Computational modelling of reversal learning

in rodents self-administering cocaine

Modelling Choices:

Q-Learning (model-free) aka Rescorla Wagner model:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha * (\, r - Q_t(c_t))$$

Policy-based π learning (model-based):

$$\pi_{t+1}(c_t) = \pi_t(c_t) + (r_t - \bar{r})$$

Softmax decision rule: explore vs exploit based on value learned above:

$$P(c_t = L | Q_t(L), Q_t(R)) \; = \frac{exp(\, Q_t(L)/\beta)}{exp(Q_t(L)/\beta) + exp(\, Q_t(R)/\beta)}$$

$$P(c_t = L | \pi_t(L), \pi_t(R)) \; = \frac{exp(\, \pi_t(L)/\beta)}{exp(\, \pi_t(L)\, /\beta) + \; exp(\, \pi_t(R)/\beta)}$$

Value representation

Decision probability

Model 1: Policy based learning

Policy-based π learning (model-based):

[1] Policy/value representations for each choice c_t at trial *t*:

$$\pi_{t+1}(c_t) = \pi_t(c_t) + (r_t - \bar{r})$$

[2] Probability of choosing c_t at trial *t* (softmax):

$$P(c_t = L|\pi_t(L), \pi_t(R)) = \frac{exp(\pi_t(L)/\beta)}{exp(\pi_t(L)/\beta) + exp(\pi_t(R)/\beta)}$$

[3] Probability of observing *data D* (a sequence of choices and rewards) = product of the individual probabilities from [2]

$$P(Data\ D|Model\ M, parameters\ \theta) = P(D|M,\theta) = \prod P(c_t|Q_t(L), Q_t(R))$$

[4] Fitting 1 parameter (β = θ) to achieve maximum likelihood of *data D*

$$\underset{\theta}{argmax}\ P(D|M,\theta)$$

Model 2: Q-learning | 2 parameters

Q-Learning (model-free) aka Rescorla Wagner model:

[1] Q value representations for each choice c_t at trial *t*:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha * (r - Q_t(c_t))$$

[2] Probability of choosing c_t at trial *t* (softmax):

$$P(c_t = L | Q_t(L), Q_t(R)) = \frac{exp(Q_t(L)/\beta)}{exp(Q_t(L)/\beta) + exp(Q_t(R)/\beta)}$$

[3] Probability of observing *data D* (a sequence of choices and rewards) = product of the individual probabilities from [2]

$$P(Data\ D | Model\ M, parameters\ \theta) = P(D|M,\theta) = \prod P(c_t | Q_t(L), Q_t(R))$$

[4] Fitting 1 parameter (β = θ) to achieve maximum likelihood of *data D* given

$$\underset{\theta}{argmax}\ P(D|M,\theta)$$

Model 3: Q-learning | 3 parameters

Q-Learning (model-free) aka Rescorla Wagner model:

[1] Q value representations for each choice c_t at trial *t*:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha * (\, r - Q_t(c_t))$$

Model-free Q-learning with 2 learning parameters: α_{REWARD} and $\alpha_{NO\ REWARD}$

[2] Probability of choosing c_t at trial *t* (softmax):

$$P(c_t = L | Q_t(L), Q_t(R)) \; = \frac{exp(\, Q_t(L)/\beta)}{exp(Q_t(L)/\beta) + exp(\, Q_t(R)/\beta)}$$

[3] $P(D|M,\theta)$: [$\alpha_{REWARD}, \alpha_{NO\ REWARD}\beta$] = θ

[4] $\underset{\theta}{argmax}\, P(D|M,\theta)$

as before

Model 4: Q-learning | 3 parameters

Q-Learning (model-free) aka Rescorla Wagner model:

[1] Q value representations for each choice c_t at trial *t*:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha * (\,r - Q_t(c_t))$$

[2] Probability of choosing c_t at trial *t* (softmax): include choice autocorrelation by modelling *kappa* κ; $-1 < \kappa < 1$
$L_{t-1} = 1$ if previous choice was *Left* otherwise $L_{t-1} = 0$

$$P(c_t = L|Q_t(L), Q_t(R), L_{t-1}, R_{t-1}) = \frac{exp(\,Q_t(L)/\beta + \kappa * L_{t-1})}{exp(Q_t(L)/\beta + \kappa * L_{t-1}) + exp(\,Q_t(R)/\beta + \kappa * R_{t-1})}$$

"perseveration" for $0 < \kappa < 1$; "switching" for $-1 < \kappa < 0$

[3] $P(D|M, \theta)$: $[\alpha, \beta, \kappa] = \theta$

[4] $\underset{\theta}{argmax}\, P(D|M, \theta)$

as before

Model comparison: Model 1 vs 2

Model-based policy π learning vs model-free Q-Learning:

Different number of parameters fitted: only β in policy learning vs α, β in Q-learning

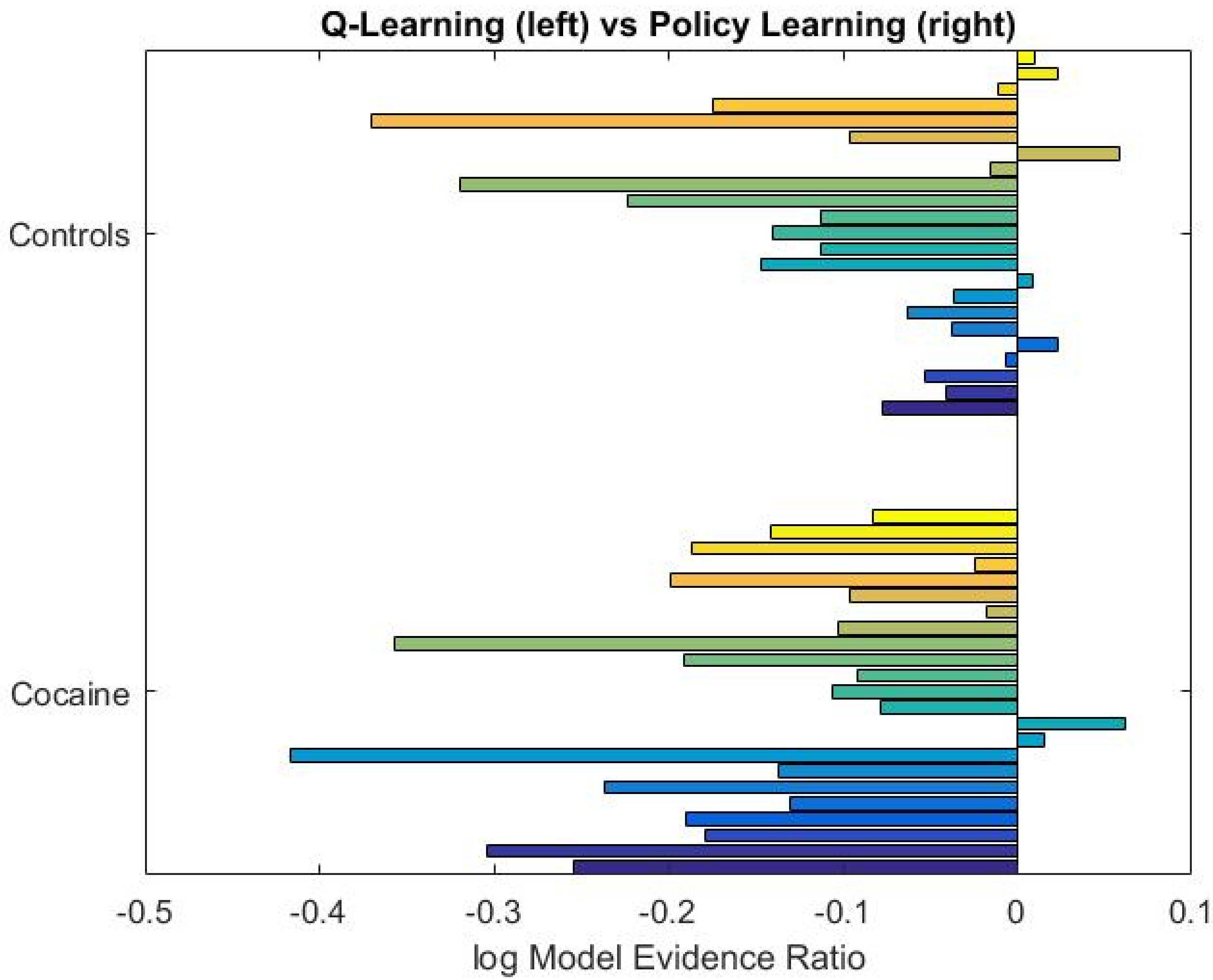
Bayes Factor: $$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1) * P(M_1)}{P(D|M_2) * P(M_2)}$$

$$where\ model\ evidence\ P(D|M)\ is\ computed\ as\ the\ average\ over\ P(D|M,\theta)$$

$$for\ each\ parameter\ probed, P(\theta|M)$$

Model comparison: Model 1 vs 2

Model-based policy π learning vs model-free Q-Learning:



Model comparison: Model 2 vs 3 & 4

Comparing the log-likelihoods of two models with the set of model parameters $\hat{\theta}_M$ that maximise the likelihood of observing *data D* given model M.

Q-Learning model M_2 with $\hat{\theta}_M$=[α, β] vs

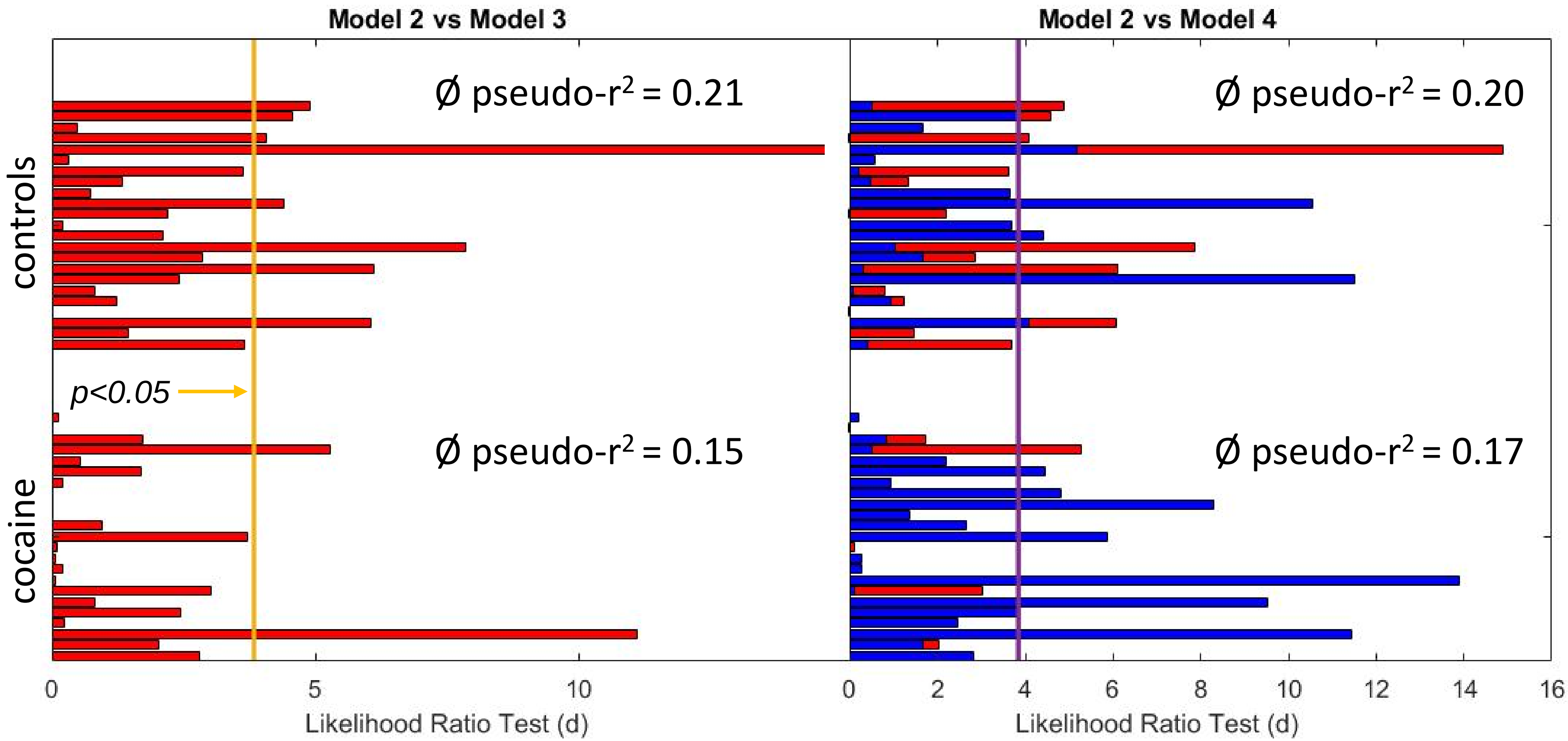
Q-Learning model M_3 with $\hat{\theta}_M$=[α_{Reward}, $\alpha_{No\ Reward}$, β]

Q-Learning model M_3 with $\hat{\theta}_M$=[α, β, κ]

$$d = 2 * \left[\log P(D|M_3, \hat{\theta}_{M_3}) - \log P(D|M_2, \hat{\theta}_{M_2})\right]$$

Since d follows a *Chi-square* distribution, we can obtain *p-values* for each of these likelihood ratios, where *Chi-square*$_{p<0.05}$=3.84

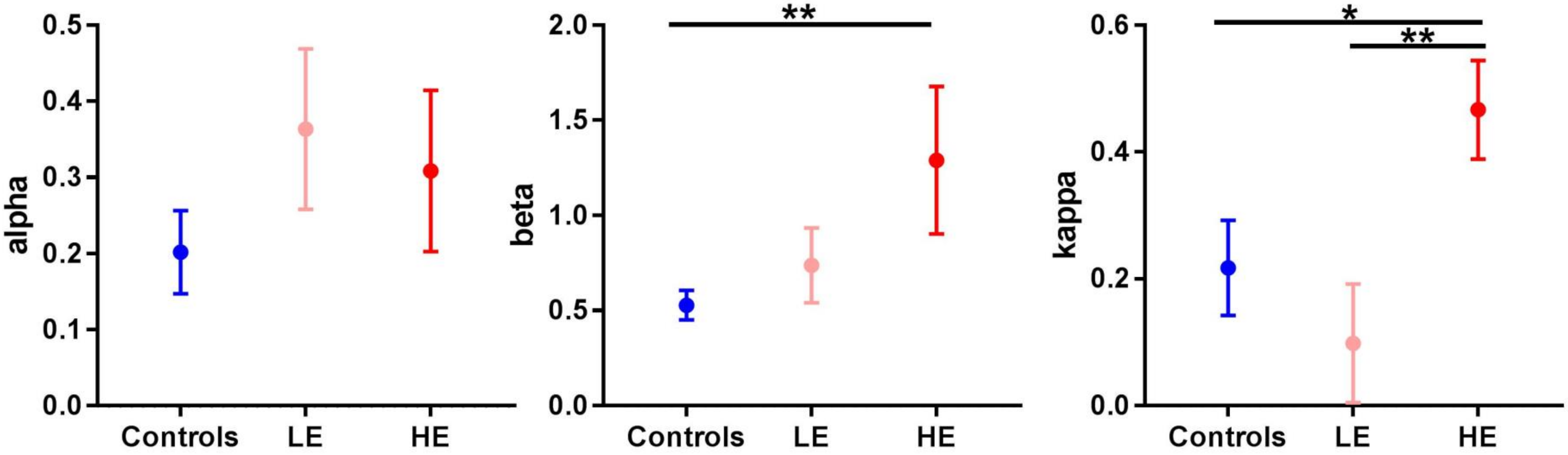
Model comparison: Model 2 vs 3 & 4



Both models with 3 parameters improve fit in many subjects, however Model 4 (with choice autocorrelation) fits the cocaine group data more accurately than Model 3 (with 2 learning parameters *alpha*)

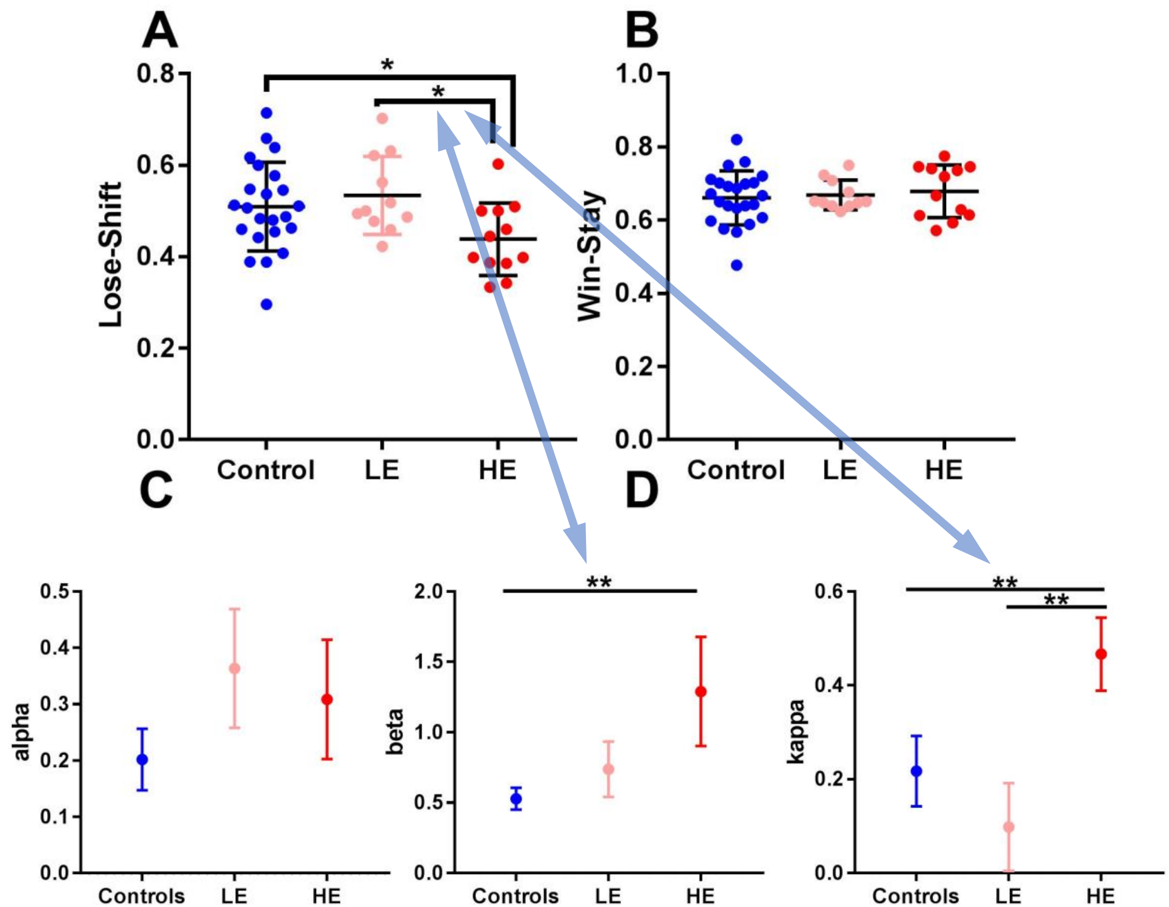
Application: Cocaine effects on reversal ability in rats

- High escalation animals are not exploiting what they learn about the choice (Q) values: large *beta* indicates random switching between responses rather than sticking with the highest Q value response
- HE animals also perseverate more, sticking with previous response rather than switching: large *kappa* indicates choice at trial *t* is influenced more by choice at trial *t-1*
- No significant differences in learning rate: *alpha*



Data are Mean ± SEM; **p<0.01; *p<0.05, LSD were used as post-hoc tests

Cocaine effects on reversal ability in rats: Computational Modelling vs Lose-Shift



Data are Mean ± SEM; **p<0.01; *p<0.05, LSD were used as post-hoc tests