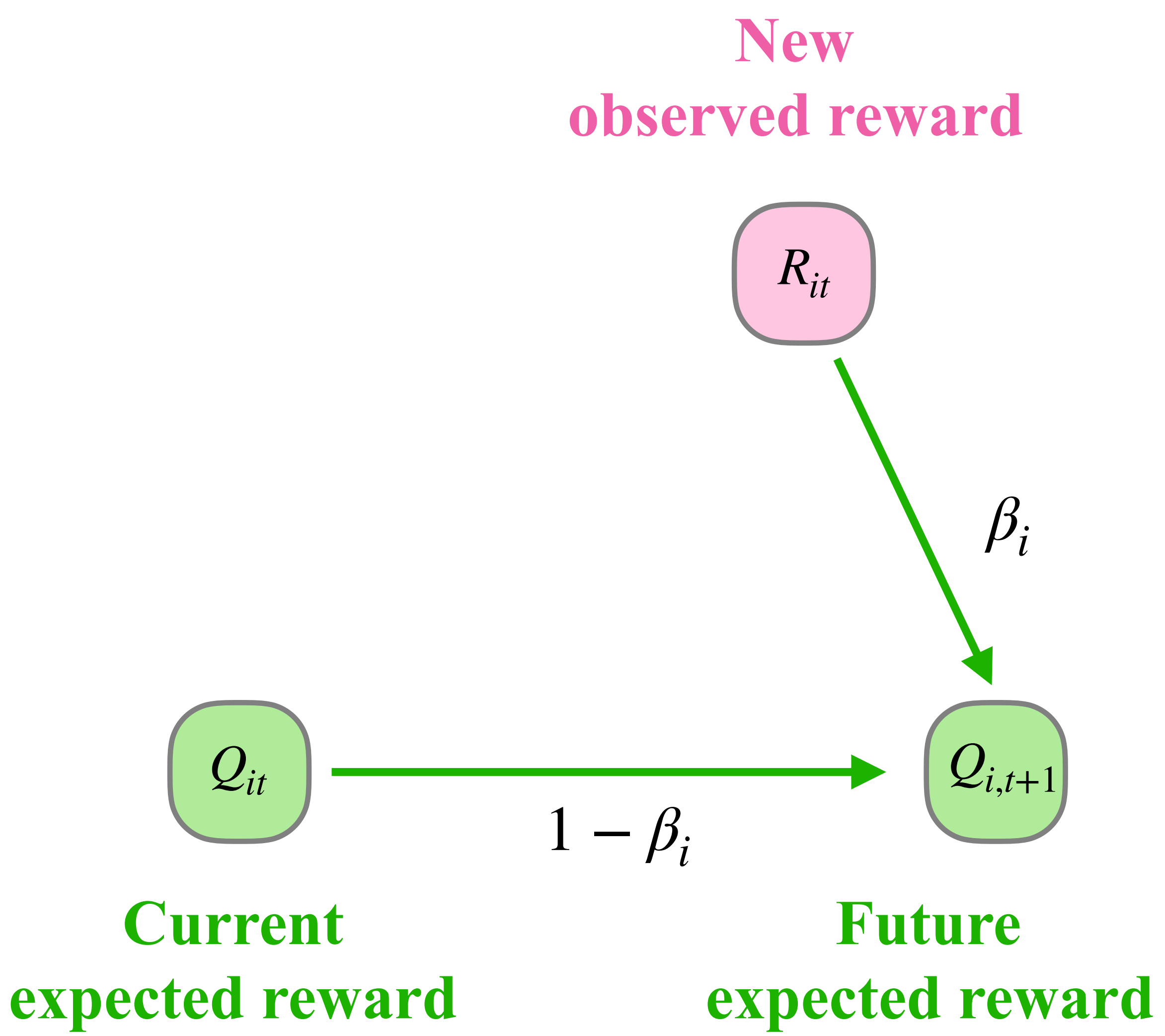
Learning rate: $\beta_i \in (0,1)$

Another view (weighted sum):

$$Q_{i,t+1}(a,s) = \left(1-\beta_i\right) Q_{it}(a,s) + \beta_i R_{it}$$

$\beta_i \to 0$, no update,

$\beta_i \to 1$, no memory



Classical RL models *(Huys et al. 2013)*

Decision making model

Contrast of expected rewards for action 1 and 0 at

the given state: $Z_{it} = Q_{it}(1, S_{it}) - Q_{it}(0, S_{it})$

weighing between two actions

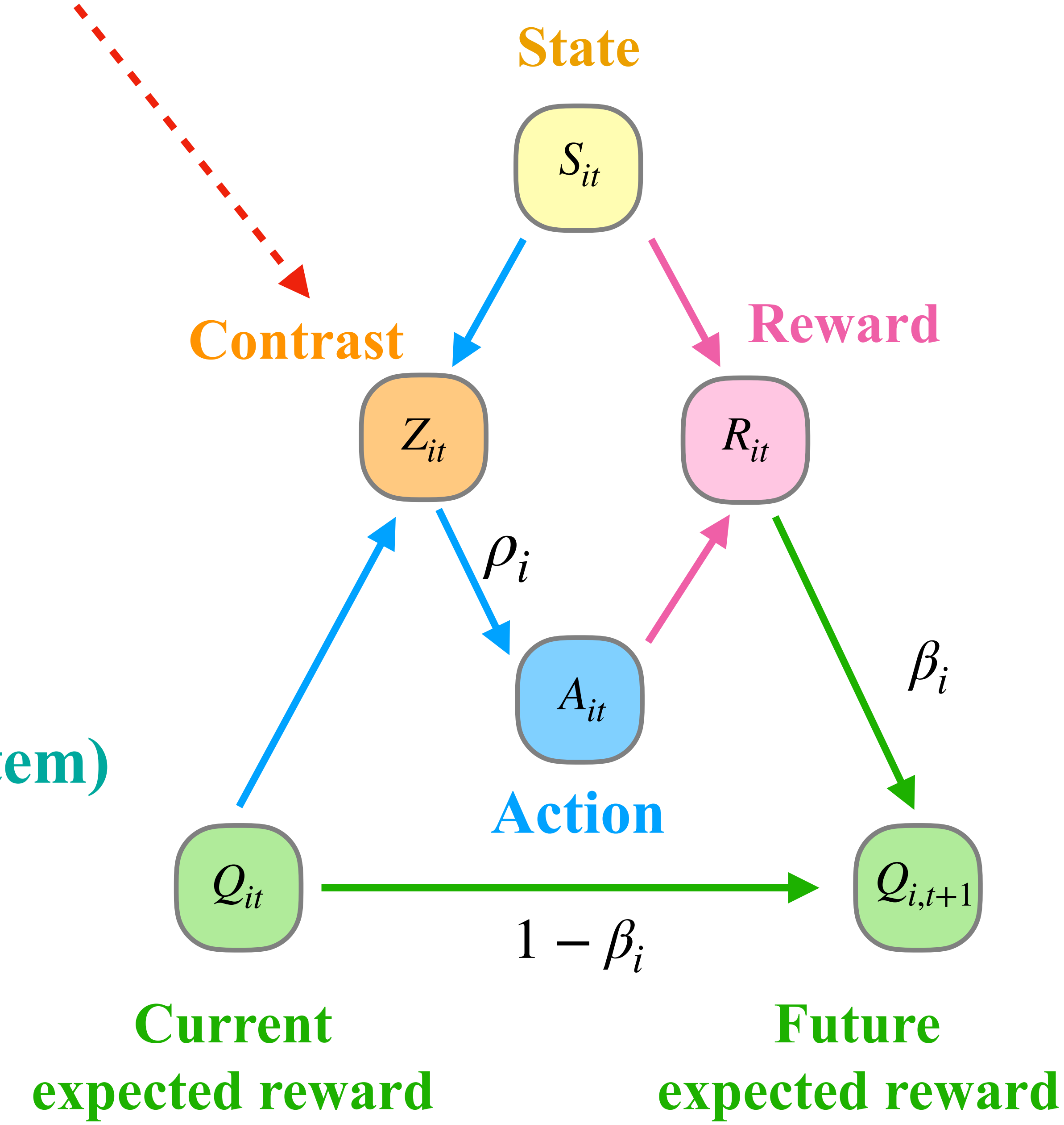
Conditional probability of taking action 1:

$$\text{logit } P\left(A_{it} = 1 \mid Z_{it}\right) = \rho_i Z_{it}$$

Reward sensitivity: $\rho_i > 0$:

if $\rho_i \to \infty$, $P(A_{it} = 1 \mid Z_{it} = 1) \to 1$,

if $\rho_i \to 0$, $P(A_{it} = 1 \mid Z_{it} = 1) \to 0.5$.



Reward generating model (from PRT system)

$$P(R_{it} = 1 \mid S_{it} = A_{it} = 1) = 0.75$$

$$P(R_{it} = 1 \mid S_{it} = A_{it} = 0) = 0.3$$

$$P(R_{it} = 1 \mid S_{it} \neq A_{it}) = 0$$

Semiparametric RL model

Guo, X., Zeng, D., Wang, Y. (2024). A Semiparametric Inverse Reinforcement Learning Approach to Characterize Decision Making for Mental Disorders. *Journal of the American Statistical Association.*

Semiparametric RL model

Decision making model (Our contribution)

Contrast of expected rewards for action 1 and 0 at

the given state: $Z_{it} = Q_{it}(1, S_{it}) - Q_{it}(0, S_{it})$

Conditional probability of taking action 1:

$$\text{logit } P\left(A_{it} = 1 \mid Z_{it}\right) = f(\,\rho_i Z_{it}\,)$$

Reward sensitivity function: $f(\,\cdot\,)$

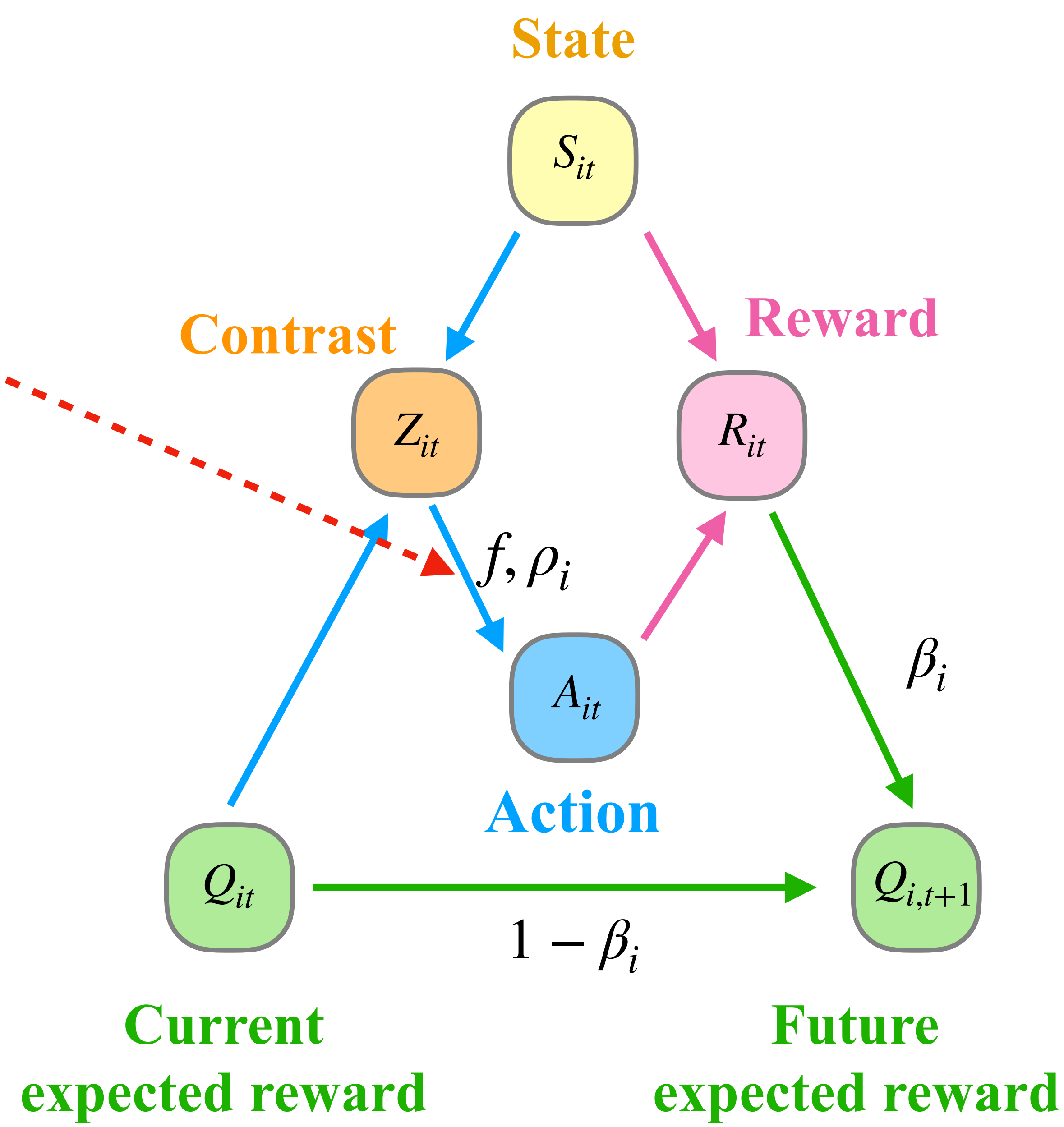
We further assume:

(i). $f(\,\cdot\,)$ **non-decreasing; (ii)** $f(0) = 0$

Properties:

(i). $P(A_{it} = 1 \,|\, Z_1) \geq P(A_{it} = 1 \,|\, Z_2)$, if $Z_1 \geq Z_2$

(ii). $P(A_{it} = 1 \,|\, Z_{it} = 0) = 0.5$



Semiparametric RL model

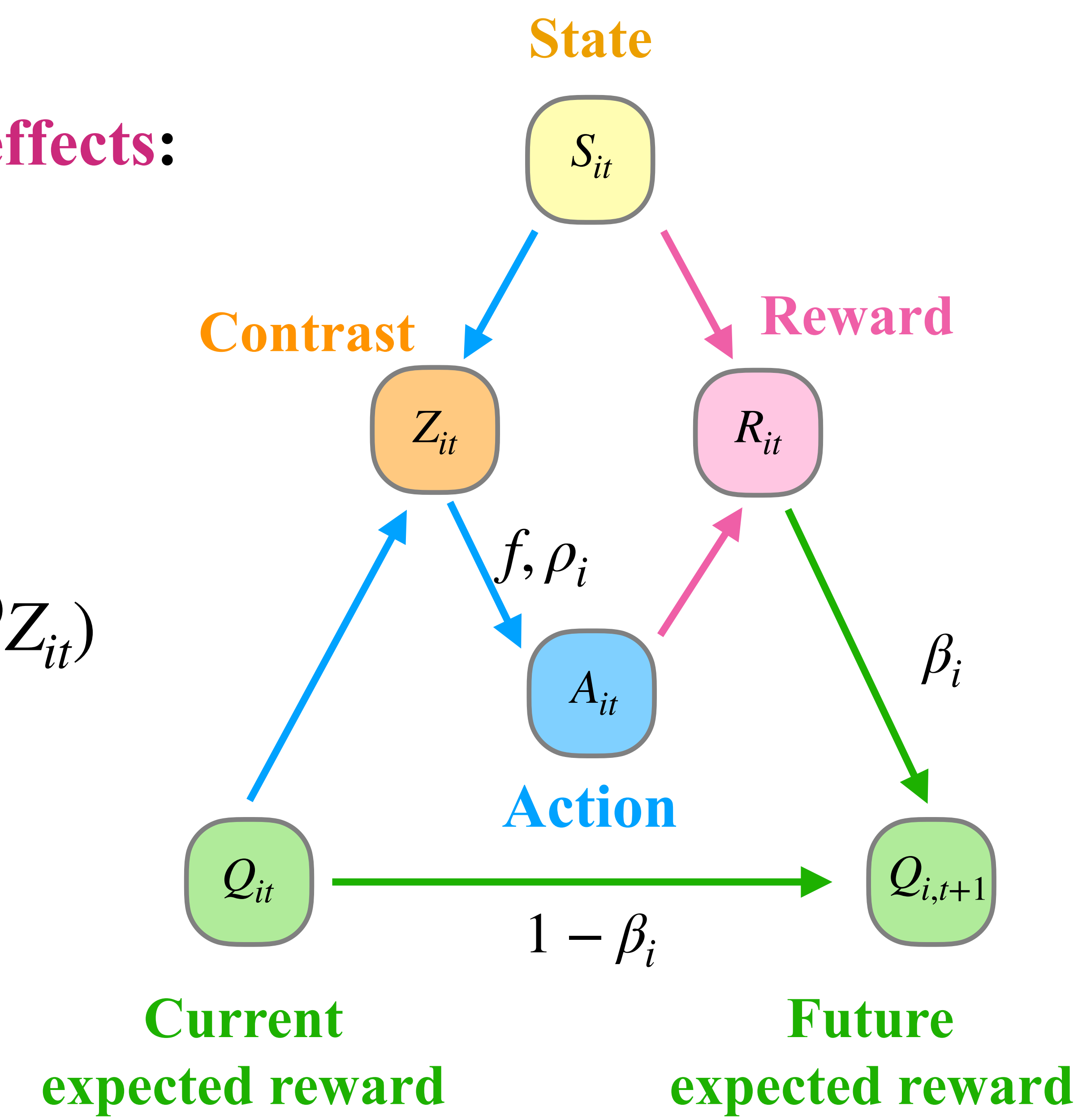
Jointly modeling all subjects (Our contribution)

Map learning rate and reward sensitivity to real line:

$$\nu_i = \text{logit}(\beta_i); \;\; \gamma_i = \log(\rho_i)$$

Subject-specific heterogeneity as random effects:

$$(\nu_i, \gamma_i) \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = (\mu_\nu, \mu_\gamma)^\top$$



Scale identifiability issue:

$$\text{logit } P\left(A_{it} = 1 \mid Z_{it}\right) = f(\rho_i Z_{it}) = f^{(c)}(\rho_i^{(c)} Z_{it})$$

where $f^{(c)}(x) = f(cx), \quad \rho_i^{(c)} = \rho_i / c$

Solution: fix the mean effect: $\mu_\gamma = 1$

Model implementation

Maximum likelihood estimation

Joint likelihood function:

$$\prod_{i=1}^{n}\prod_{t=1}^{T} \underbrace{P\left(S_{it}\right)}_{\text{by PRT system}}^{\text{state generating}} \underbrace{P\left(A_{it} \mid Z_{it}; \nu_i, \gamma_i, f\right)}_{\text{by subject}}^{\text{decision making}} \underbrace{P\left(R_{it} \mid S_{it}, A_{it}\right)}_{\text{by PRT system}}^{\text{reward generating}}$$

Parameter of interest:

Group-level: **learning rate** μ_ν **, reward sensitivity function** $f(\cdot)$.

Subject-level: **learning rate** ν_i (or β_i) **, reward sensitivity** γ_i (or ρ_i) .

Only need to focus on:

$$L\left(\{\nu_i, \gamma_i\}_i, f\ ;\ \{S_{it}, A_{it}, R_{it}\}_{i,t}\right) \propto \prod_{i=1}^{n}\prod_{t=1}^{T} P\left(A_{it} \mid Z_{it}\ ;\ \nu_i, \gamma_i, f\right)$$

Integrate the random effects:

$$L\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, f\right) \propto \underbrace{\prod_{i=1}^{n}}_{\text{Parallel computing}} \left[\underbrace{\iint}_{\text{Gauss-Hermite quadrature}} \overbrace{\phi(\nu_i, \gamma_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}^{\text{multivariate normal PDF}} \prod_{t=1}^{T} P\left(A_{it} \mid Z_{it}\ ;\ \nu_i, \gamma_i, f\right) d\nu_i d\gamma_i \right]$$

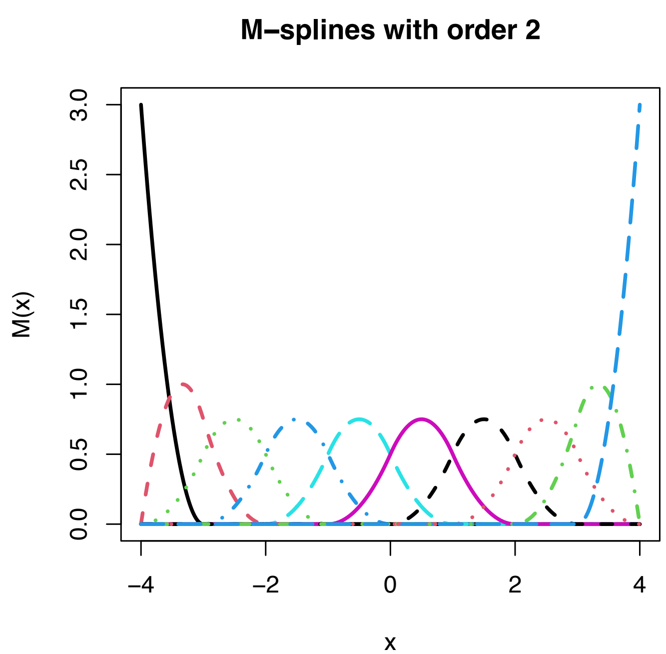
Model implementation

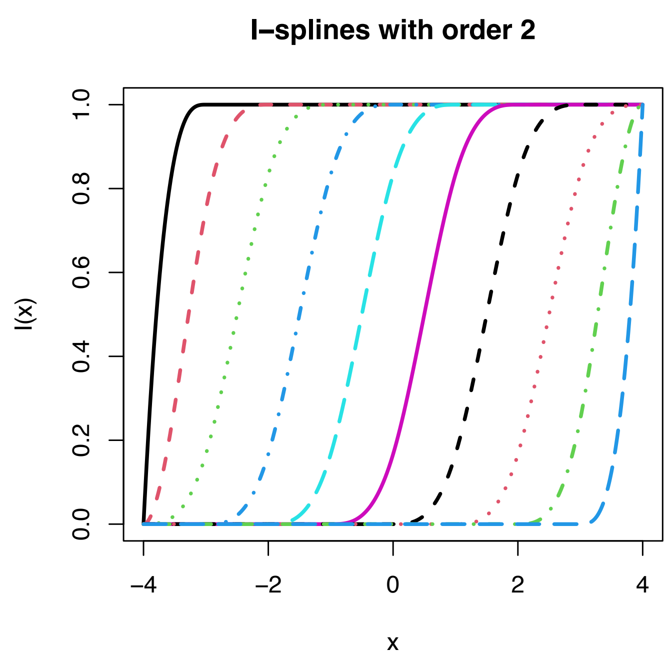
Nonparametric function modeling

Recall

(i). $f(\cdot)$ **non-decreasing; (ii)** $f(0) = 0$

We use I-spline to model nondecreasing nonlinear functions (*Ramsay 1988*).





- **M-spline: nonnegative spline functions (properties similar to B-spline).**
- **I-spline: integral of M-spline, hence nondecreasing.**

We approximate:

the *k*-th I-spline function

$$\tilde{f}(x) = \sum_{k=1}^{K} \left\{ I_k(x) - I_k(0) \right\} b_k, \quad b_k \geq 0.$$

Model implementation

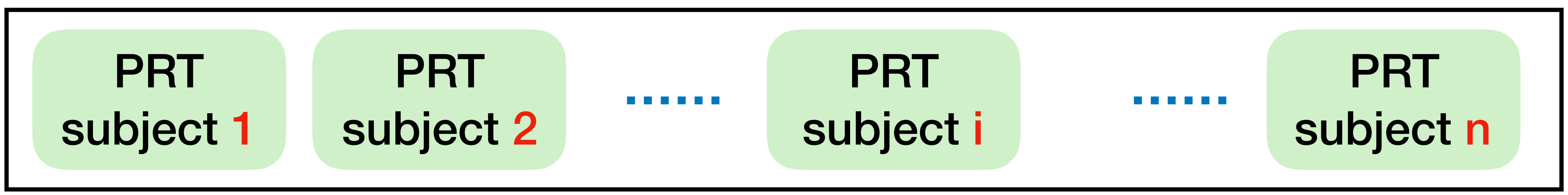
Model inference

- **Parametric bootstrap** is **not** applicable because **state/ reward generating function** is **unknown**.

- **Nonparametric bootstrap** is applied.

$\{S_{it}, A_{it}, R_{it}\}_{t=1}^{T}$

RL data



treat each subject's data as a unit when resampling.

- **Bootstrap confidence intervals/bands are constructed using normal approximation.**

- **Extensive simulation studies show strong performance in estimation and inference.**

Simulation study

Compare semiparametric and linear RL (200 replicates, 50 bootstrap samples)

T	n		Semiparametric				Linear	
			RB	SD	SE	CP	RB	SD
100	100	μ_ν	0.014	0.316	0.346	97	0.103	0.301
		$\sigma^2_{\nu,\nu}$	−0.119	0.208	0.301	98	−0.577	0.235
		$\sigma^2_{\gamma,\gamma}$	−0.154	0.132	0.132	98	0.533	0.037
		$\sigma^2_{\nu,\gamma}$	0.163	0.119	0.135	98	−0.251	0.070
		α	−0.053	0.454	0.445	95	−0.055	0.233
		ω	−0.011	0.052	0.057	96	−0.062	0.061

T	n		Semiparametric				Linear	
			RB	SD	SE	CP	RB	SD
100	100	$f(-1.0)$	−0.021	0.188	0.221	98	−0.191	0.153
		$f(-0.5)$	−0.028	0.171	0.177	97	−0.341	0.077
		$f(0.5)$	−0.001	0.188	0.185	97	0.341	0.077
		$f(1.0)$	0.006	0.167	0.181	96	0.191	0.153
		$f(1.5)$	0.010	0.179	0.194	98	0.047	0.230
		$f(2.0)$	−0.025	0.277	0.284	97	−0.090	0.307

Asymptotic theory

$\boldsymbol{\theta}$ **is the collection of all parameter of interests except** f

Consistency (T fixed, $n \to \infty$)

Theorem 1. *Under Conditions 1–4,* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \to 0$, $\|\widehat{f} - f_0\|_{\mathcal{L}_2} \to 0$ *in probability. Furthermore,* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2 + \|\widehat{f} - f_0\|_{\mathcal{L}_2}^2 = o_p(n^{-1/2})$.

f **converges in** $\mathscr{L}_2$

Asymptotic normality (T fixed, $n \to \infty$)

Theorem 2. *Under Conditions 1–4,* $n^{1/2}\{\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0,\ \widehat{f} - f_0\}$ *converges in distribution to a zero-mean and tight Gaussian process in the metric space* $l^{\infty}(\mathcal{O}_\theta \times \mathcal{F}_f)$ *as* $n \to \infty$.

The linear functional of f **coverages in distribution.**

Conditions 1-4 in the Appendix

Application to EMBARC Study

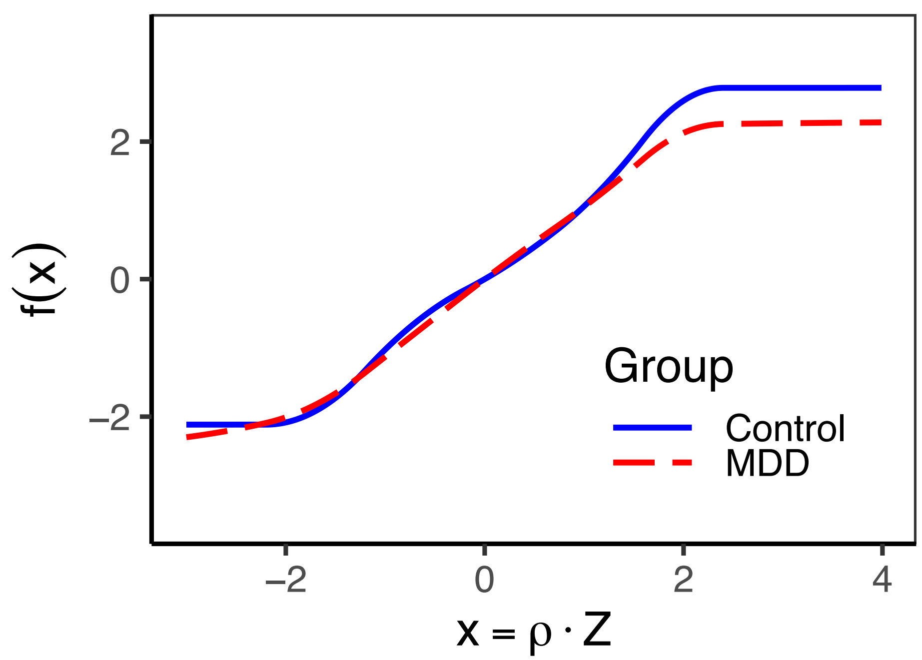
Results: MDD vs Control

Learning Rate:

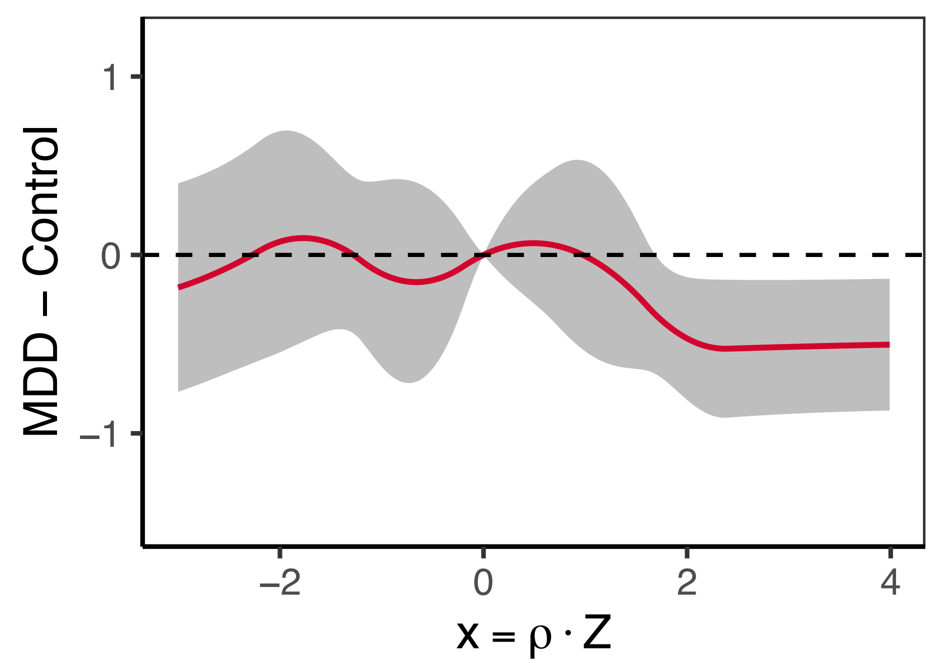
The difference of learning rate between MDD group and Control group is not significant.

Reward sensitivity function $f(\cdot)$:

$f(\cdot)$ for MDD and Control



$f(\cdot)$ difference between MDD and Control



- **Nonlinear (a floor and ceiling effect).**
- **The Control group has a larger reward sensitivity function compared to the MDD group when the contrast is a large positive value.**

What does the floor and ceiling effect of f(.) tell us?

Consider 3 decision-making models:

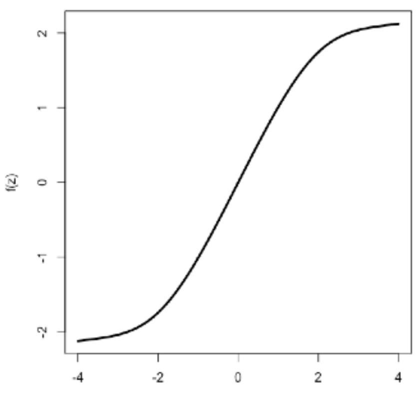
Classical RL:

$$P(A = 1 \mid Z) = \frac{1}{1 + \exp(-Z)}$$

Semiparametric RL:

$$P(A = 1 \mid Z) = \frac{1}{1 + \exp(-f(Z))}$$

$f(z)$:

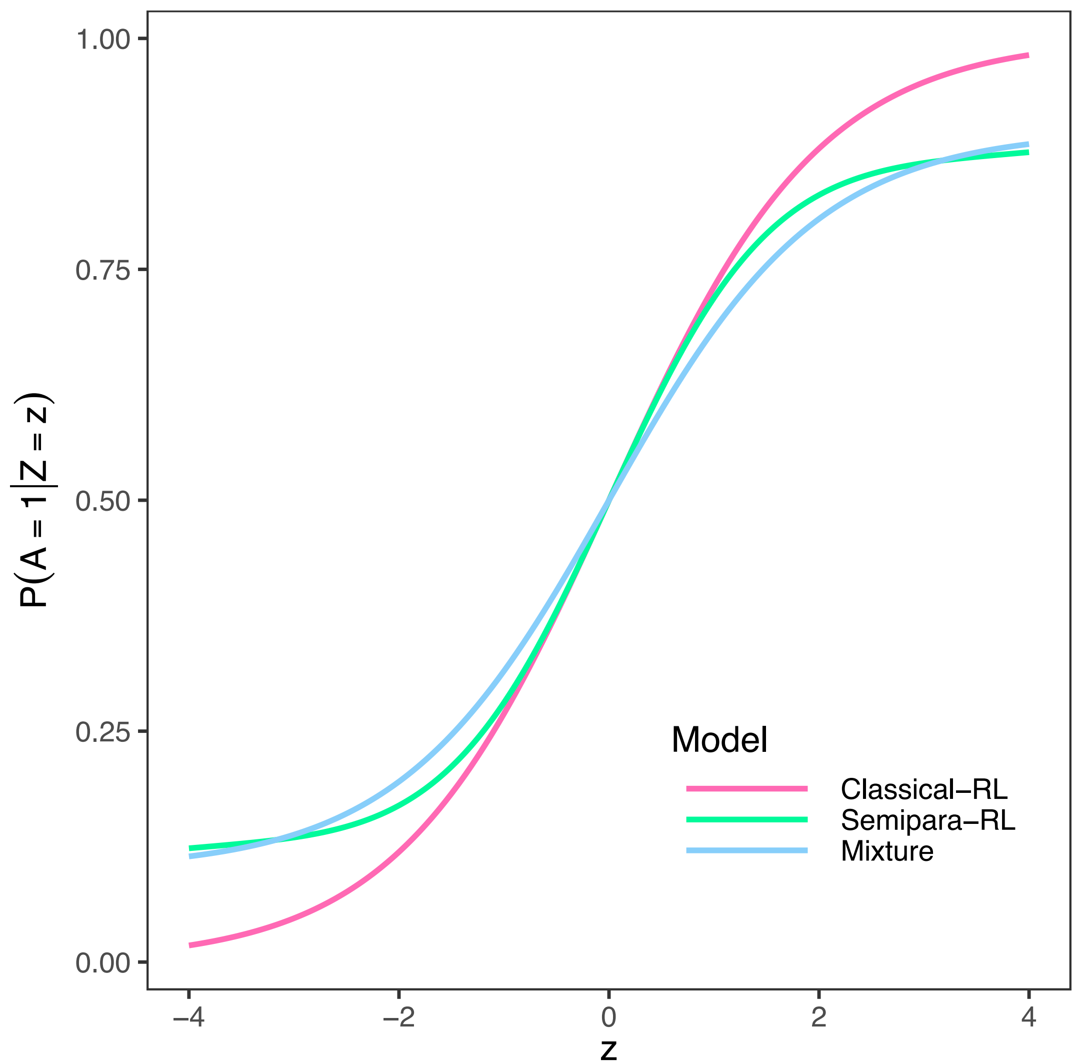


Mixture (Classical RL and random):

$$P(A = 1 \mid Z, U = 1) = \frac{1}{1 + \exp(-Z)}$$

$$P(A = 1 \mid Z, U = 0) = 0.5, \quad P(U = 1) = 0.8$$

Visualize $P(A = 1 \mid Z)$



Question: Is decision-making more complex than a single RL model?

(*Iigaya et al., 2018*; *Ashwood et al., 2022*) provide evidence that subjects employ multiple learning strategies for decision-making.

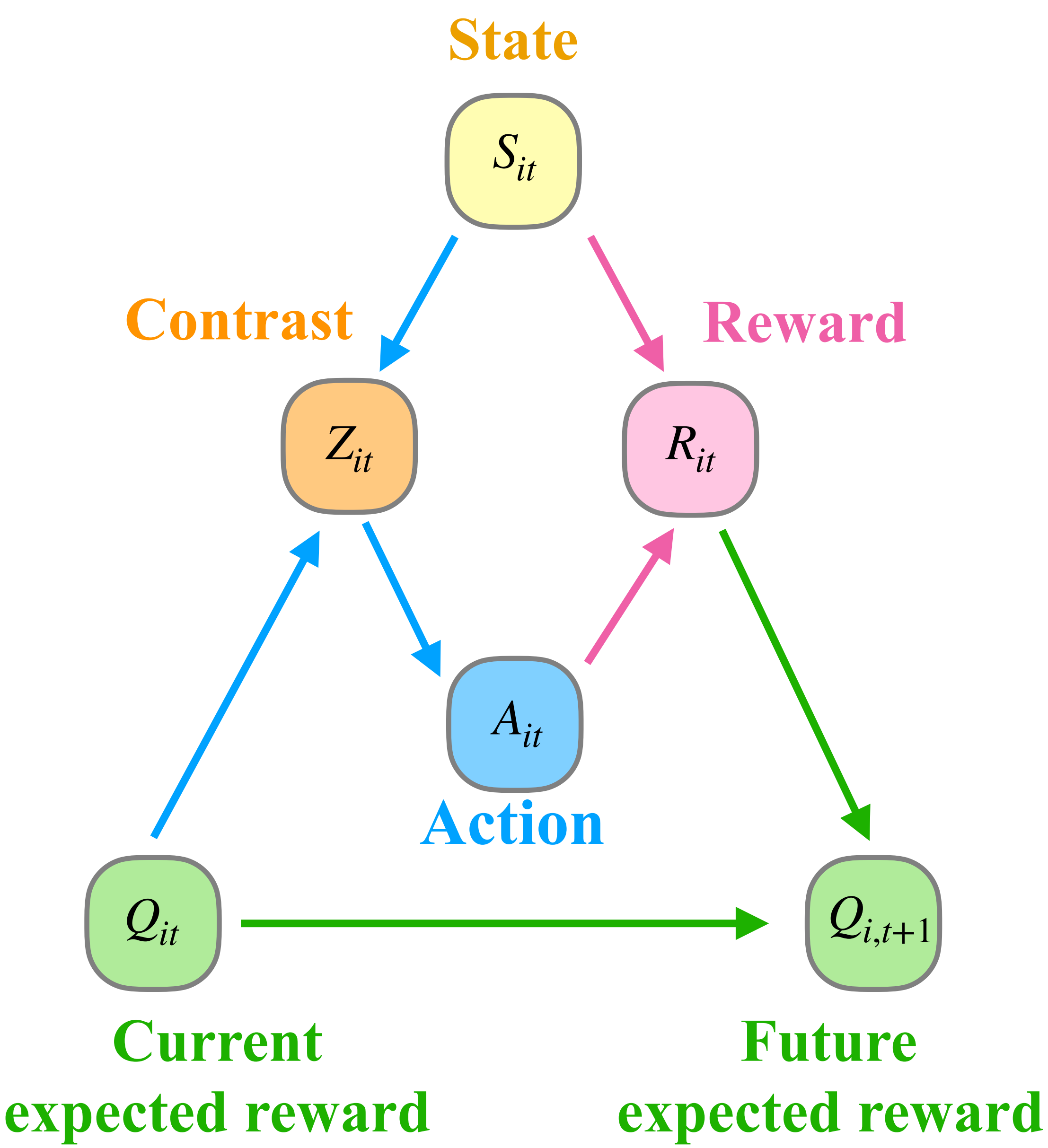
RL-HMM framework

Guo, X., Zeng, D., Wang, Y. (2024). HMM for Discovering Decision-Making Dynamics Using Reinforcement Learning Experiments. *Accepted by Biostatistics, arXiv:2401.13929*

RL-HMM framework

engaged vs lapse

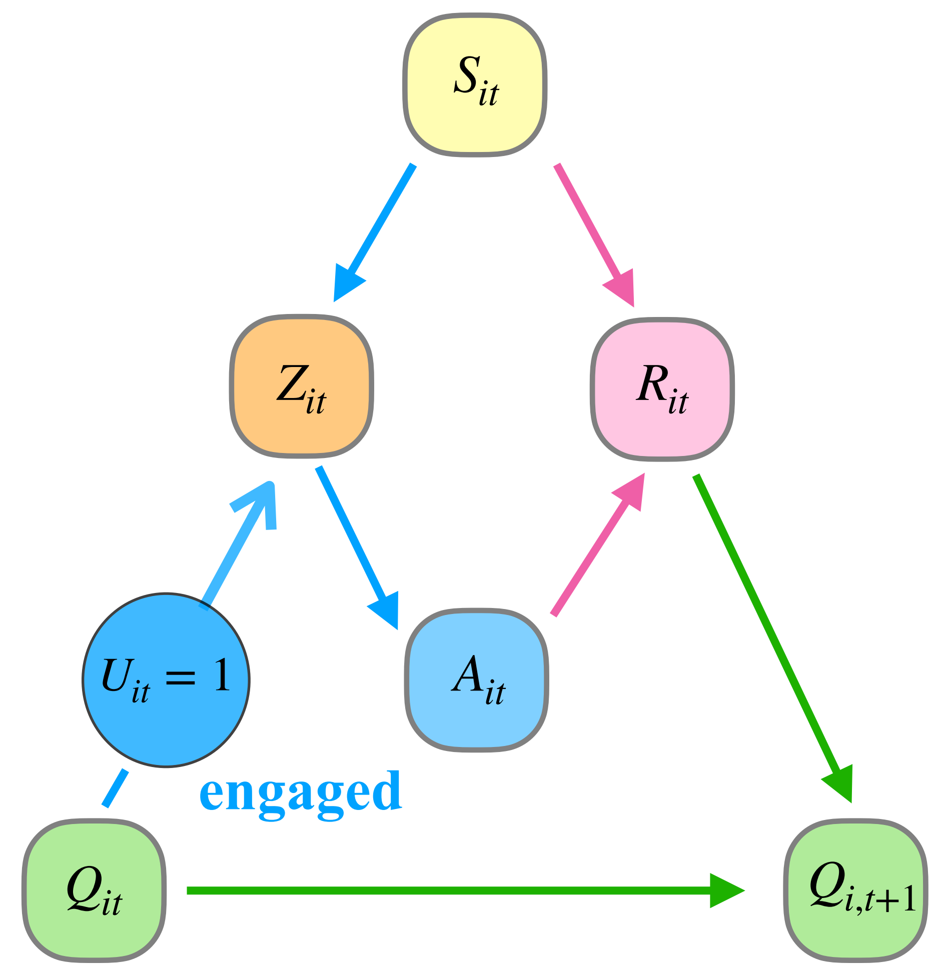
RL framework



RL-HMM framework

Learning strategy: engaged

$U_{it} = 1$

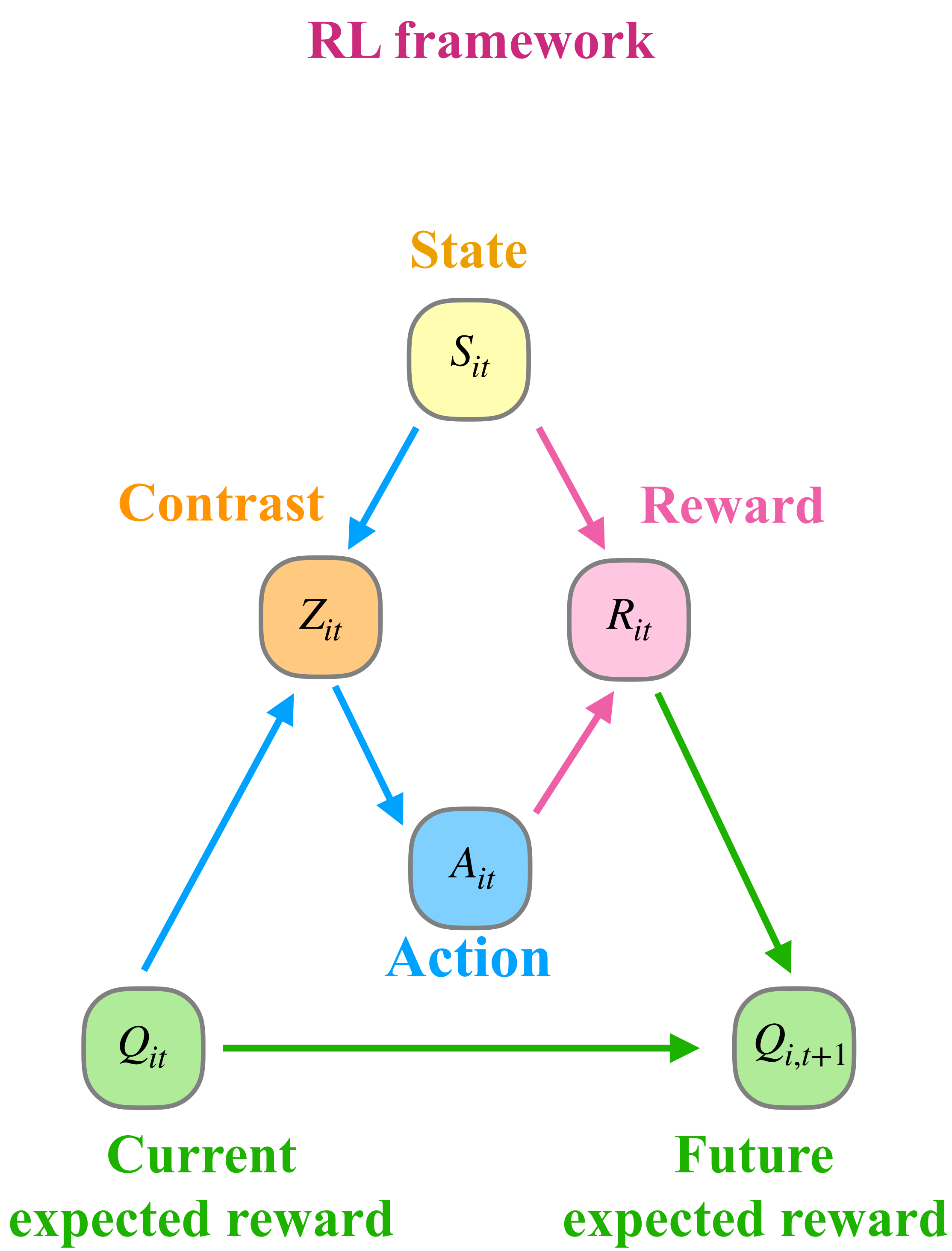


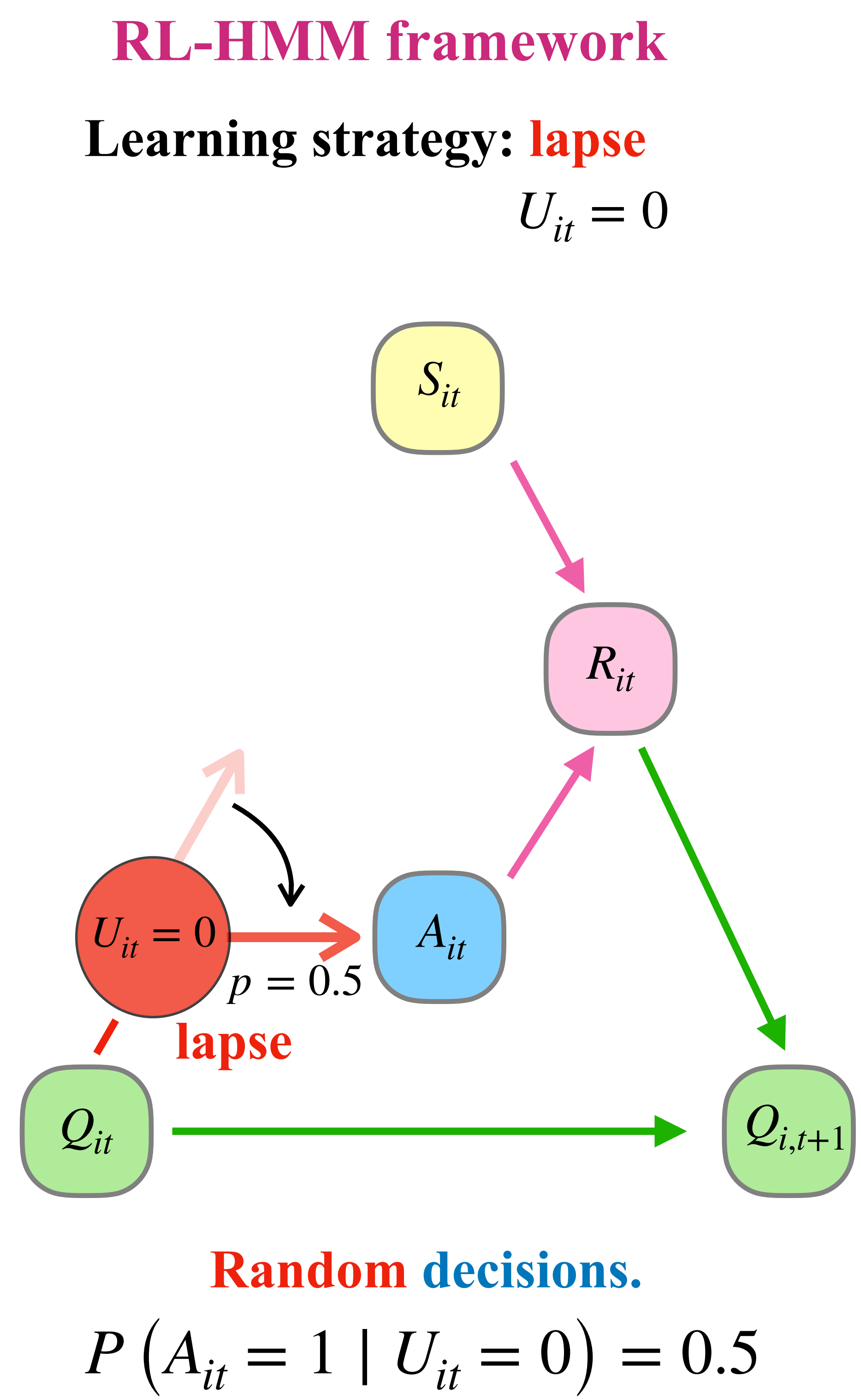
The same decision-making model as the RL framework.

$$\operatorname{logit} P\left(A_{it} = 1 \mid U_{it} = 1, Z_{it}\right) = \rho Z_{it}$$

RL-HMM framework

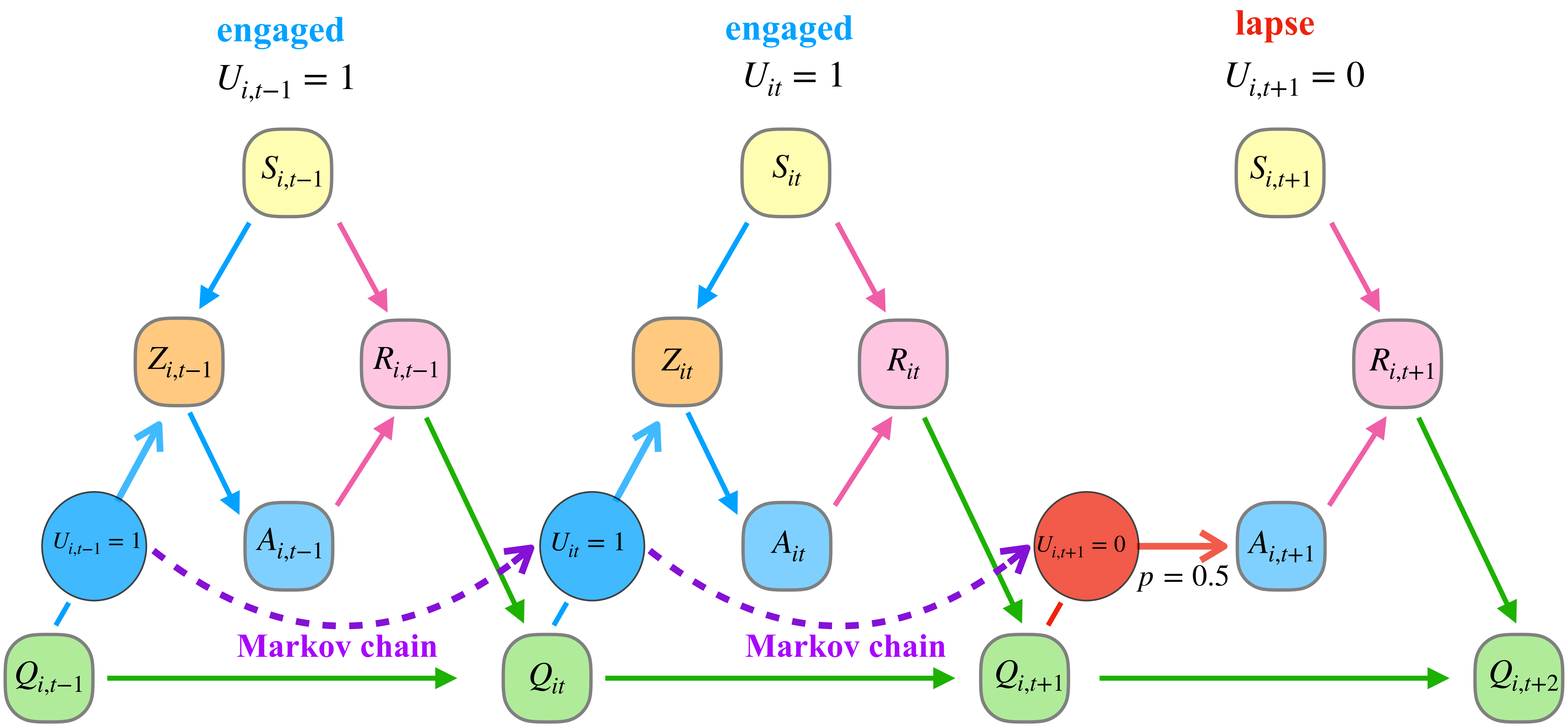
engaged vs lapse





RL-HMM framework

State switching between engaged vs lapse



State switching: $\text{logit } P\left(U_{i,t+1} = 1 \mid U_{it} = j\right) = \zeta_j(t)$

Nonparametric function to allow non-stationarity

Model implementation

EM algorithm

Joint log-likelihood:

Initial state ↓ state transition ↓ decision making ↓

$$\mathscr{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\log P(U_{i1}) + \sum_{t=1}^{T-1} \log P\left(U_{i,t+1} \mid U_{it}\right) + \sum_{t=1}^{T} \log P\left(A_{it} \mid U_{it}, S_{it}, Z_{it}\right) \right).$$

- **E-step:** **take the expected value of $\mathscr{L}_n(\boldsymbol{\theta})$, denoted by $\mathscr{J}_n\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}}\right)$, in terms of $P(U_{it} \mid A_{i1}, \ldots, A_{iT})$ and $P(U_{it}, U_{it-1} \mid A_{i1}, \ldots, A_{iT})$, where the above two probabilities can be computed by the forward-backward algorithm *(Baum et al., 1970)*.**

- **M-step:** **minimize the objective function: $-\mathscr{J}_n\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}}\right) + \text{Pen}\left(\zeta_0, \zeta_1\right)$, where the penalty of the Markov transition functions can be fused-lasso or trend filtering *(Tibshirani, 2014)*.**

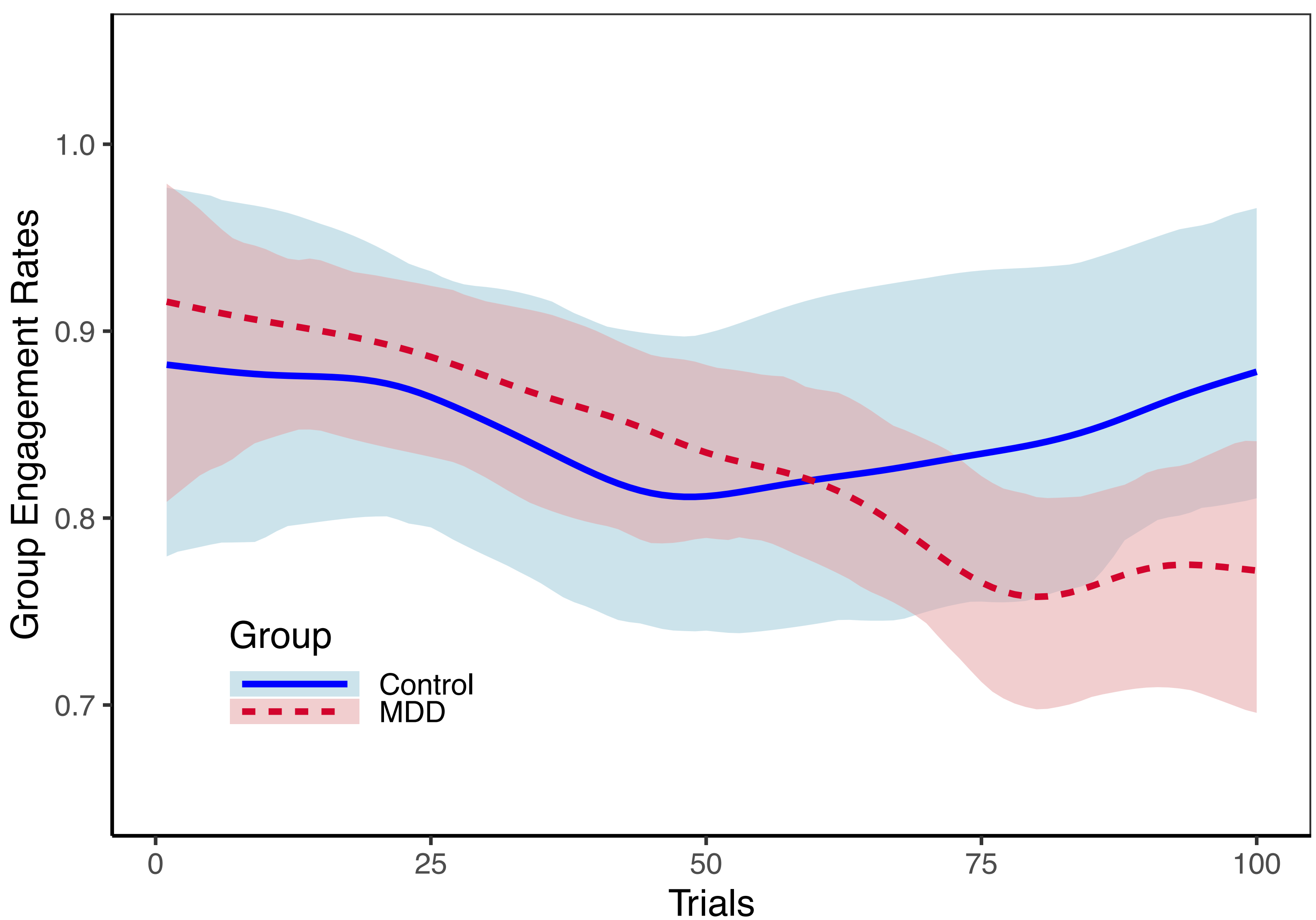
Application to EMBARC Study

Results: MDD vs Control

Individual engaged probability at trial t: $H_i(t) = P\left(U_{it} = 1 \mid A_{i[1:T]}\right)$ → **posterior probability for subject i being engaged at trial t.**

Group engaged rate at trial t: $\bar{H}(t) = n^{-1} \sum_{i=1}^{n} H_i(t)$

Group engaged rates (MDD vs Control)



MDD group potentially experiences greater difficulty in concentration compared to the control group at the second half of the task.

Application to EMBARC Study

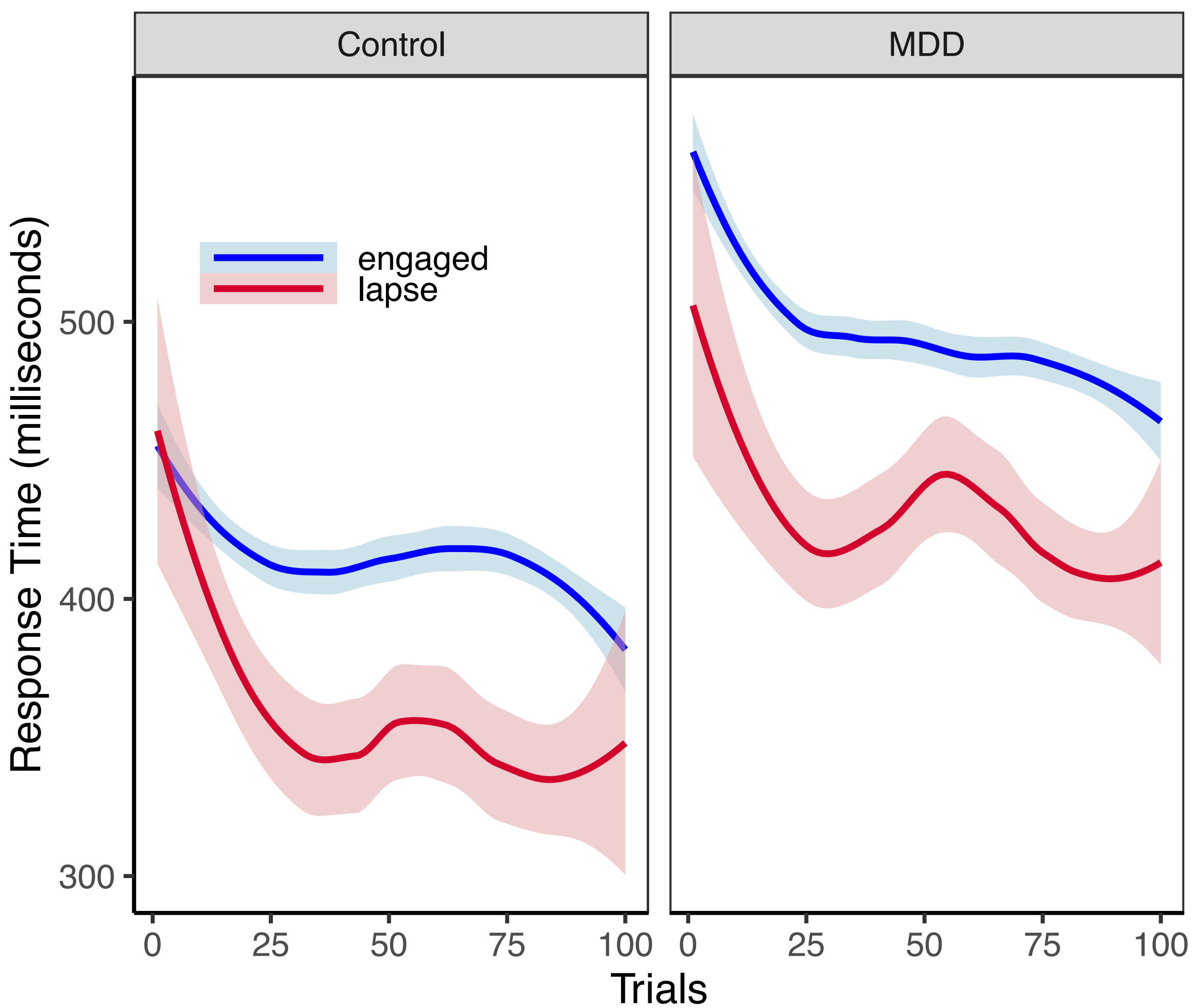
Results: MDD vs Control

Individual engaged probability at trial *t*: $H_i(t) = P\left(U_{it} = 1 \mid A_{i[1:T]}\right)$

Identify the learning strategies: **engaged**, if $H_i(t) \geq 0.5$ **lapse**, if $H_i(t) < 0.5$

Response time (decision making time): time between **state-showing** and **action-taking**.

Response time vs Trials



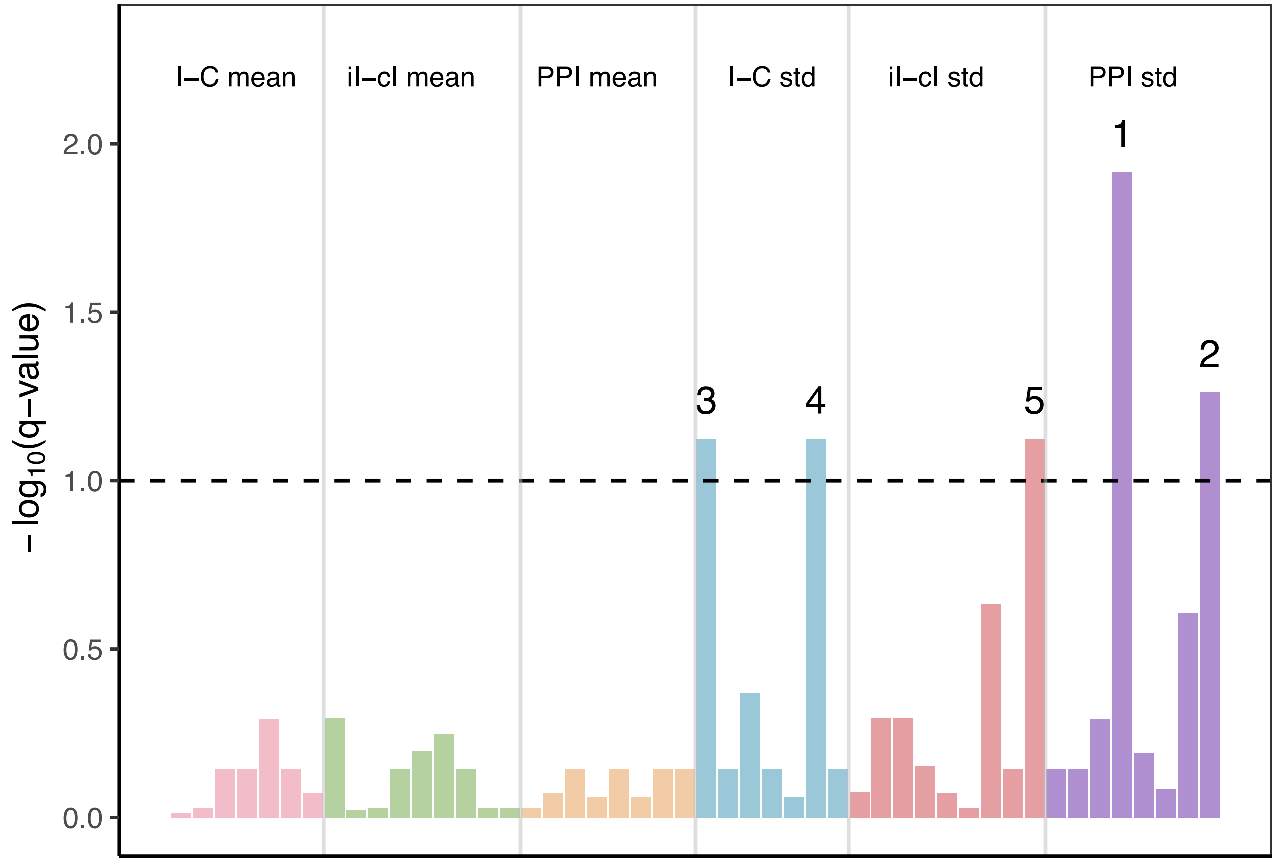
- **'Engaged' strategy takes more time to make decisions compared to the 'lapse' strategy.**
- **Control group takes less time to make decisions than the MDD group.**

Brain-behavior association

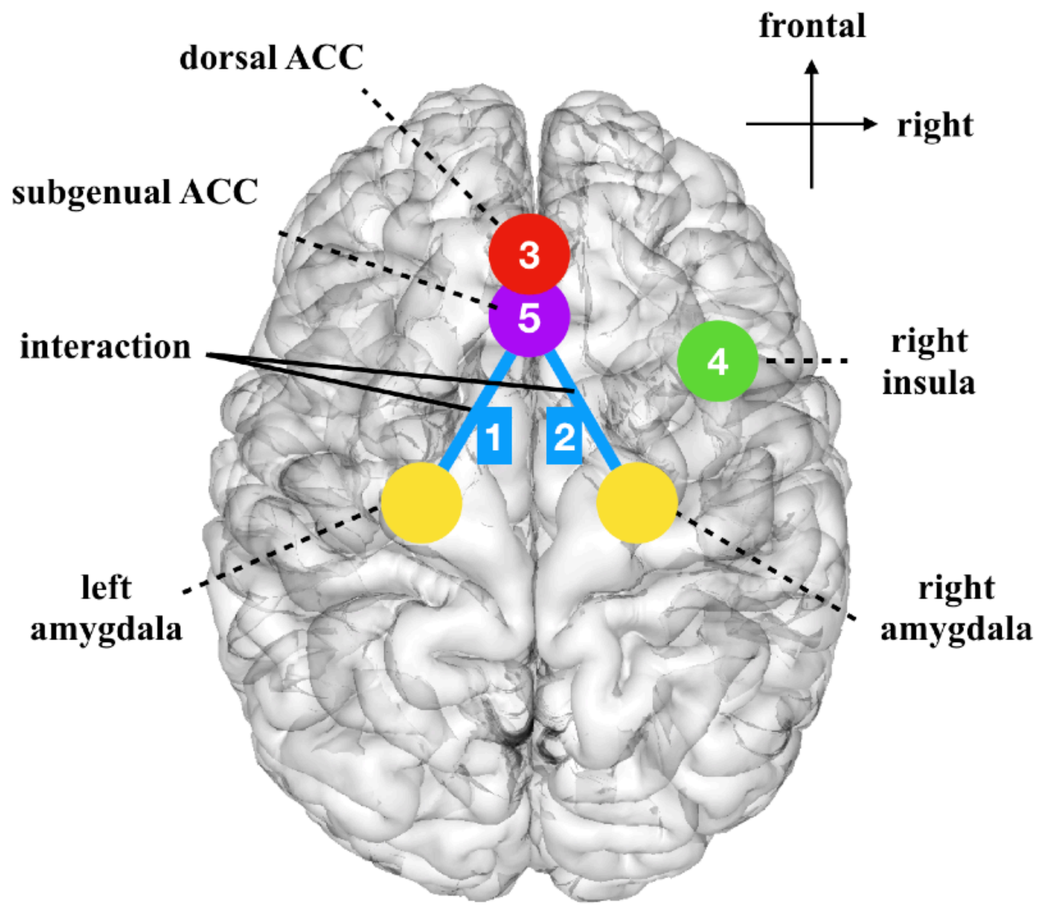
We focus on fMRI measures in an Emotional Conflict Task *(Etkin et al., 2006)* assessing amygdala-anterior cingulate (ACC) circuitry.

I-C: activation conflicts
iI-cI: activation conflict adaptations
PPI: psychophysiological interaction

fMRI measures vs IES significance



Visualization



An increased engagement in reward learning tasks corresponds to a decreased variability in brain activity during an emotional conflict task.

Discussion

Propose Semiparametric inverse RL and RL-HMM frameworks to characterize reward-based decision-making with an application of probabilistic reward tasks in the EMBARC study.

Semiparametric inverse RL

- **The Control group has a larger reward sensitivity function compared to the MDD group when receiving enough rewards.**
- **The reward sensitivity function is nonlinear with a floor and ceiling effect.**

RL-HMM

- **Humans employ multiple decision-making strategies in reward learning.**
- **MDD group potentially experiences greater difficulty in concentration compared to the control group.**

Extensions

- **Jointly modeling RL process and response time.**
- **Brain-behavior association.**
- **Jointly modeling multiple human tasks.**

Acknowledgement

Postdoctoral Fellow: **Xingche Guo (incoming Assistant Professor at University of Connecticut)**

Key References

Guo, X., Zeng, D., & Wang, Y. (2024). A Semiparametric Inverse Reinforcement Learning Approach to Characterize Decision Making for Mental Disorders. ***Journal of the American Statistical Association*****.**

Guo, X., Zeng, D., Wang, Y. (2024). HMM for Discovering Decision-Making Dynamics Using Reinforcement Learning Experiments. ***Accepted by Biostatistics, arXiv:2401.13929***

Reference support: **NS073671, GM124104, and MH123487**

Thank you