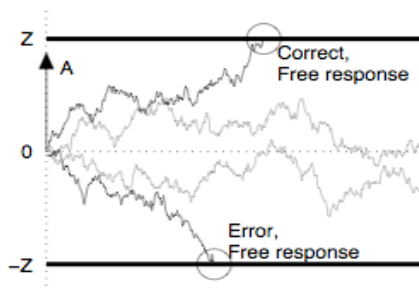
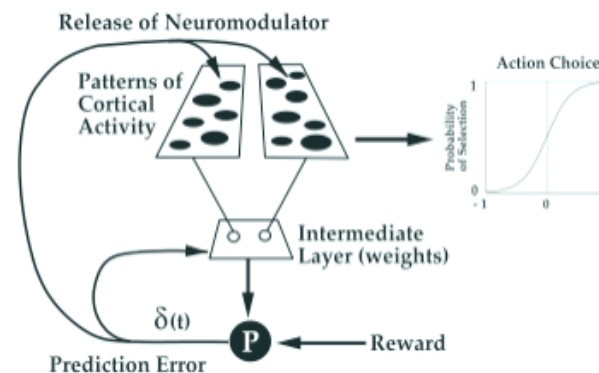
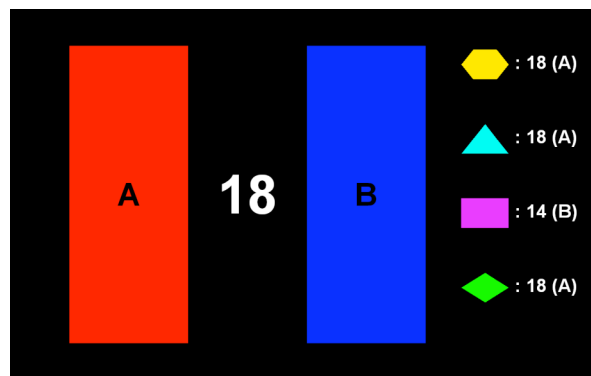
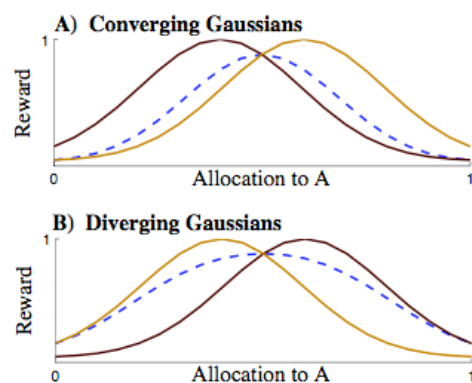


# Stochastic models for decisions in social groups, or, gambling with partial feedback.

Philip Holmes,

Andrea Nedic, Damon Tomlin, Deborah Prentice & Jonathan D. Cohen,  
Princeton Neuroscience Institute & EE, Psychology, MAE, Applied Mathematics.



$$p_A = \frac{1}{1 + e^{\mu(w_B - w_A)}}$$

AFOSR MURI-16 review, Alexandria, VA, Nov 12th, 2009.

# Contents

- I:** Gambling in groups against a two-armed bandit as a model for decisions in a social context.
- II:** Extension of a simple choice model to include social feedback. Model comparisons and fits.
- III:** A sample of behavioral results: the pleasures and perils of peeking at others' rewards.
- IV:** Summary.

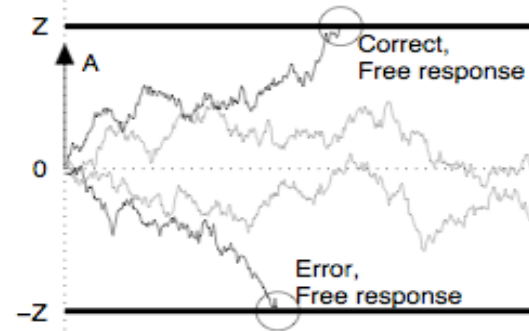
**A modest advertisement for stochastic ODEs.**

# Preamble: optimal choices

The drift-diffusion (DD) process, a cornerstone of 20th century physics, and perhaps the simplest stochastic ODE, is an **optimal decision maker** for two-alternative forced choice perception tasks with noisy data:

$$dx = A dt + c dW$$

drift rate      noise strength



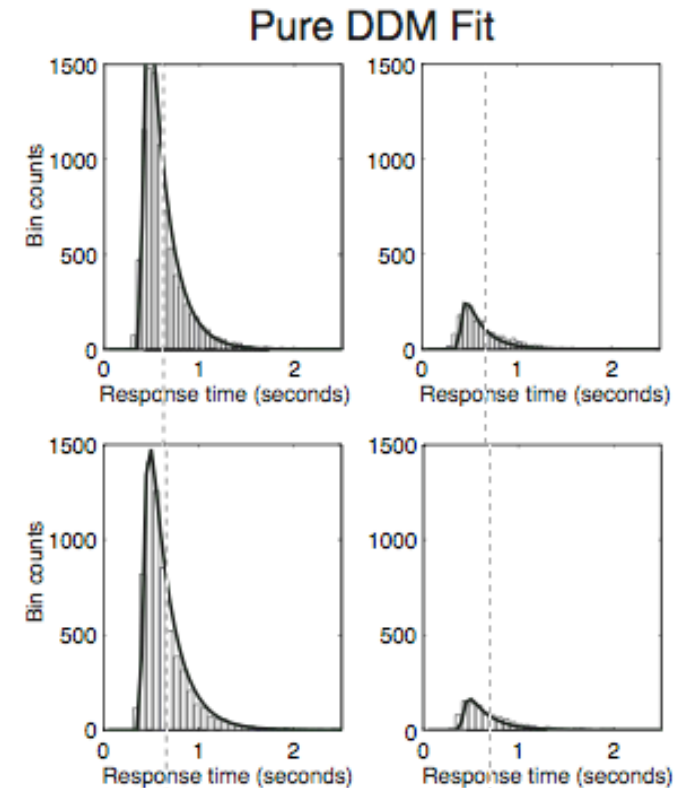
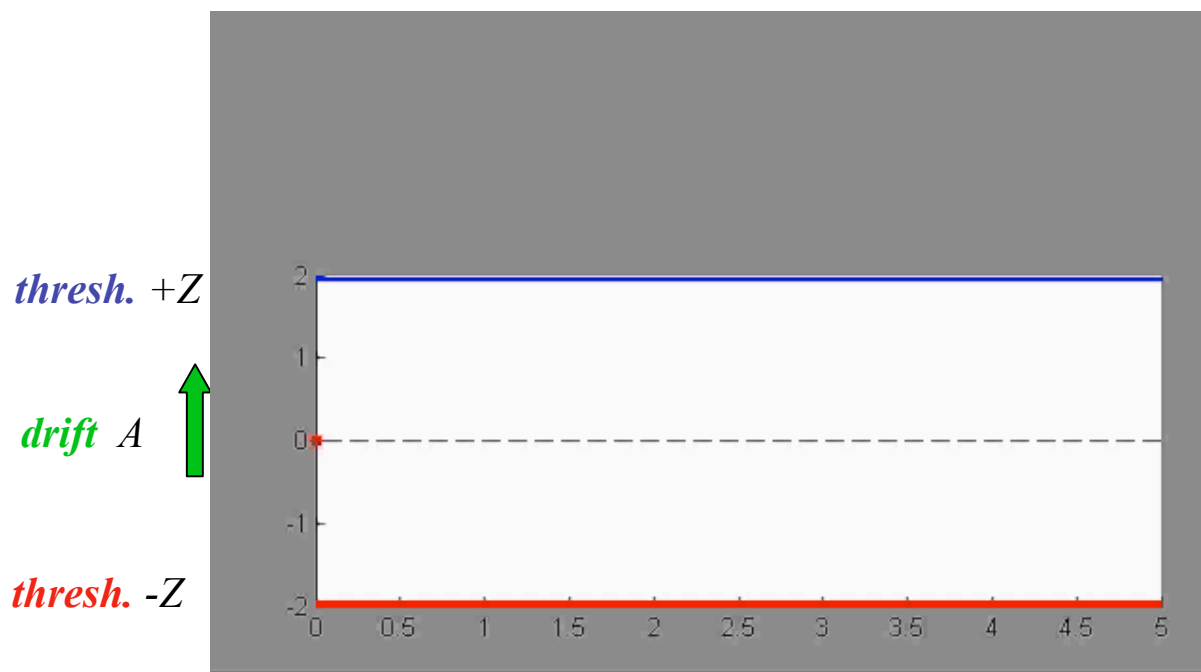
Here  $x(t)$  is the accumulated evidence (the log likelihood ratio): when it reaches the threshold  $Z$  or  $-Z$ , declare **R** or **L** the winner. DD is a continuum limit of SPRT (Wald, 1947). The model has only 2 parameters:

$$\alpha = (Z/A) \text{ (TDR)} \quad \text{Avg. Accuracy} = \frac{1}{1 + \exp(\alpha\beta)}$$
$$\beta = (A/c)^2 \text{ (SNR)}$$

**Behavioral data, neural recordings in primates, fMRI and EEG in humans (MURI 16) and spiking neuron models support the contention that evidence is accumulated in this manner by cortical networks.**

# Behavioral evidence: RT distributions

Human reaction time data in free response mode can be fitted to the first passage threshold crossing times of a DD process.



Prior or bias toward one alternative can be implemented by setting starting point  $x(0) \neq 0$ . Extended DD: variable drift rates & starting points.

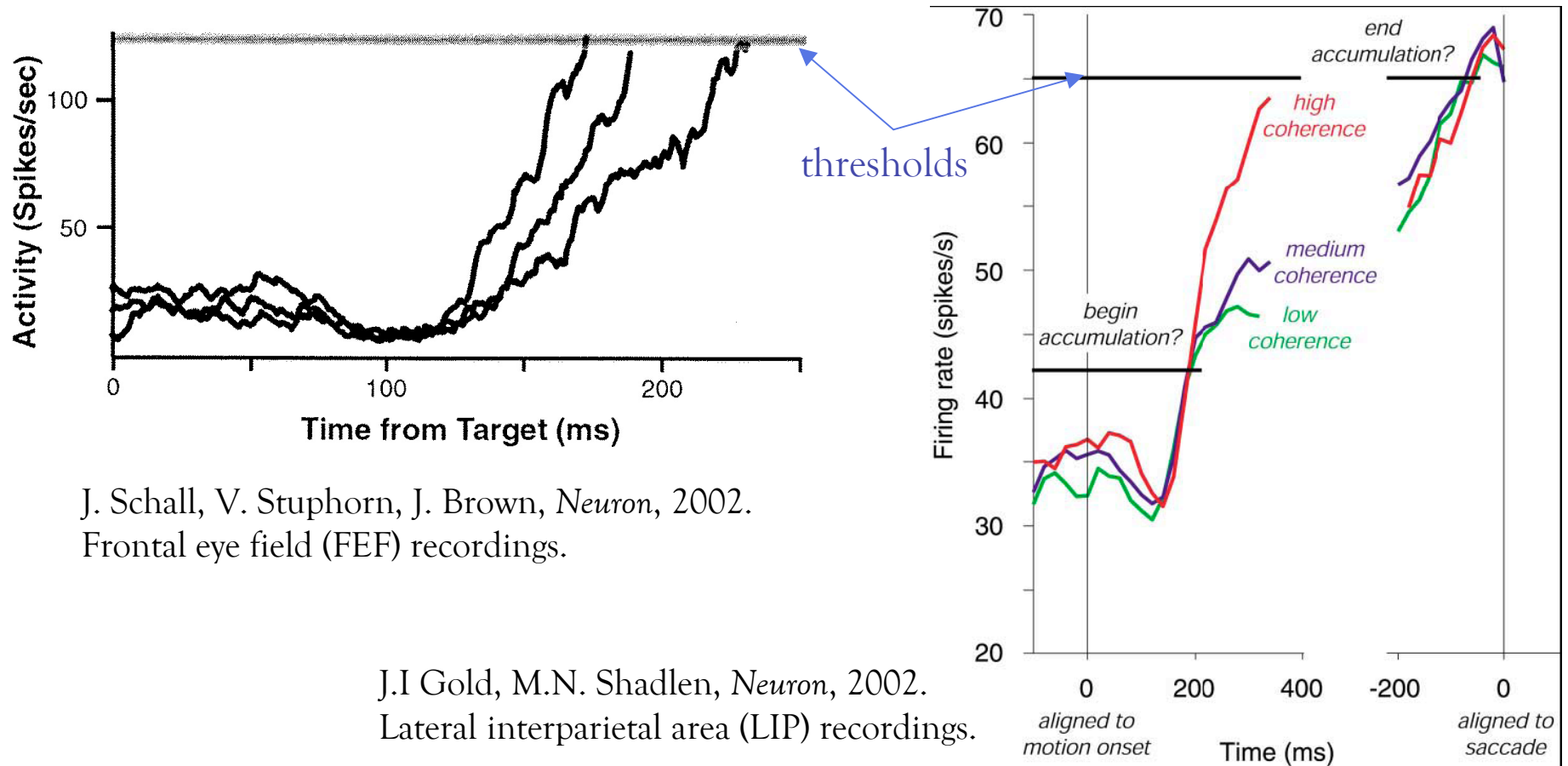
Simple expressions for mean RT, accuracy.

Ratcliff et al., *Psych Rev.* 1978, 1999, 2004

Simen et al., *J. Exp. Psych.:HPP*, in press, 2009.

# Neural evidence: firing rates

Spike rates of neurons in cortical areas rise during stimulus presentation, monkeys signal their choice after a threshold is crossed (like DD process).

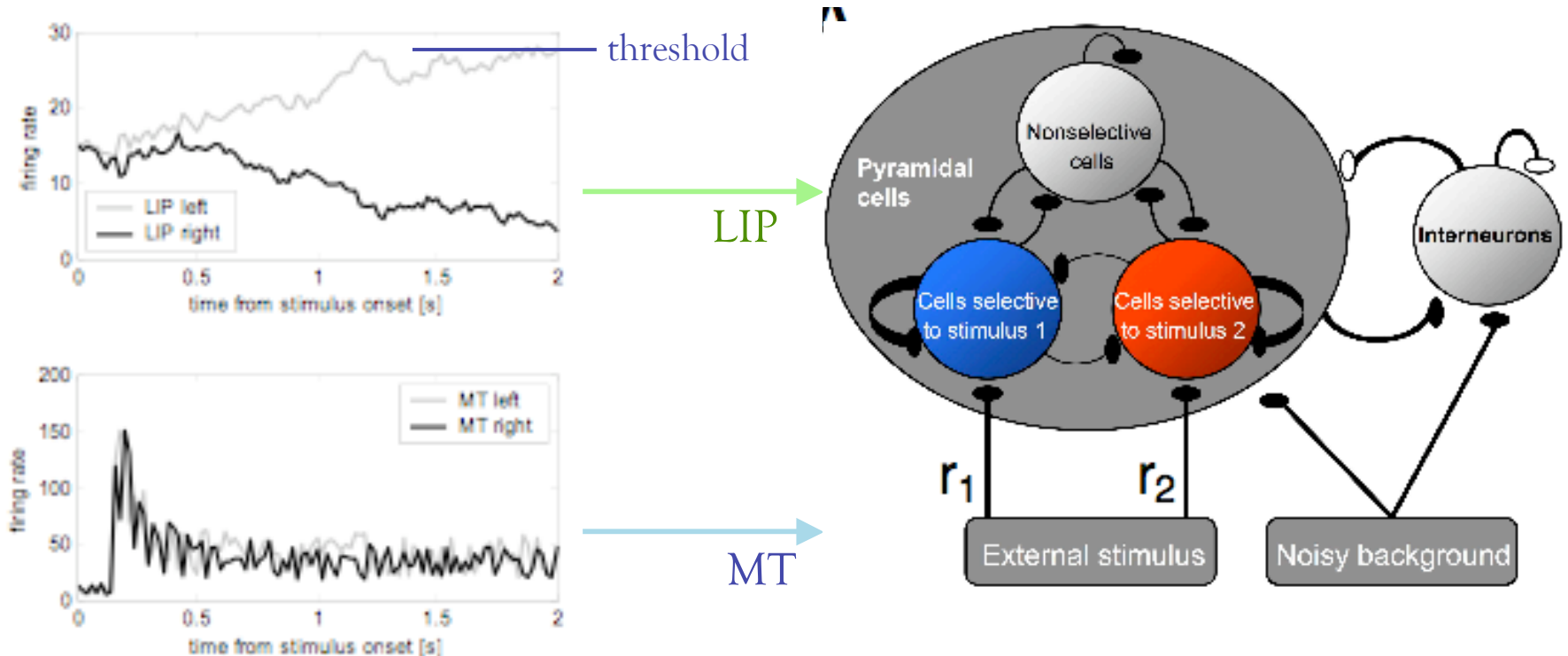


J. Schall, V. Stuphorn, J. Brown, *Neuron*, 2002.  
Frontal eye field (FEF) recordings.

J.I Gold, M.N. Shadlen, *Neuron*, 2002.  
Lateral interparietal area (LIP) recordings.

# Neural evidence: spiking neuron models

In decisions based on visual perception, motion sensitive cells in visual cortex (MT) pass noisy signals on to LIP, FEF, where integration occurs.



## Experimental observations:

K.H. Britten, M.N. Shadlen, W.T. Newsome, J.D. Schall & A. Movshon, various papers, 1992-2004.

## Related work on MURI 15

## Simulation & analysis of spiking neurons:

X.-J. Wang, *Neuron*, 2002;

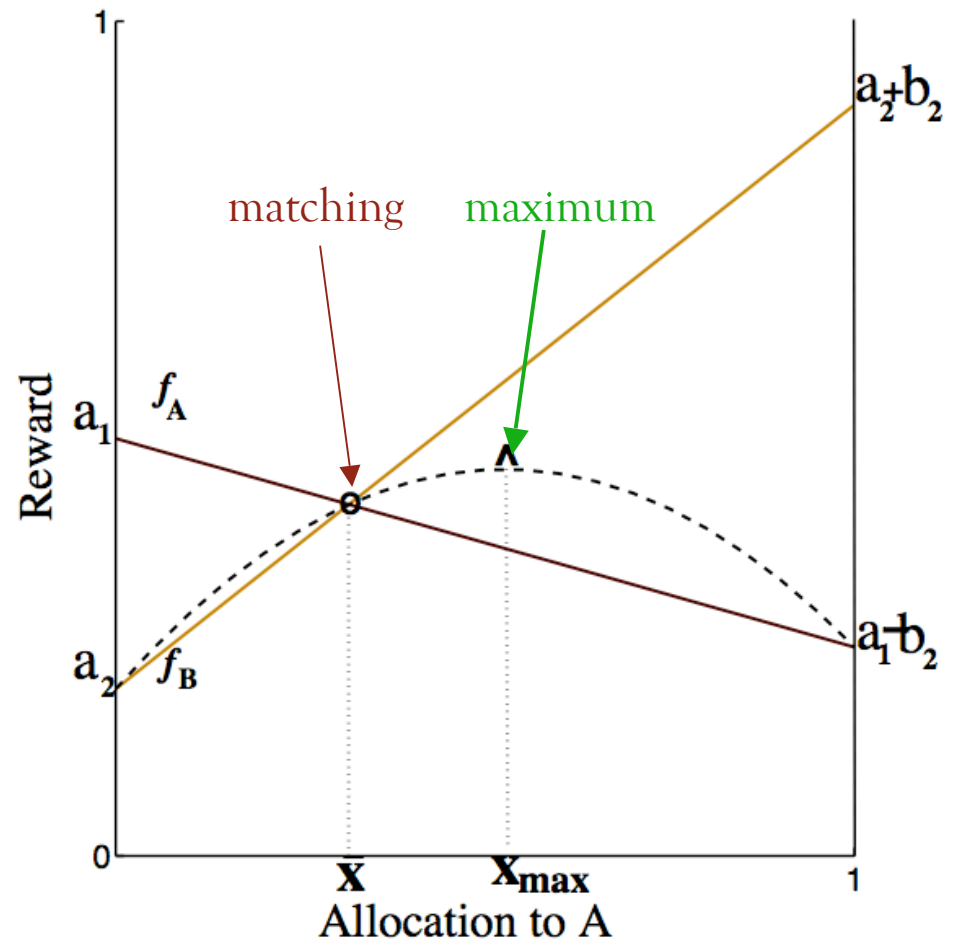
K-F. Wong & X.-J. Wang, *J. Neurosci.*, 2006.

Ongoing work extending to model NE modulation of LIP:  
P. Eckhoff, K-F. Wong, H. *J. Neurosci*, 29 (13), 2009; mean field reduction to 4- and 2-population models (in review); stochastic averaging over populations: A. Saxe.

# Preamble: preference matching

R. Herrnstein studied pigeons and people choosing freely between two alternatives with rewards that decreased (linearly) the more frequently they were chosen ( $f_A, f_B$ ).

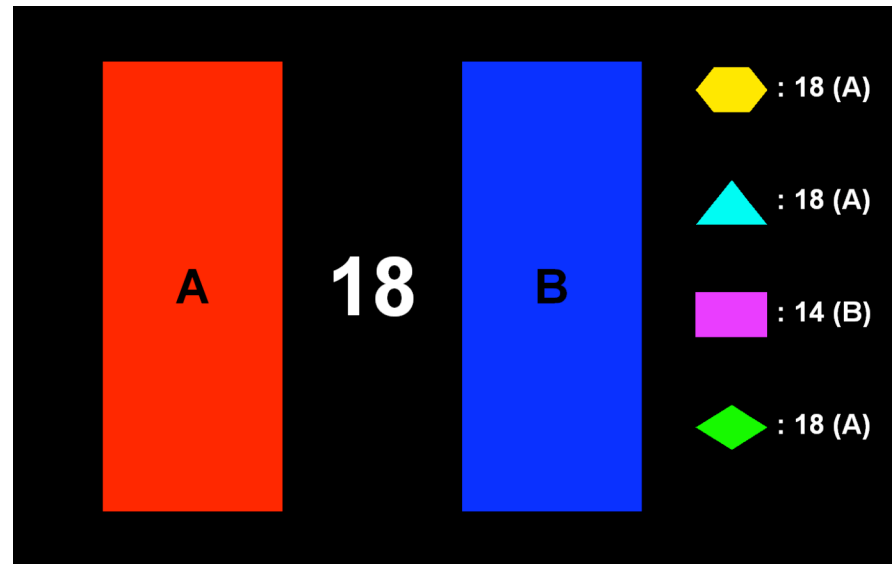
He found that subjects tended to settle at the matching point, where Allocations to A and B imply equal rewards for A and B, even if those are not the allocations that yield the maximum average rewards (dashed curve).



R.J. Herrnstein, 1961, 1982, 1990, 1991.

# I. The games: gambling in groups

A two-armed bandit game, motivated by Herrnstein, via Egelman & Montague, with limited social feedback.



Feedback  
on last choice

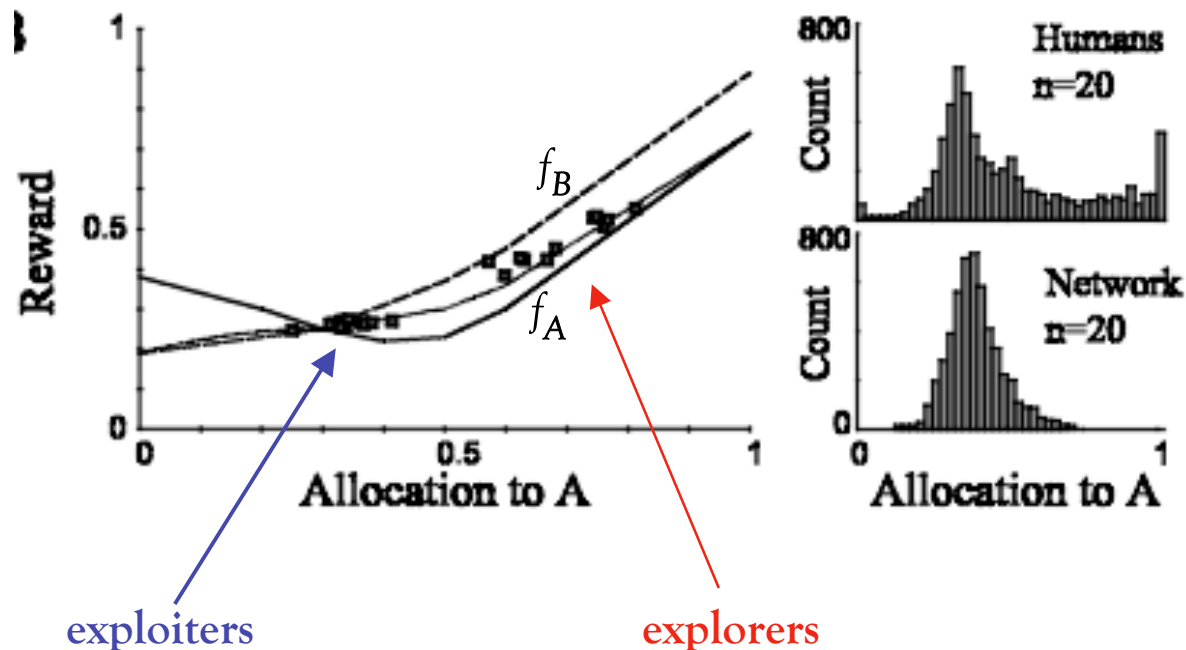
none (alone),  
choices,  
rewards,  
both.

- **128 participants:** Latin square design, all subjects played every game, all experienced every feedback condition at least once.
- **Data:** Free choices **A** or **B** on each trial in 2 sec window followed by delay and feedback; subjects performed 150 synchronized trials (choices) in each session (reaction times collected, not yet analyzed).

**Experimental design and data collection: Damon Tomlin:  
Behavioral and fMRI imaging data (hyperscanning x5).**

# The original E-P-M rising optimum task

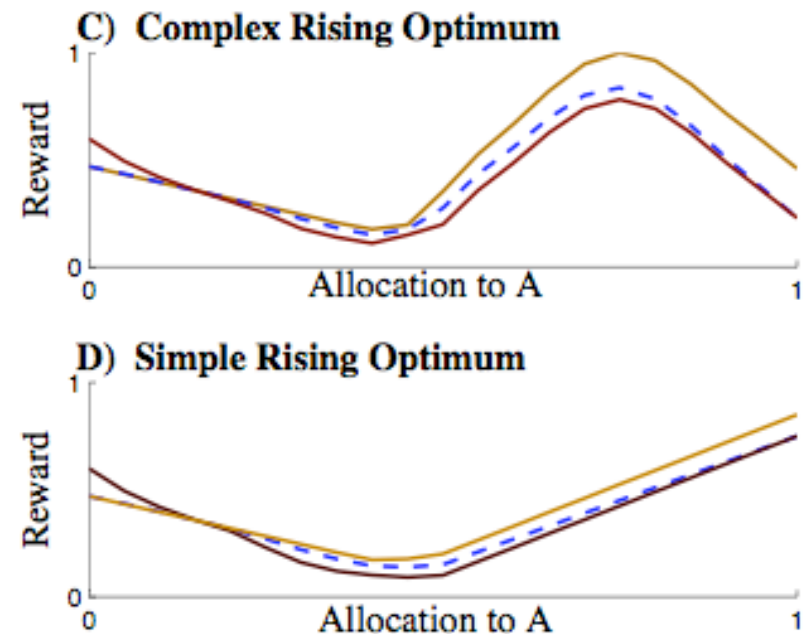
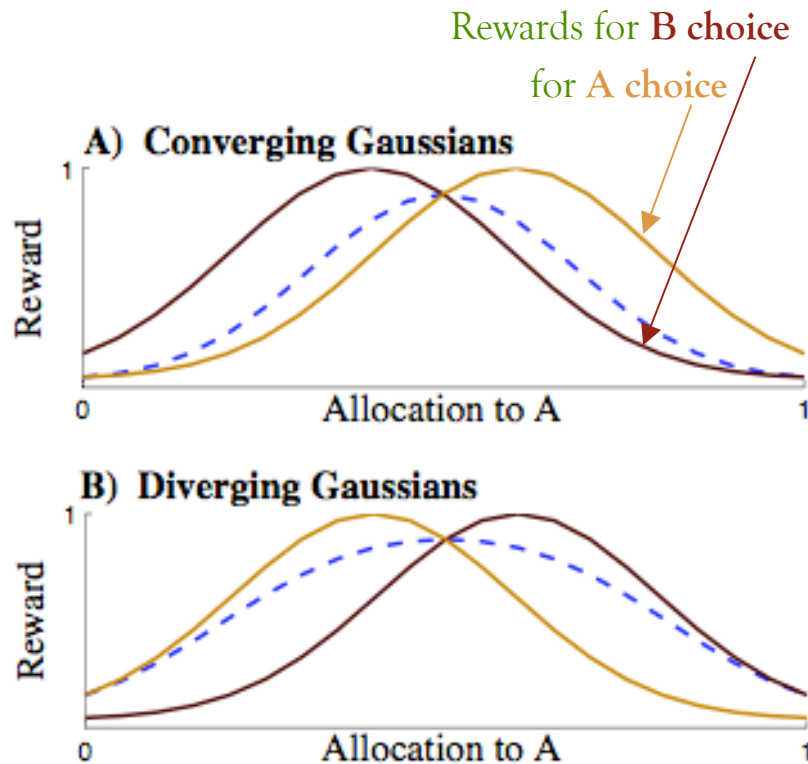
Egelman, Person & Montague found that individual subjects playing a rising optimum task divided into two types: “conservatives” (**exploiters**) remained near the matching point; “risk-takers” (**explorers**) crossed the low rewards region and approached closer to the optimal (all A) strategy.



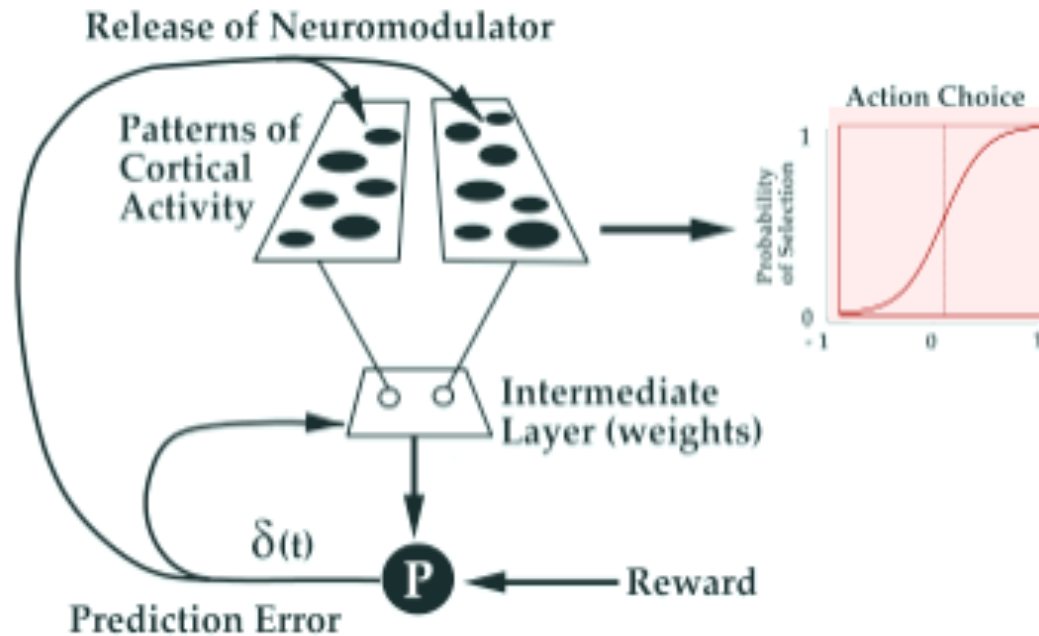
Egelman et al., *J. Cog. Neurosci.* 10, 1998; Montague & Berns, *Neuron* 36, 2002.

# Our deterministic nonlinear reward schedules

Subjects play four games, choosing freely and obtaining rewards based on their 20 previous choices, with nonlinear schedules from “simple” converging gaussians (A), “tricky” diverging gaussians (B), to more complicated rising optima cases with multiple maxima (C,D).



# A model for choice informed by experience 1



$$p_A = \frac{1}{1 + e^{\mu(w_B - w_A)}}$$

Softmax rule: identical to correct choice probability predicted by DD process!

$\mu \sim Z/A \sim$  threshold,

$w_B - w_A \sim (A/c)^2 \sim$  SNR.

Reward = input from pathways representing reward stimuli

$$r_* = [r_A \text{ or } r_B]$$

P = linear unit representing midbrain neurons (prediction error). Reinforcement learning by Rescorla-Wagner rule:

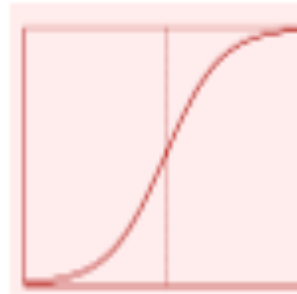
$$\delta = r_* - w_*$$

$$w_* \longleftarrow w_* + (\lambda \delta)$$

[Temporal-Difference Reinforcement Learning (TD RL)]

# A model for choice informed by experience 2

$$P(A) = \frac{1}{1 + e^{-\mu(w_A - w_B)}}$$



**Softmax function**  $P(A)$   
determines next choice; drift  
corresponds to difference in  
expected rewards for A and B

**Reinforcement learning:** update expected rewards as follows:

If A is chosen on the  $n_{\text{th}}$  trial  $w_A(n+1) = (1-\lambda)w_A(n) + \lambda r$ ,  $w_B(n+1) = w_B(n)$ ;

If B is chosen on the  $n_{\text{th}}$  trial  $w_B(n+1) = (1-\lambda)w_B(n) + \lambda r$ ,  $w_A(n+1) = w_A(n)$ ;

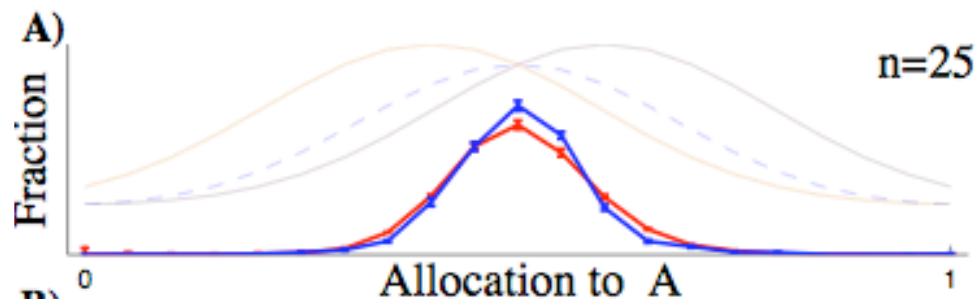
**2 parameters:**  $\mu$  specifies slope of sigmoid:  $\mu \rightarrow \infty =$  deterministic limit.

Learning rate  $\lambda$  specifies a fading memory of previous choices.

$\lambda=0$ : no learning;  $\lambda=1$ : memory of previous choice erased.

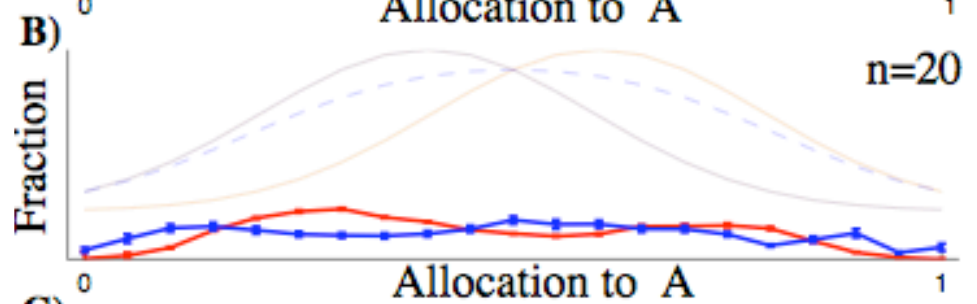
# Model fits to choice allocations without feedback

Conv.  
Gaussian



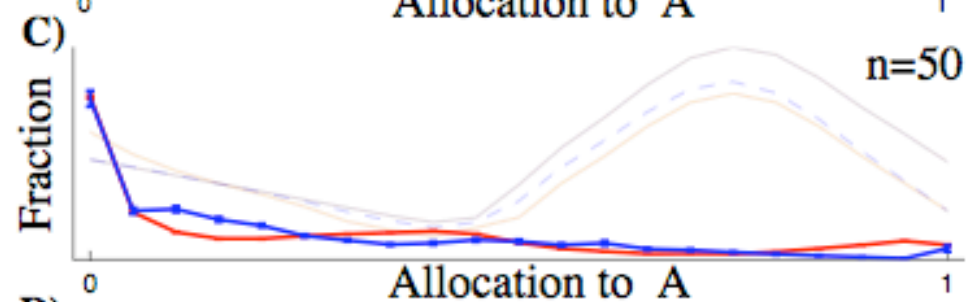
$\lambda = 0.98, \mu = 2.5$   
Fit err = 0.061

Div.  
Gaussian



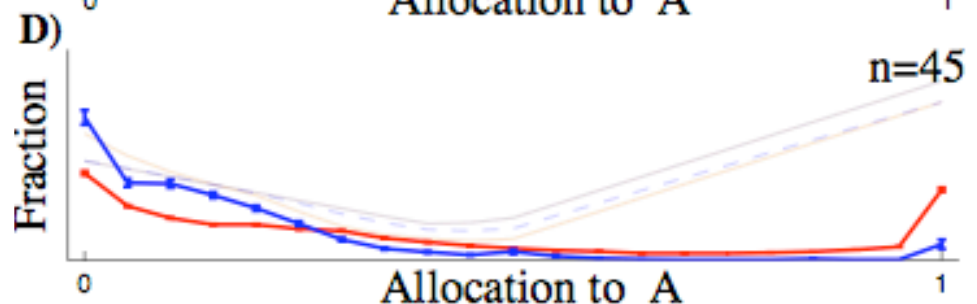
$\lambda = 0.14, \mu = 2.9$   
Fit err = 0.119

Complex  
RO



$\lambda = 0.11, \mu = 11.3$   
Fit err = 0.085

Simple  
RO



$\lambda = 0.06, \mu = 11.0$   
Fit err = 0.188

Fit by minimizing  
mean square error.

Blue: data averages across subjects, with SEs; Red: model fits.

“Group baselines”

# Parameter values tell stories

CG:  $\lambda = 0.98$ ,  $\mu=2.5$

Fit err = 0.061

DG:  $\lambda = 0.14$ ,  $\mu=2.9$

Fit err = 0.119

CRO:  $\lambda = 0.11$ ,  $\mu=11.3$

Fit err = 0.085

SRO:  $\lambda = 0.06$ ,  $\mu=11.0$

Fit err = 0.188

Combine Gaussian tasks

CDG:  $\lambda = 0.91$ ,  $\mu=2.5$

Fit errs = 0.07 & 0.12

$\lambda$  approx 1 for CG => favor current reward information;  $\lambda$  approx 0.1 for ROs => use longer history of recent rewards.

$\mu$  approx 1-2 for CG,DG;  $\mu$  approx 11 for ROs => steeper softmax, more deterministic choices in more complicated tasks.

Same model works for all games: reward schedules play major roles in determining distributions & parameter values.

More on this later.

## II. Propose feedback rules and test them

First consider **choice** feedback: I know what they are doing, but not **how well** they're doing. Follow the **majority** (herding)? Go it alone (contrarian)? Add a **bias to the difference in expected rewards** (argument of softmax  $P(A)$ ):

$$w_A(n) - w_B(n) + \nu_c f(n) \cdot \begin{cases} +1, & \text{if } \# \text{ A's} > \# \text{ B's;} \\ -1, & \text{if } \# \text{ B's} > \# \text{ A's;} \\ 0, & \text{otherwise.} \end{cases} ,$$

where  $f(n) = \pm 1$  if indiv followed or countered majority recently.

Now for **rewards**: I know they're doing better or worse than I am, but just **what** are they doing? Promote **exploration**: bias at random in favor of either choice:

$$w_A(n) - w_b(n) + u(n), \text{ with} \\ u(n) = u(n-1) + \nu_r b(n) |r(n) - \max(\text{all rewards on trial } n)| , \\ \text{where } b(n) \text{ is a binary random variable.}$$

We have have proposed and tried many other feedback rules, mostly with **one social feedback “strength” parameter** (e.g., modify slope  $\mu$ ). Test them by having the model predict each subject's choice sequences, given feedback from other group members.

*\* Here's where fMRI neuroimaging results help! \**

A. Nedic et al., Proc 47th IEEE-CDC 2008.

# A maximum likelihood test

To assess the ability of competing models to fit binary choice data, we compare individual's choice sequences  $\{d(n)\} = \text{AABABBBAA}$  ... with model predictions, given priors  $\{d(j)\}_{j=0}^{n-1}$  on individual and (in case of feedback) on other group members:

$$L(d|p) = \prod_{\{n|d(n)=A\}} P_n(A) \times \prod_{\{n|d(n)=B\}} [1 - P_n(A)]$$
$$\text{Avg } L(d|p) = \exp \left\{ \frac{1}{K} \left[ \sum_{\{n|d(n)=A\}} \log P_n(A) + \sum_{\{n|d(n)=B\}} \log [1 - P_n(A)] \right] \right\}.$$

Average likelihood values are interpreted as follows: 1 implies perfect prediction, 0.5 implies prediction at chance level (e.g. if  $P(A) = 0.5$ ), and 0.0 implies perfect **anti**-prediction.

L. Corrado et al., *J. Exp. Anal. Behav.* 84, 2005.

Akaike information criterion used to compare quality of fit for models with different numbers of parameters.

H. Akaike *J. Econometrics* 16, 1981.

# Social feedback model fits

Choice feedback: boost  $w_A - w_B$  toward majority

$+\nu_c f(n)g(n)$ , where  $f(n) = \pm 1$  or 0 if followed majority or not,  $g(n) = \pm 1$  if  $\#A's > \#B's$  or not, and  $f(n), g(n) = 0$  if no majorities.

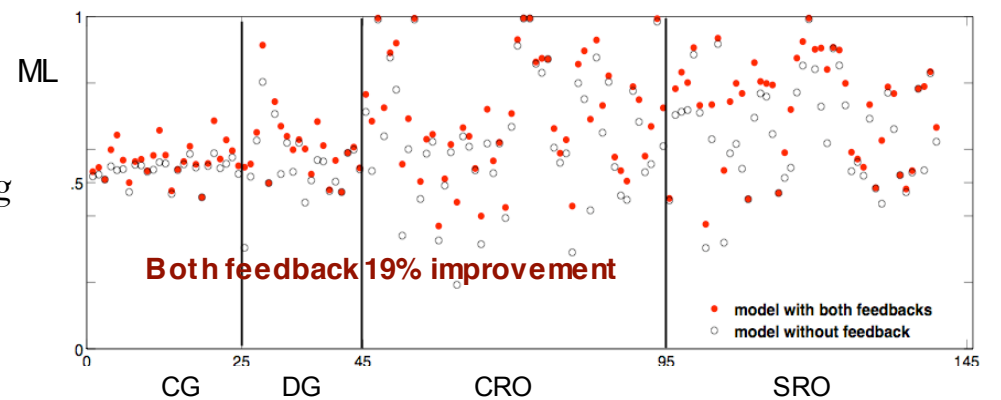
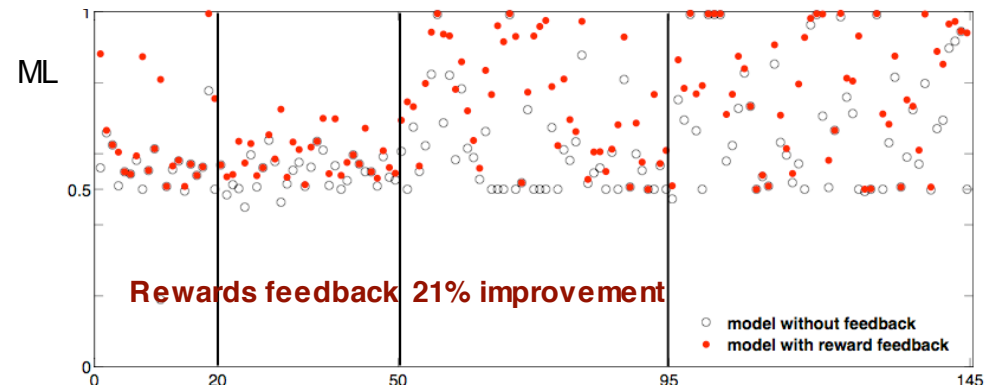
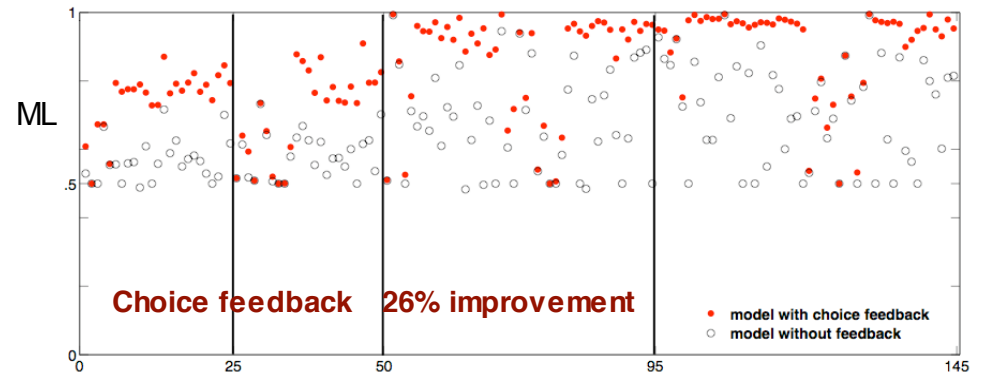
Rewards feedback: bias  $P(\text{choose A})$  randomly

$+\nu_r b(n) |r(n) - \max(\text{all rewards on trial } n)|$   
where  $b(n)$  is a binary random variable.

Assess via max. likelihood of model, given the probability predicted for each choice conditioned on feedback from rest of group.

Including feedback yields significant improvement in model predictions, notably for choice feedback in in tasks CRO, SRO.

Figures show ML values for fits to all subjects playing each game under appropriate feedback condition, ML = 1: perfect prediction; ML = 0.5: chance level; ML = 0: perfect anti-prediction. Vertical bars divide games as follows: CG | DG | CRO | SRO.

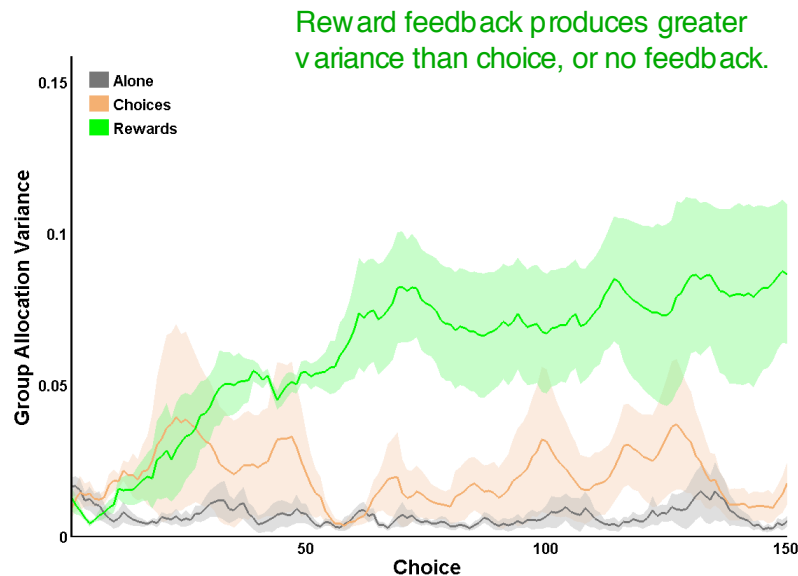
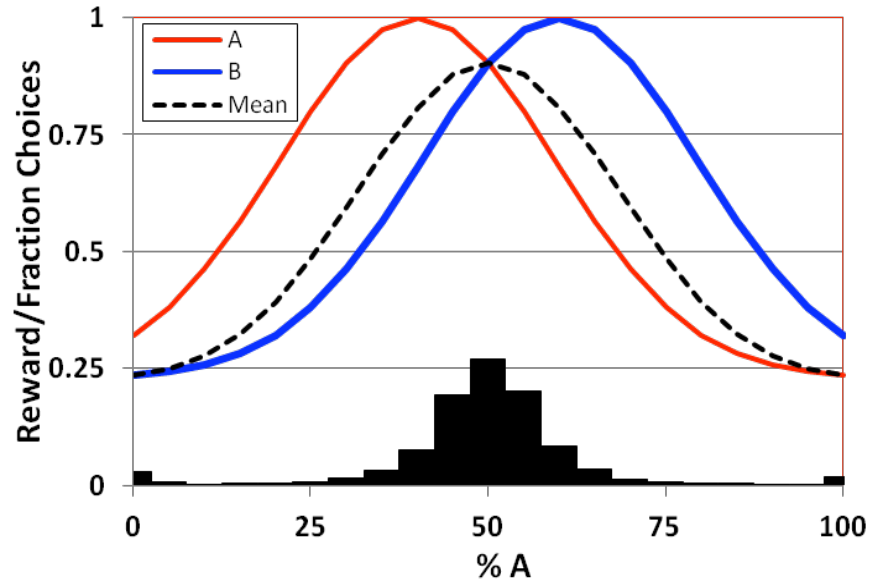


# III. Samples of new results

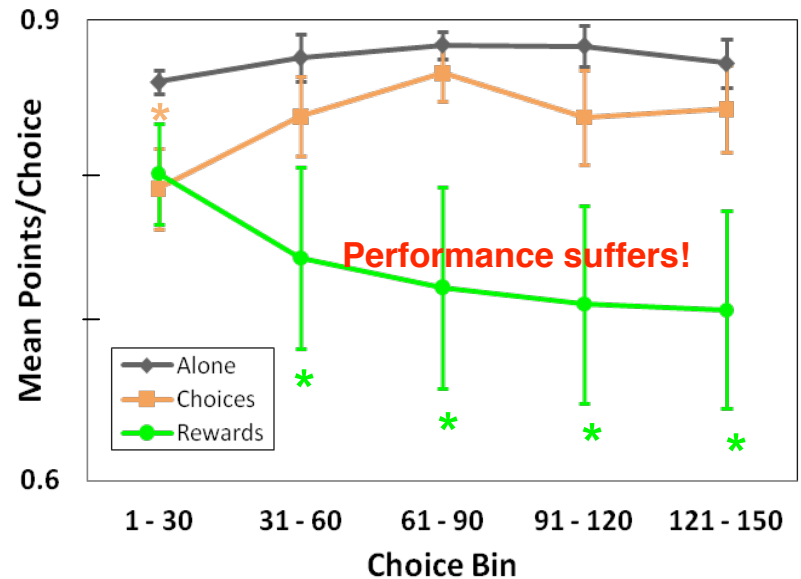
## Reward feedback can hurt you!

In the converging Gaussian game, with stable matching point at max reward, information on other group members' rewards causes greater exploration, causing **progressive decline in one's own rewards**.

On rising optimum games with multiple maxima, reward feedback can allow more subjects to explore and thereby discover global maxima (not shown here).

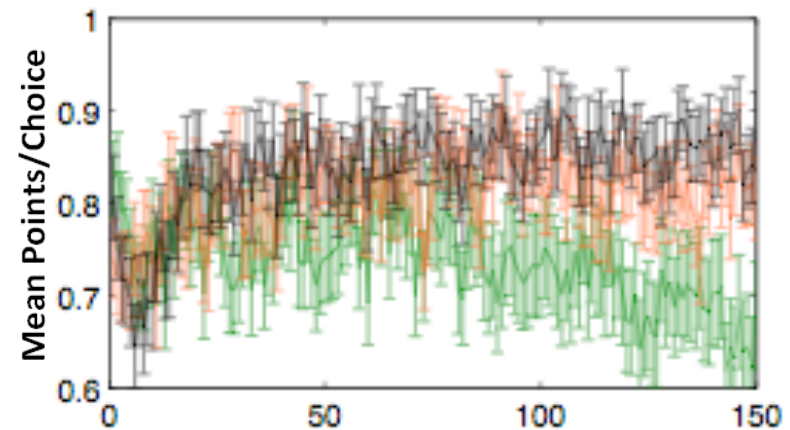
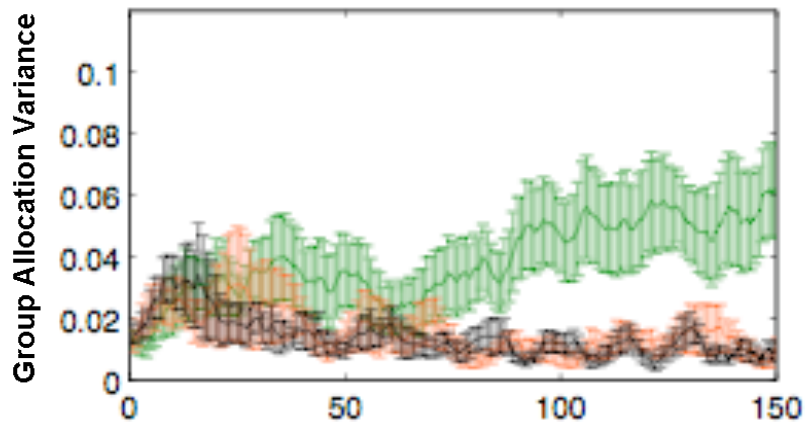
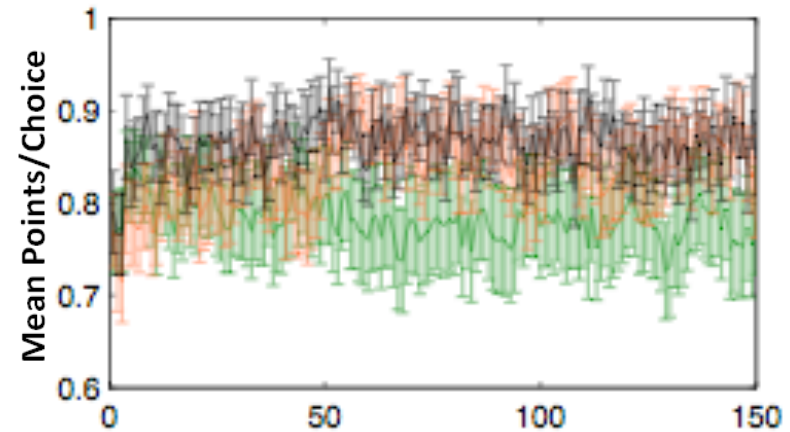
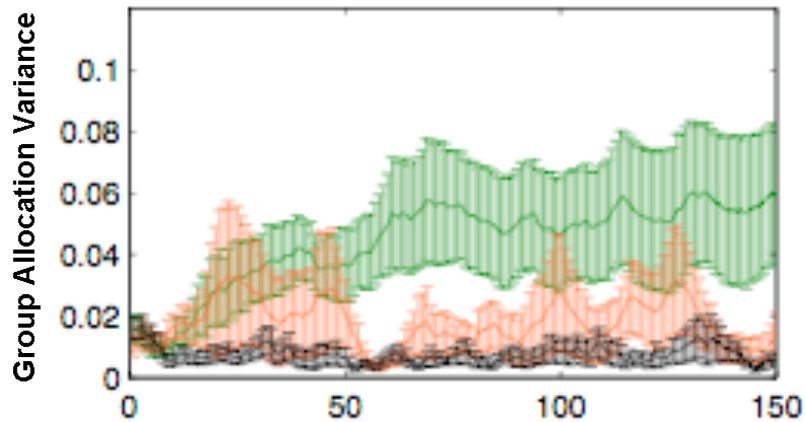


Reward feedback produces greater variance than choice, or no feedback.



# The model reproduces this group behavior

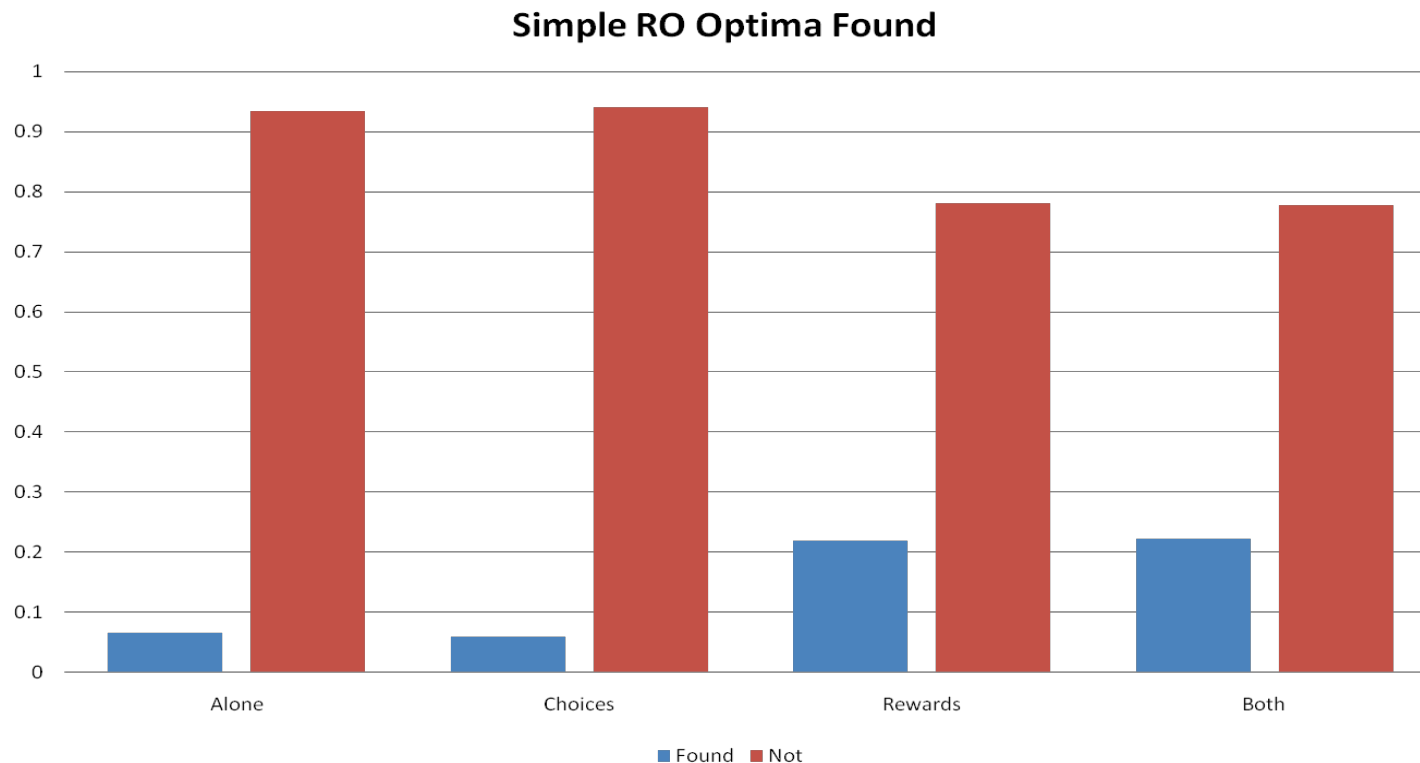
Reward feedback data above, model results below.



Green: rewards feedback; orange: choice feedback; black: alone (no feedback).

# Rewards feedback helps SRO performance

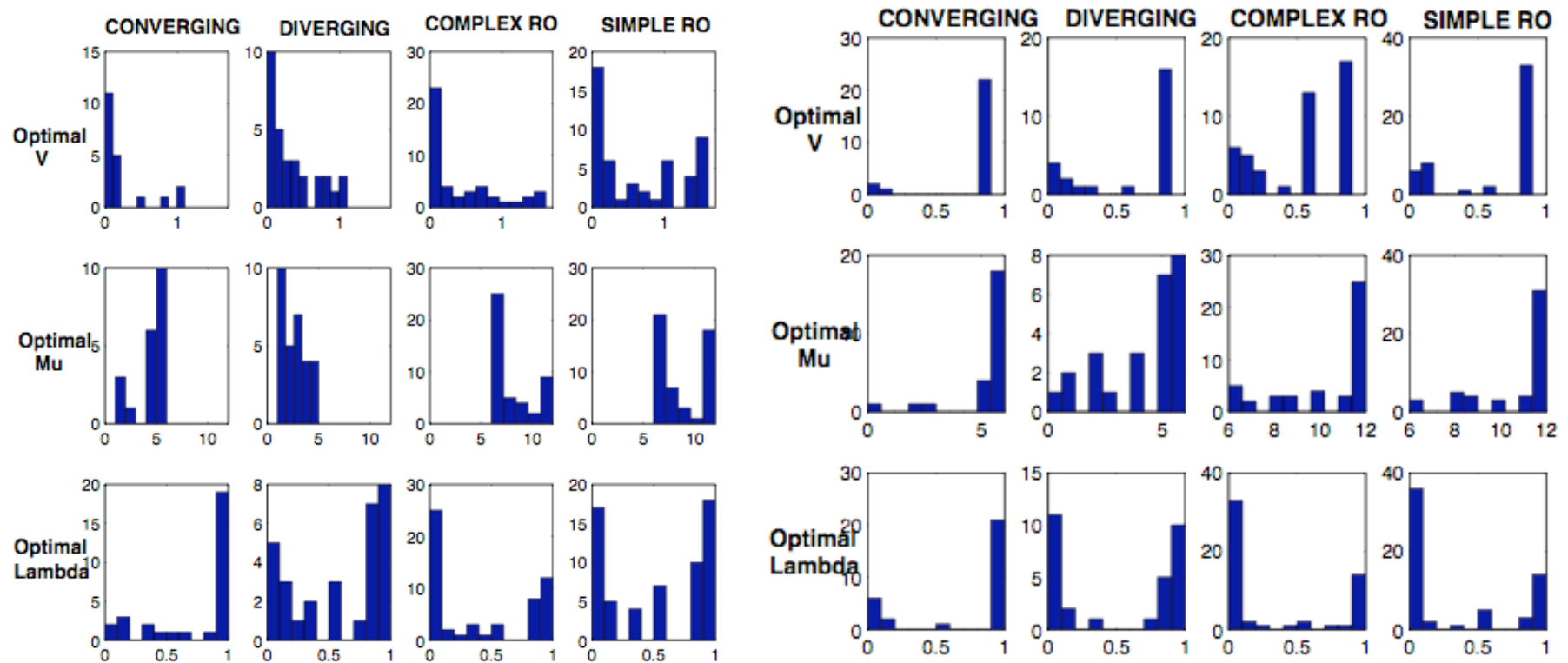
Under rewards and choice & rewards conditions subjects are more likely to discover the global optimum, although their rewards are no higher than those who remain at the local maximum.



Exploration is good, but too much exploration can hurt.

# The model reveals individual differences 1

Histograms of individual parameter values under reward (left) and choice (right) feedback. Note low/high  $\mu$ 's for 2 kinds of games. Do bimodal  $\mu$  and  $\lambda$ 's  $\Rightarrow$  2 kinds of folks?



Parameter fits to individuals reveals systematic differences among tasks and feedback conditions. Plan to use these to further analyze neuroimaging data.

# The model reveals individual differences 2

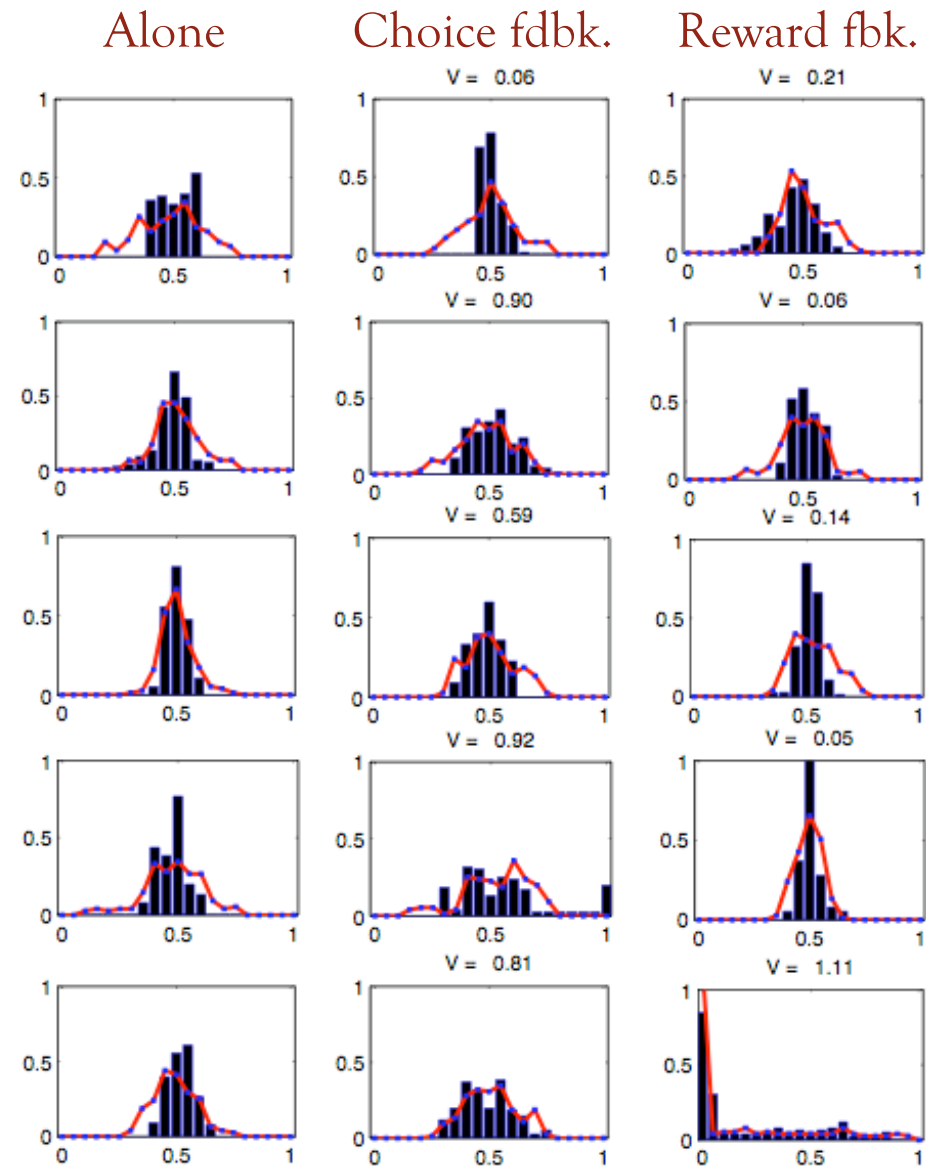
histograms: data;  
red curves: model.

Histograms of individual choice allocations within groups. CG game, with and without social feedback.

Allocation distributions reveal that behaviors are more variable than in alone condition.

Social interaction parameters  $\nu_C, \nu_R$  span the range  $[0,1]$ .

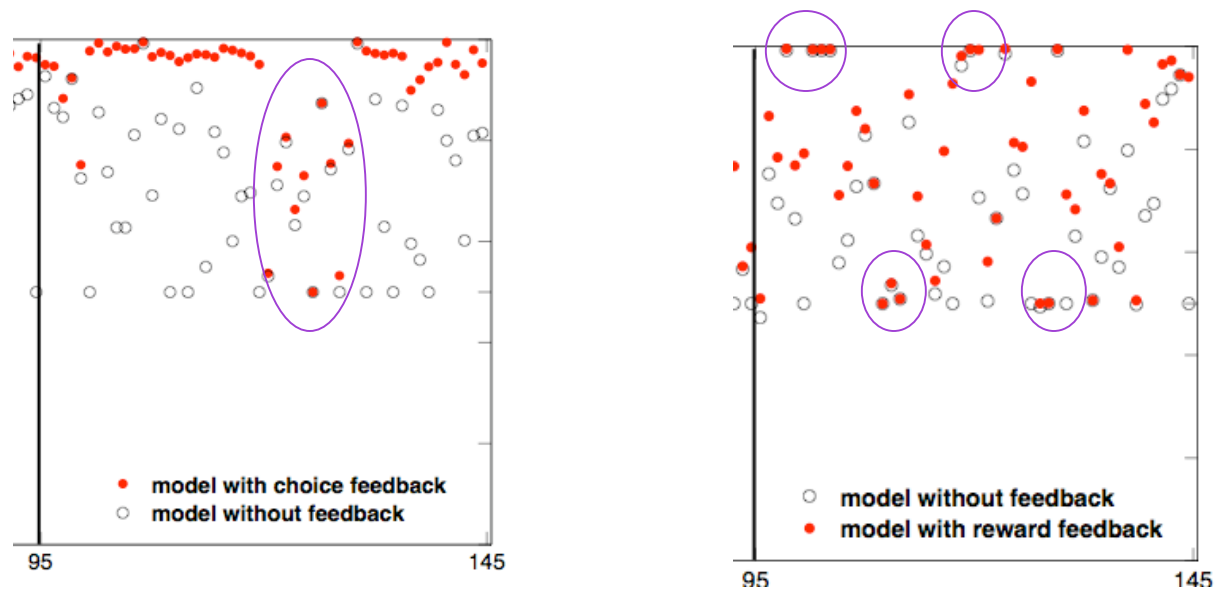
Model captures outlier subject.



Nedic et al., *in prep.*, 2009.

# The model distinguishes group and individual differences

Model fit comparisons suggest different individual responses to feedback. Is lack of improvement on adding  $\nu_C, \nu_T$  due to ignoring feedback? Are some subjects overwhelmed by info? E.g. (SRO, CRO):



Examine **brain images** to seek differential activity (or lack thereof) in relevant cortical areas responsive to reward expectation, social feedback (PFC, DLPFC, OFC, putamen, cingulate, ... regions typically more active in feedback conditions).

# Summary

- Neural activity in simple decisions resembles a DD/OU process.
- We generalize a TD RL model for choice probabilities based on recent rewards in a two-armed bandit task to study collective and individual behaviors in groups receiving information on other players' choices, rewards, and both.
- Model fits are reasonably good: feedback models provide significant improvements in predicting choices, highlight individual differences.
- Data and model show that reward feedback can degrade performance, but can also promote exploration.
- Trying to understand what models reveal on individual differences.
- Currently trying to use model as regressor for neuro-imaging results.

Good mathematical models are not just (reasonably) **faithful**; they're also **simple** (but not too simple), and (approximately) **soluble**. They focus attention, help to sharpen questions.

Thanks for your attention!