Outcome valence and prediction error sign invariance of the reinforcement learning models

Kamil Bonna[1,2], Karolina Finc[1], David Meder[3], Kristoffer Madsen[3], Włodzisław Duch[1,2], Oliver Hulme[3]

[1] Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University, Toruń, Poland
[2] Institute of Physics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University in Toruń, Poland
[3] Danish Research Centre for Magnetic Resonance, Copenhagen University, Copenhagen, Denmark

Probabilistic reversal learning task

During the fMRI scanning session participant carried out the PRL task in two conditions: reward-seeking and punishment-avoiding. Participants were instructed to repeatedly choose between yellow and blue boxes in order to collect as many points as possible (reward condition) or loose as little points as possible (punishment condition). Participants had to learn reward probabilities from experience. Task parameters:

- box probabilities were set to 0.8 and 0.2
- reward contingency changed 4 times throughout the task
- single run consisted of 110 trials

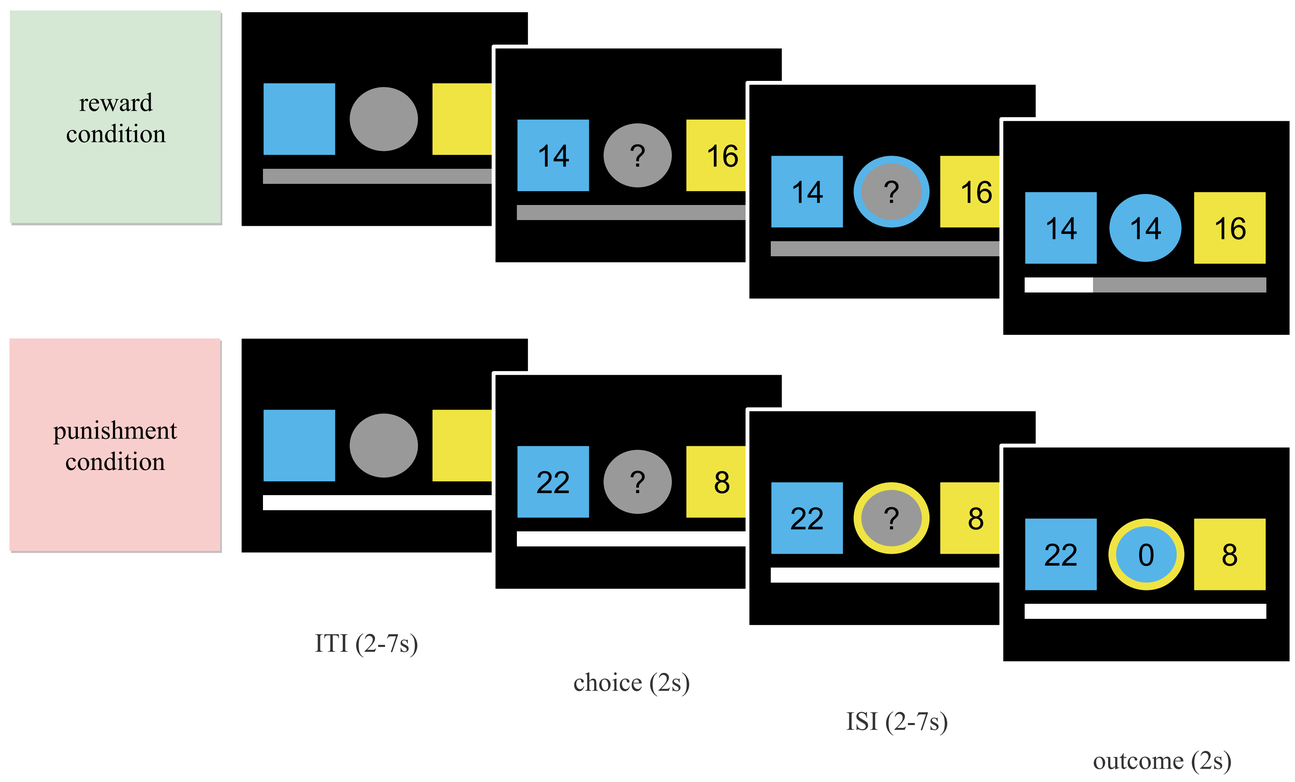
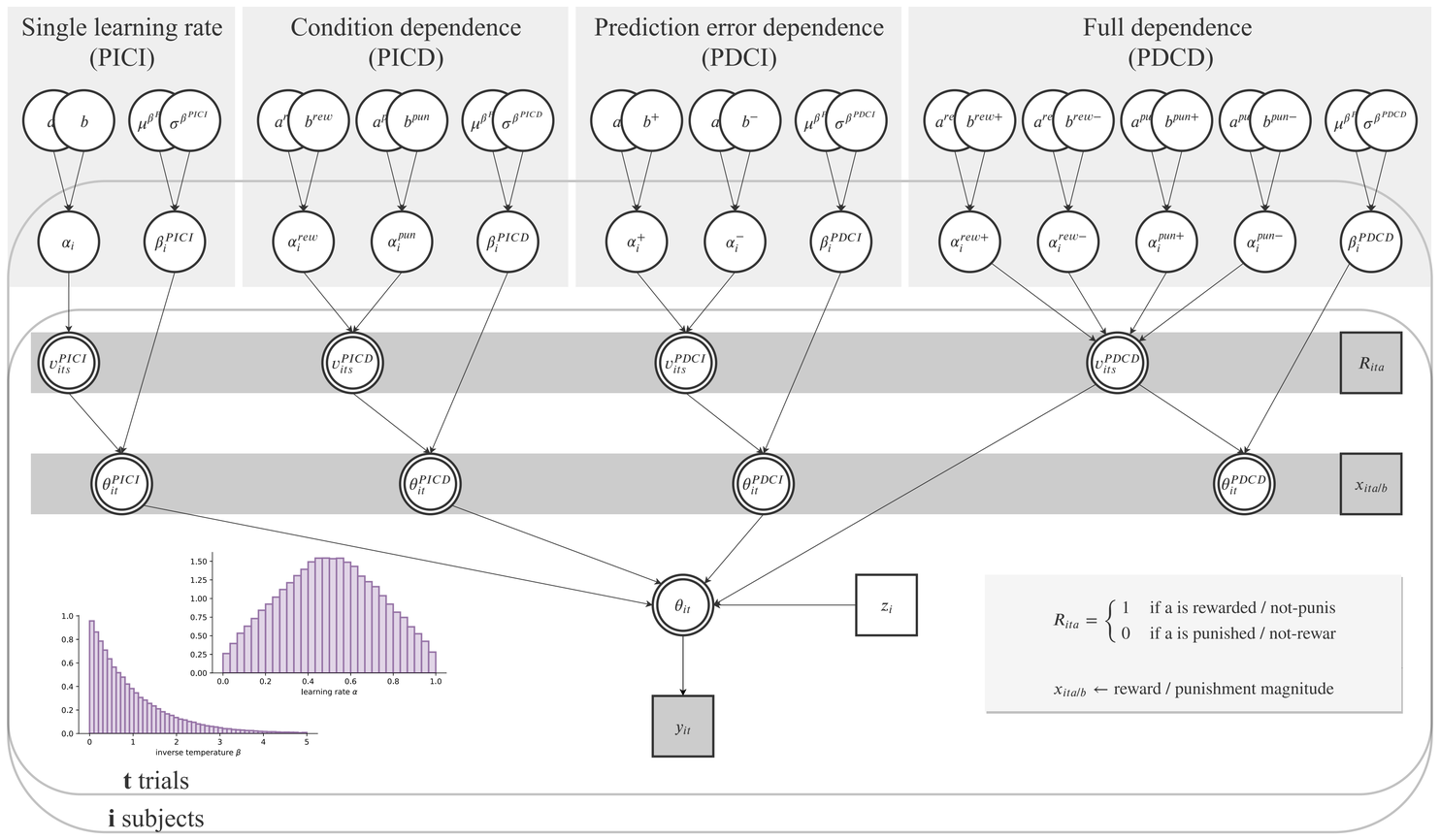


Figure 1. Example probabilistic reversal learning task structure.

Hierarchical latent-mixture model

	Single learning rate (PICI)	Condition dependence (PICD)	Prediction error dependence (PDCI)	Full dependence (PDCD)
hyperpriors	$a \sim \text{Uniform}(1, 10)$ $b \sim \text{Uniform}(1, 10)$	$\mu^{\beta} \sim \text{Uniform}(-2.3, 3.4)$ $\sigma^{\beta} \sim \text{Uniform}(0.01, 1.6)$		
priors	$\alpha_i \sim \text{Beta}(a, b)$	$\beta_i \sim \text{Lognormal}(\mu^{\beta}, \sigma^{\beta})$		
success probability	$v^{PD}_{itc} = \begin{cases} v^{PD}_{it-1c} + \alpha_i^{rew+/pun+}(1 - v^{PD}_{it-1c}) & \text{if} \quad R_{itc} = 1 \\ v^{PD}_{it-1c} + \alpha_i^{rew-/pun-}(0 - v^{PD}_{it-1c}) & \text{if} \quad R_{itc} = 0 \end{cases}$			$v^{PD}_{itu} = 1 - v^{PD}_{itc}$
	$v^{PI}_{ita} = v^{PI}_{it-1a} + \alpha_i^{rew/pun}(R_{ita} - v^{PI}_{it-1a})$			$v^{PI}_{itb} = 1 - v^{PI}_{ita}$
choice probability	$\theta_{it} = (1 + \exp(\beta_i(v_{ita}x_{ita} - v_{itb}x_{itb})))^{-1}$			
model indicator	$z_i \sim Categorical(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$			
behavioral response	$y_{it} \sim Bernoulli(\theta_{it})$			



Model selection

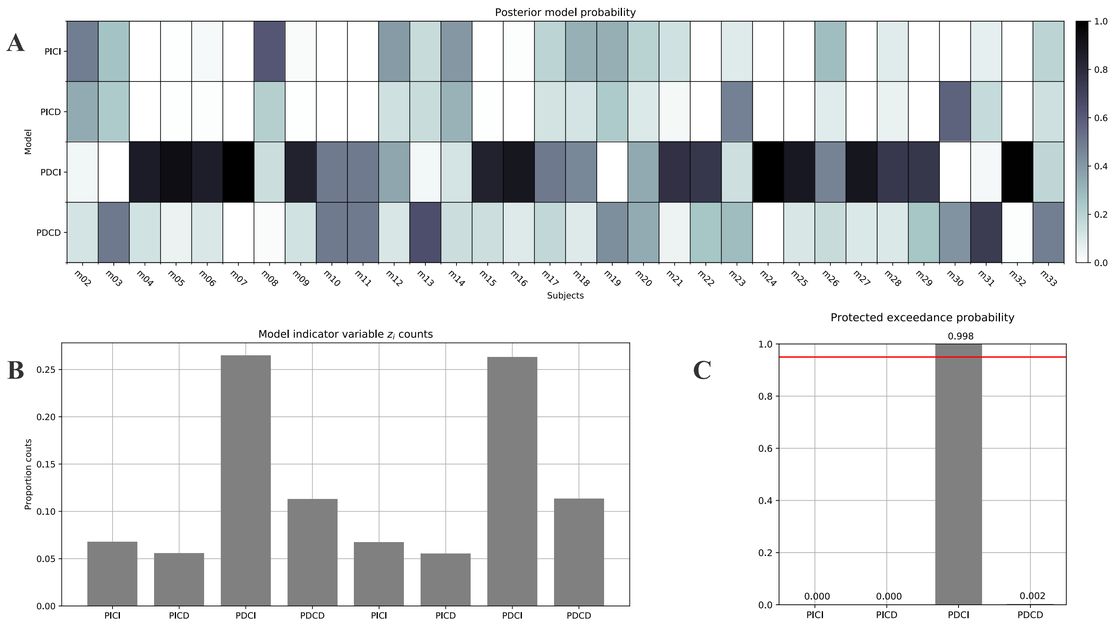


Figure 2. Model selection results. Posterior distribution of model indicator variable for (A) individual subjects and (B) marginalized over subjects. (C) Protected exceedance probabilities for all four model.

Parameter recovery

We found that model with separate learning rates for positive and negative PEs (PDCI) was the best model (PEP > 95%) for the observed behavioral data. However, for some subjects other models had highest posterior model probability. We evaluated PDCI model separately to recover model parameters. For each subject, we calculated mean of the posterior for learning rates and inverse temperature. We found that:

- almost all subjects had extremely high values (close to 1) of learning rate for positive prediction error
- learning rates for negative prediction errors had moderate values
- difference between two learning rates was significantly correlated with mean number of reversals (r=-0.76, p<0.001)

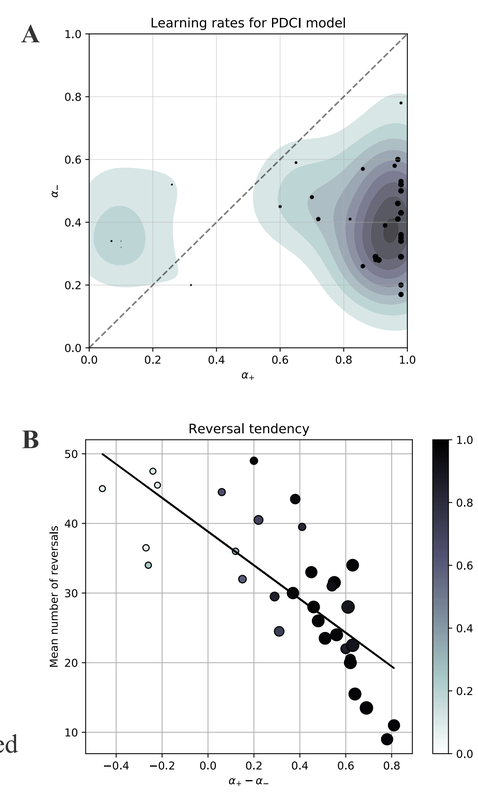


Figure 3. Parameter recovery for the winning model. (A) Estimated learning rates. (B) Correlation between estimated parameters and reversal tendency.

e-mail: bonna@doktorant.umk.pl

Code available on: https://github.com/kbonna/decidenet