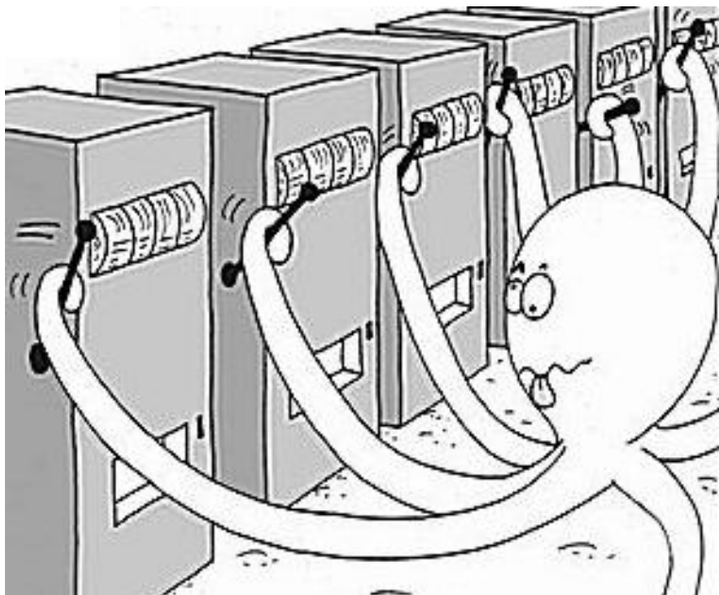
Learning and Decision Making under Uncertainty: From Basic Science to Psychiatry and Back Again

Angela Yu

Cognitive Science &
Halıcıoğlu Data Science Institute
UC San Diego



Learning & Decision-Making under Uncertainty

Learning & Decision-Making under Uncertainty

Repeated Decisions ⇒ Stochastic Outcomes

Learning & Decision-Making under Uncertainty

Repeated Decisions ⇒ Stochastic Outcomes

Restaurants, research projects, dating

Learning & Decision-Making under Uncertainty

Repeated Decisions ⇒ Stochastic Outcomes

Restaurants, research projects, dating

Each choice ⇒ immediate return + information gain

Outline

Outline

- **Computational modeling**

Outline

- **Computational modeling**
- **Applications to psychiatry**

Outline

- **Computational modeling**
- **Applications to psychiatry**
  - ✳ **Addiction**

Outline

- **Computational modeling**
- **Applications to psychiatry**
  - ✳ **Addiction**
  - ✳ **Depression/anxiety**

Outline

- **Computational modeling**
- **Applications to psychiatry**
  - ✳ **Addiction**
  - ✳ **Depression/anxiety**
- **Discussion**

Outline

- **Computational modeling**
- **Applications to psychiatry**
  - ✳ **Addiction**
  - ✳ **Depression/anxiety**
- **Discussion**

Conceptual Framework

Learning
(Information Processing)

- inferring state of the world
- predicting outcomes
  (short-term & long-term)
- learning environmental statistics

Conceptual Framework

Learning
(Information Processing)

Decision Making
(Action Selection)

- inferring state of the world
- predicting outcomes
  (short-term & long-term)
- learning environmental statistics

Conceptual Framework

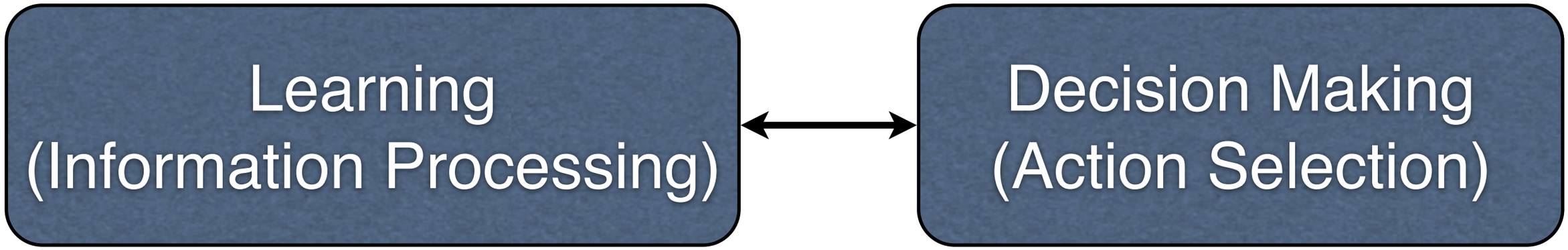
Learning (Information Processing)

- inferring state of the world
- predicting outcomes (short-term & long-term)
- learning environmental statistics

Decision Making (Action Selection)

- deciding among options based on incomplete/noisy information
- taking into account context/goal

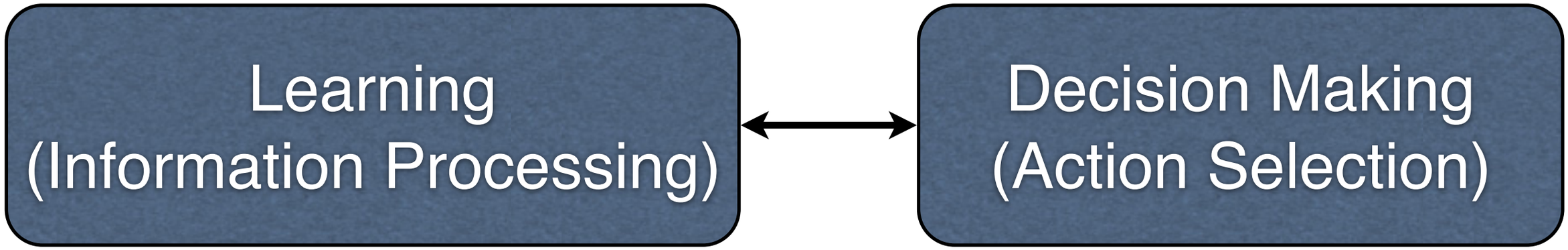
Conceptual Framework



- inferring state of the world
- predicting outcomes (short-term & long-term)
- learning environmental statistics

- deciding among options based on incomplete/noisy information
- taking into account context/goal

Conceptual Framework

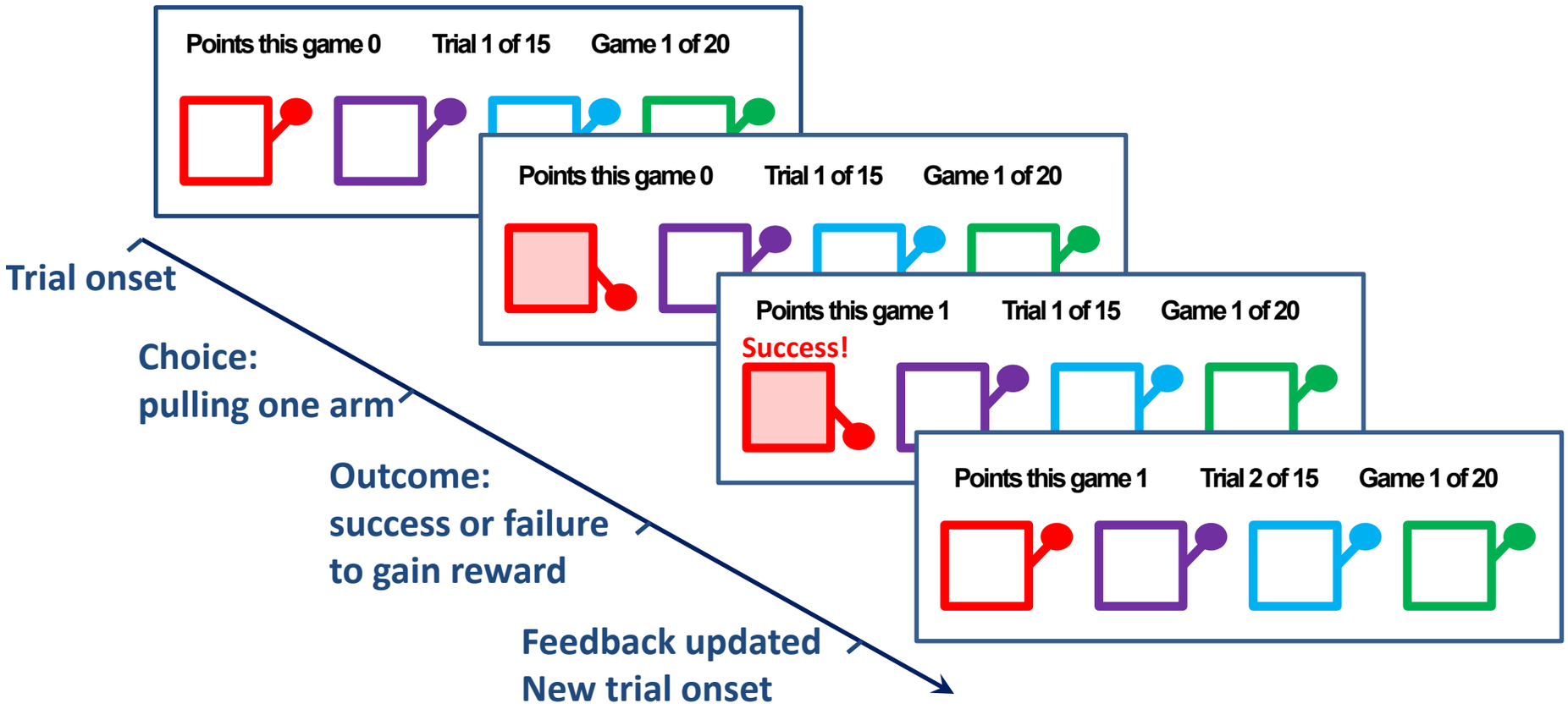


- inferring state of the world
- predicting outcomes (short-term & long-term)
- learning environmental statistics

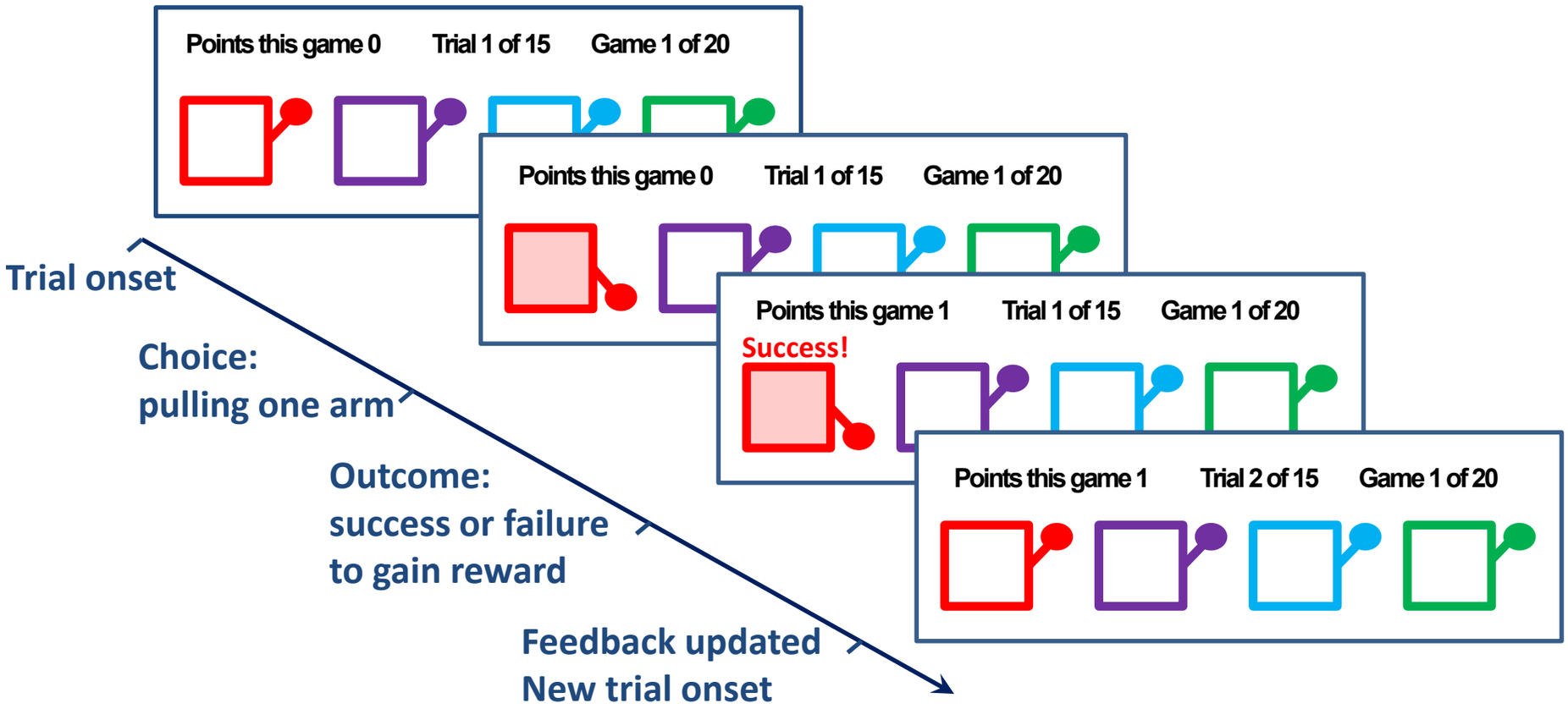
- deciding among options based on incomplete/noisy information
- taking into account context/goal

Canonical experimental paradigm: multi-armed bandit task

Multi-Arm Bandit Task

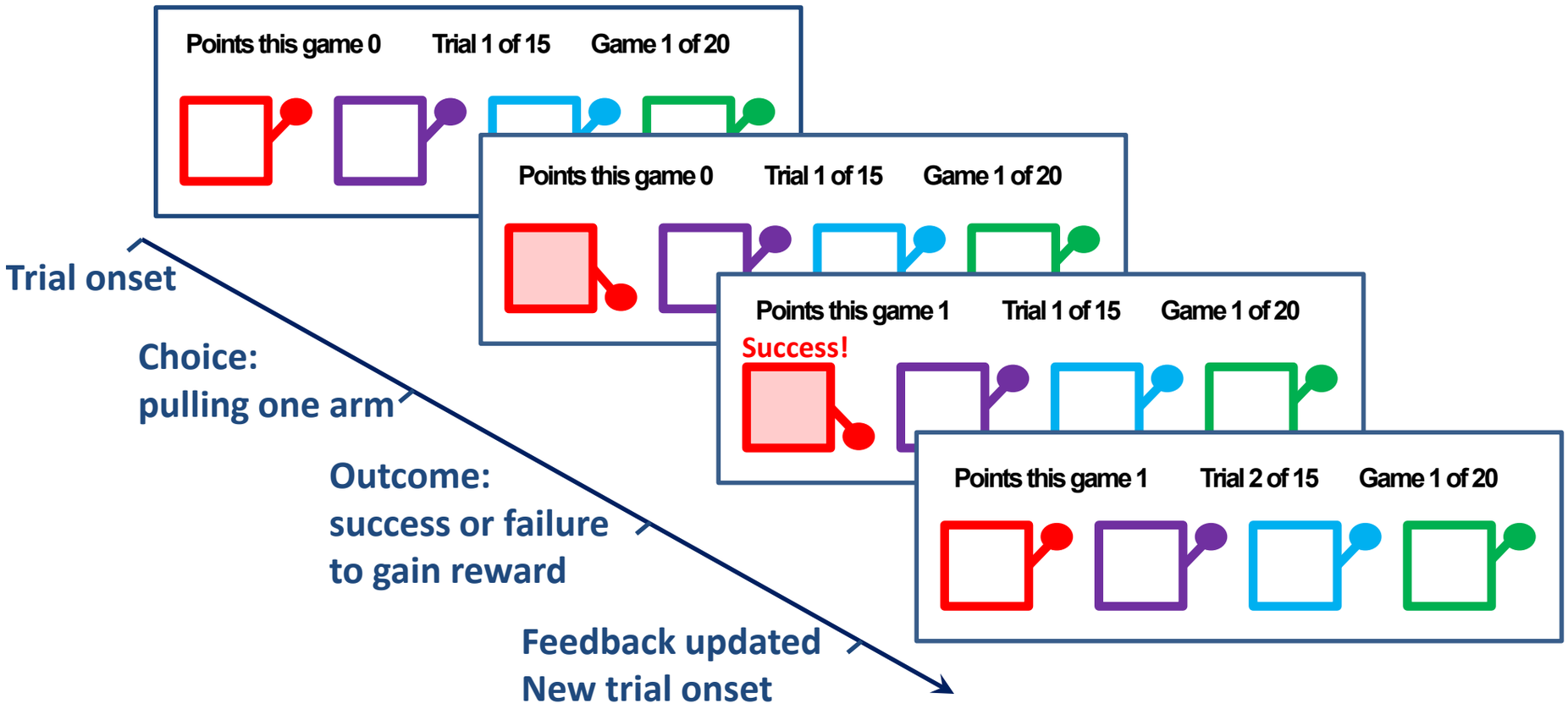


Multi-Arm Bandit Task



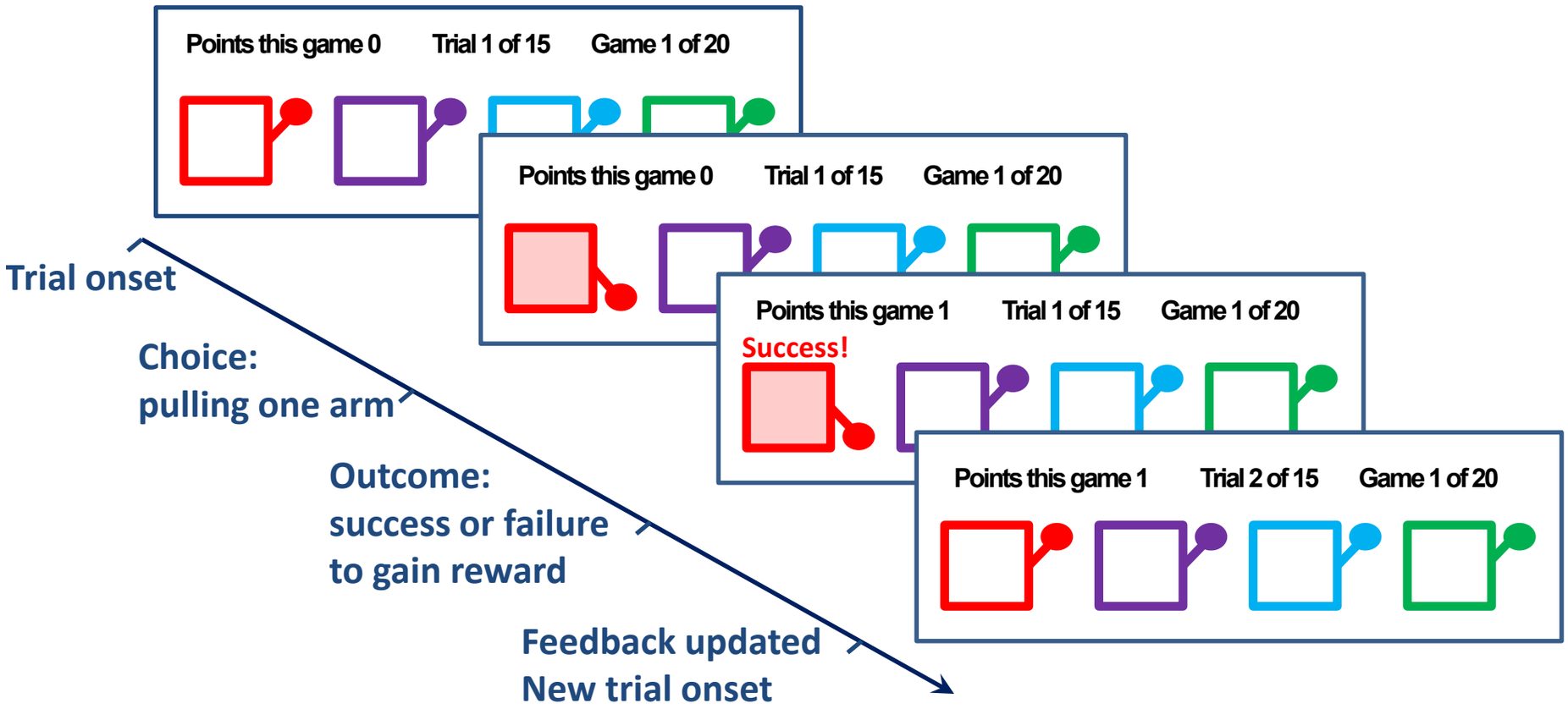
- only see chosen option outcome

Multi-Arm Bandit Task



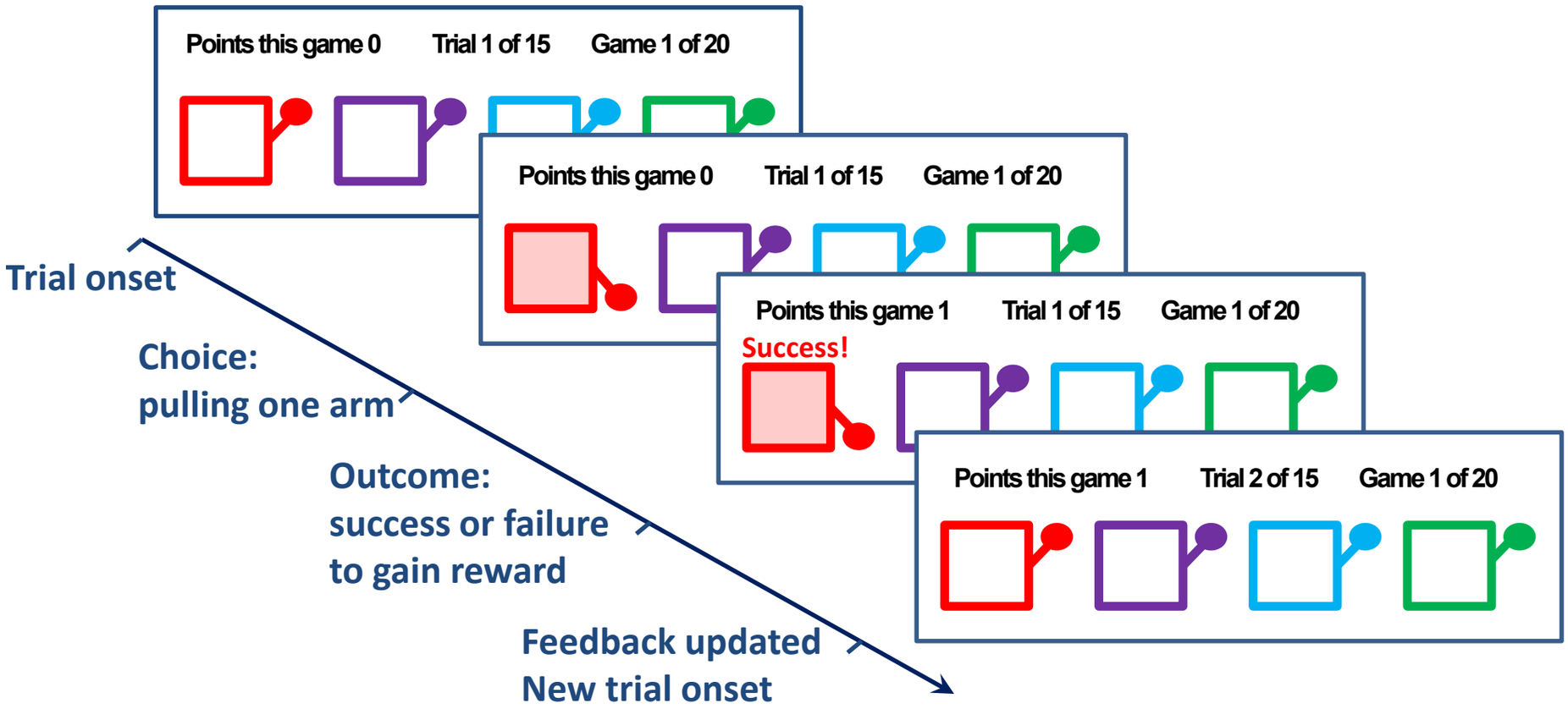
- only see chosen option outcome
- **exploitation:** choose option with highest estimated reward rate

Multi-Arm Bandit Task



- only see chosen option outcome
- **exploitation:** choose option with highest estimated reward rate
- **exploration:** explore unknown options (information gain)

Multi-Arm Bandit Task



- only see chosen option outcome
- **exploitation:** choose option with highest estimated reward rate
- **exploration:** explore unknown options (information gain)
- need to balance between exploitation and exploration

Experimental Design

Experimental Design

- 4-armed bandit task (n=107)

Experimental Design

- 4-armed bandit task (n=107)
- subjects paid according to total “rewards”

Experimental Design

- 4-armed bandit task (n=107)
- subjects paid according to total “rewards”
- binary rewards, 50 games, 15 trials/game

Experimental Design

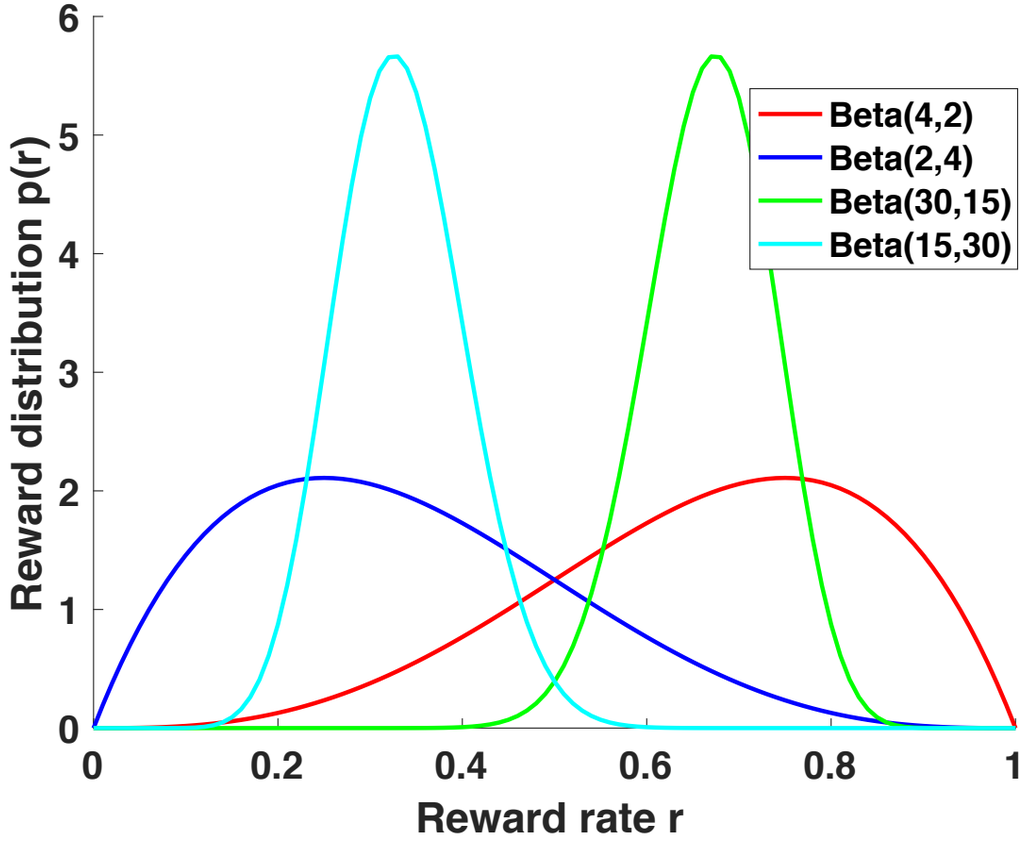
- 4-armed bandit task (n=107)
- subjects paid according to total “rewards”
- binary rewards, 50 games, 15 trials/game
- 4 conditions: mean x variance

	High μ	Low μ
High σ	Beta(4, 2)	Beta(2, 4)
Low σ	Beta(30, 15)	Beta(15, 30)

Experimental Design

- 4-armed bandit task (n=107)
- subjects paid according to total “rewards”
- binary rewards, 50 games, 15 trials/game
- 4 conditions: mean x variance

	High μ	Low μ
High σ	Beta(4, 2)	Beta(2, 4)
Low σ	Beta(30, 15)	Beta(15, 30)

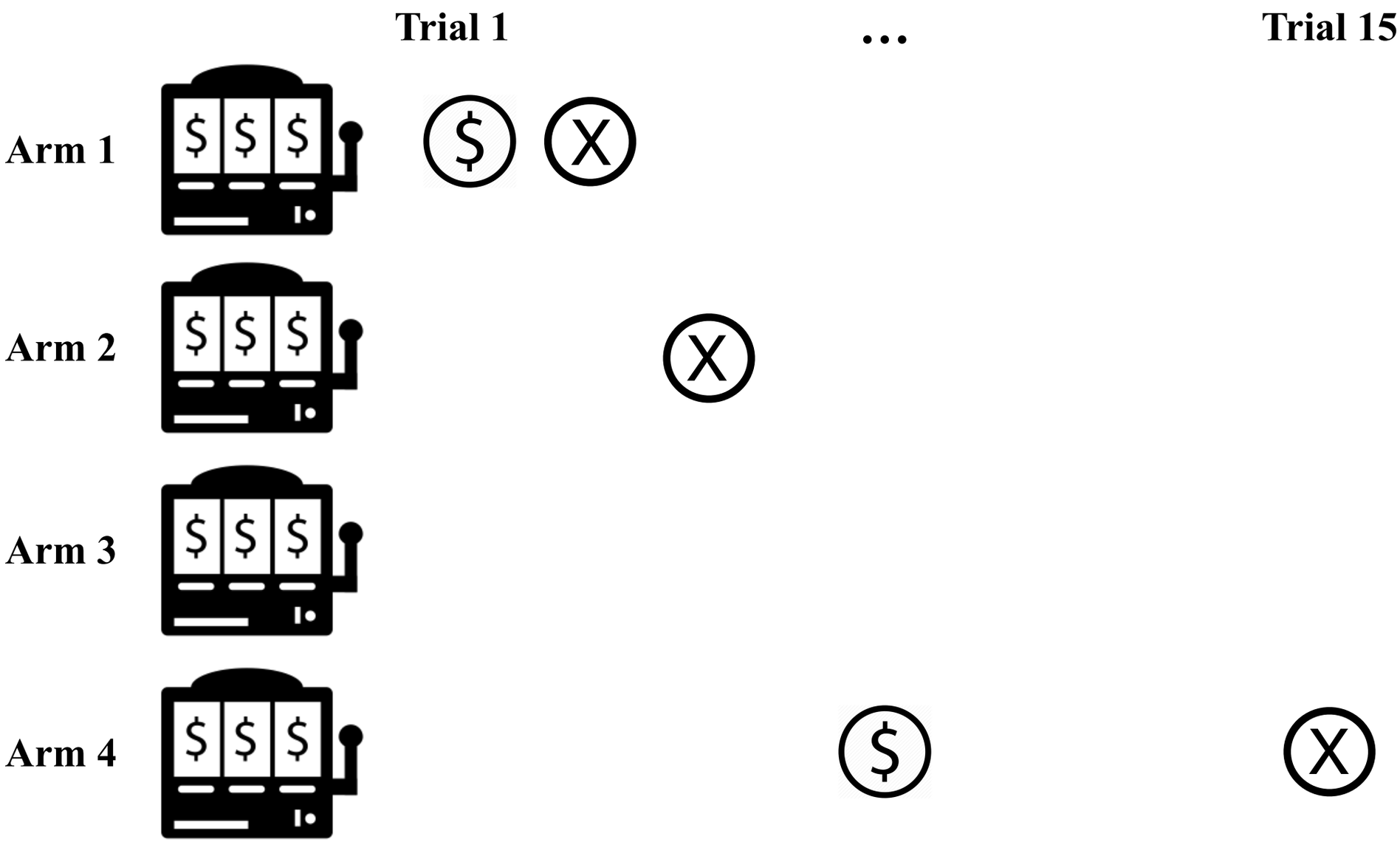


Accurate Representation of Prior?

Subjects report E[reward] for unseen arms

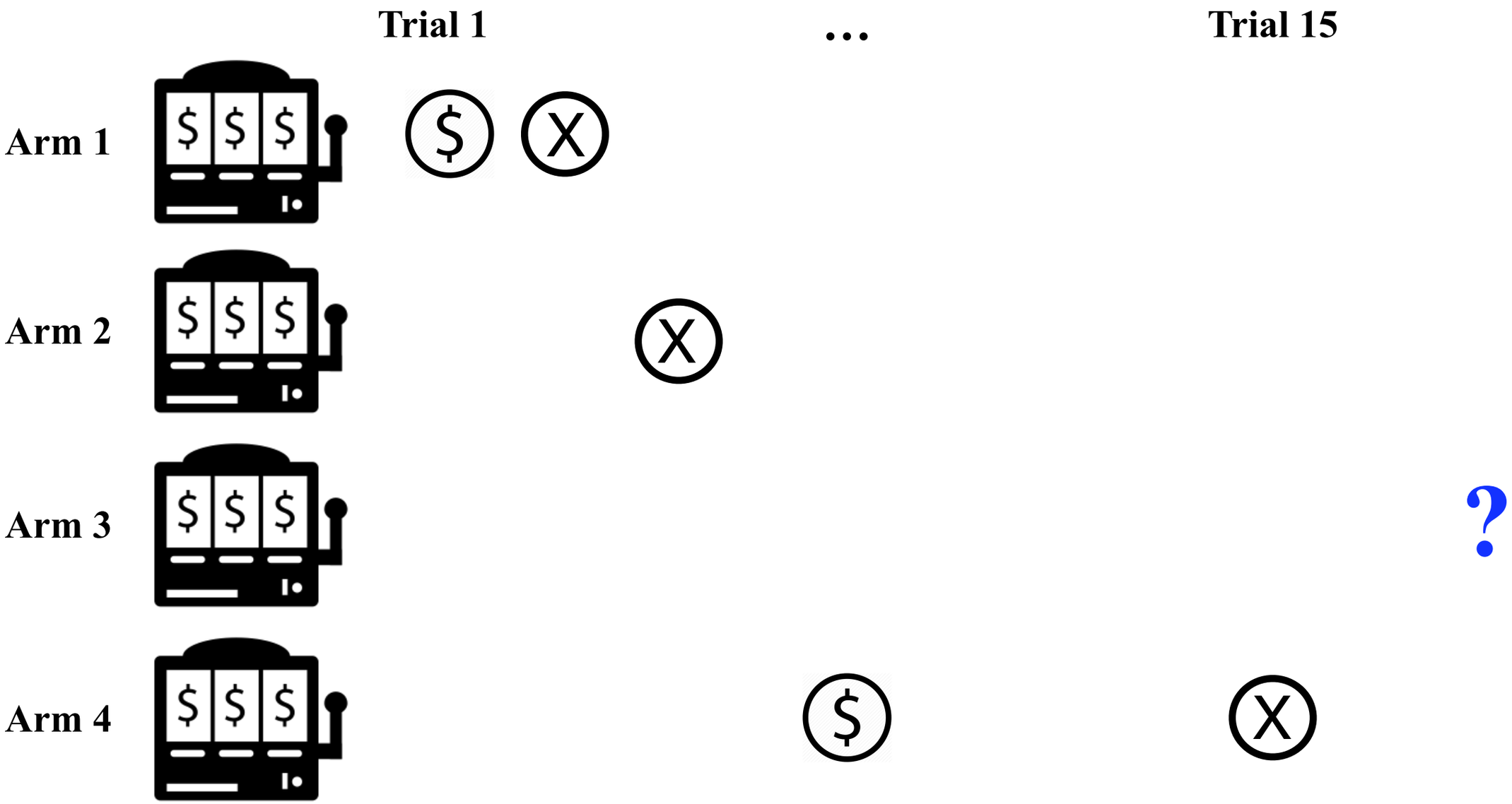
Accurate Representation of Prior?

Subjects report E[reward] for unseen arms



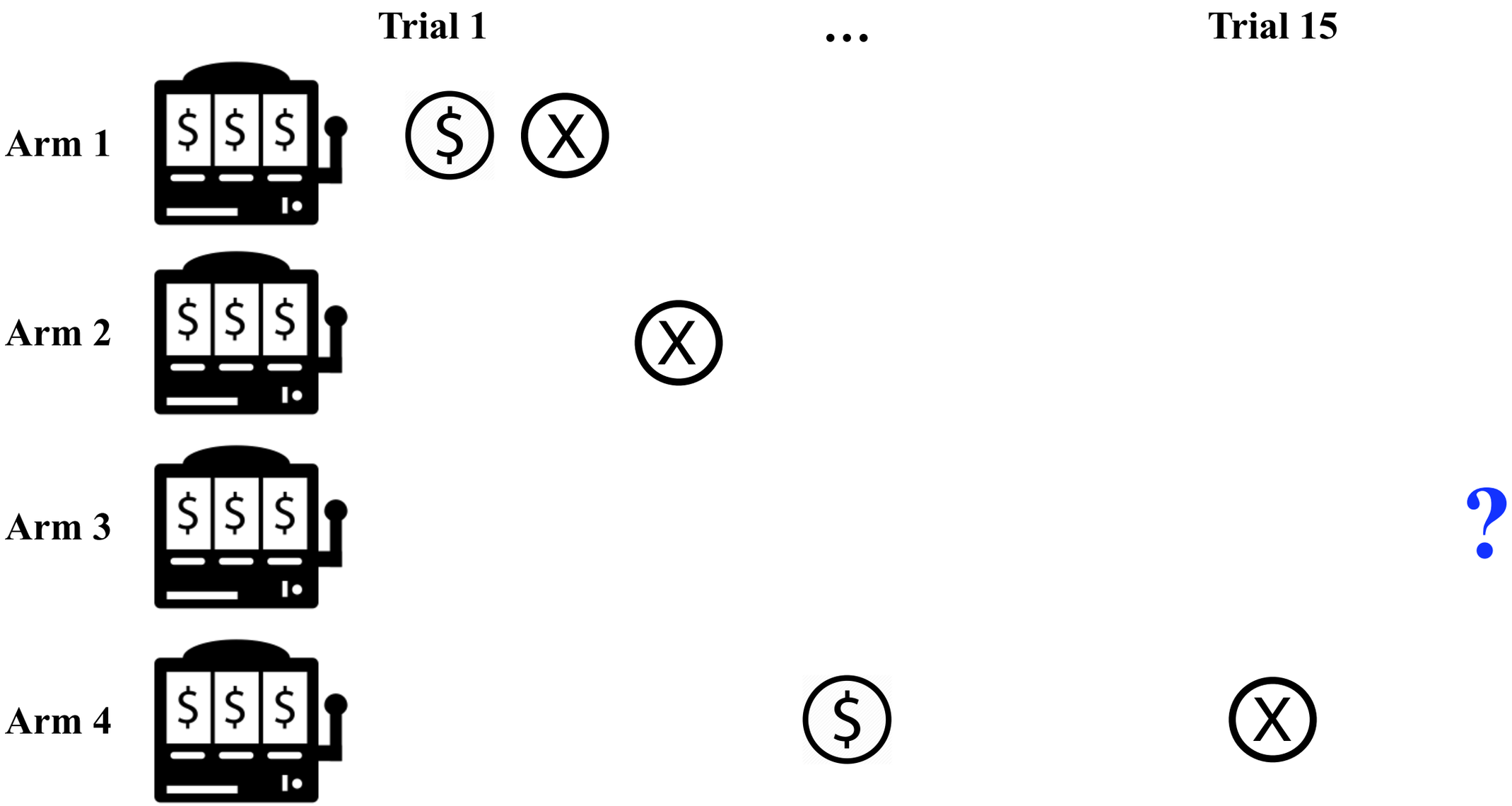
Accurate Representation of Prior?

Subjects report E[reward] for unseen arms



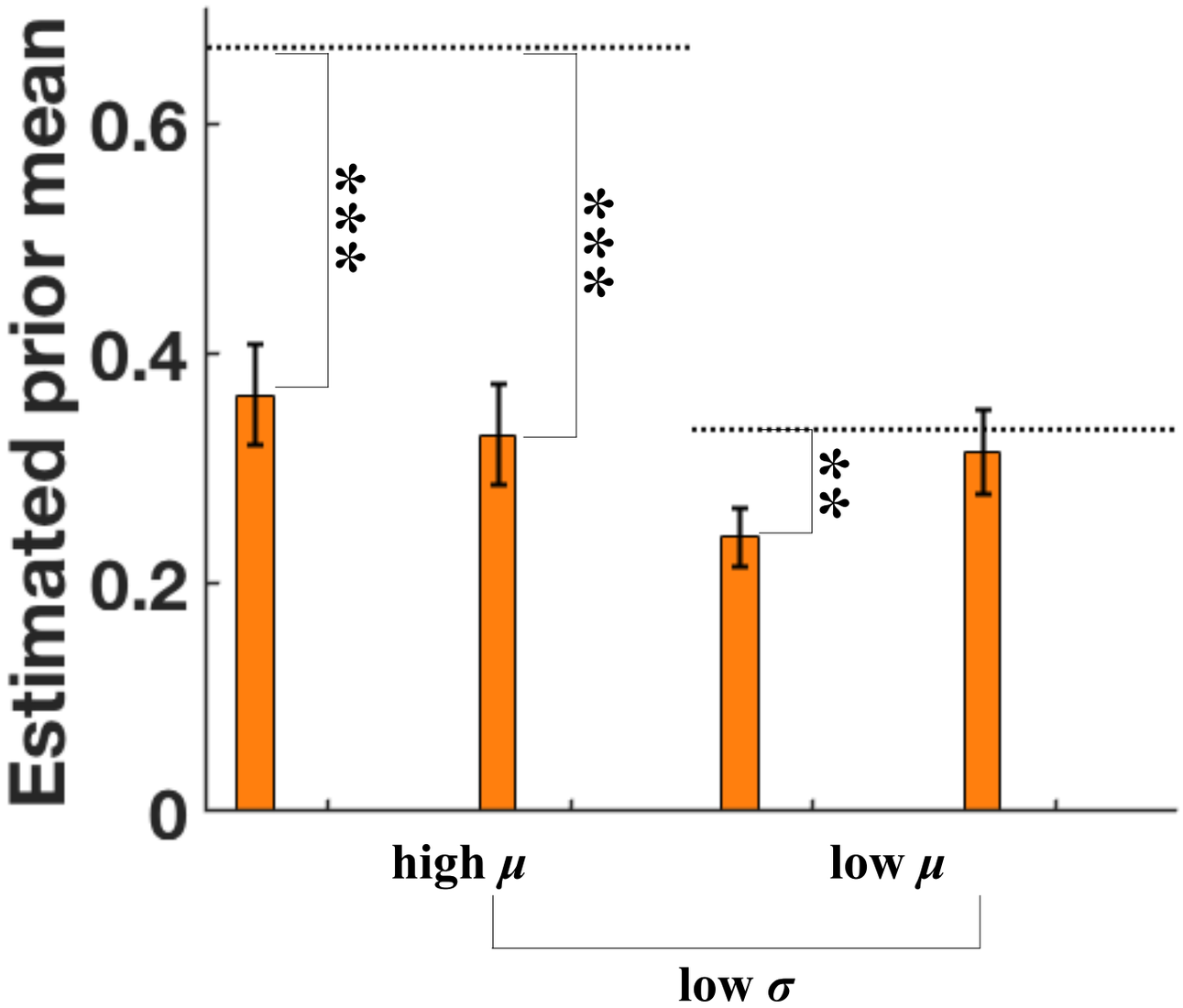
Accurate Representation of Prior?

Subjects report E[reward] for unseen arms

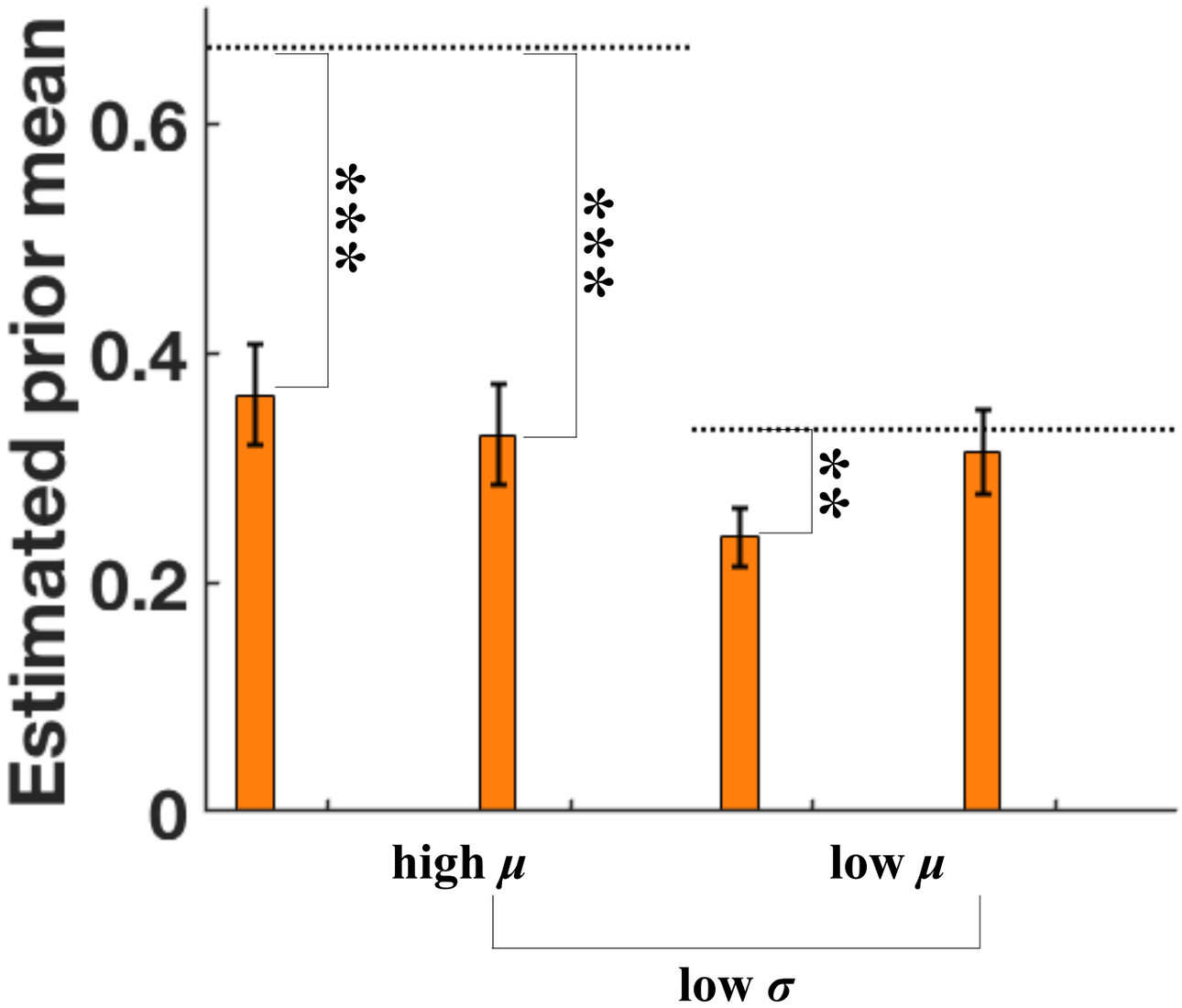


*"You did not choose arm **3** in this game. If you were to choose it 10 times, how many fish would you expect to catch?"*

Self-Report: Reward Underestimation

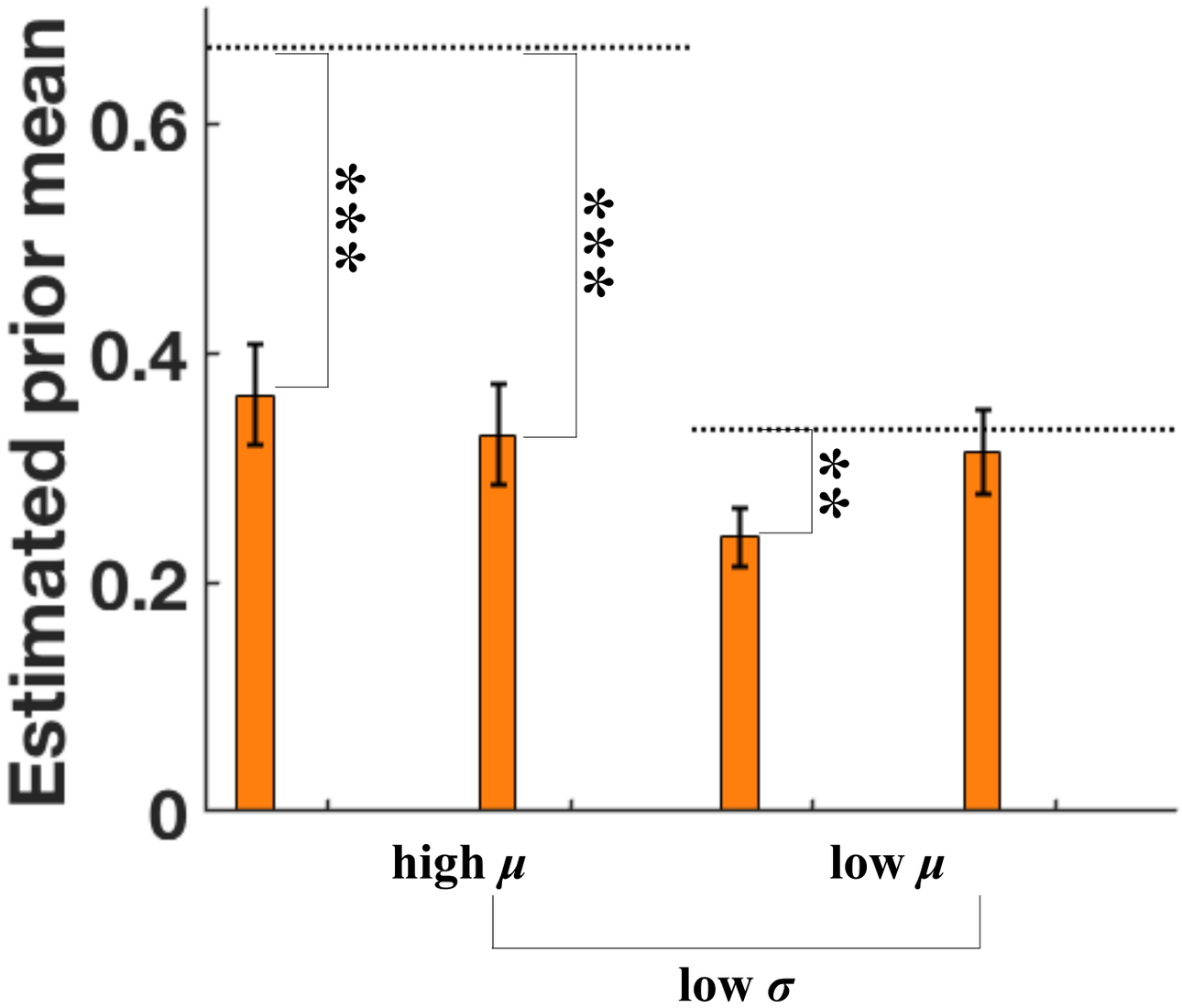


Self-Report: Reward Underestimation



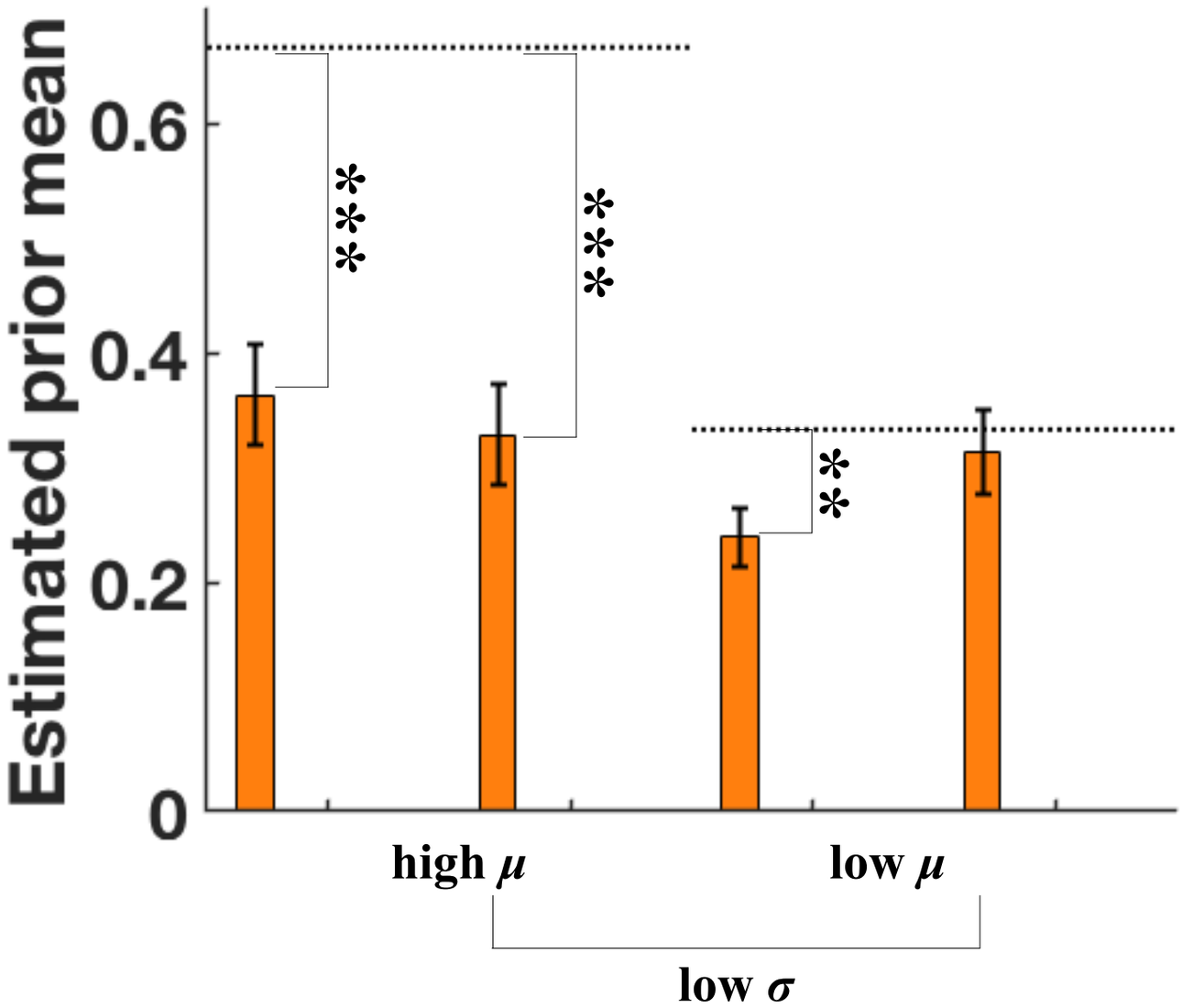
- Subjects underestimate E[reward] (especially *abundant* env.)

Self-Report: Reward Underestimation



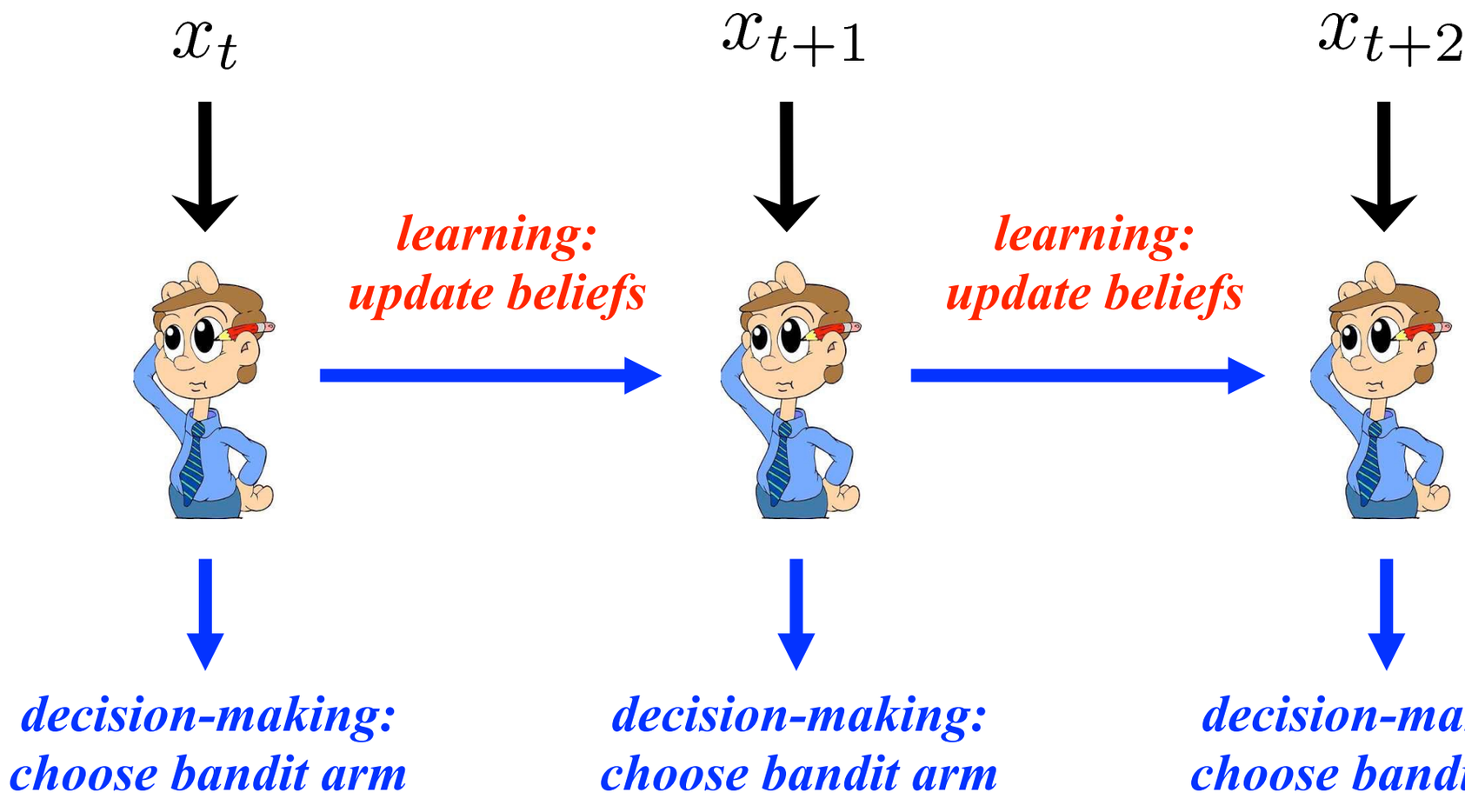
- Subjects underestimate E[reward] (especially *abundant* env.)
- Is this real? Or *selection bias? Confirmation bias?*

Self-Report: Reward Underestimation



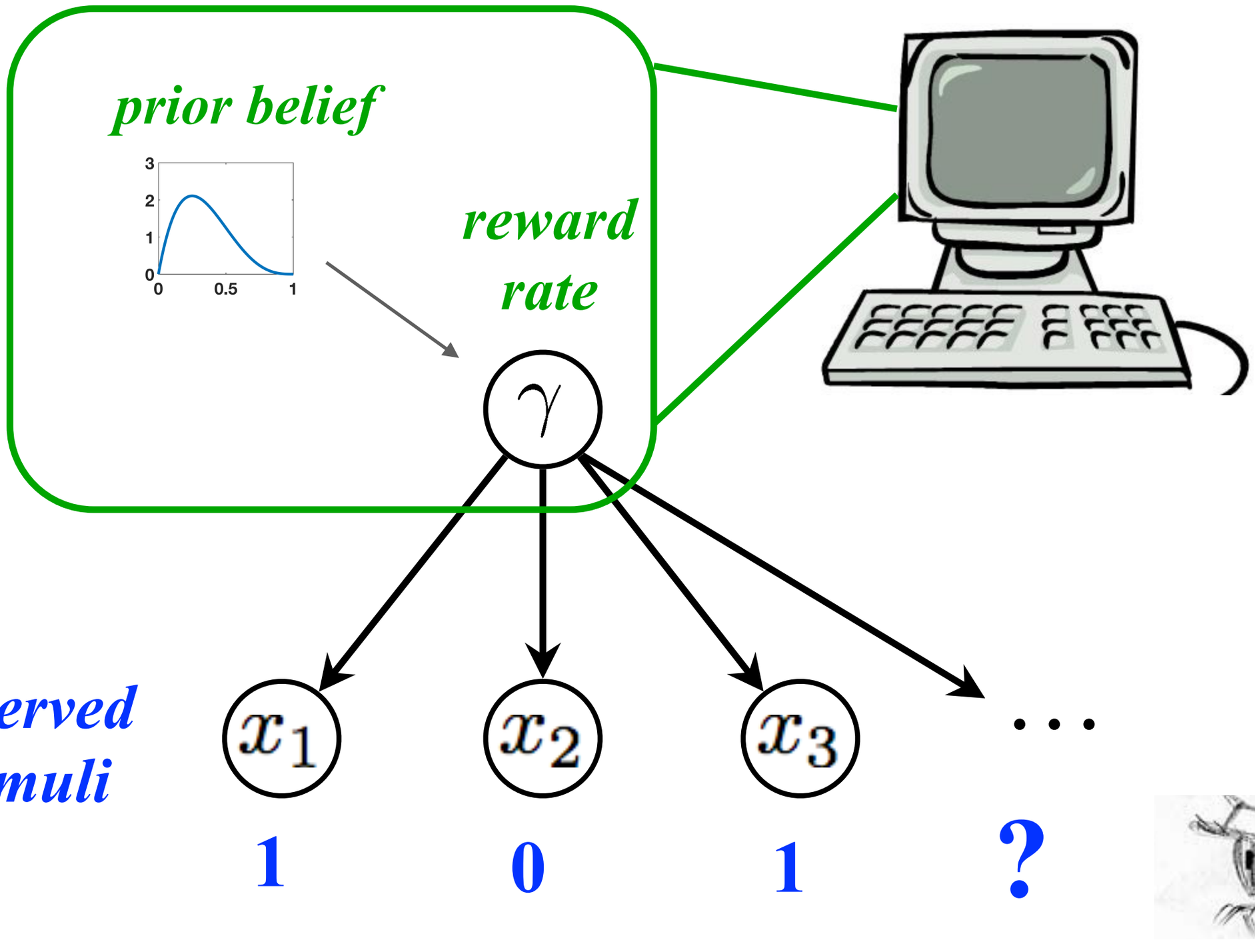
- Subjects underestimate E[reward] (especially *abundant* env.)
- Is this real? Or *selection bias? Confirmation bias?*
- Alternative approach: model behavior ⇒ still under-estimation?

Sequential Learning & Decision-Making

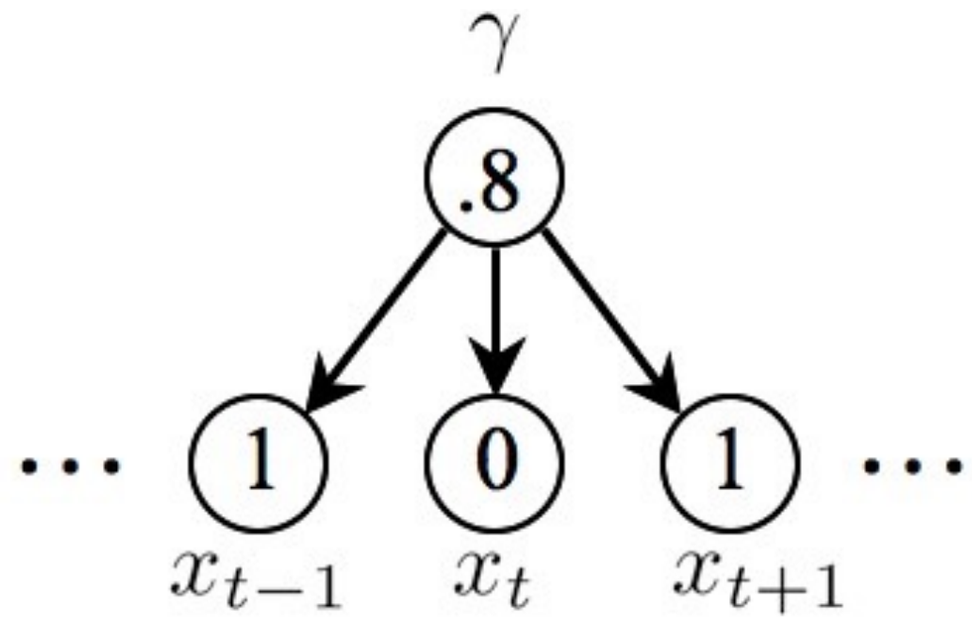


Learning Model I: Fixed Belief Model (FBM)

(Yu & Cohen, *NIPS*, 2009; Zhang & Yu, *NIPS,* 2013)



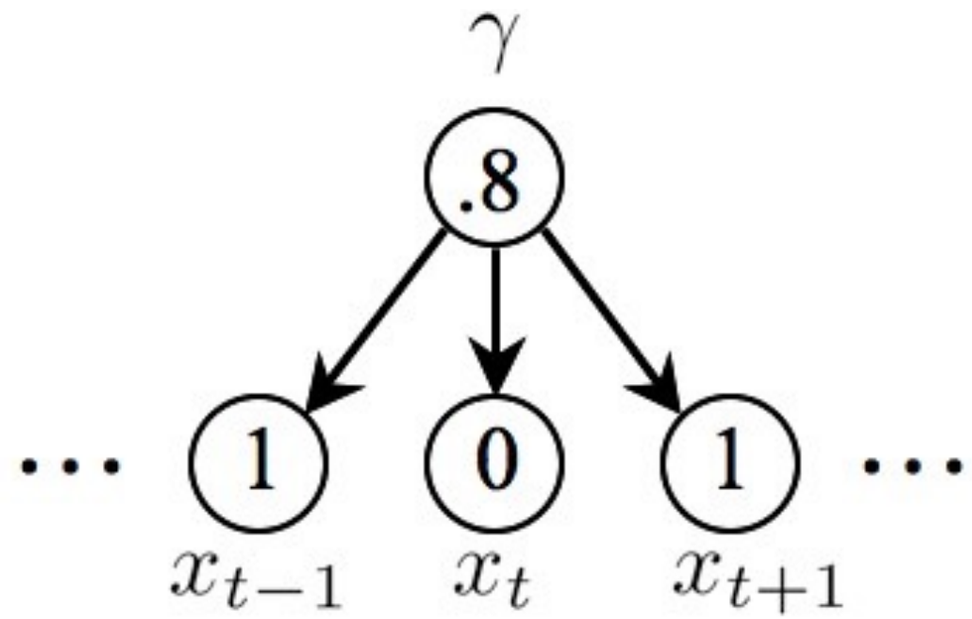
Bayesian Inference



Generative Model
(what subject "knows")

$$P(x_t|\gamma) = Bernoulli(x_t;\gamma)$$

Bayesian Inference

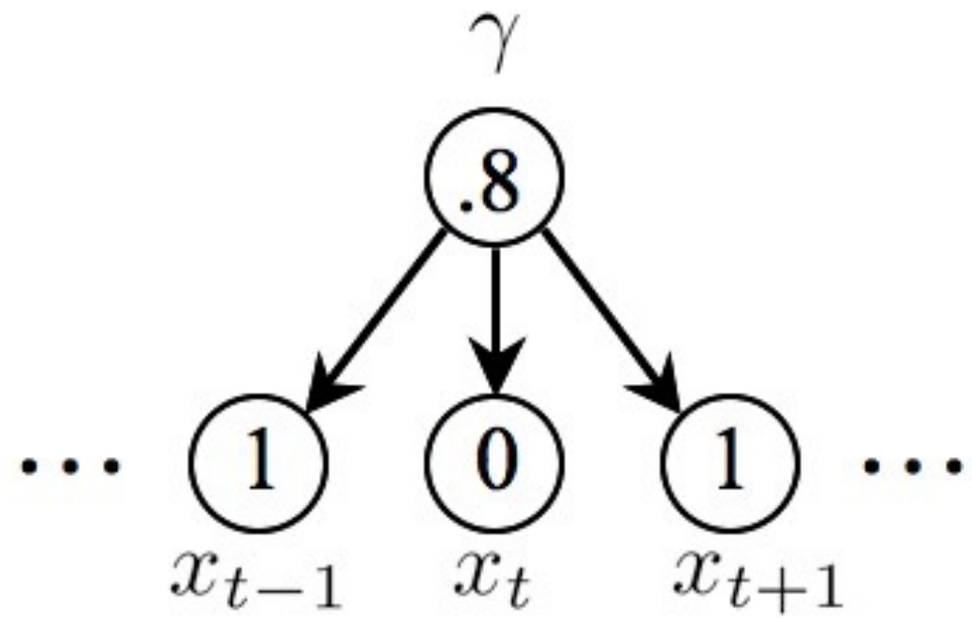


Generative Model
(what subject "knows")

$$P(x_t|\gamma) = Bernoulli(x_t;\gamma)$$

$$p(\gamma) = Beta(\gamma;a,b)$$

Bayesian Inference



Generative Model
(what subject "knows")

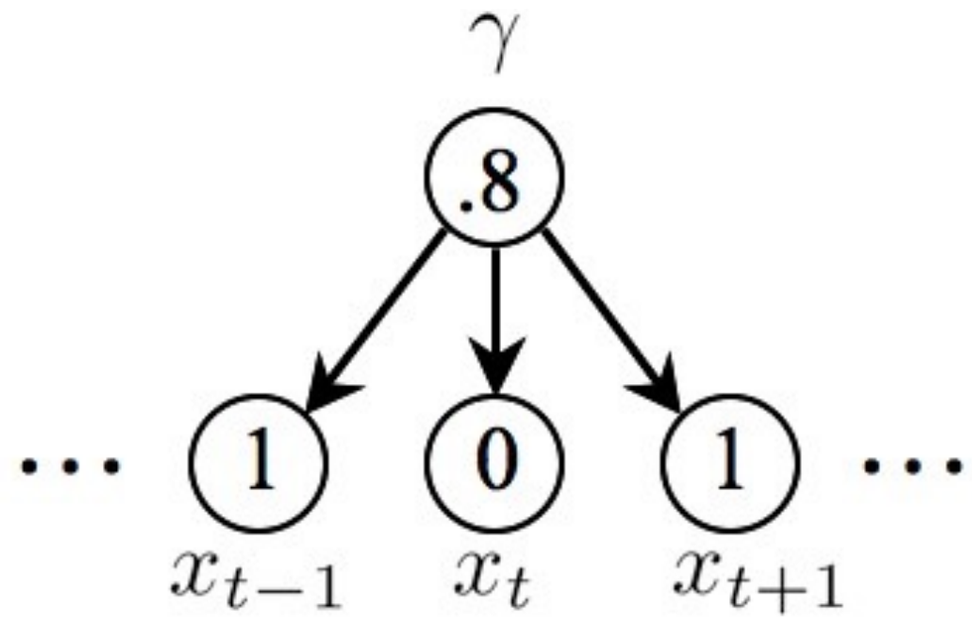
$$P(x_t|\gamma) = Bernoulli(x_t;\gamma)$$

$$p(\gamma) = Beta(\gamma;a,b)$$

Belief Updating
(Bayes' Rule)

$$p(\gamma|\mathbf{x}_1^t) = \frac{P(x_t|\gamma)p(\gamma|\mathbf{x}_1^{t-1})}{P(x_t|\mathbf{x}_1^{t-1})}$$

Bayesian Inference



Generative Model
(what subject "knows")

$$P(x_t|\gamma) = Bernoulli(x_t;\gamma)$$

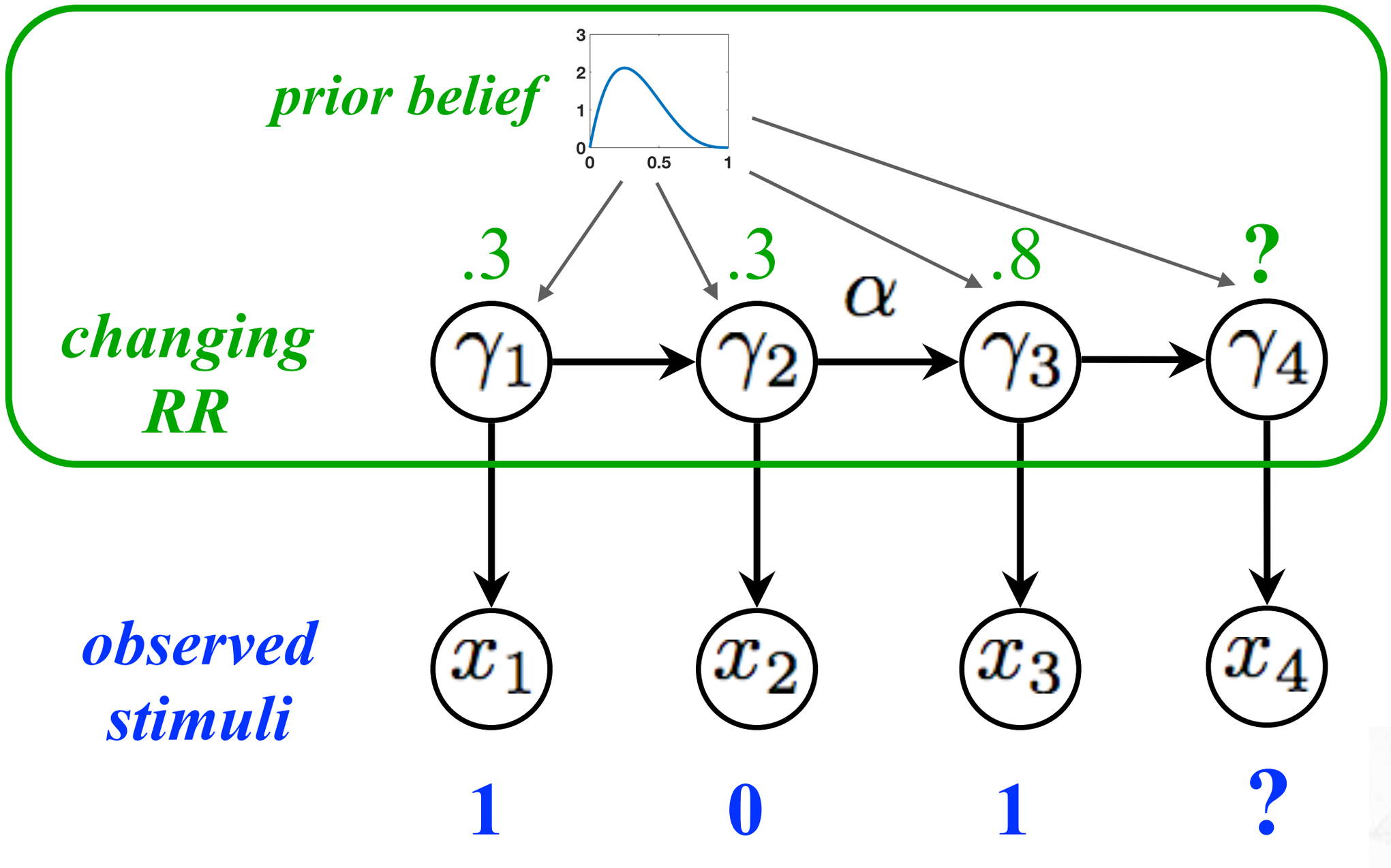
$$p(\gamma) = Beta(\gamma;a,b)$$

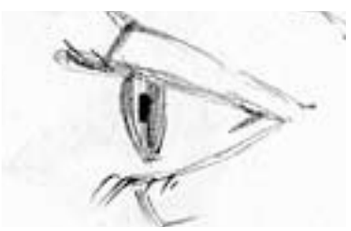
Belief Updating
(Bayes' Rule)

$$p(\gamma|\mathbf{x}_1^t) = \frac{P(x_t|\gamma)p(\gamma|\mathbf{x}_1^{t-1})}{P(x_t|\mathbf{x}_1^{t-1})}$$

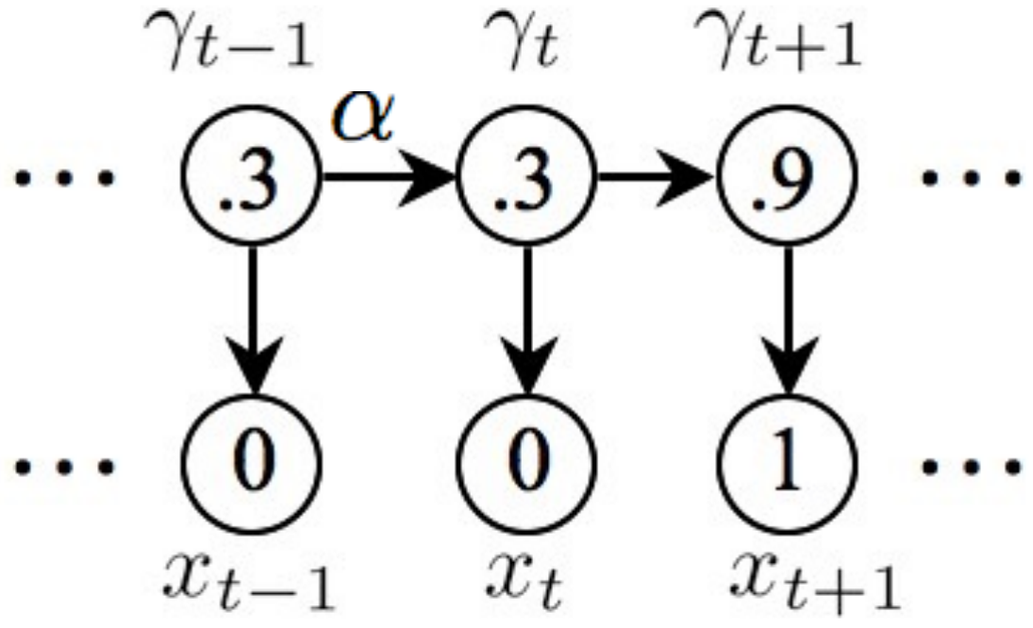
$$P(x_{t+1}=1|\mathbf{x}_1^t) = \mathbb{E}[\gamma|\mathbf{x}_1^t]$$

Learning Model II: Dynamic Belief Model (DBM)

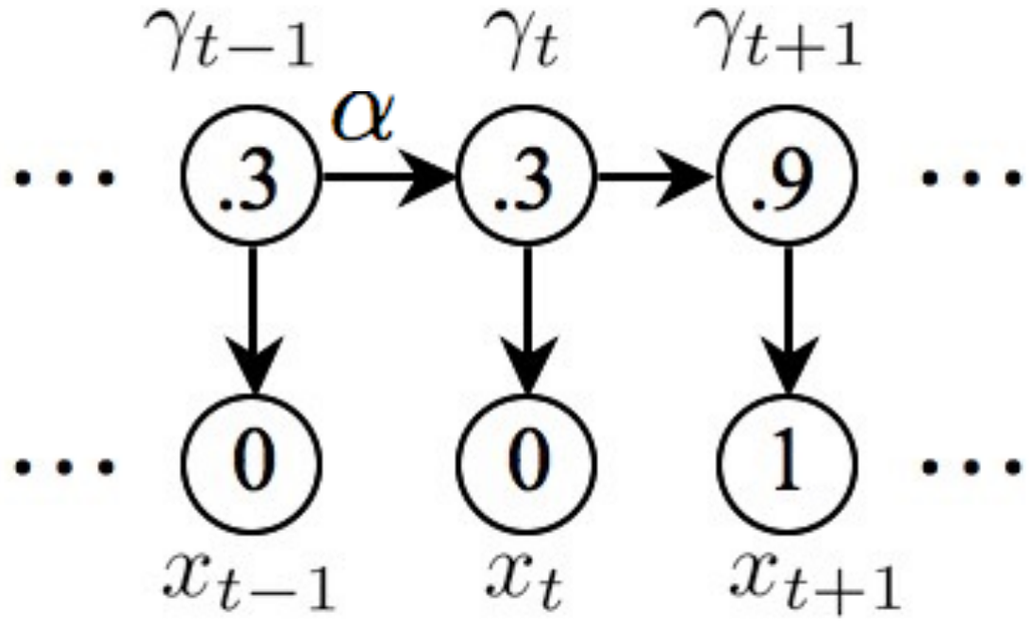




Bayesian Inference



Bayesian Inference

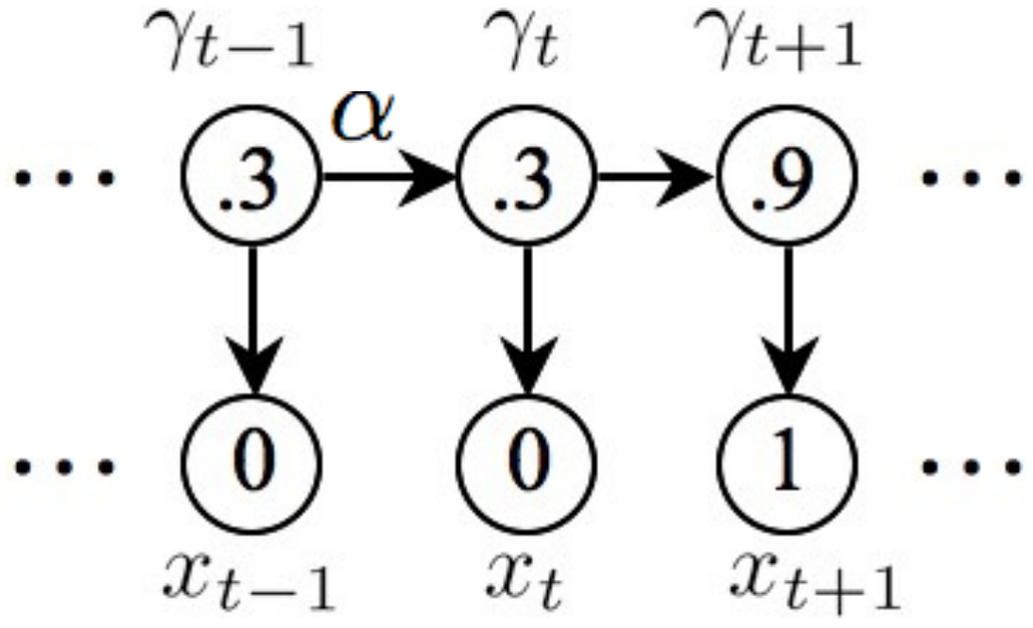


Generative Model
(what subject "knows")

$$P(x_t|\gamma_t) = Bernoulli(x_t; \gamma_t)$$

$$p_0(\gamma) = Beta(a, b)$$

Bayesian Inference



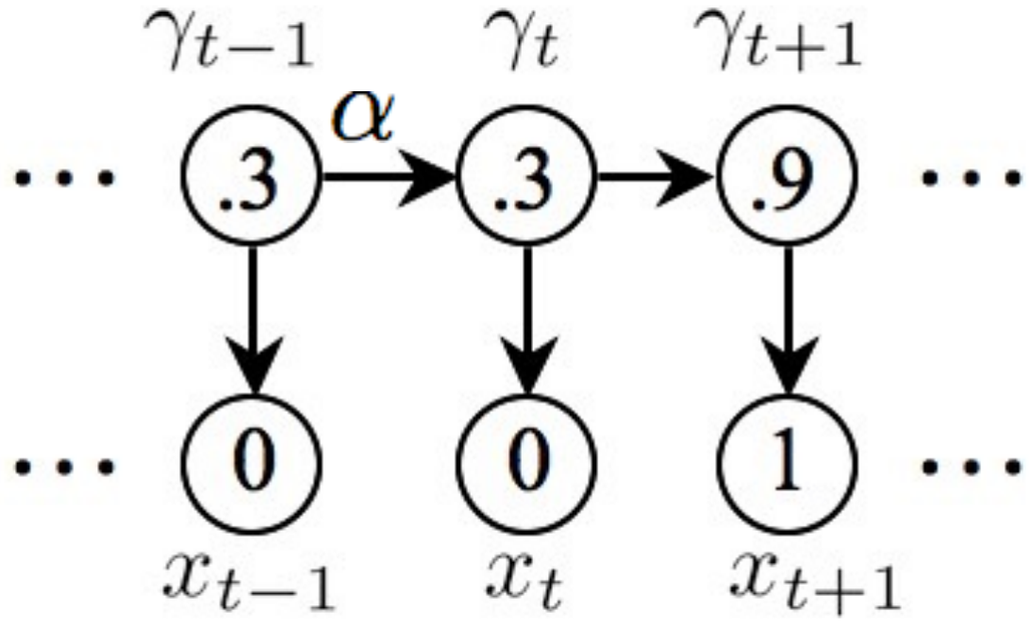
Generative Model
(what subject "knows")

$$P(x_t|\gamma_t) = Bernoulli(x_t;\gamma_t)$$

$$p_0(\gamma) = Beta(a,b)$$

$$p(\gamma_t|\gamma_{t-1}) = \alpha\delta(\gamma_t - \gamma_{t-1}) + (1-\alpha)p_0(\gamma_t)$$

Bayesian Inference



Generative Model
(what subject "knows")

$$P(x_t|\gamma_t) = Bernoulli(x_t;\gamma_t)$$

$$p_0(\gamma) = Beta(a,b)$$

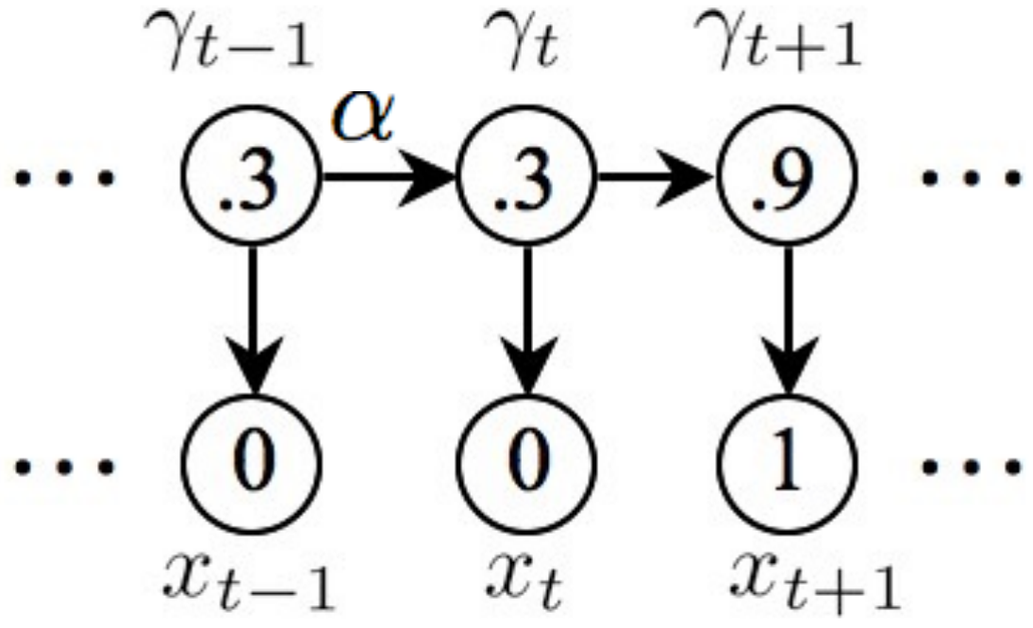
$$p(\gamma_t|\gamma_{t-1}) = \alpha\delta(\gamma_t - \gamma_{t-1}) + (1-\alpha)p_0(\gamma_t)$$

Belief Updating
(Bayes' Rule)

$$p(\gamma_t|\mathbf{x}_1^t) = \frac{P(x_t|\gamma_t)p(\gamma_t|\mathbf{x}_1^{t-1})}{P(x_t|\mathbf{x}_1^{t-1})}$$

$$P(x_{t+1}=1|\mathbf{x}_1^t) = \mathbb{E}[\gamma_{t+1}|\mathbf{x}_1^t]$$

Bayesian Inference



Generative Model
(what subject "knows")

$$P(x_t|\gamma_t) = Bernoulli(x_t;\gamma_t)$$

$$p_0(\gamma) = Beta(a,b)$$

$$p(\gamma_t|\gamma_{t-1}) = \alpha\delta(\gamma_t - \gamma_{t-1}) + (1-\alpha)p_0(\gamma_t)$$

Belief Updating
(Bayes' Rule)

$$p(\gamma_t|\mathbf{x}_1^t) = \frac{P(x_t|\gamma_t)p(\gamma_t|\mathbf{x}_1^{t-1})}{P(x_t|\mathbf{x}_1^{t-1})}$$

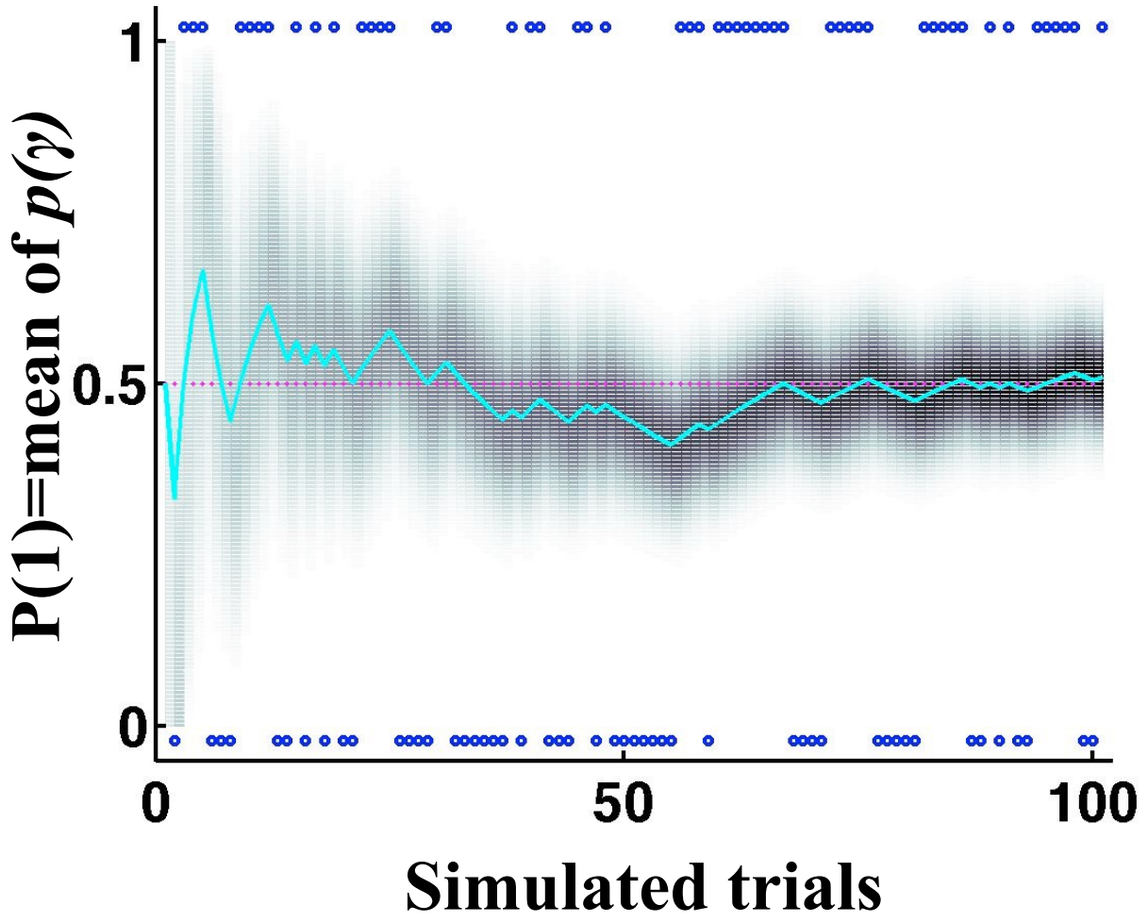
$$P(x_{t+1}=1|\mathbf{x}_1^t) = \mathbb{E}[\gamma_{t+1}|\mathbf{x}_1^t]$$

$$p(\gamma_{t+1}|\mathbf{x}_1^t) = \alpha p(\gamma_t|\mathbf{x}_1^t) + (1-\alpha)p_0(\gamma_{t+1})$$

Sequential Learning: FBM vs. DBM

Given a sequence of Bernoulli data from *fixed* reward rate (γ = .5) …

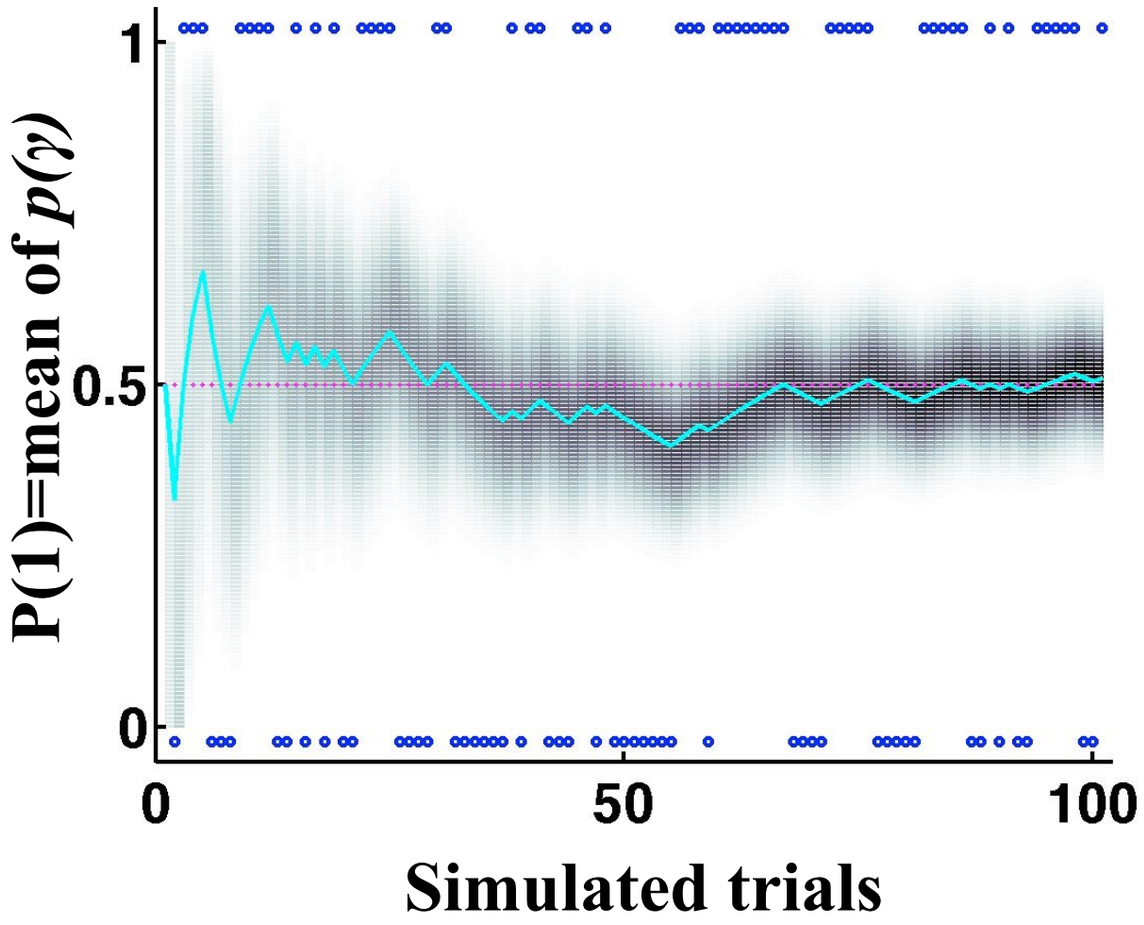
FBM: posterior *p*(γ)



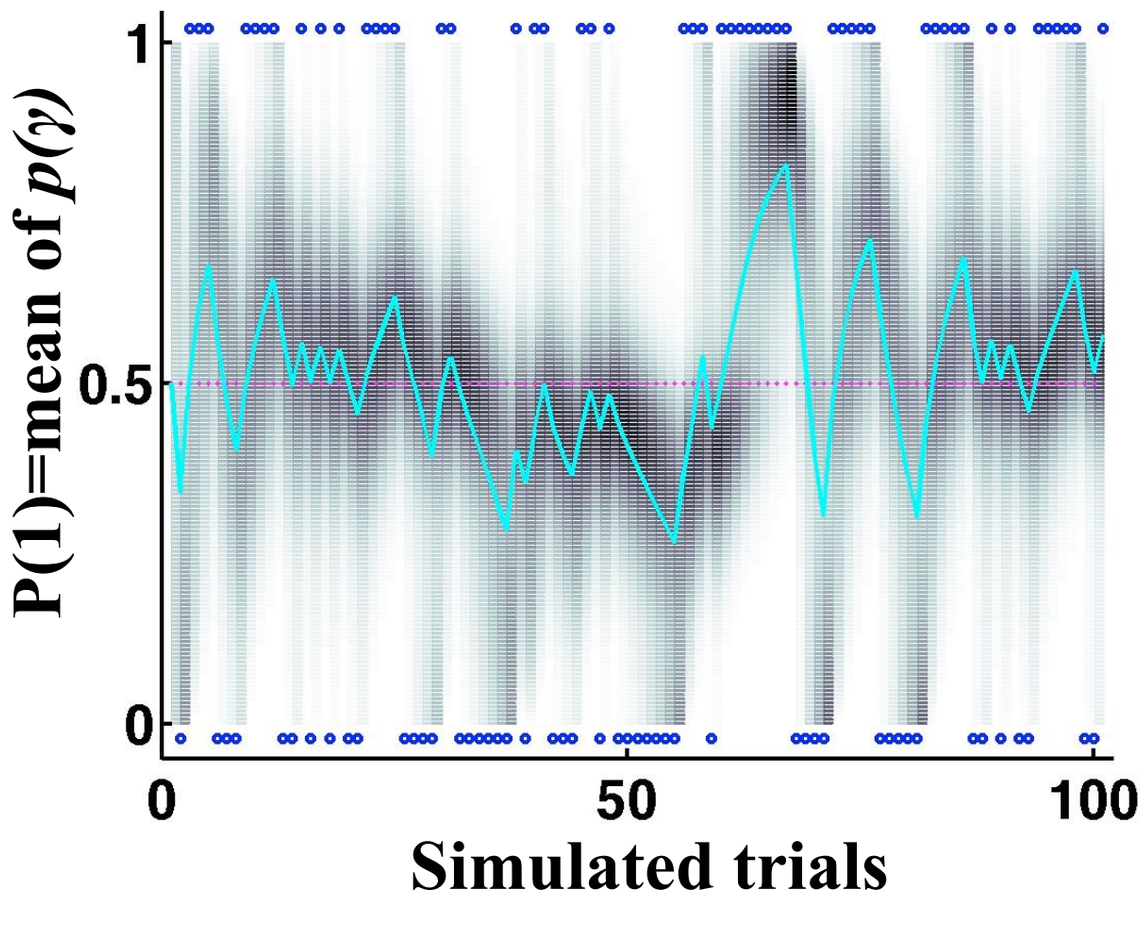
Sequential Learning: FBM vs. DBM

Given a sequence of Bernoulli data from *fixed* reward rate (γ = .5) …

FBM: posterior *p*(γ)



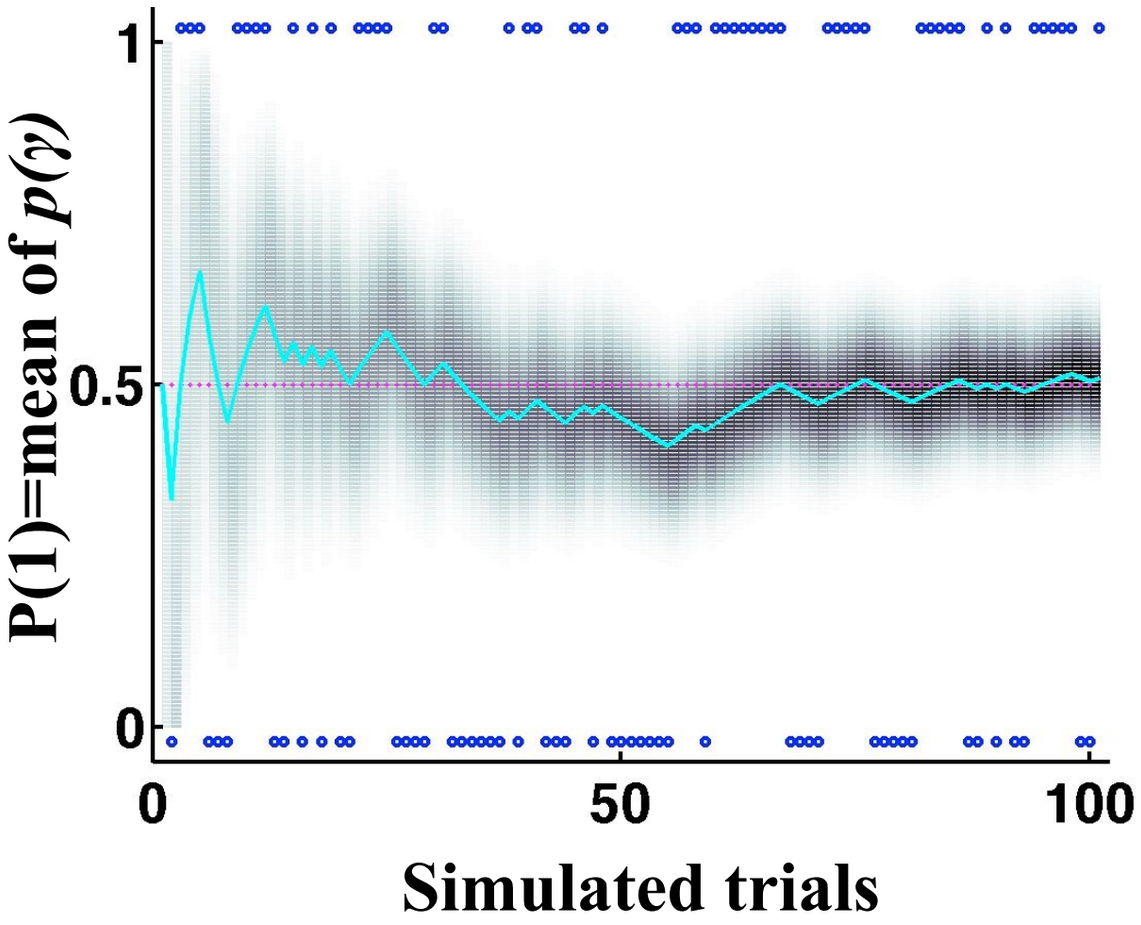
DBM: posterior *p*(γ)



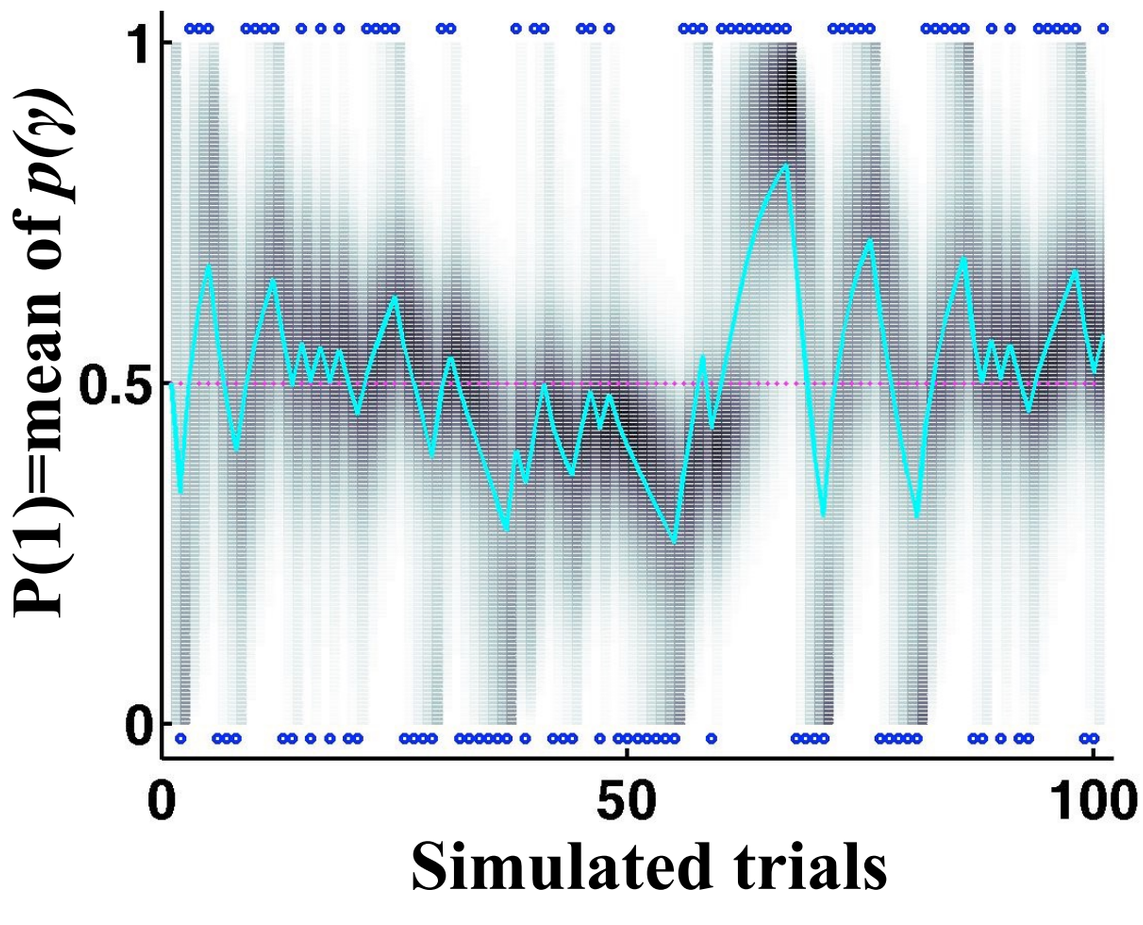
Sequential Learning: FBM vs. DBM

Given a sequence of Bernoulli data from *fixed* reward rate (γ = .5) …

FBM: posterior $p(\gamma)$



DBM: posterior $p(\gamma)$

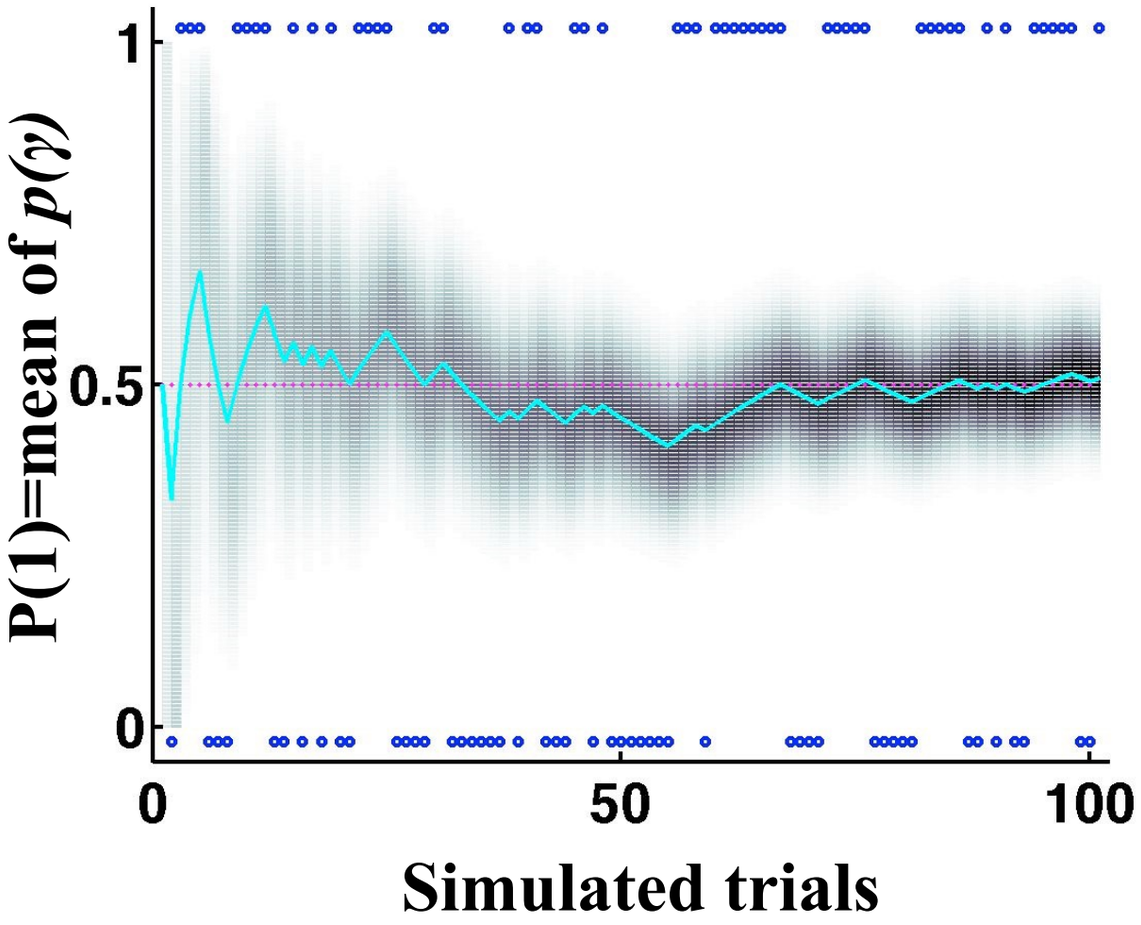


Driven by *long-term* average

Sequential Learning: FBM vs. DBM

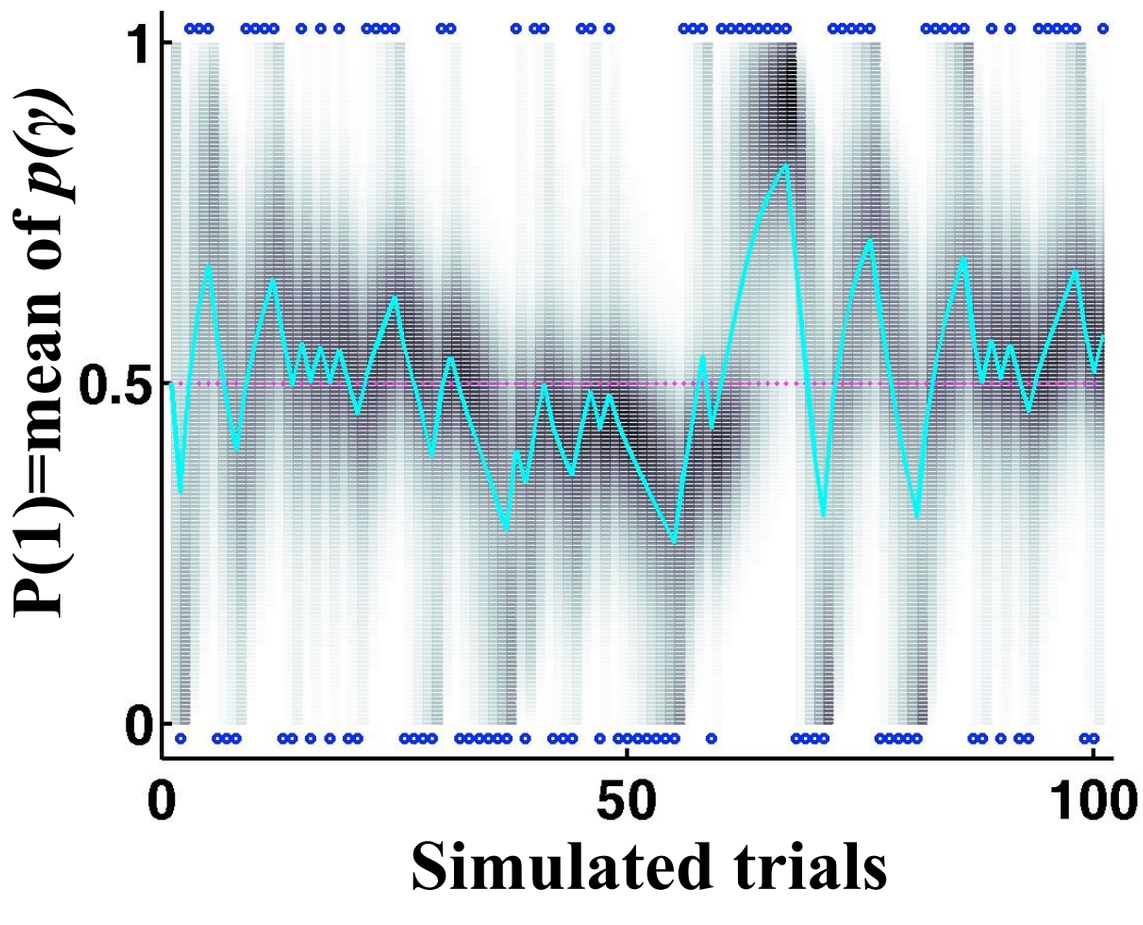
Given a sequence of Bernoulli data from *fixed* reward rate (γ = .5) …

FBM: posterior $p(\gamma)$



Driven by *long-term* average

DBM: posterior $p(\gamma)$



Driven by *local* patterns

Why DBM if Rewards Truly Fixed?

Why DBM if Rewards Truly Fixed?

- DBM (Yu & Cohen, 2009) proposed to capture sequential effect in human 2AFC behavior (Yu & Cohen, 2009) — expectancy driven by local chance patterns in stimulus statistics

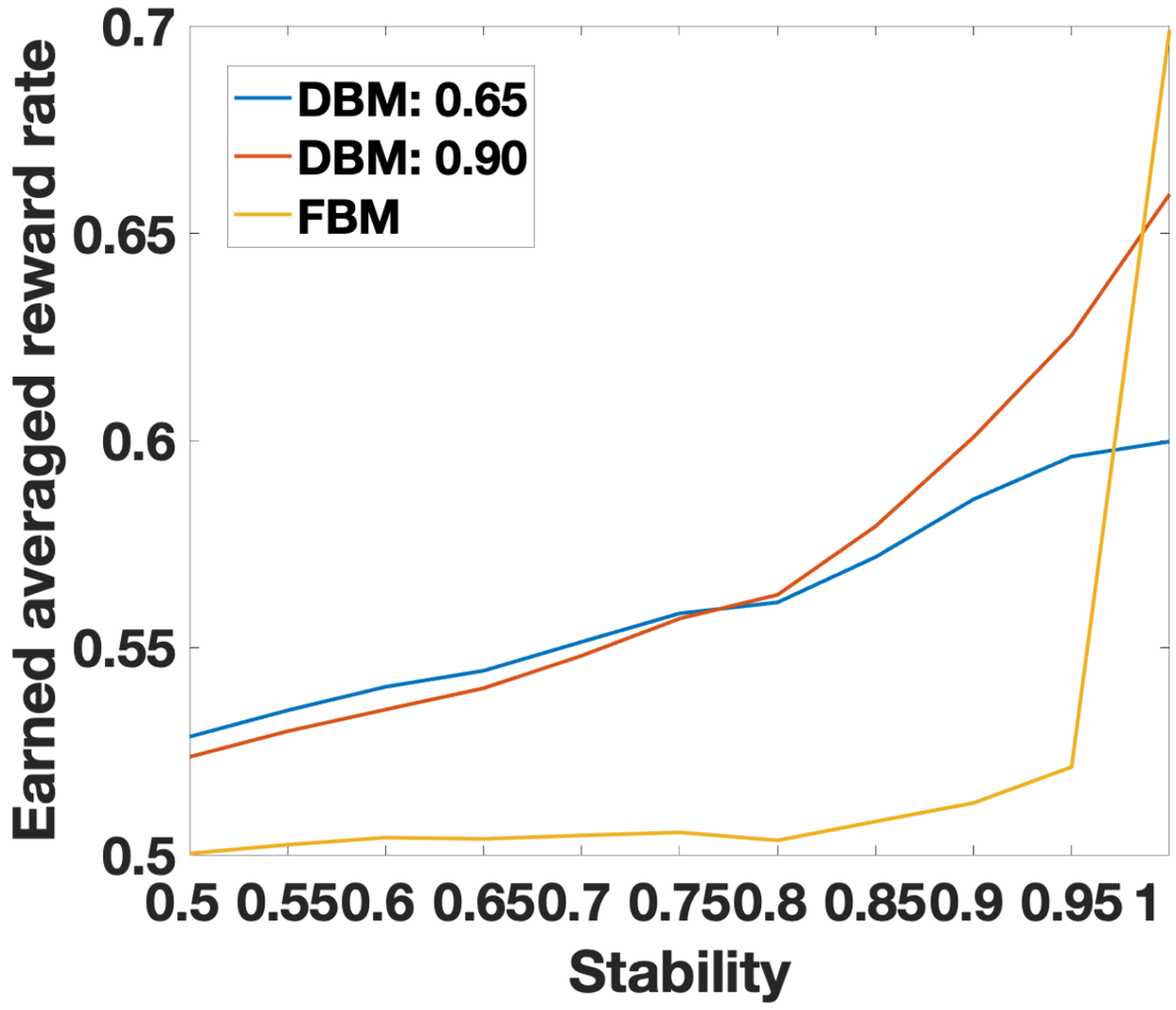
Why DBM if Rewards Truly Fixed?

- DBM (Yu & Cohen, 2009) proposed to capture sequential effect in human 2AFC behavior (Yu & Cohen, 2009) — expectancy driven by local chance patterns in stimulus statistics
- DBM > FBM in explaining human behavior
  - 2-alternative forced choice (Yu A, Cohen J., 2009)
  - inhibitory control (Ide J.S. et al, 2013)
  - visual search (Yu A. J., Huang H., 2014)
  - bandit task (Zhang S., Yu J. A., 2013)

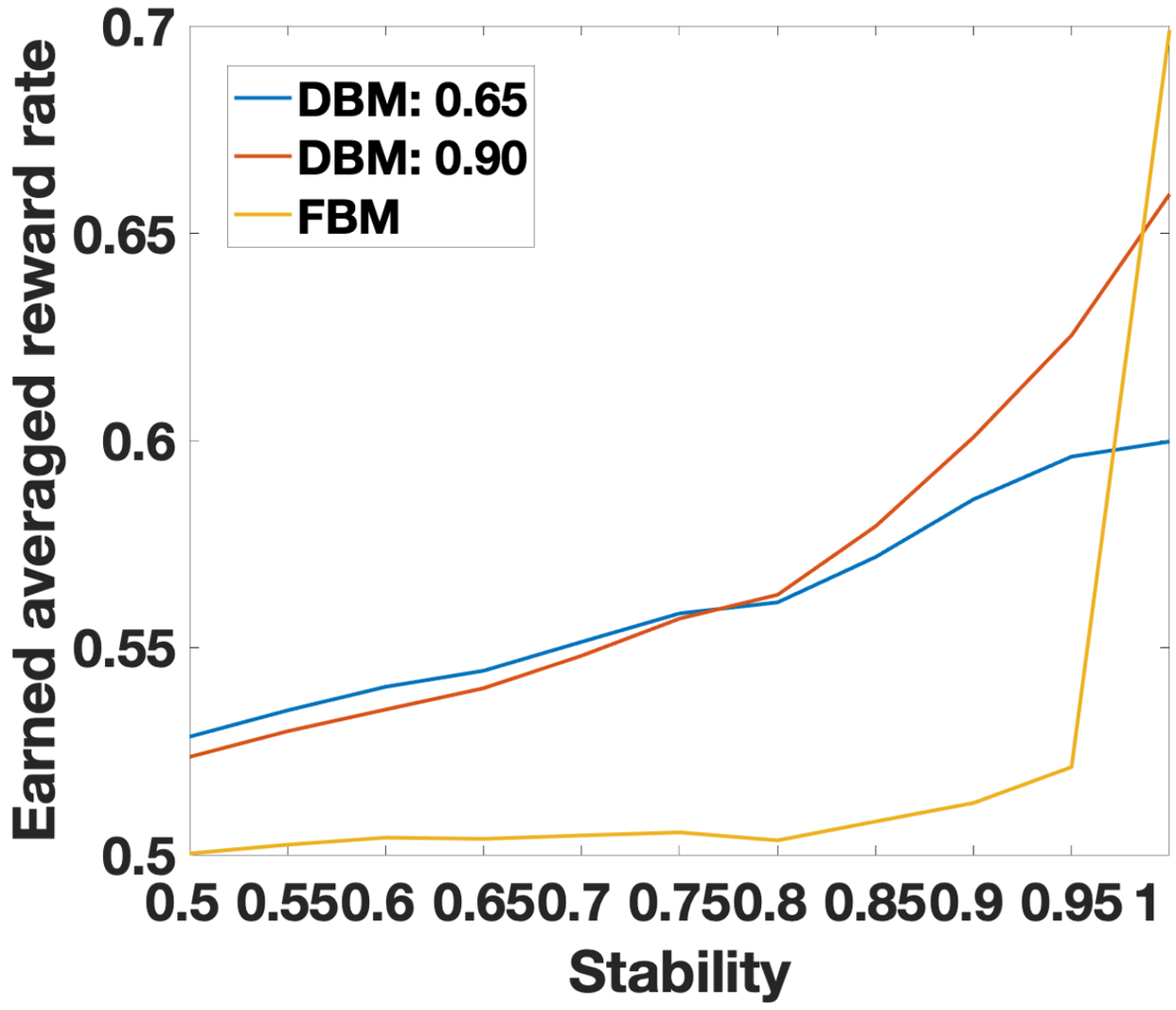
Why DBM if Rewards Truly Fixed?

- DBM (Yu & Cohen, 2009) proposed to capture sequential effect in human 2AFC behavior (Yu & Cohen, 2009) — expectancy driven by local chance patterns in stimulus statistics
- DBM > FBM in explaining human behavior
  - 2-alternative forced choice (Yu A, Cohen J., 2009)
  - inhibitory control (Ide J.S. et al, 2013)
  - visual search (Yu A. J., Huang H., 2014)
  - bandit task (Zhang S., Yu J. A., 2013)
- Implies human subjects assume environmental statistics to be changeable even though they are truly fixed in the experiment

Why DBM if Rewards Truly Fixed?

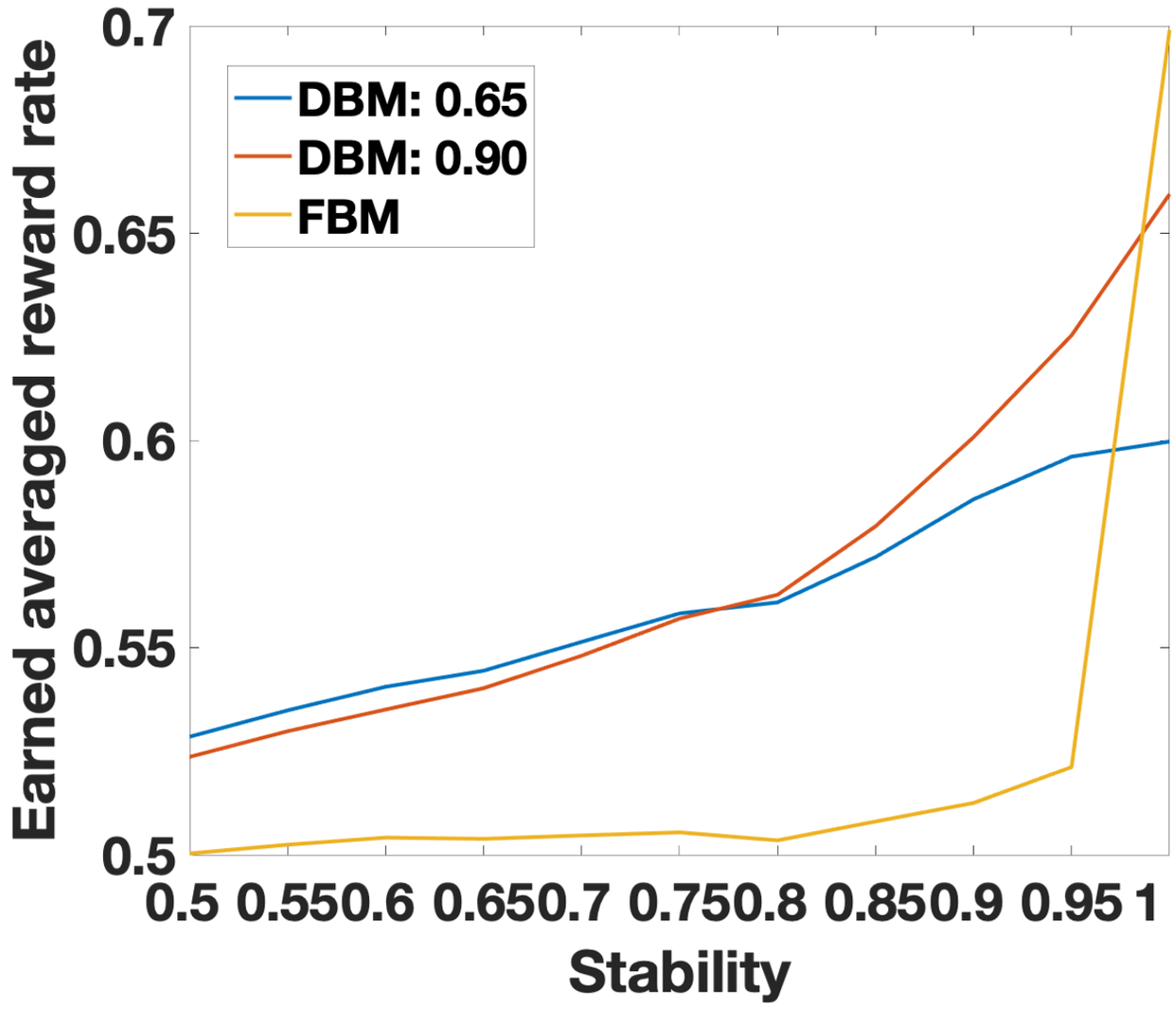


Why DBM if Rewards Truly Fixed?



Simulation Results:

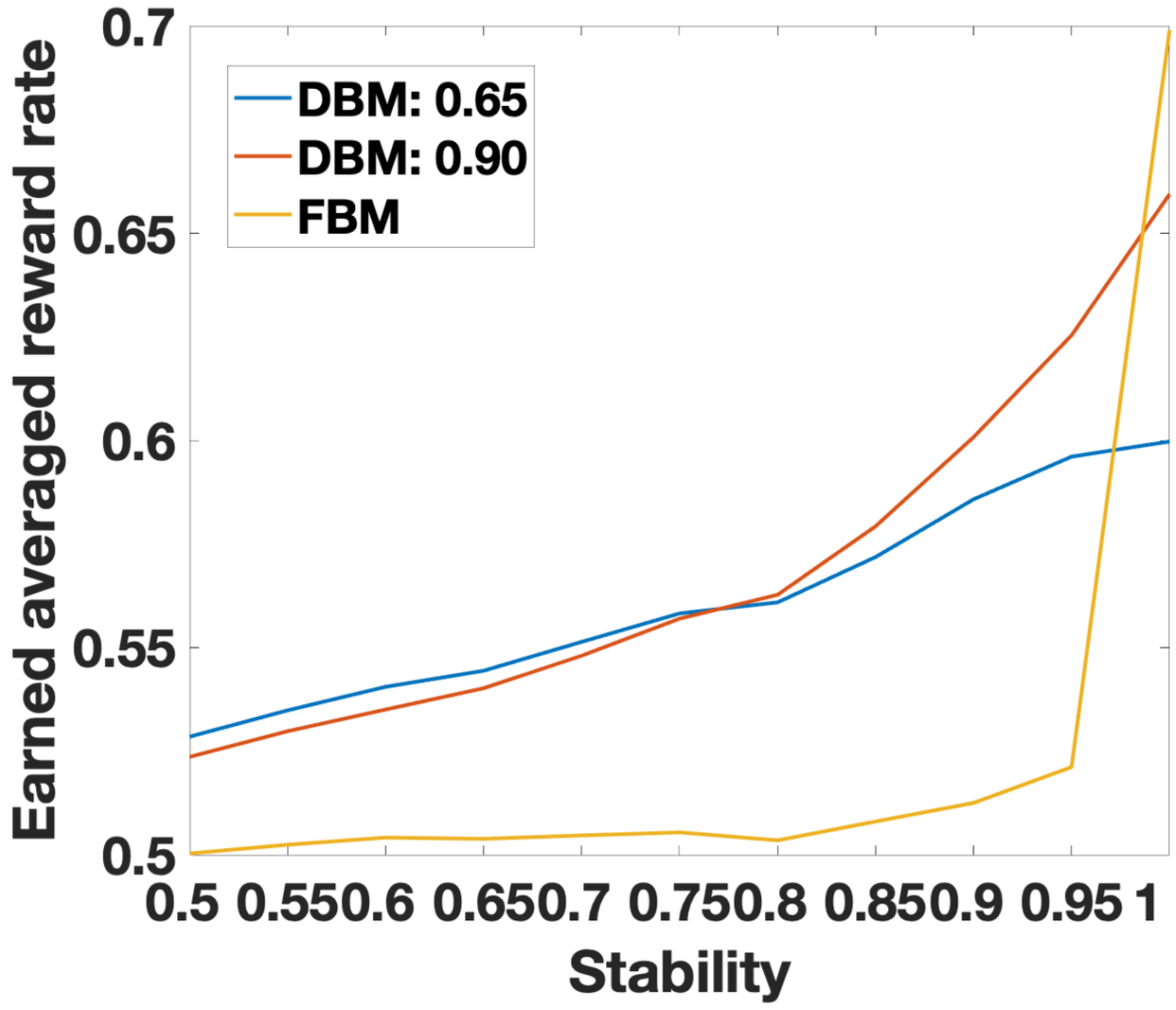
Why DBM if Rewards Truly Fixed?



Simulation Results:

- Assuming stationarity terrible for non-stationary environment

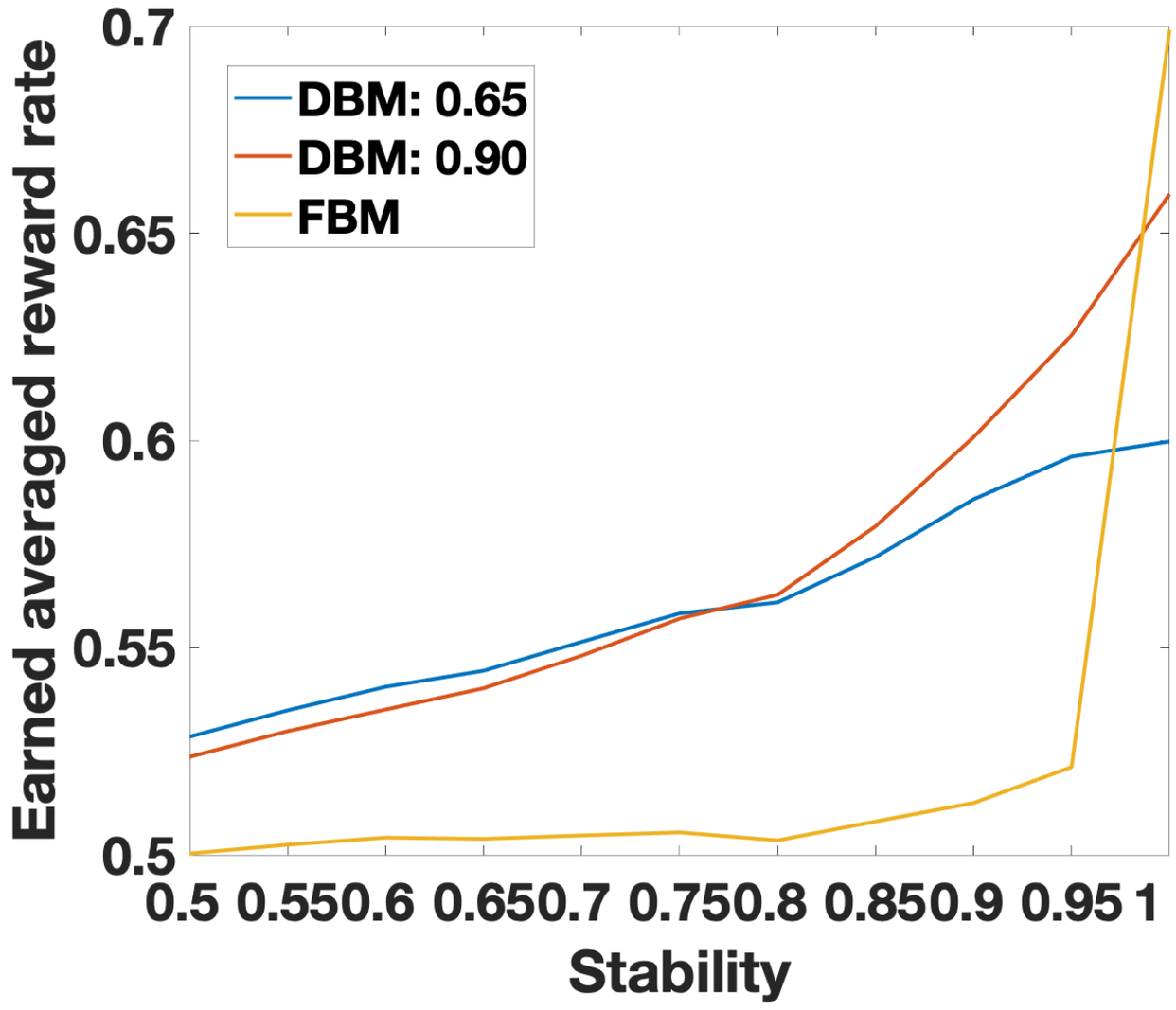
Why DBM if Rewards Truly Fixed?



Simulation Results:

- Assuming stationarity terrible for non-stationary environment
- Matching exact value of stability parameter not so important when the environment is highly volatile

Why DBM if Rewards Truly Fixed?



Simulation Results:

- Assuming stationarity terrible for non-stationary environment
- Matching exact value of stability parameter not so important when the environment is highly volatile
- ⇒ "Sweet spot" for stability (indeed we always find estimated human stability parameter to be around 0.75-0.90)

Relationship b/t DBM & Reinforcement Learning

Relationship b/t DBM & Reinforcement Learning

- Rescorla-Wagner/Q-learning (Rescorla & Wagner, 1972):

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \epsilon(\underbrace{R_t - \hat{\theta}_k^{t-1}}_{\delta_t})$$

δ_t - *prediction error*

Relationship b/t DBM & Reinforcement Learning

- Rescorla-Wagner/Q-learning (Rescorla & Wagner, 1972):

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \epsilon(\underbrace{R_t - \hat{\theta}_k^{t-1}}_{\delta_t})$$

δ_t - *prediction error*

- DBM related to RW, but has a persistent prior bias that pushes belief toward prior mean θ_0 (Ryali, Reddy, Yu, *NIPS*, 2018)

Relationship b/t DBM & Reinforcement Learning

- Rescorla-Wagner/Q-learning (Rescorla & Wagner, 1972):

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \epsilon(\underbrace{R_t - \hat{\theta}_k^{t-1}}_{\delta_t})$$

δ_t - *prediction error*

- DBM related to RW, but has a persistent prior bias that pushes belief toward prior mean θ_0 (Ryali, Reddy, Yu, *NIPS*, 2018)

- Chosen arm: $\hat{\theta}_k^t = \underbrace{(1-\alpha)\theta_0}_{\text{persistent prior bias}} + \alpha(\hat{\theta}_k^{t-1} + \underbrace{g_t(R_t - \hat{\theta}_k^{t-1})}_{\delta_t \text{ - prediction error}})$

Relationship b/t DBM & Reinforcement Learning

- Rescorla-Wagner/Q-learning (Rescorla & Wagner, 1972):

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \epsilon(\underbrace{R_t - \hat{\theta}_k^{t-1}}_{\delta_t})$$

δ_t - *prediction error*

- DBM related to RW, but has a persistent prior bias that pushes belief toward prior mean θ_0 (Ryali, Reddy, Yu, *NIPS*, 2018)

- Chosen arm: $\hat{\theta}_k^t = \underbrace{(1-\alpha)\theta_0}_{\text{persistent prior bias}} + \alpha(\hat{\theta}_k^{t-1} + \underbrace{g_t(R_t - \hat{\theta}_k^{t-1})}_{\delta_t \text{ - prediction error}})$

- Unchosen arm (related to "forgetting" Q-learning):

$$\hat{\theta}_j^t = (1-\alpha)\theta_0 + \alpha\hat{\theta}_j^{t-1}$$

Decision Model: Softmax (SM)

Decision Model: Softmax (SM)

- Commonly used in bandit task (e.g. Daw et al., 2006, Schönberg et al, 2007, Dezza et al., 2017)

Decision Model: Softmax (SM)

- Commonly used in bandit task (e.g. Daw et al., 2006, Schönberg et al, 2007, Dezza et al., 2017)
- Probability of choosing arm *k* given estimated reward rates:

$$Pr(D_t = k) = \frac{(\hat{\theta}_k^t)^b}{\sum_j^K (\hat{\theta}_j^t)^b}$$

Decision Model: Softmax (SM)

- Commonly used in bandit task (e.g. Daw et al., 2006, Schönberg et al, 2007, Dezza et al., 2017)
- Probability of choosing arm *k* given estimated reward rates:

$$Pr(D_t = k) = \frac{(\hat{\theta}_k^t)^b}{\sum_j^K (\hat{\theta}_j^t)^b}$$

- b $\rightarrow \infty$: choose option with highest current E[reward]

Decision Model: Softmax (SM)

- Commonly used in bandit task (e.g. Daw et al., 2006, Schönberg et al, 2007, Dezza et al., 2017)
- Probability of choosing arm *k* given estimated reward rates:

$$Pr(D_t = k) = \frac{(\hat{\theta}_k^t)^b}{\sum_j^K (\hat{\theta}_j^t)^b}$$

- b $\rightarrow \infty$: choose option with highest current E[reward]
- $b = 1$: exact matching (Herrnstein, 1961)

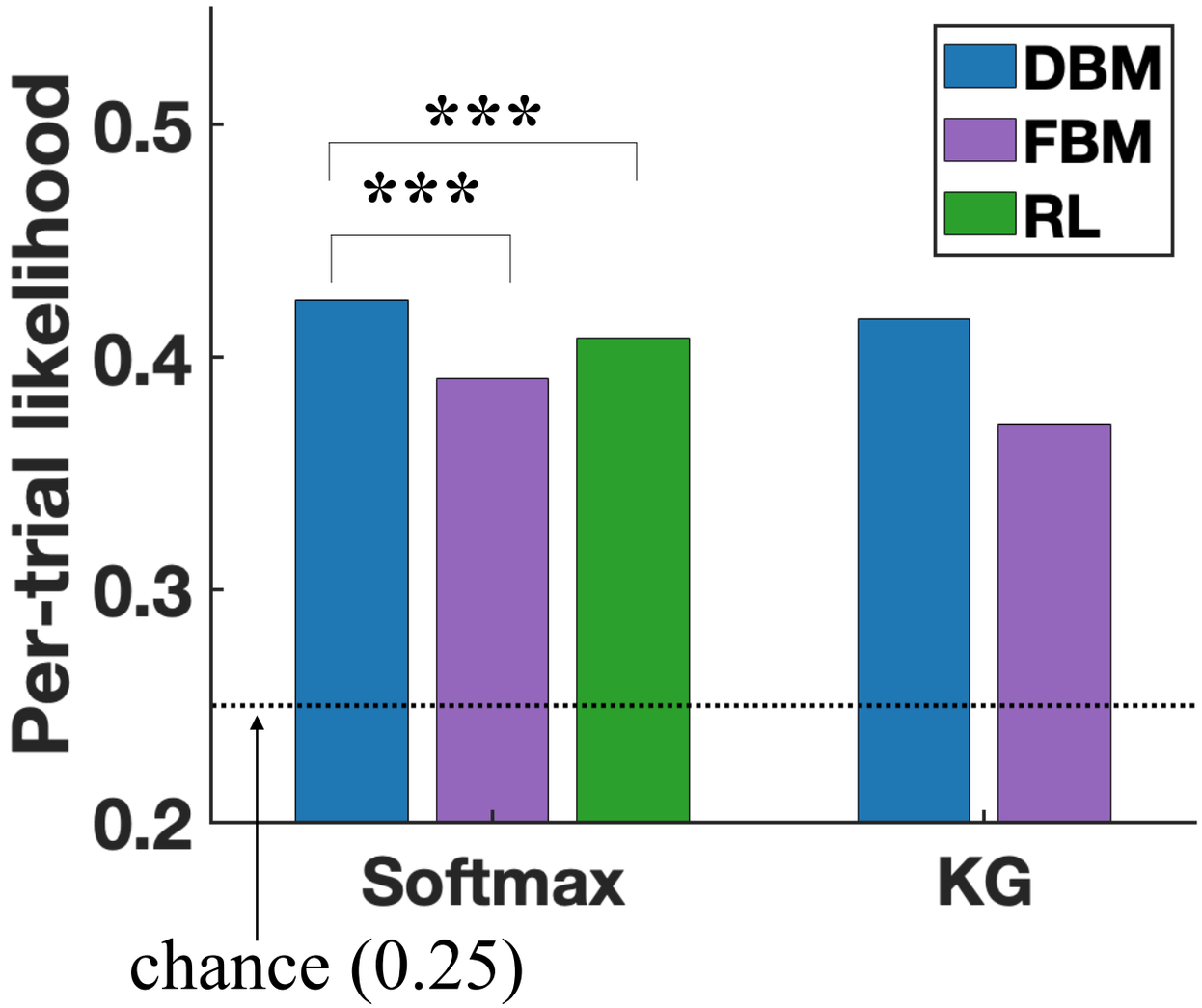
Decision Model: Softmax (SM)

- Commonly used in bandit task (e.g. Daw et al., 2006, Schönberg et al, 2007, Dezza et al., 2017)
- Probability of choosing arm *k* given estimated reward rates:

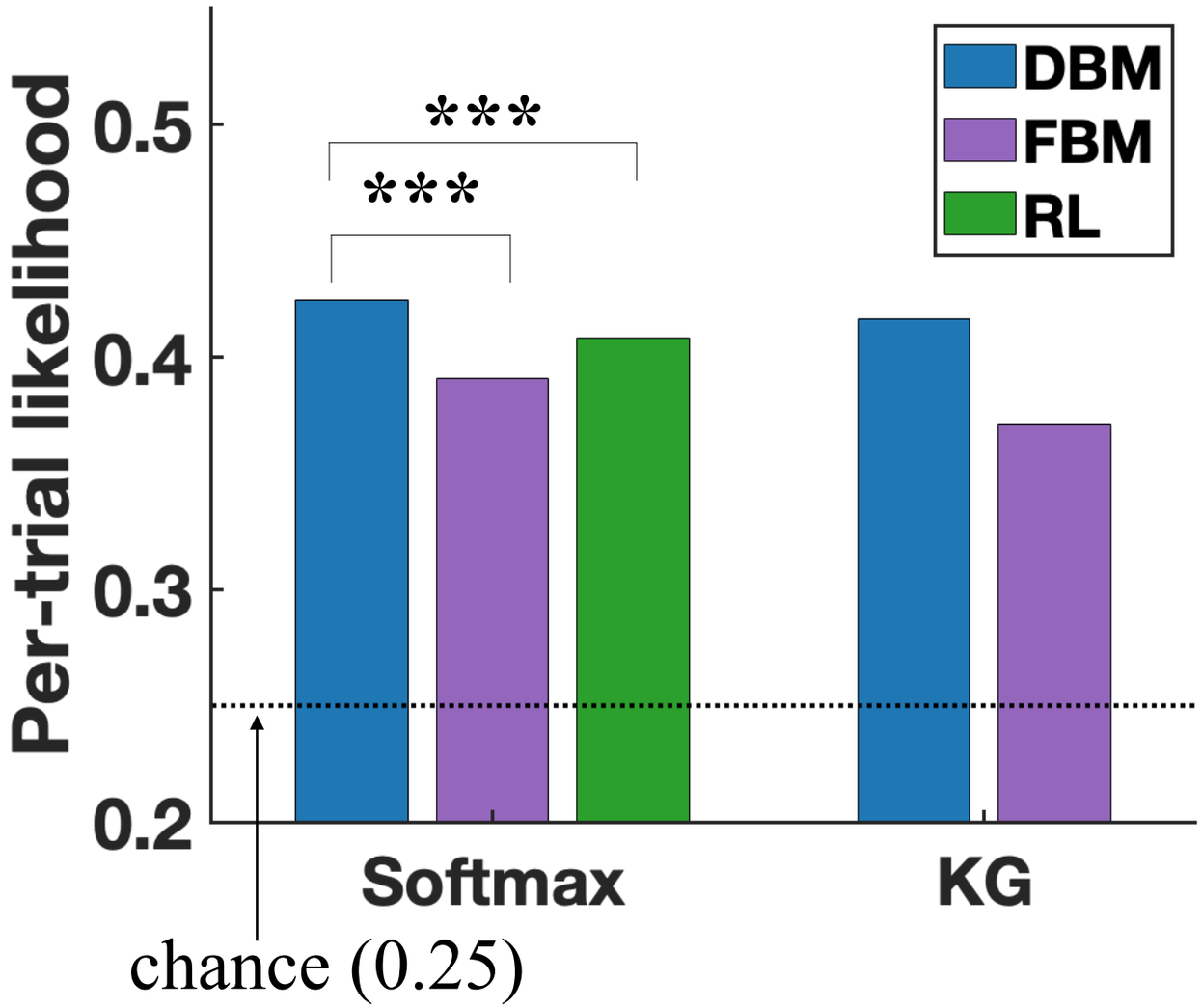
$$Pr(D_t = k) = \frac{(\hat{\theta}_k^t)^b}{\sum_j^K (\hat{\theta}_j^t)^b}$$

- b $\rightarrow \infty$: choose option with highest current E[reward]
- $b = 1$: exact matching (Herrnstein, 1961)
- b $\rightarrow$ 0: random policy

Model Comparison

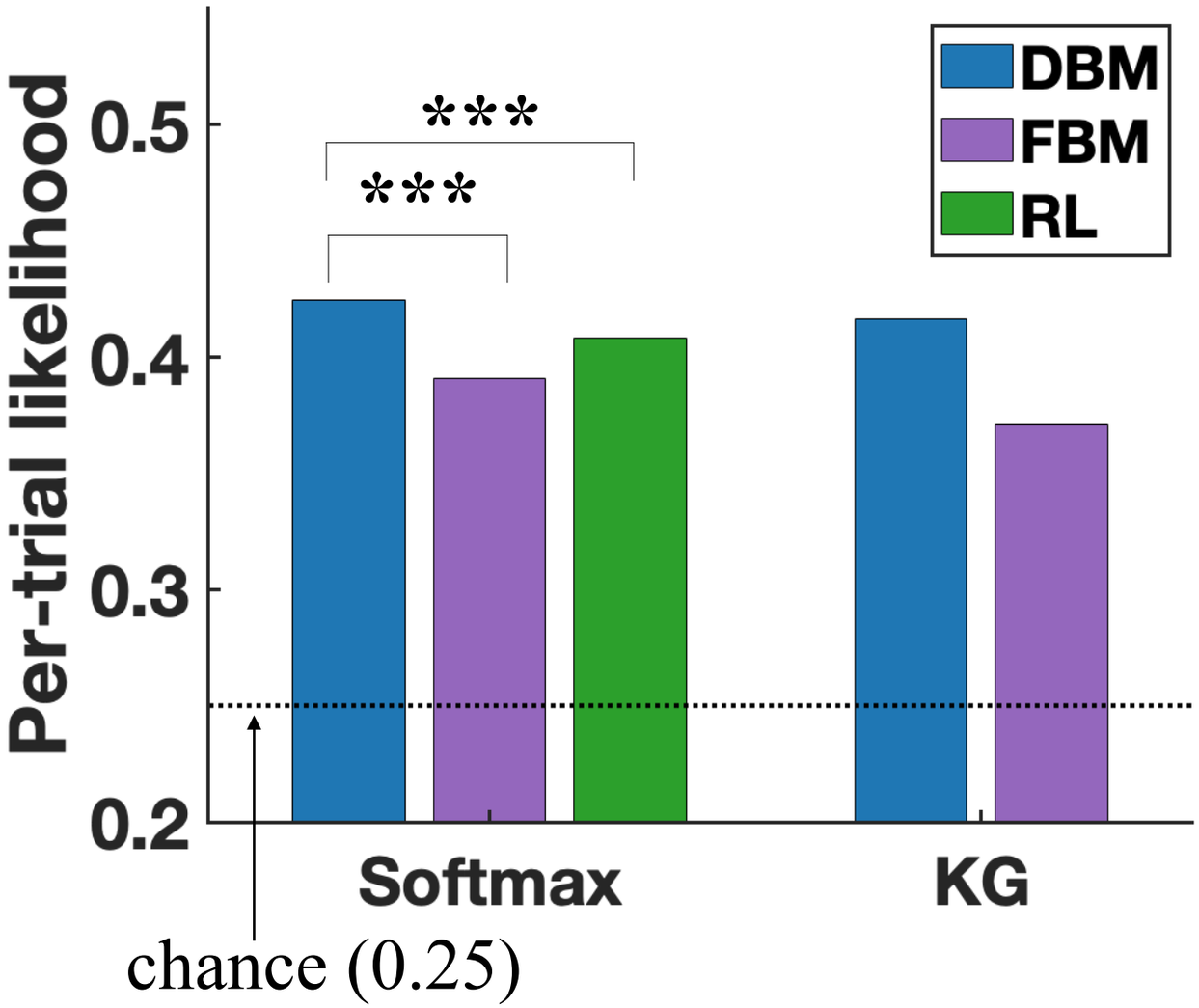


Model Comparison



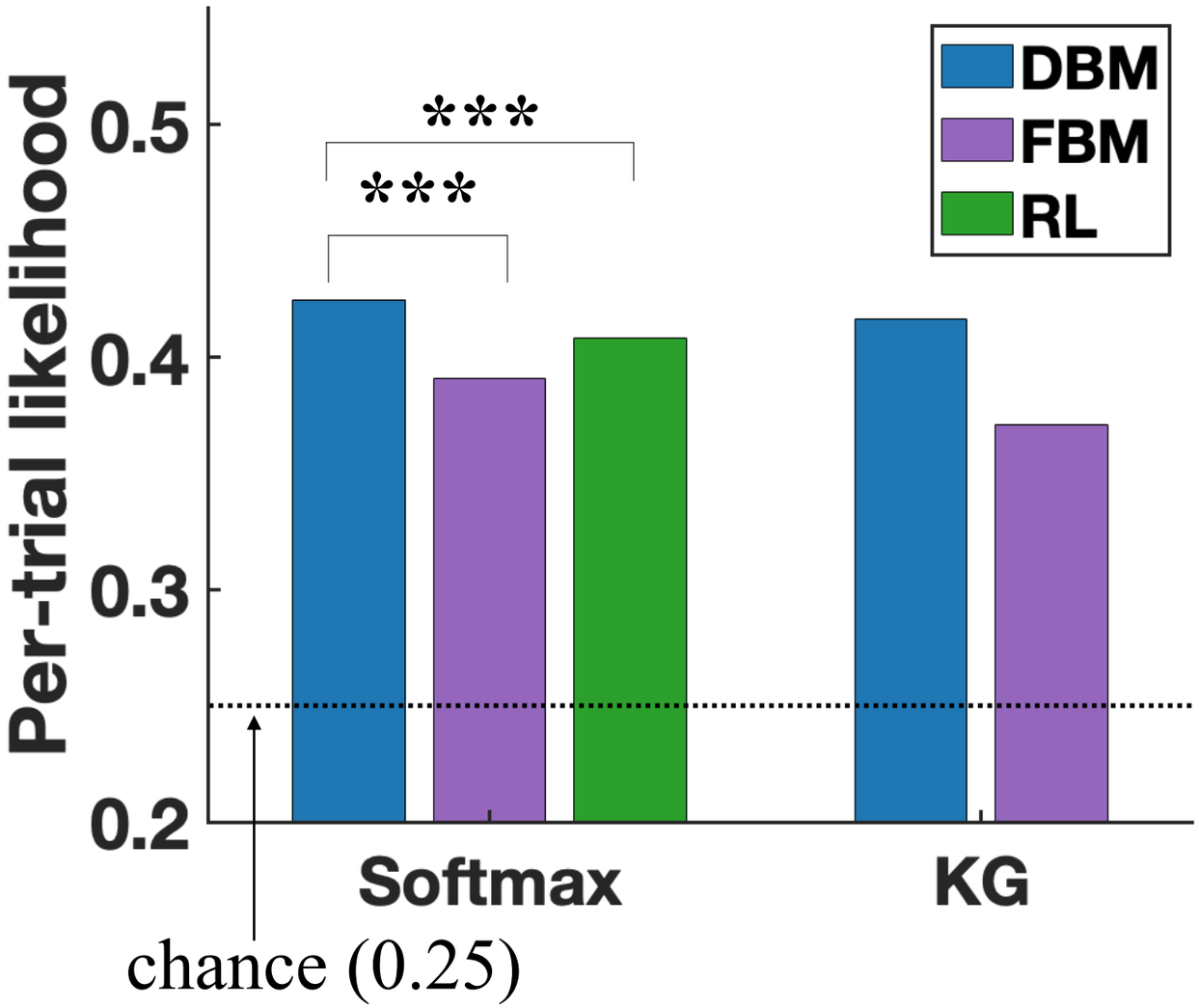
- 10-fold cross-validation (5 games held out)

Model Comparison



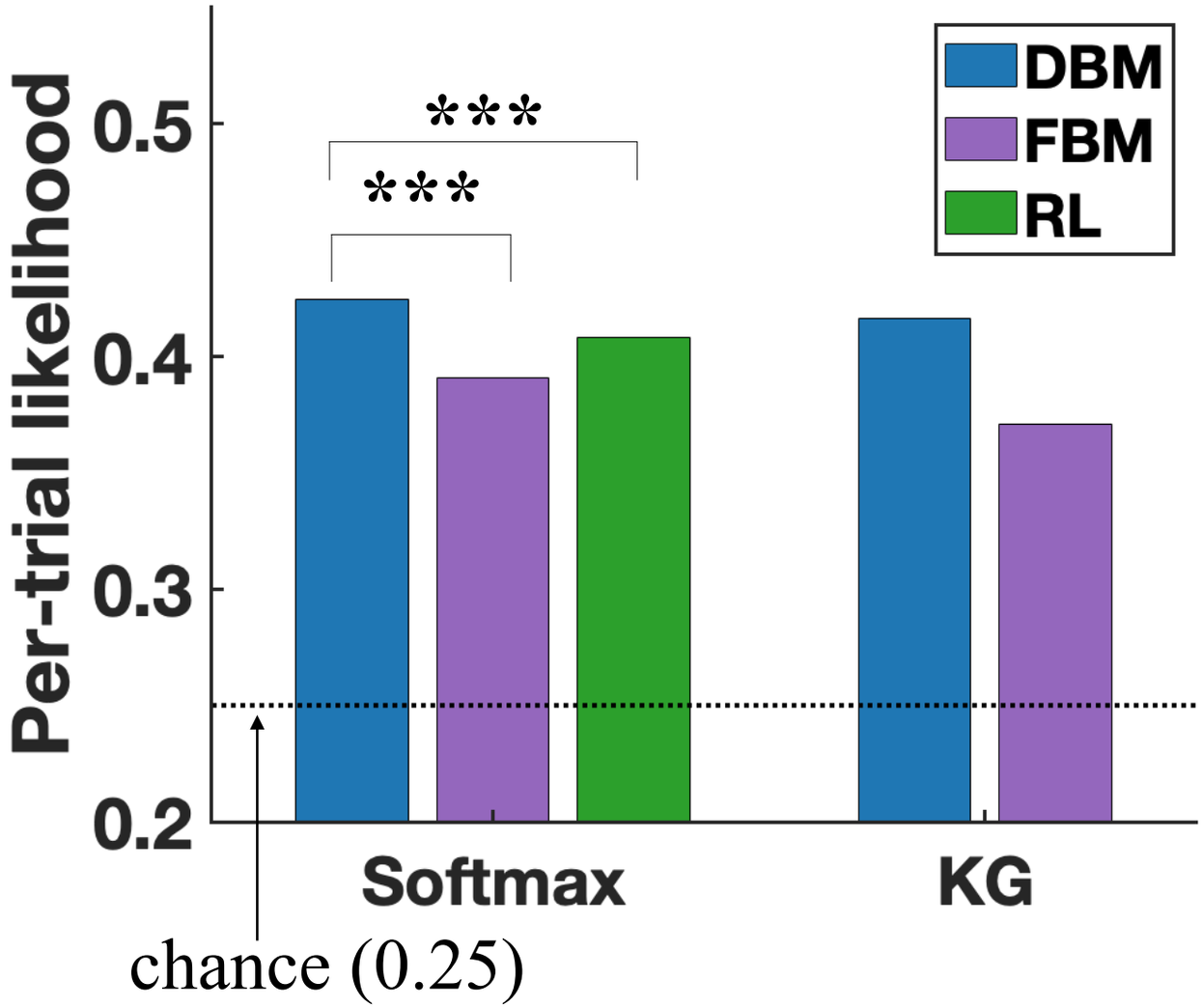
- 10-fold cross-validation (5 games held out)
- Best model: DBM+softmax

Model Comparison



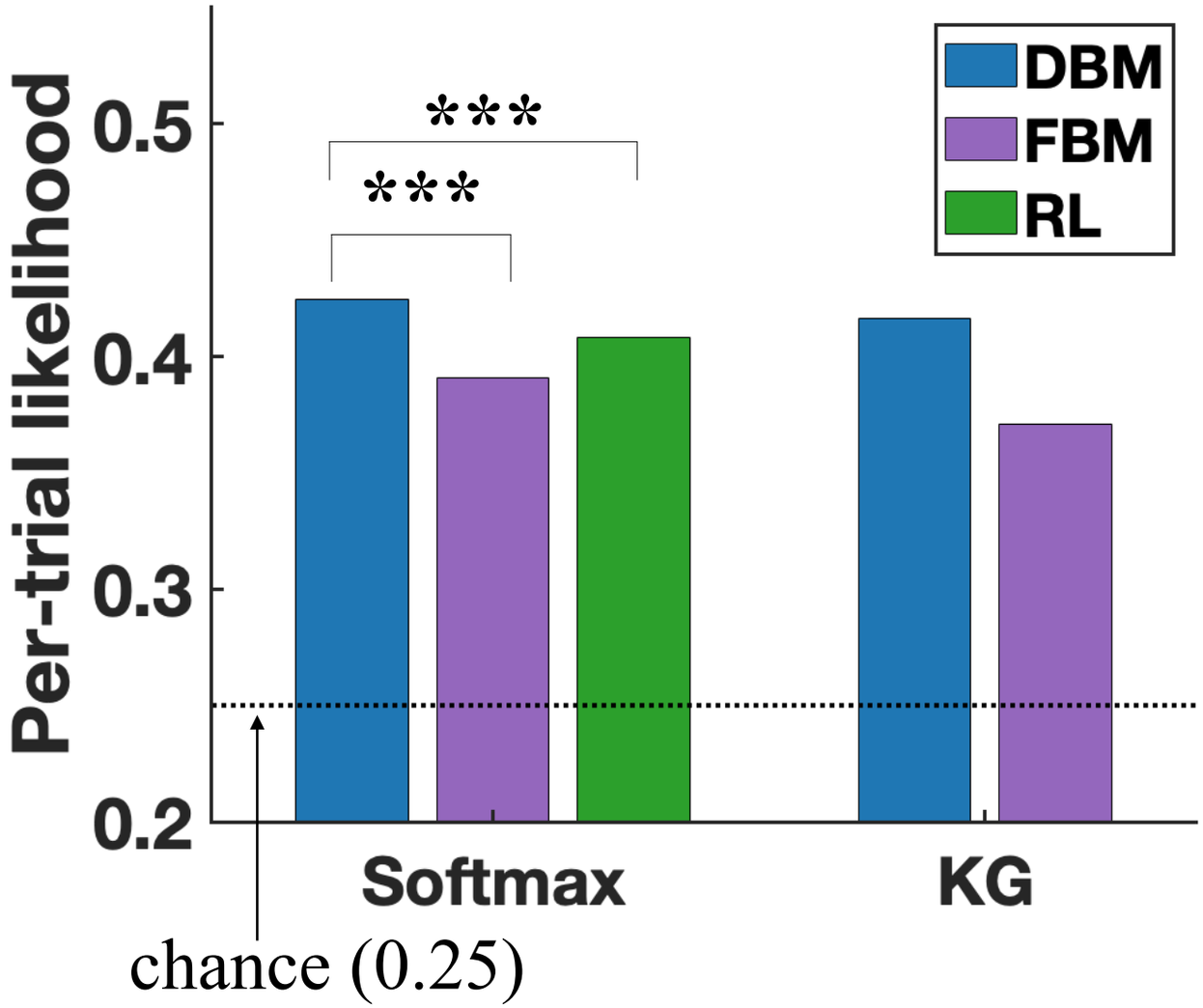
- 10-fold cross-validation (5 games held out)
- Best model: DBM+softmax
- DBM, RL > FBM: persistent sensitivity to recent reward history

Model Comparison



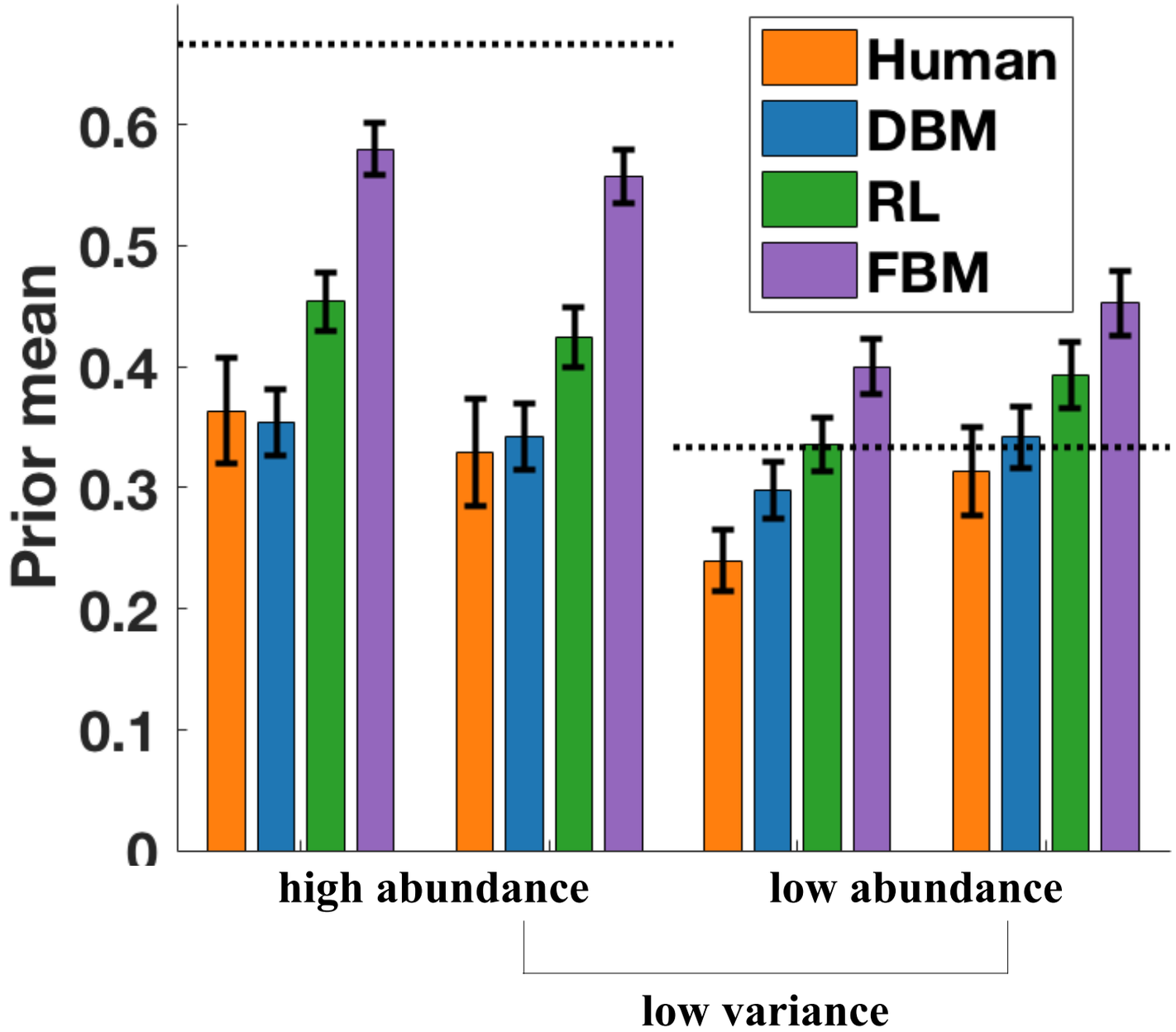
- 10-fold cross-validation (5 games held out)
- Best model: DBM+softmax
- DBM, RL > FBM: persistent sensitivity to recent reward history
- DBM > RL: both sensitive to recent reward history, but DBM also has persistent "prior bias"

Model Comparison

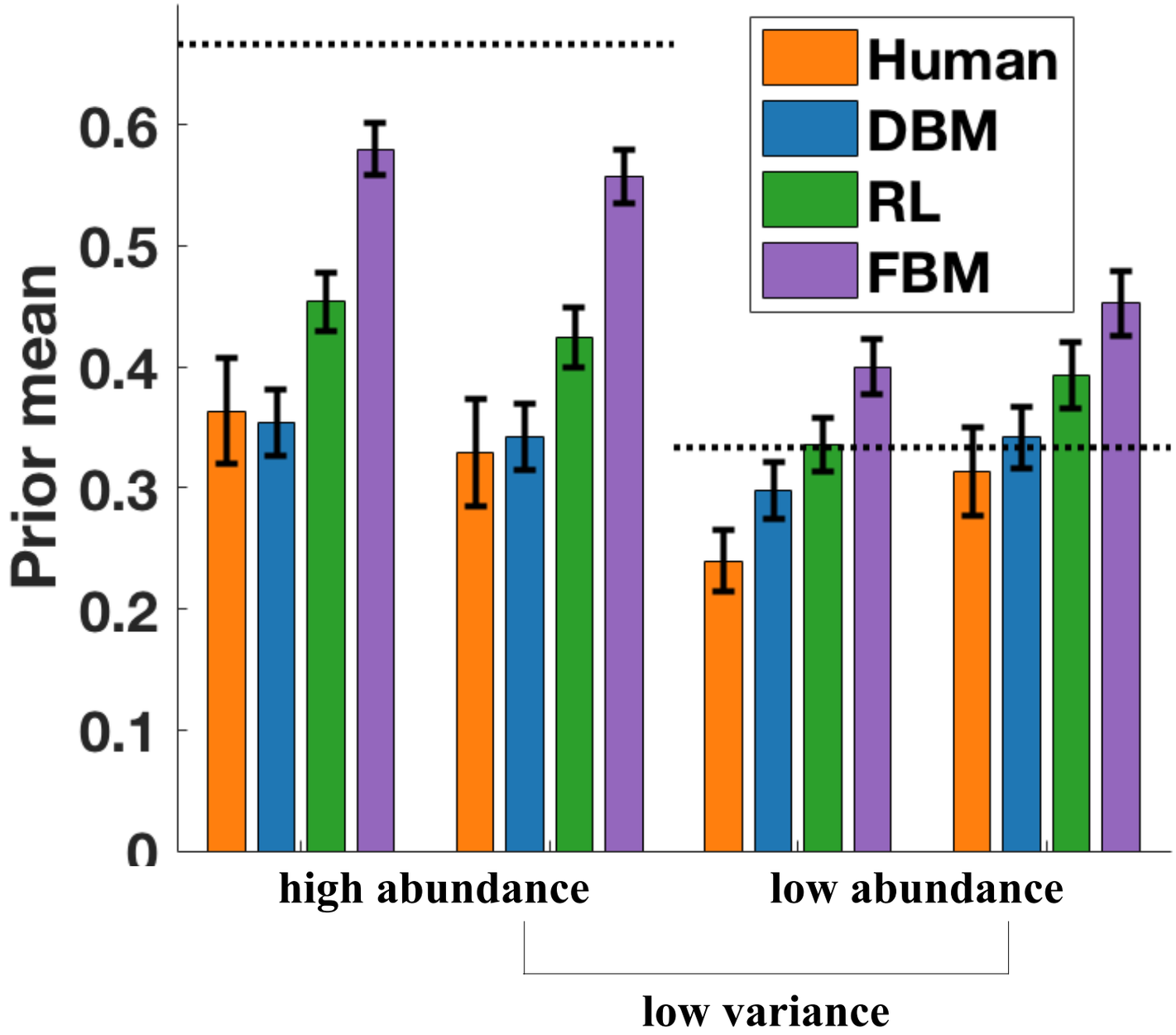


- 10-fold cross-validation (5 games held out)
- Best model: DBM+softmax
- DBM, RL > FBM: persistent sensitivity to recent reward history
- DBM > RL: both sensitive to recent reward history, but DBM also has persistent "prior bias"
- Softmax > knowledge gradient: KG explicitly computes value of exploration vs. exploitation (Zhang & Yu, 2013), but humans behave more like simpler softmax

Estimated Prior Mean of Reward

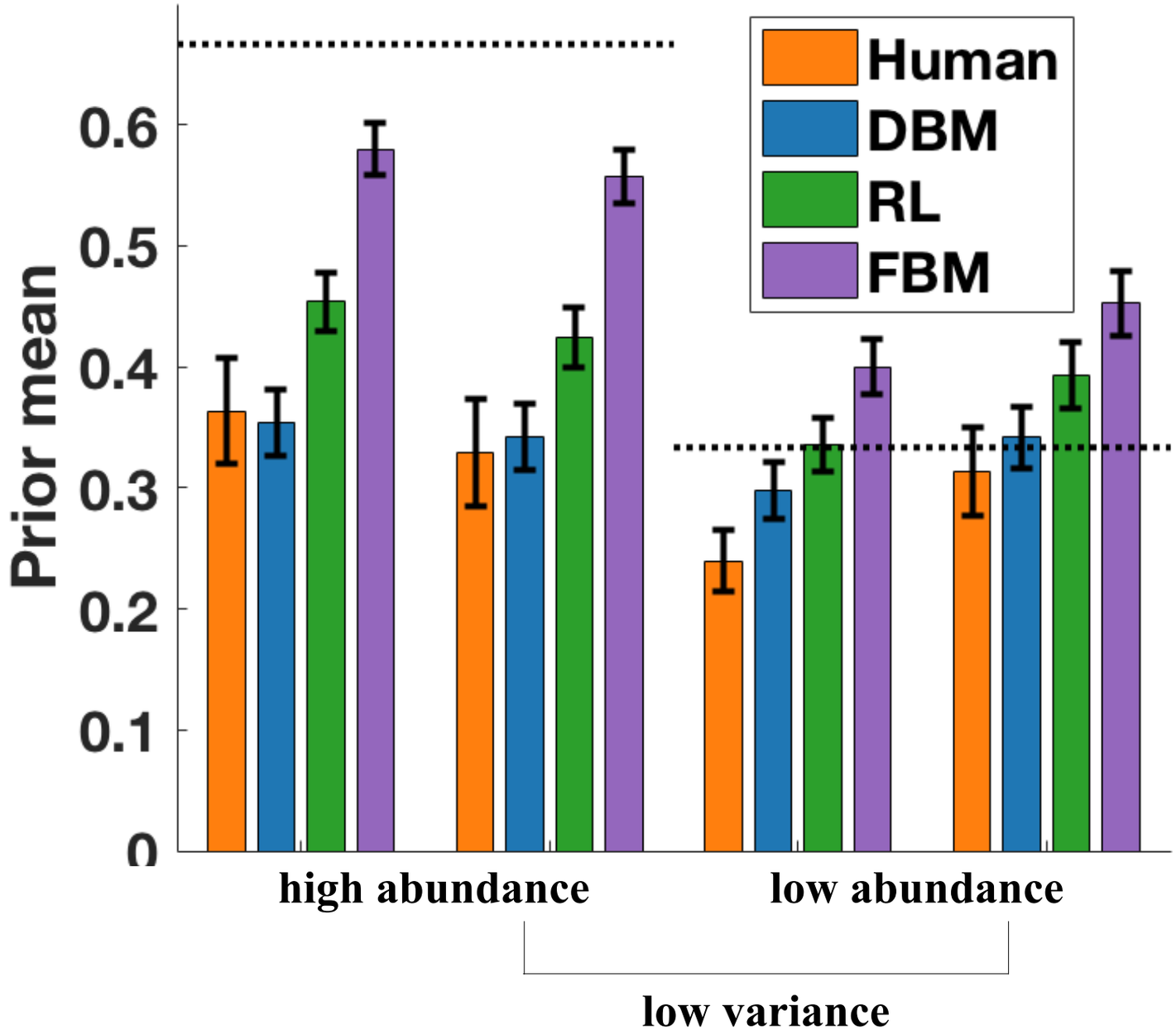


Estimated Prior Mean of Reward



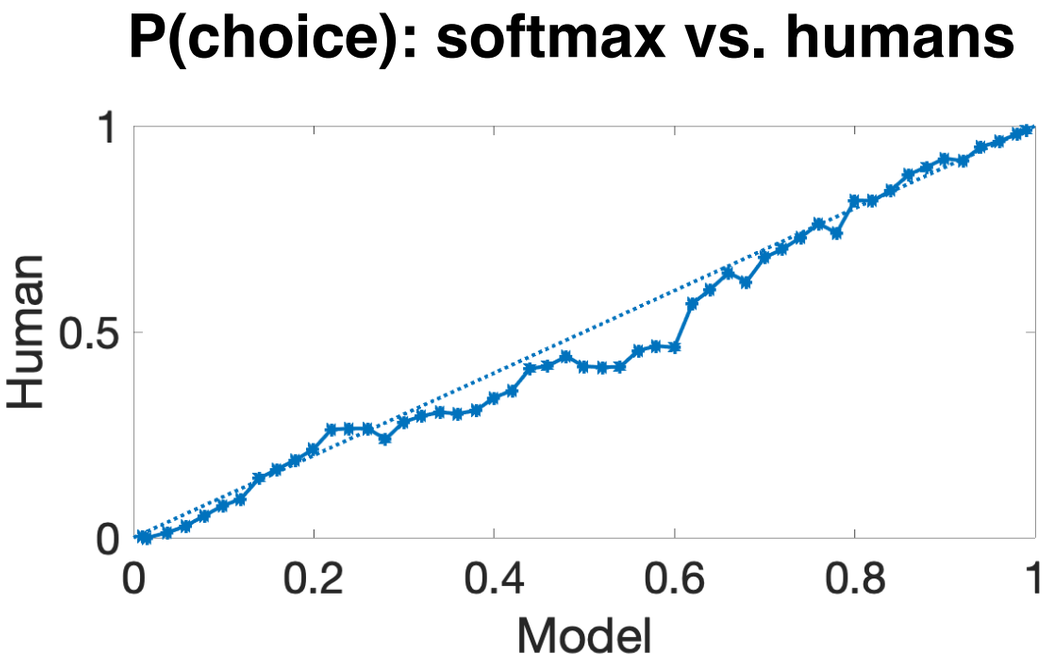
- DBM-estimated prior mean closely matches human self-report

Estimated Prior Mean of Reward

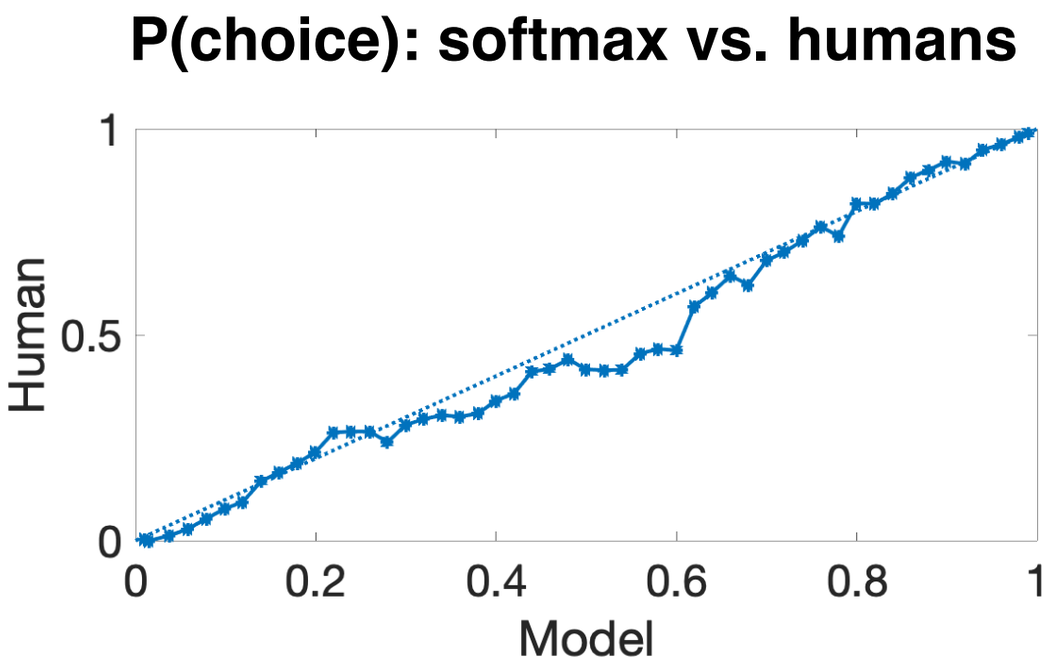


- DBM-estimated prior mean closely matches human self-report
- FBM & RL under-estimate magnitude of under-estimation

Choice of Decision Model

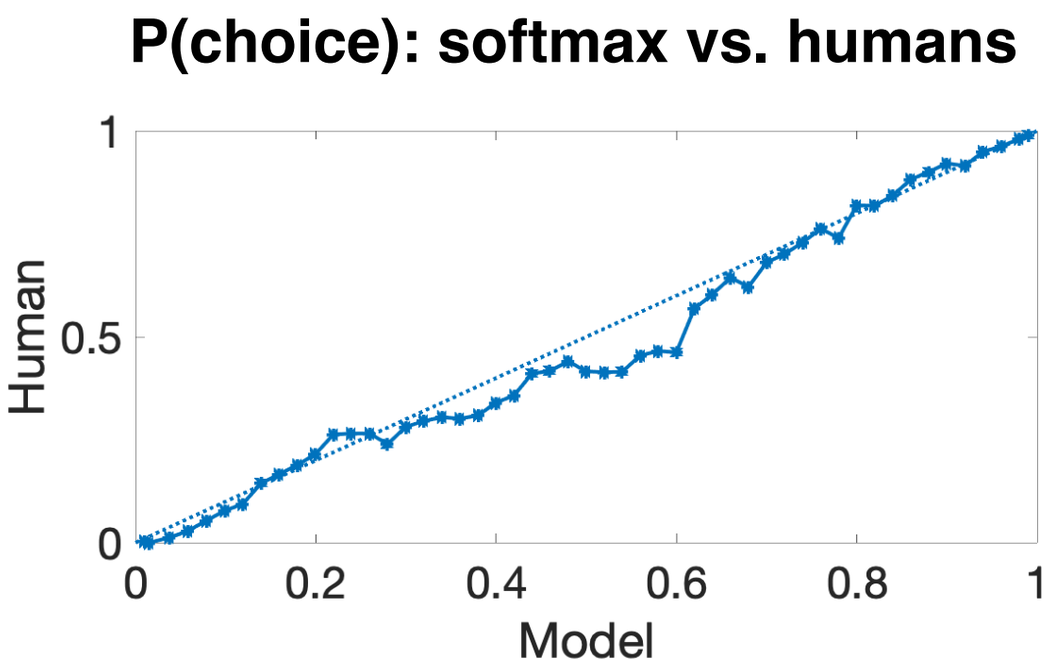


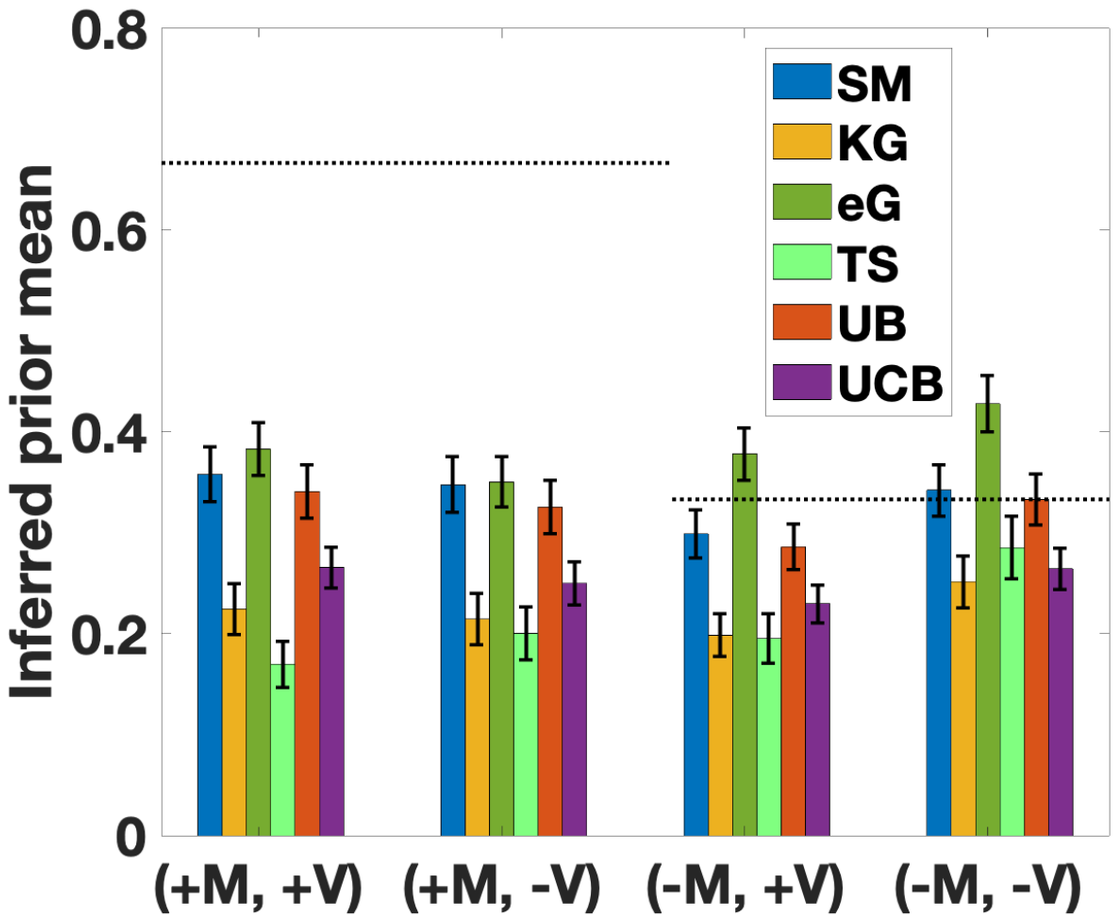
Choice of Decision Model



- Human choice probability closely correspond to (softmax) model predictions (binned data)

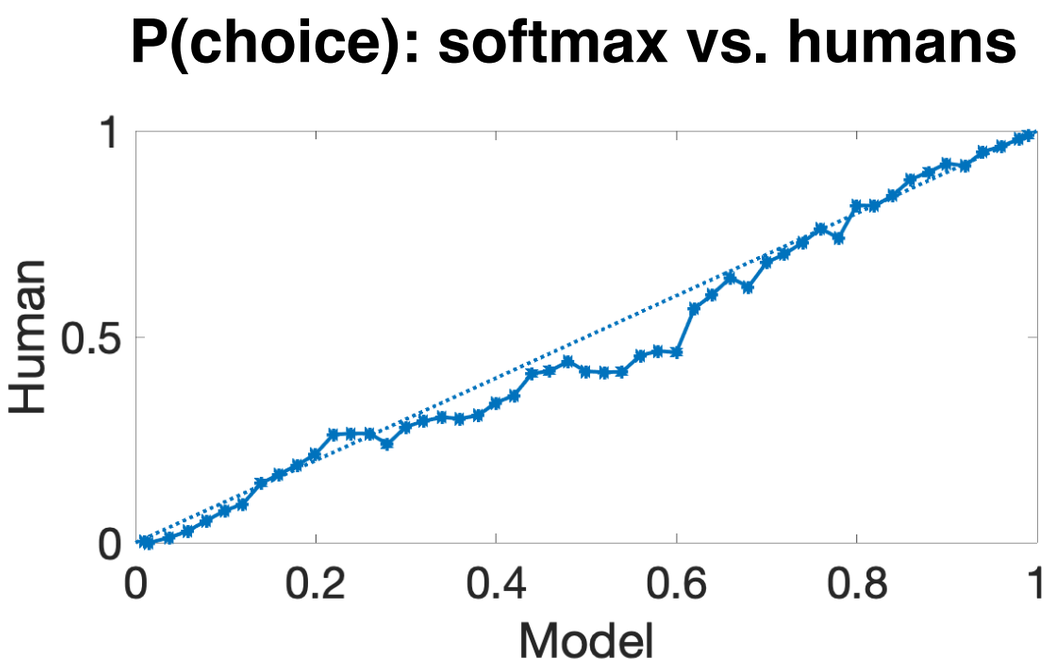
Choice of Decision Model

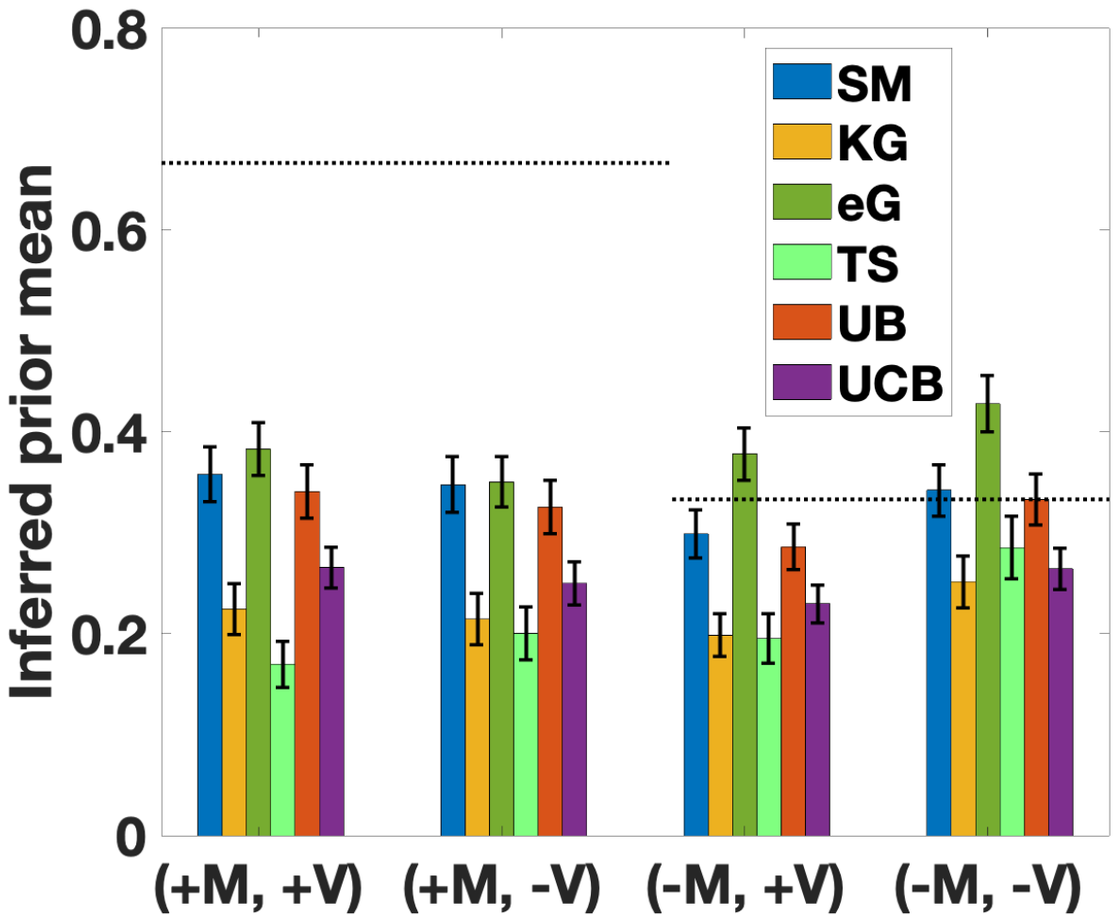




- Human choice probability closely correspond to (softmax) model predictions (binned data)
- In any case, estimated prior mean very similar across different decision policies

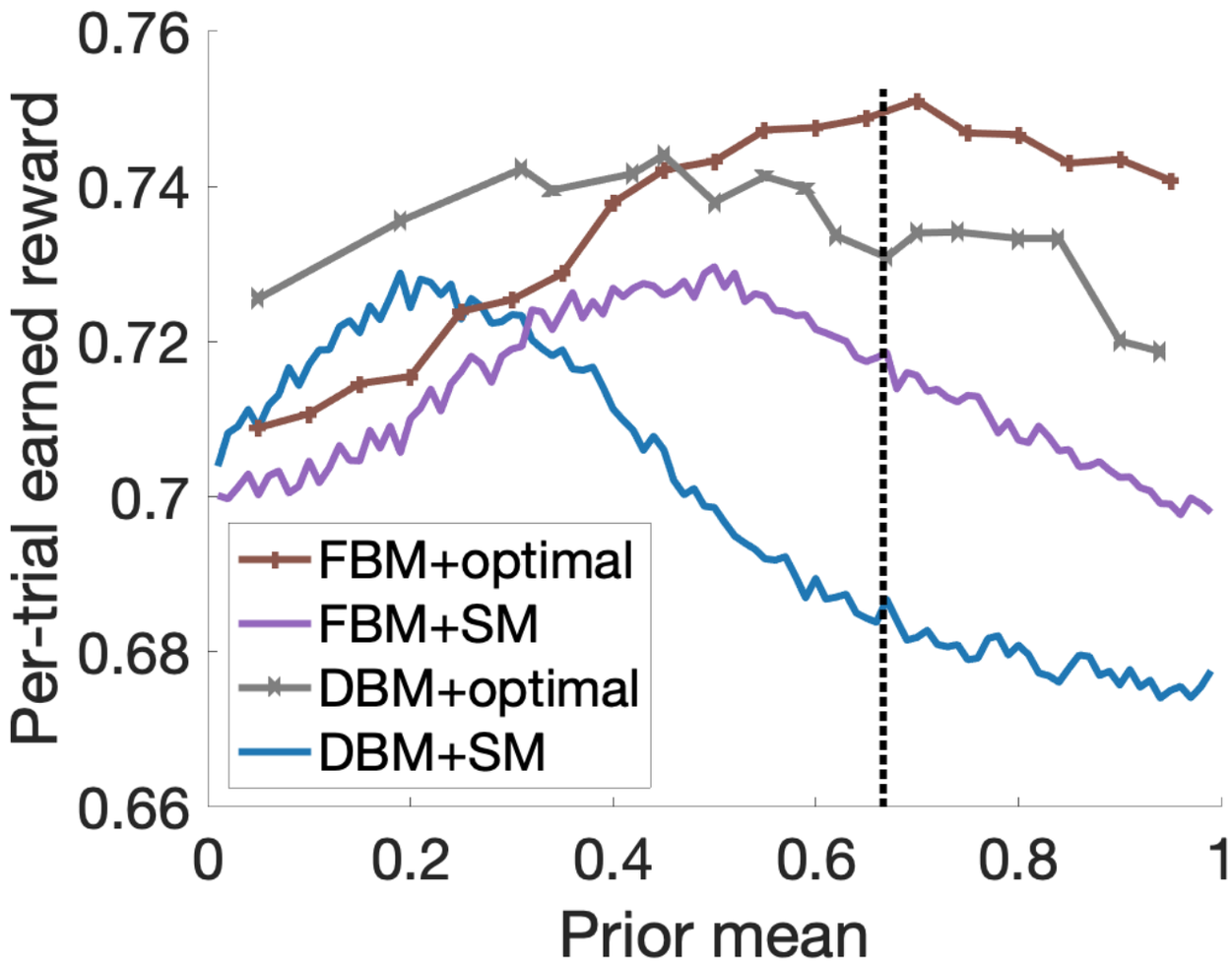
Choice of Decision Model



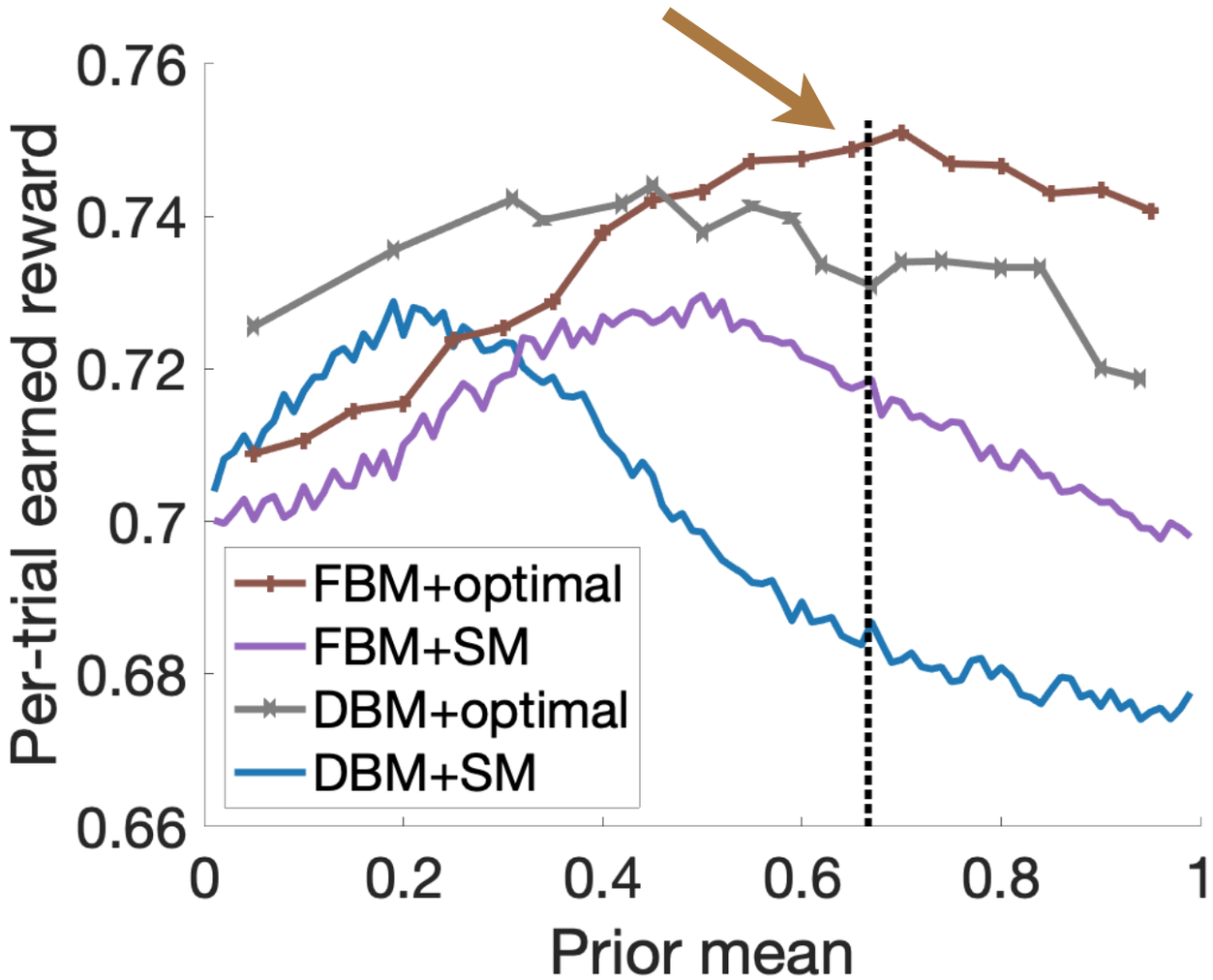


- Human choice probability closely correspond to (softmax) model predictions (binned data)
- In any case, estimated prior mean very similar across different decision policies
- Under-estimation independent of choice of decision model

Why Underestimate Reward Prior Mean?

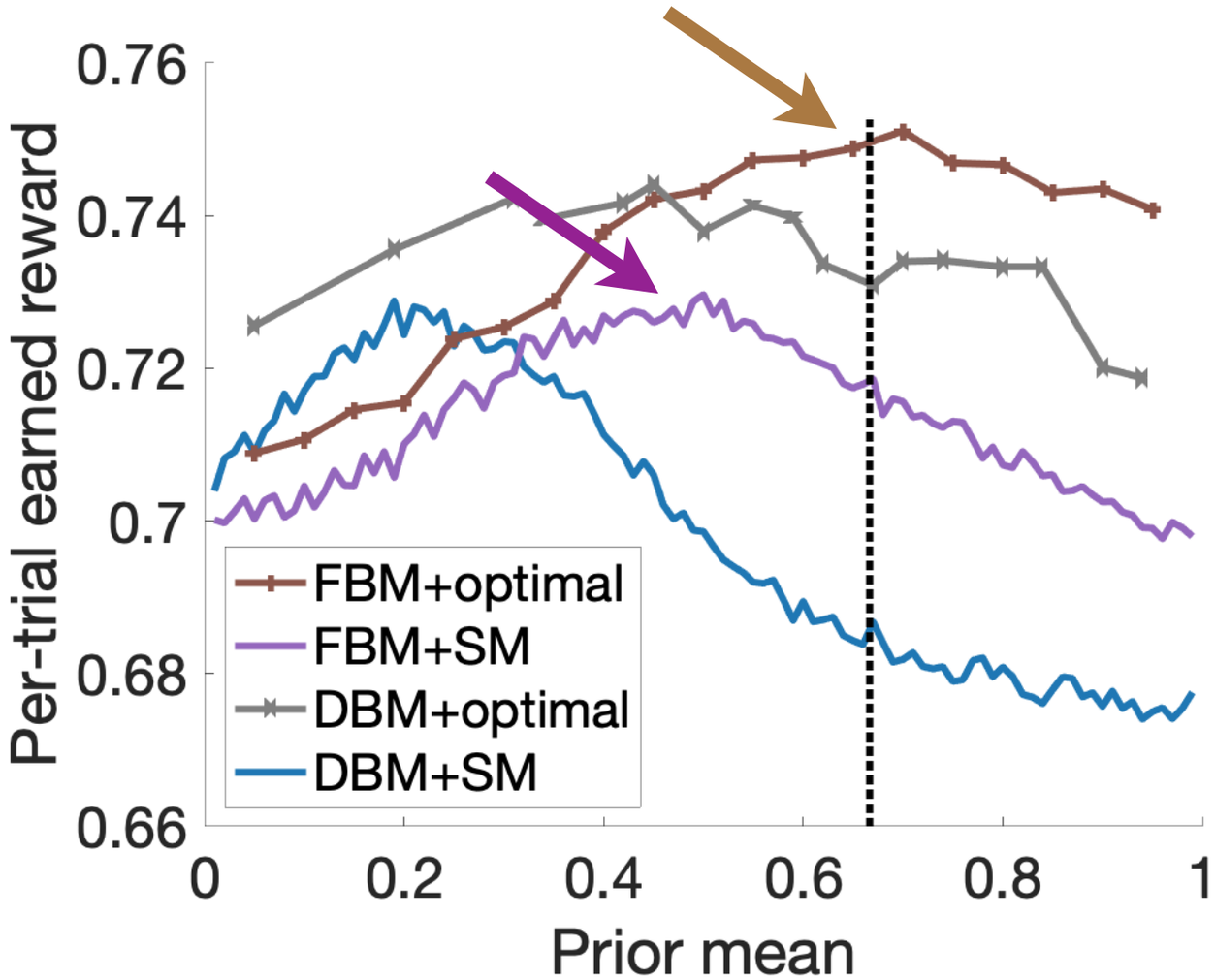


Why Underestimate Reward Prior Mean?



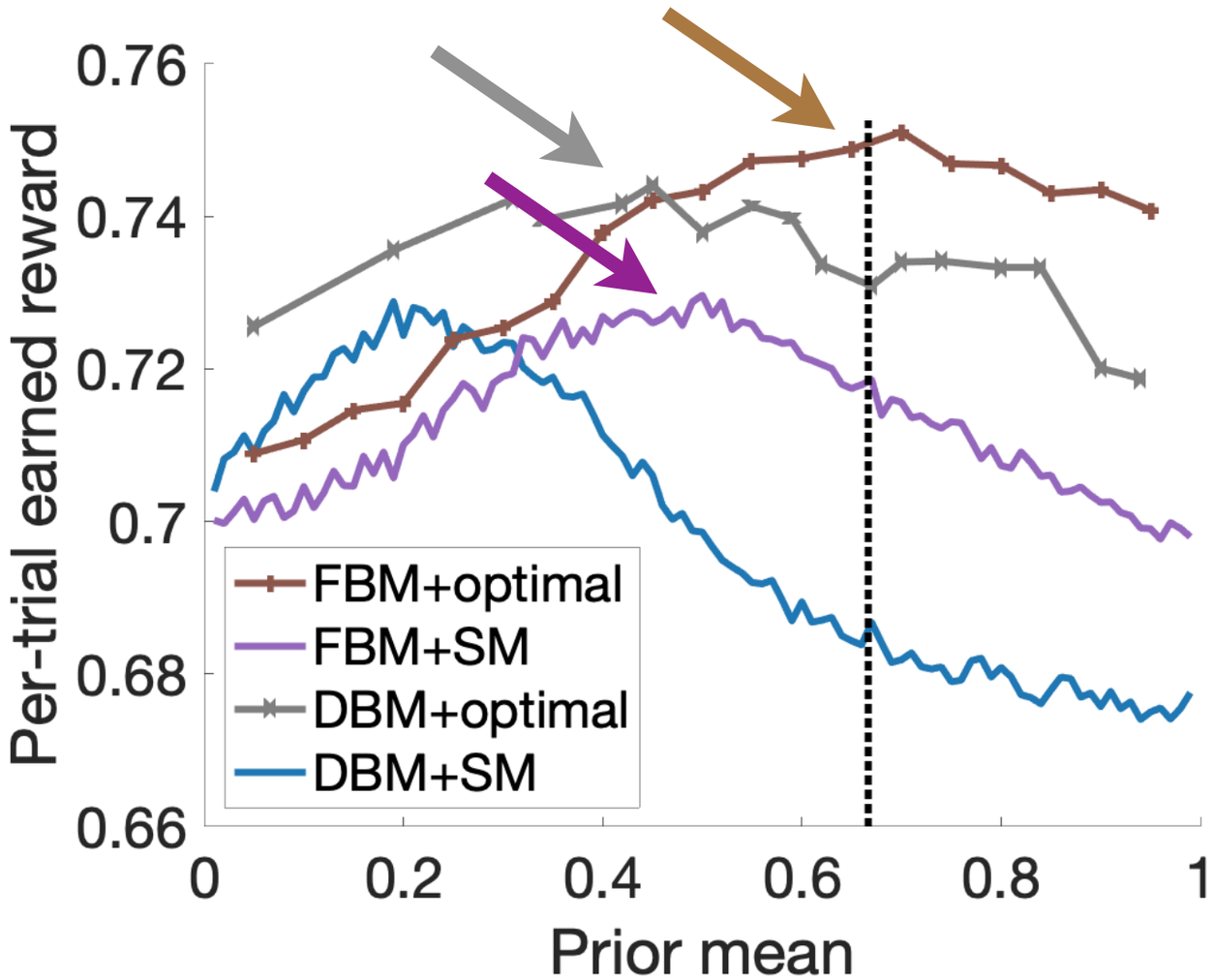
- Correct learning model (FBM) + optimal policy: true prior best

Why Underestimate Reward Prior Mean?



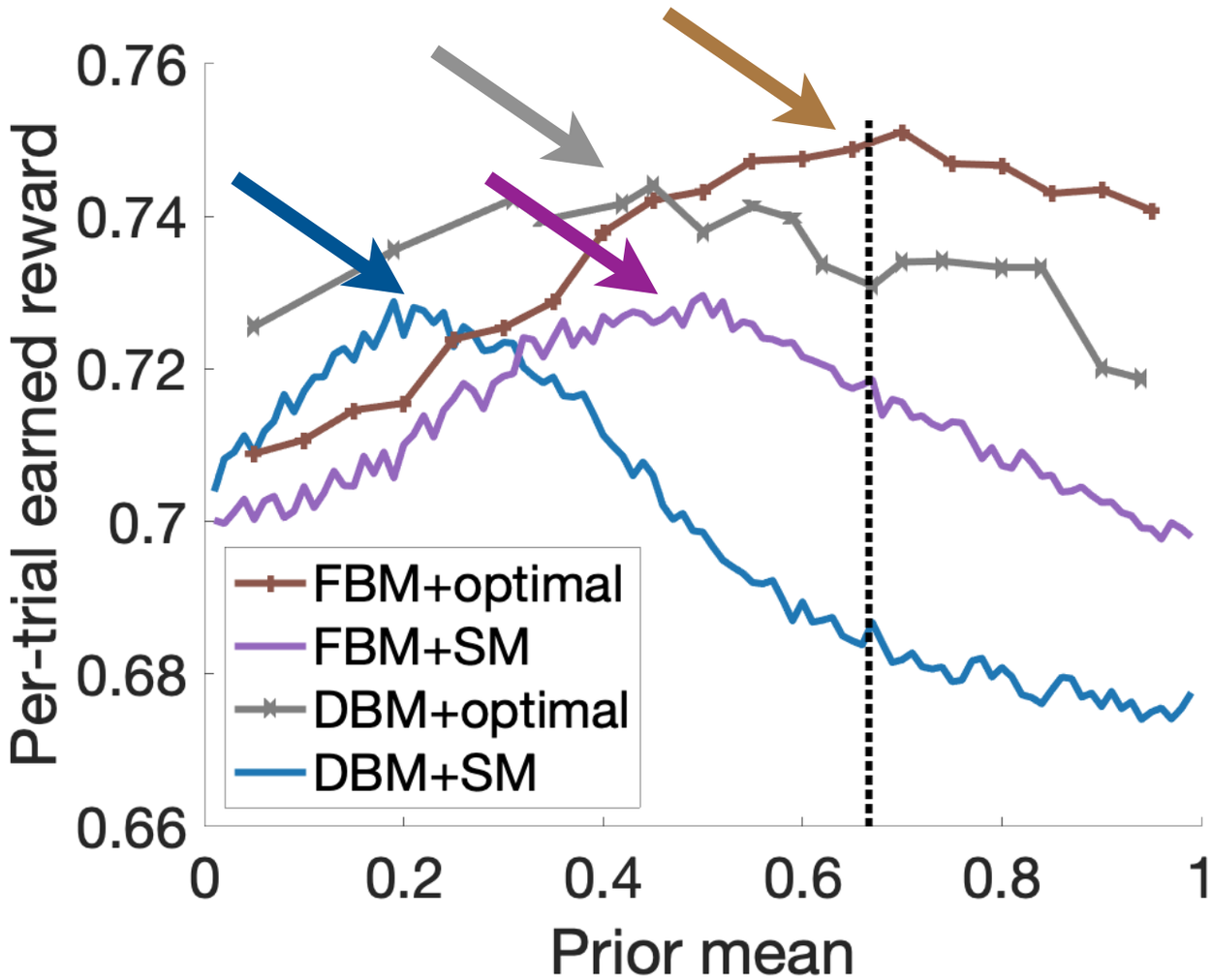
- Correct learning model (FBM) + optimal policy: true prior best
- Correct learning model (FBM) + simple policy (softmax): lower prior best

Why Underestimate Reward Prior Mean?



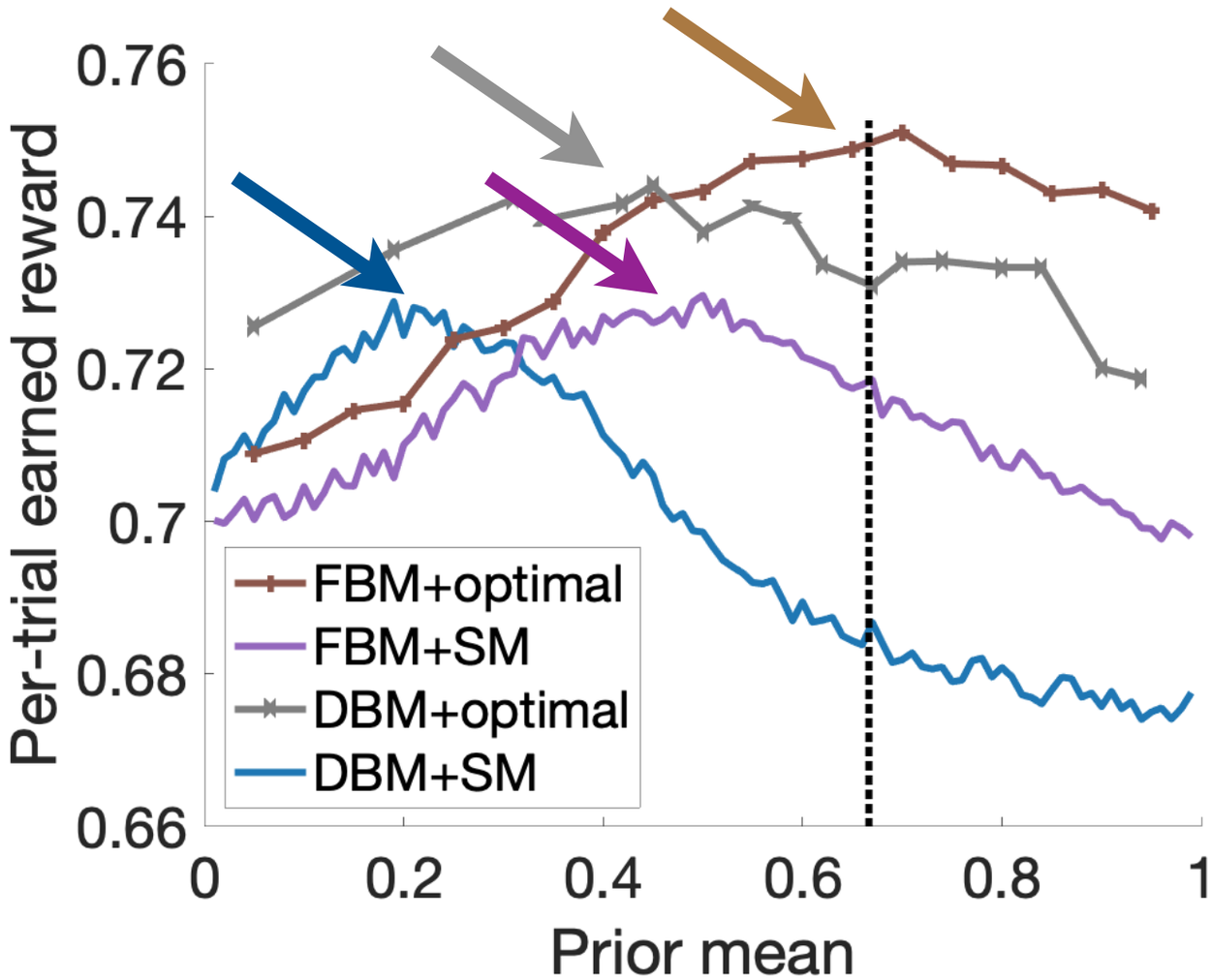
- Correct learning model (FBM) + optimal policy: true prior best
- Correct learning model (FBM) + simple policy (softmax): lower prior best
- Incorrect learning model (DBM) + optimal policy: lower prior best

Why Underestimate Reward Prior Mean?



- Correct learning model (FBM) + optimal policy: true prior best
- Correct learning model (FBM) + simple policy (softmax): lower prior best
- Incorrect learning model (DBM) + optimal policy: lower prior best
- Incorrect learning model (DBM) + simple policy (softmax): **even lower prior** best

Why Underestimate Reward Prior Mean?



- Correct learning model (FBM) + optimal policy: true prior best
- Correct learning model (FBM) + simple policy (softmax): lower prior best
- Incorrect learning model (DBM) + optimal policy: lower prior best
- Incorrect learning model (DBM) + simple policy (softmax): **even lower prior** best
- FBM vs. DBM: lower prior for DBM almost entirely compensates for false volatility assumption

Why does lower prior mean help?

Why does lower prior mean help?

- Optimal policy: never shifts away after a win

Why does lower prior mean help?

- Optimal policy: never shifts away after a win
- Softmax introduces some probability of shifting away

Why does lower prior mean help?

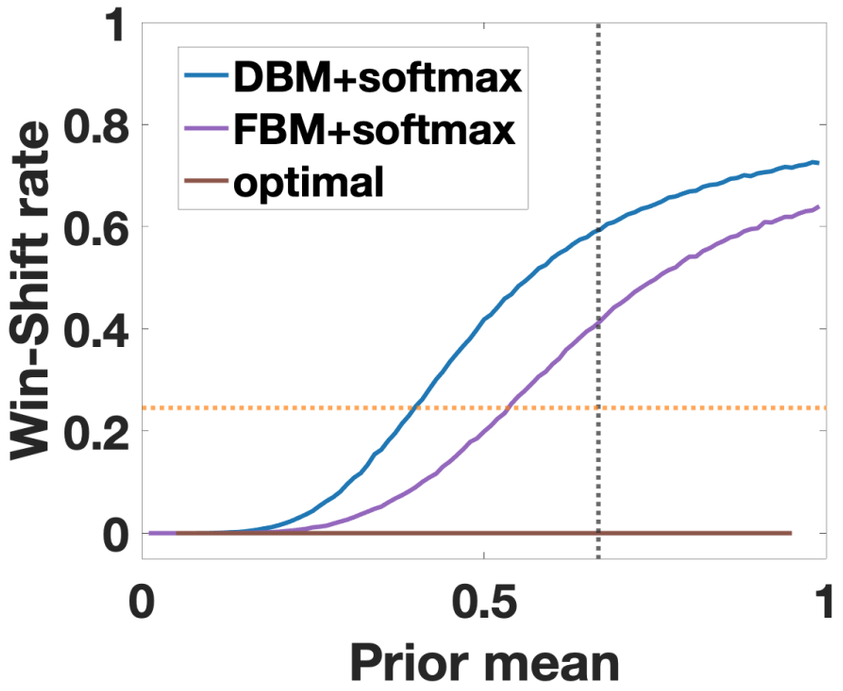
- Optimal policy: never shifts away after a win
- Softmax introduces some probability of shifting away
- Assumption of volatility (DBM) “leaks away” confidence in a good arm (and increases chance of formerly unappealing unchosen arm becoming “good” now)

Why does lower prior mean help?

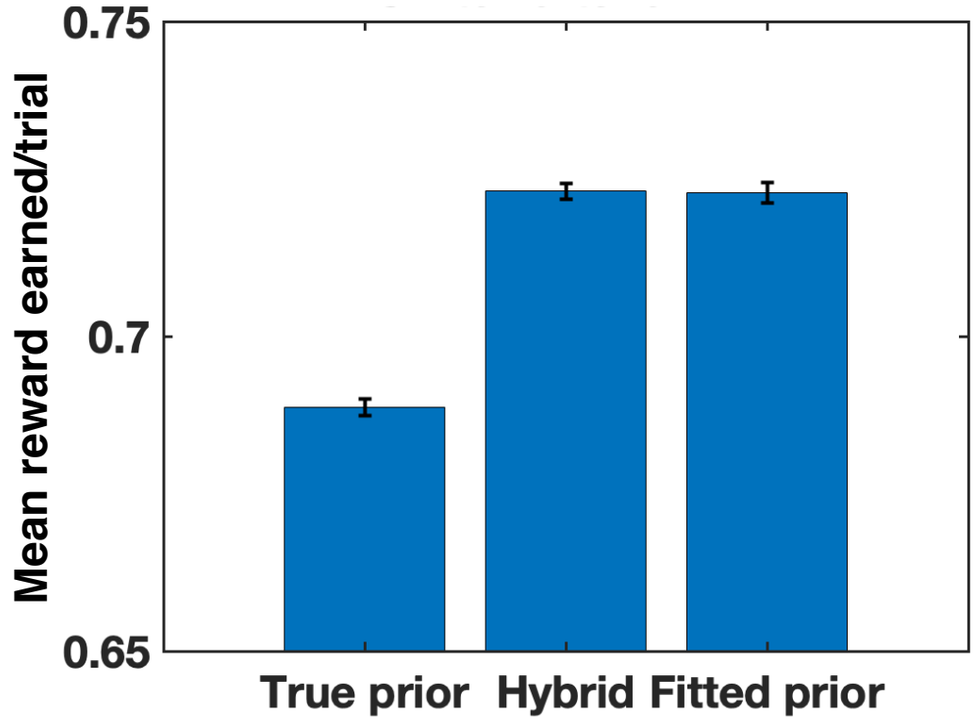
- Optimal policy: never shifts away after a win
- Softmax introduces some probability of shifting away
- Assumption of volatility (DBM) "leaks away" confidence in a good arm (and increases chance of formerly unappealing unchosen arm becoming "good" now)
- Lower prior for **unchosen** arms: discourages shifting!

Why does lower prior mean help?

Lower prior mean ⇒ ↓ P(win-shift)

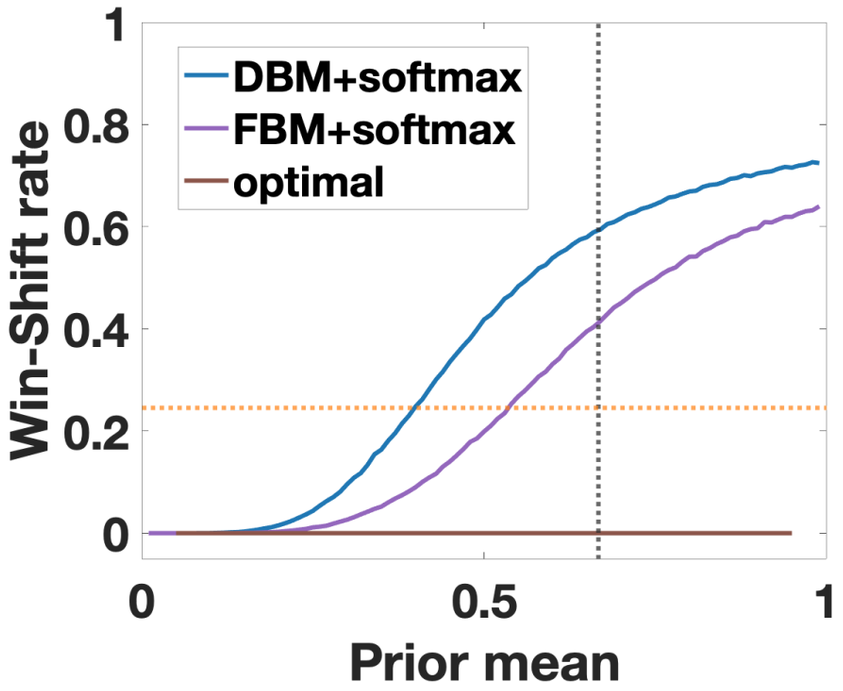


Lower prior mean ⇒ ↑ performance

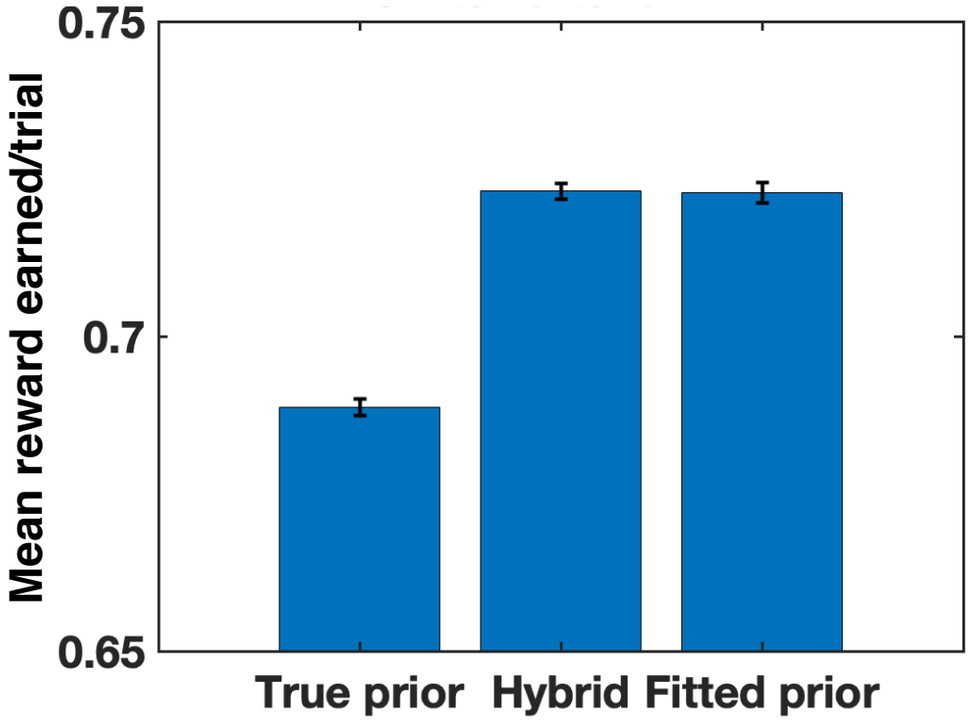


Why does lower prior mean help?

Lower prior mean ⇒ ↓ P(win-shift)



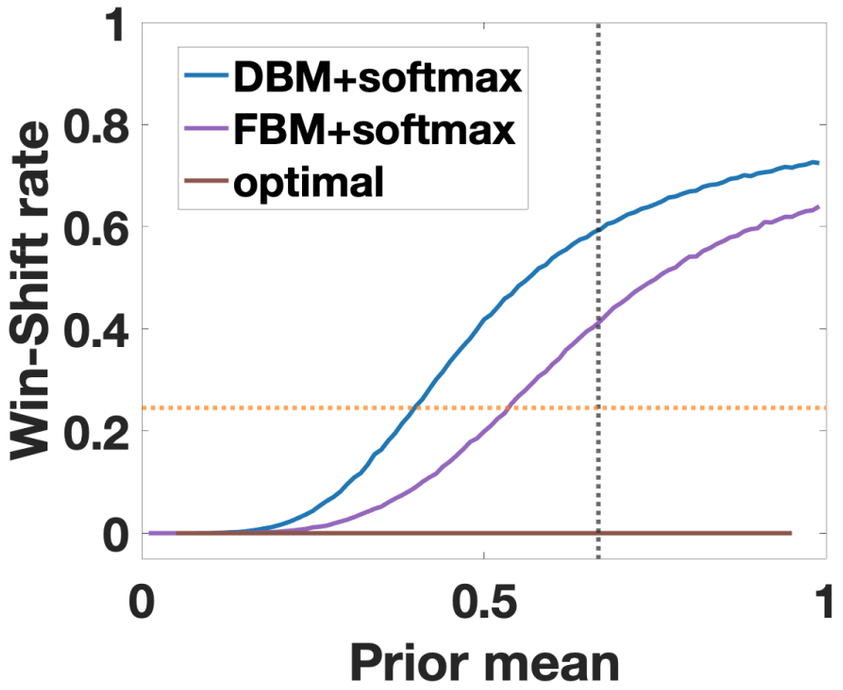
Lower prior mean ⇒ ↑ performance



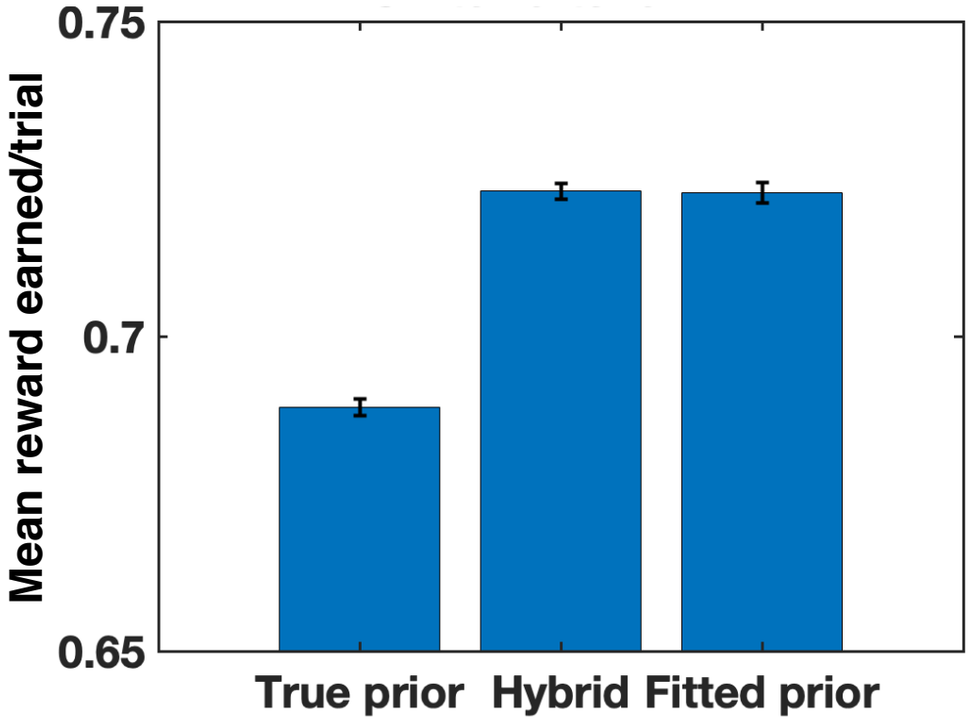
Simulation results

Why does lower prior mean help?

Lower prior mean ⇒ ↓ P(win-shift)



Lower prior mean ⇒ ↑ performance

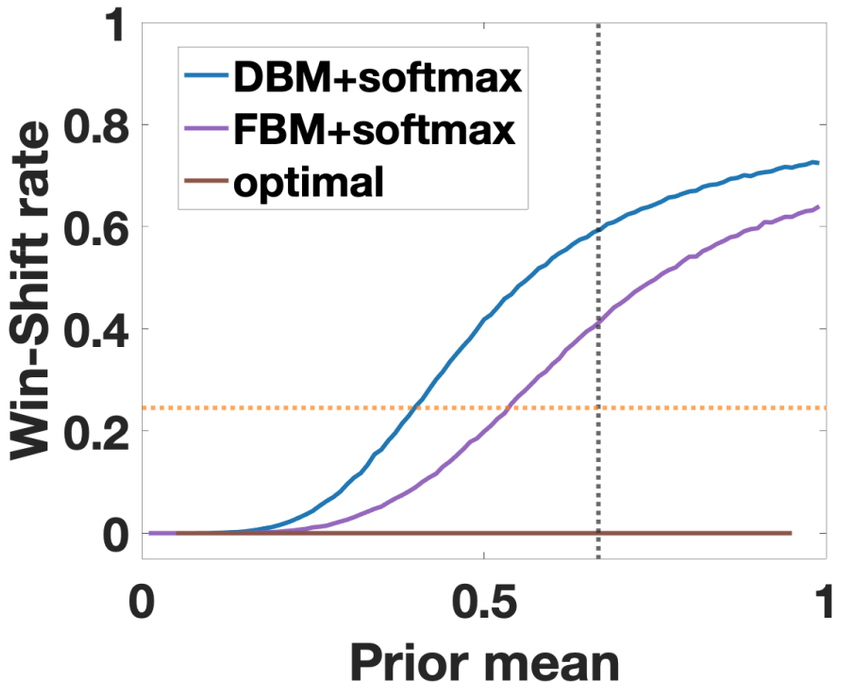


Simulation results

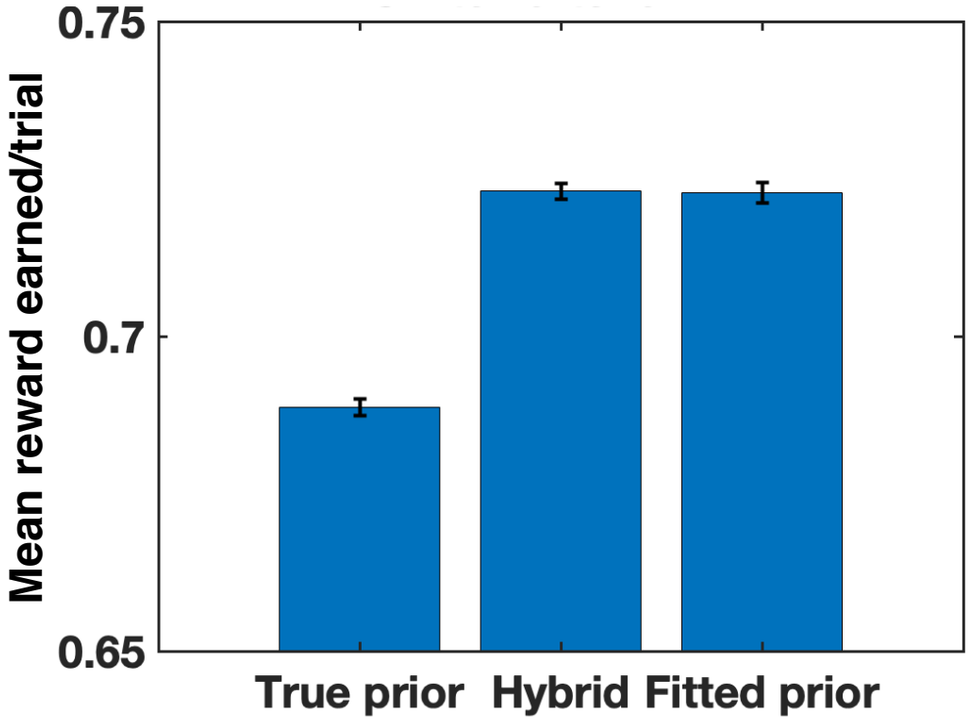
- Lower prior mean decreases P(win-shift); optimal policy has P(win-shift)=0

Why does lower prior mean help?

Lower prior mean ⇒ ↓ P(win-shift)



Lower prior mean ⇒ ↑ performance

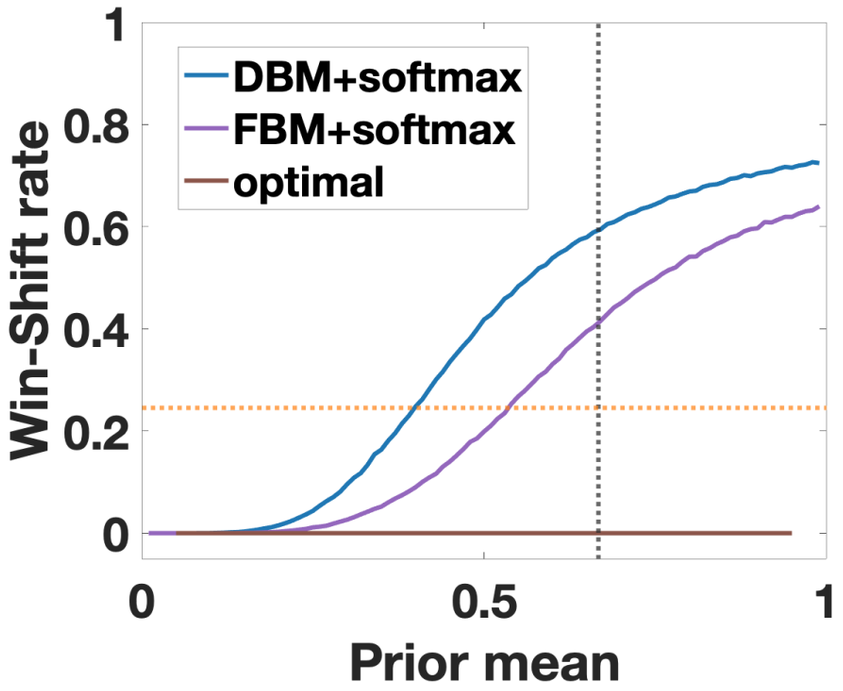


Simulation results

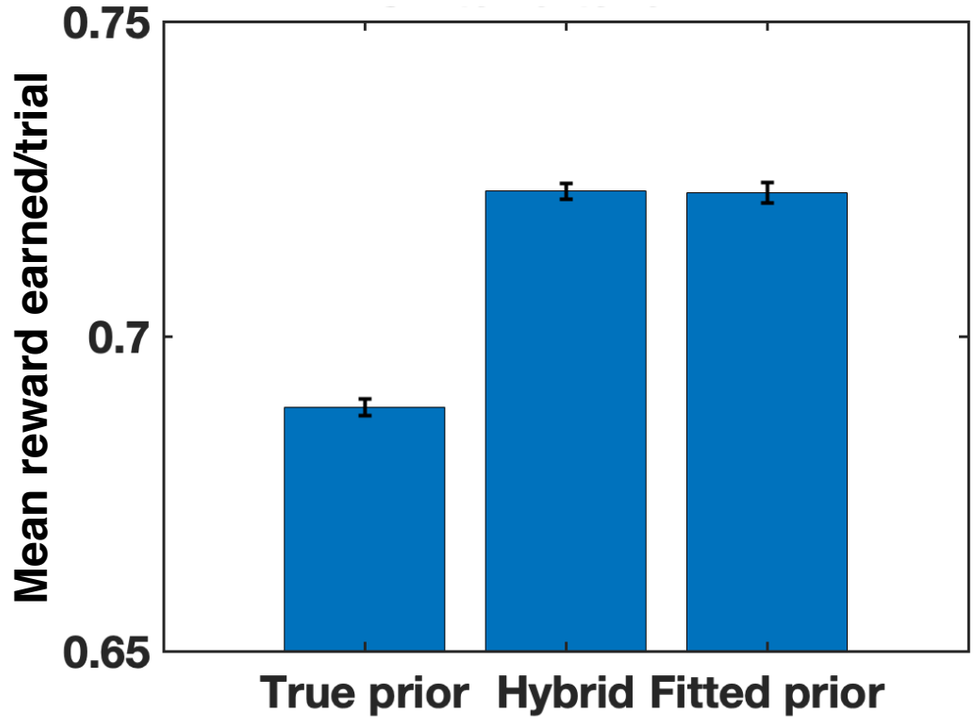
- Lower prior mean decreases P(win-shift); optimal policy has P(win-shift)=0
- Replacing true prior with fitted prior after each **win**: entirely captures benefit of lower prior

Why does lower prior mean help?

Lower prior mean ⇒ ↓ P(win-shift)



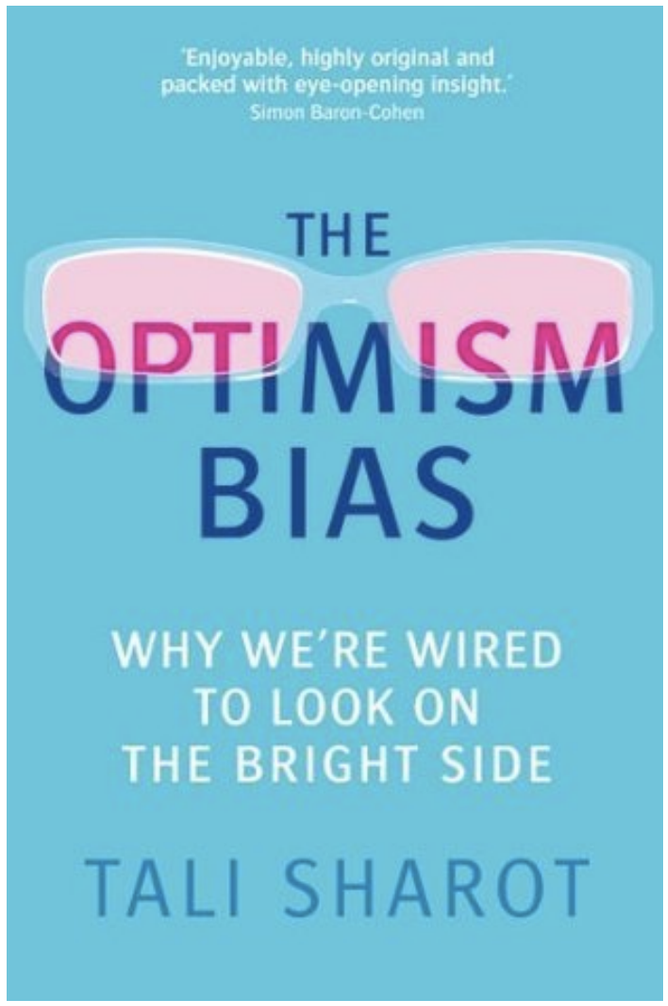
Lower prior mean ⇒ ↑ performance



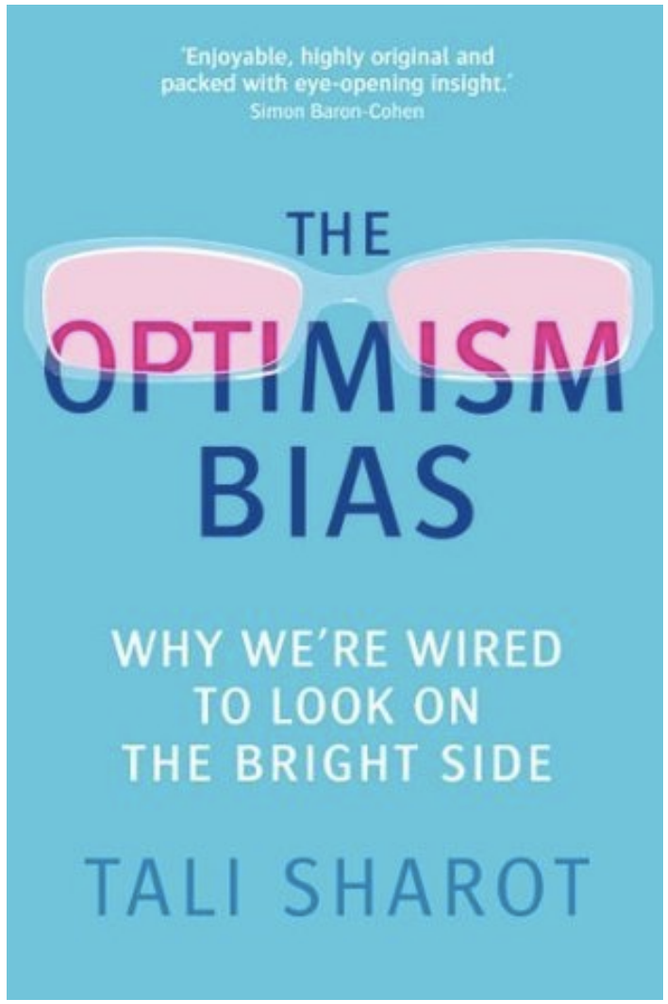
Simulation results

- Lower prior mean decreases P(win-shift); optimal policy has P(win-shift)=0
- Replacing true prior with fitted prior after each **win**: entirely captures benefit of lower prior
- Lower prior mean affects **unchosen** arm more than **chosen** arm
  ⇒ **relative optimism** about chosen arm

And What of Optimism Bias?



And What of Optimism Bias?



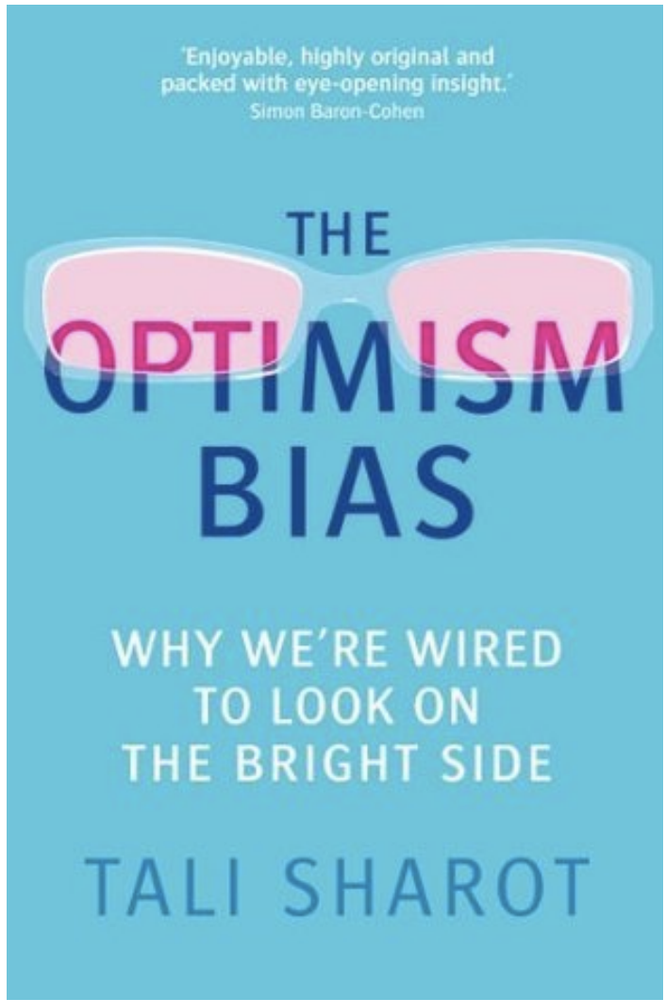
nature neuroscience

How unrealistic optimism is maintained in the face of reality

Tali Sharot[1,4], Christoph W Korn[2–4] & Raymond J Dolan[1]

Unrealistic optimism is a pervasive human trait that influences domains ranging from personal relationships to politics and finance. How people maintain unrealistic optimism, despite frequently encountering information that challenges those biased beliefs, is unknown. We examined this question and found a marked asymmetry in belief updating. Participants updated their beliefs more in response to information that was better than expected than to information that was worse. This selectivity was mediated by a relative failure to code for errors that should reduce optimism. Distinct regions of the prefrontal cortex tracked estimation errors when those called for positive update, both in individuals who scored high and low on trait optimism. However, highly optimistic individuals exhibited reduced tracking of estimation errors that called for negative update in right inferior prefrontal gyrus. These findings indicate that optimism is tied to a selective update failure and diminished neural coding of undesirable information regarding the future.

And What of Optimism Bias?



nature neuroscience

How unrealistic optimism is maintained in the face of reality

Tali Sharot[1,4], Christoph W Korn[2–4] & Raymond J Dolan[1]

Unrealistic optimism is a pervasive human trait that influences domains ranging from personal relationships to politics and finance. How people maintain unrealistic optimism, despite frequently encountering information that challenges those biased beliefs, is unknown. We examined this question and found a marked asymmetry in belief updating. Participants updated their beliefs more in response to information that was better than expected than to information that was worse. This selectivity was mediated by a relative failure to code for errors that should reduce optimism. Distinct regions of the prefrontal cortex tracked estimation errors when those called for positive update, both in individuals who scored high and low on trait optimism. However, highly optimistic individuals exhibited reduced tracking of estimation errors that called for negative update in right inferior prefrontal gyrus. These findings indicate that optimism is tied to a selective update failure and diminished neural coding of undesirable information regarding the future.

nature human behaviour

ARTICLES

PUBLISHED: 20 MARCH 2017 | VOLUME: 1 | ARTICLE NUMBER: 0067

Behavioural and neural characterization of optimistic reinforcement learning

Germain Lefebvre[1,2], Maël Lebreton[3,4], Florent Meyniel[5], Sacha Bourgeois-Gironde[2,6] and Stefano Palminteri[1,7]*

When forming and updating beliefs about future life outcomes, people tend to consider good news and to disregard bad news. This tendency is assumed to support the optimism bias. Whether this learning bias is specific to 'high-level' abstract belief update or a particular expression of a more general 'low-level' reinforcement learning process is unknown. Here we report evidence in favour of the second hypothesis. In a simple instrumental learning task, participants incorporated better-than-expected outcomes at a higher rate than worse-than-expected ones. In addition, functional imaging indicated that inter-individual difference in the expression of optimistic update corresponds to enhanced prediction error signalling in the reward circuitry. Our results constitute a step towards the understanding of the genesis of optimism bias at the neurocomputational level.

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017)

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017)
  - 2-armed bandit task, 2 x 2 design: mean x variance

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017)
  - 2-armed bandit task, 2 x 2 design: mean x variance
  - 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017)
  - 2-armed bandit task, 2 x 2 design: mean x variance
  - 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$
  - Could this actually be explained by reward rate under-estimation?

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017)
  - 2-armed bandit task, 2 x 2 design: mean x variance
  - 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$
  - Could this actually be explained by reward rate under-estimation?
    - apparent contradiction: devaluation of chosen arm = pessimism

And What of Optimism Bias?

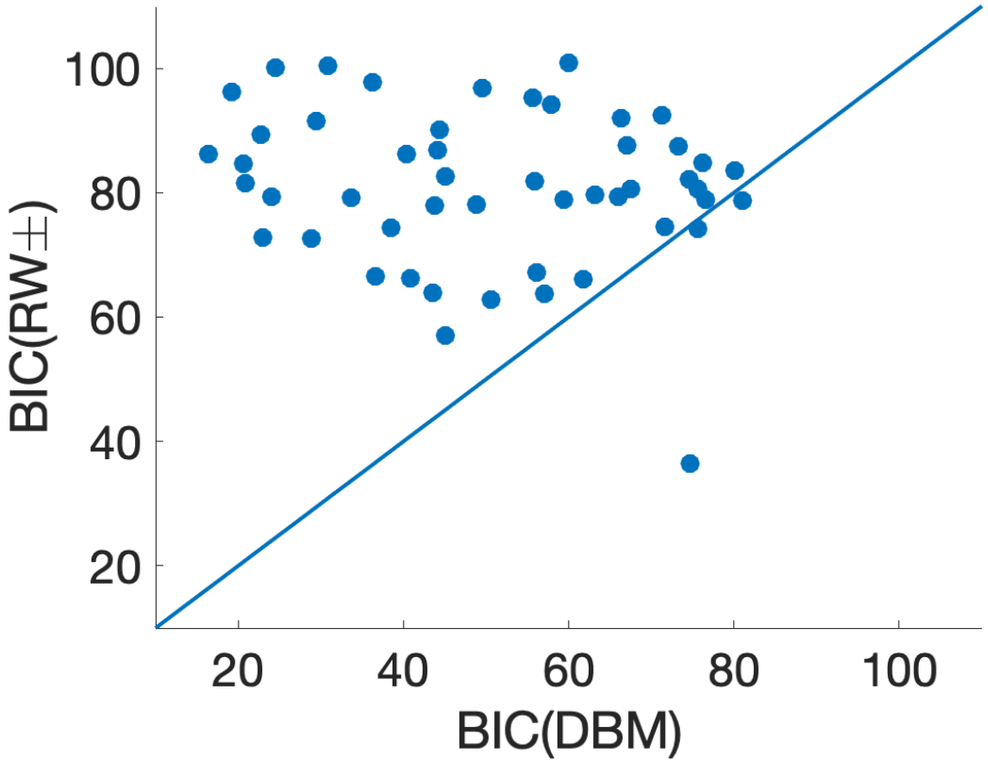
(Zhou, Guo, Yu, *Cogsci*, 2020)

$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017)
  - 2-armed bandit task, 2 x 2 design: mean x variance
  - 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$
  - Could this actually be explained by reward rate under-estimation?
    - apparent contradiction: devaluation of chosen arm = pessimism
    - However: lower prior mean affects unchosen arm more than chosen arm $\Rightarrow$ relative optimism about chosen arm

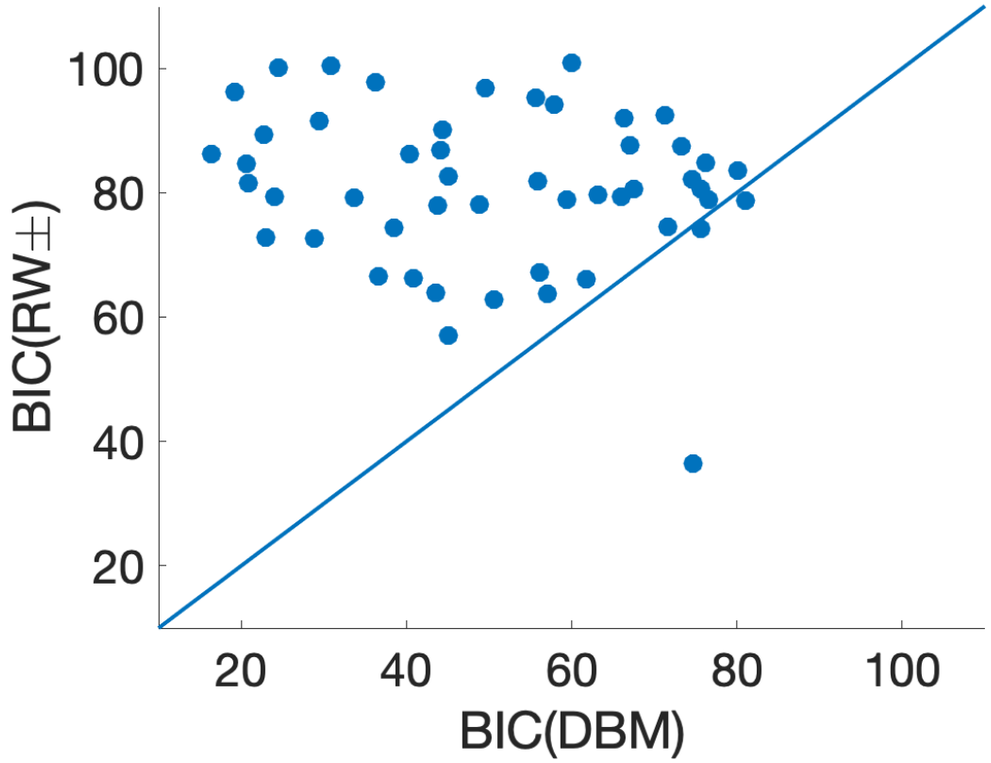
And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)



And What of Optimism Bias?

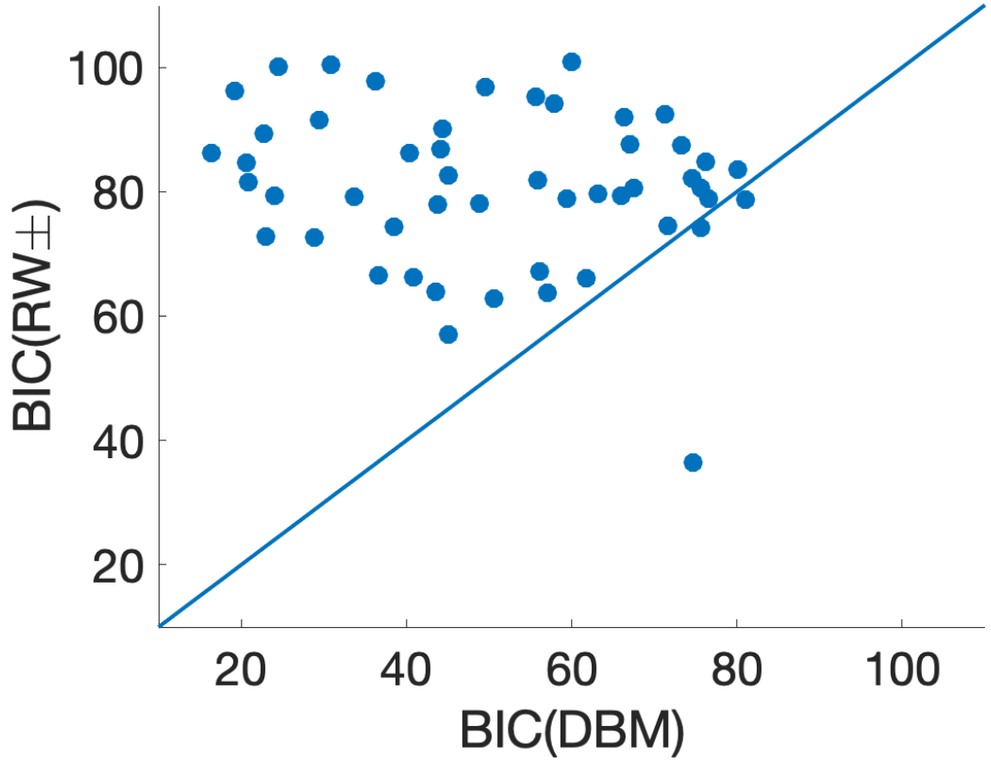
(Zhou, Guo, Yu, *Cogsci*, 2020)

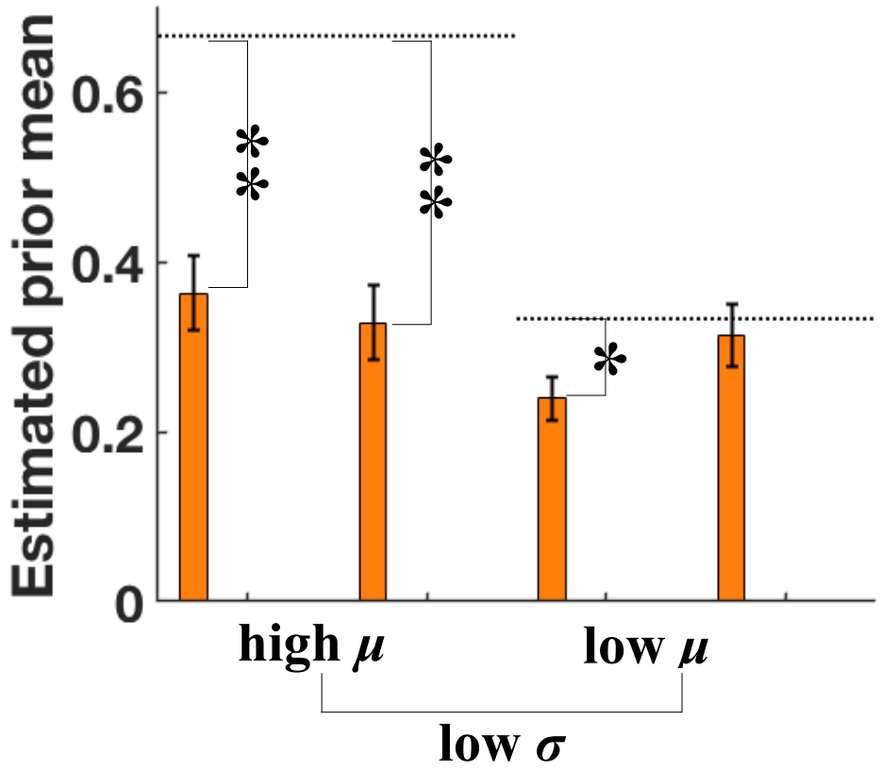


- DBM accounts for subjects' choices much better than RW±

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

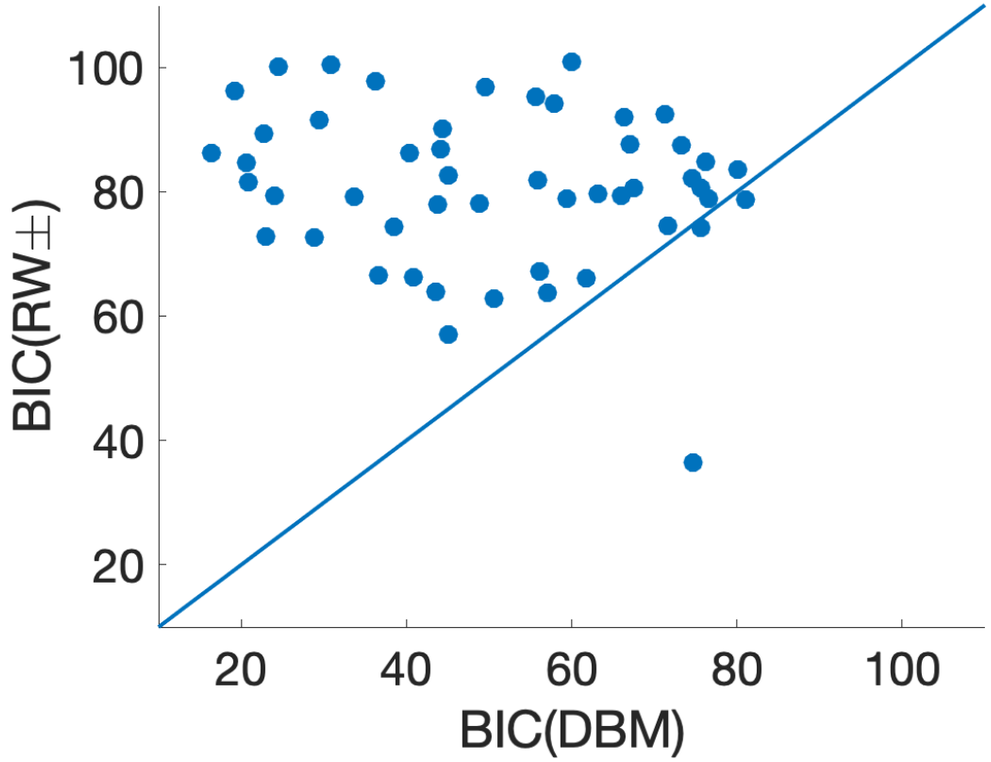


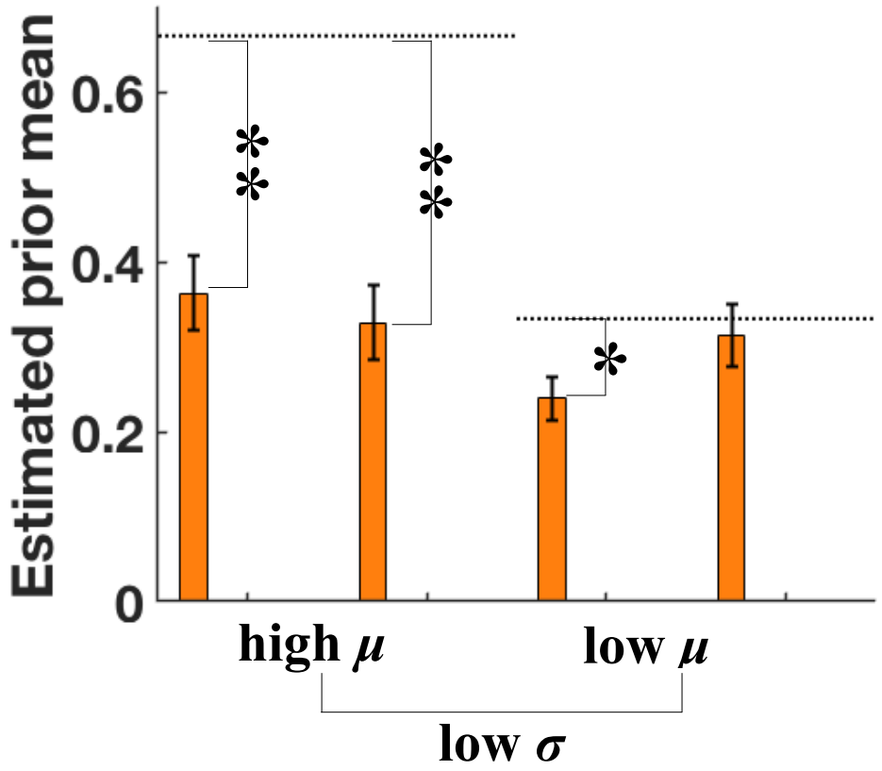


- DBM accounts for subjects' choices much better than RW±
- Estimated prior mean: 0.32 (±0.19), very similar to previous experiment — universal?

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

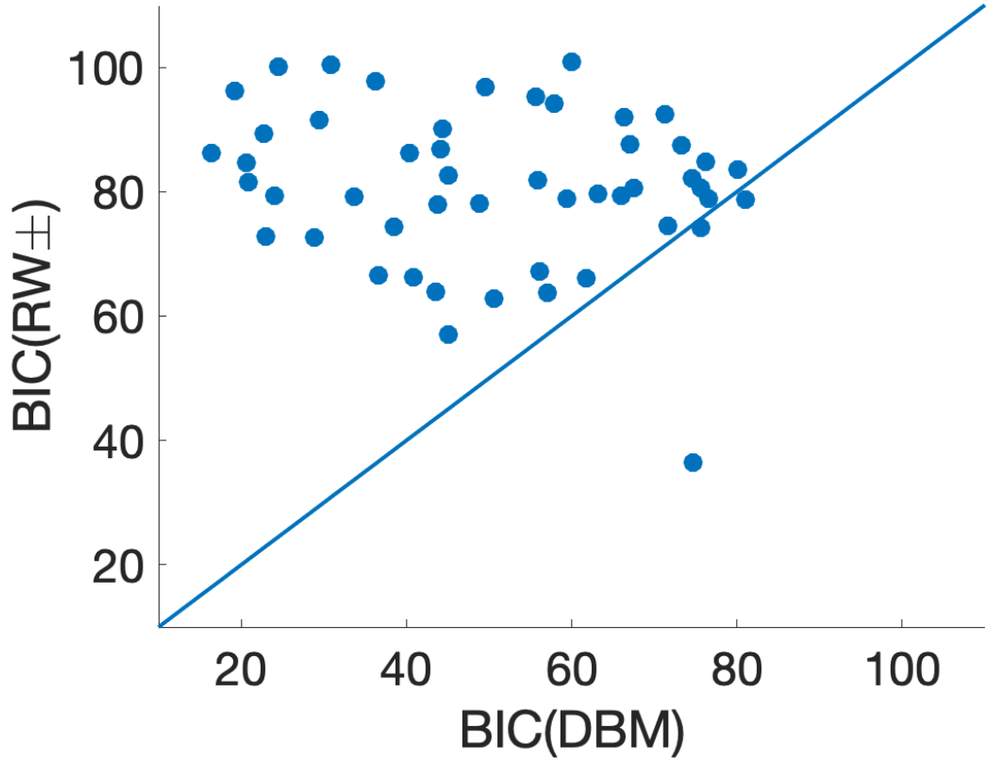


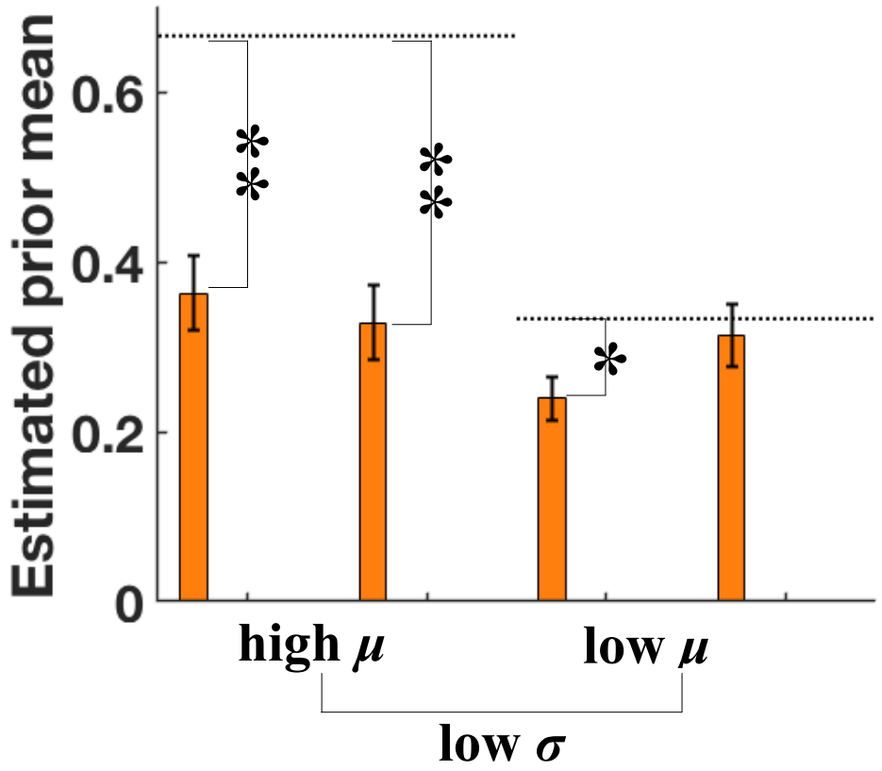


- DBM accounts for subjects' choices much better than RW±
- Estimated prior mean: 0.32 (±0.19), very similar to previous experiment — universal?
- Synthetic data generated from DBM used to fit RW± → $\varepsilon^+ > \varepsilon^-$ (not shown)

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

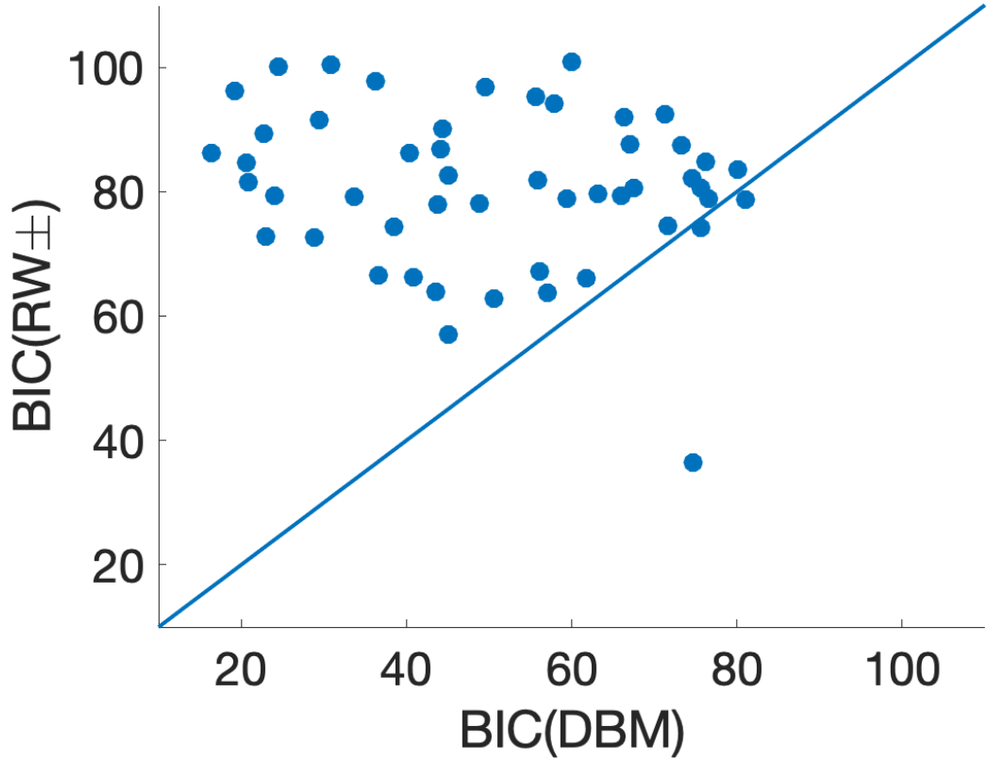


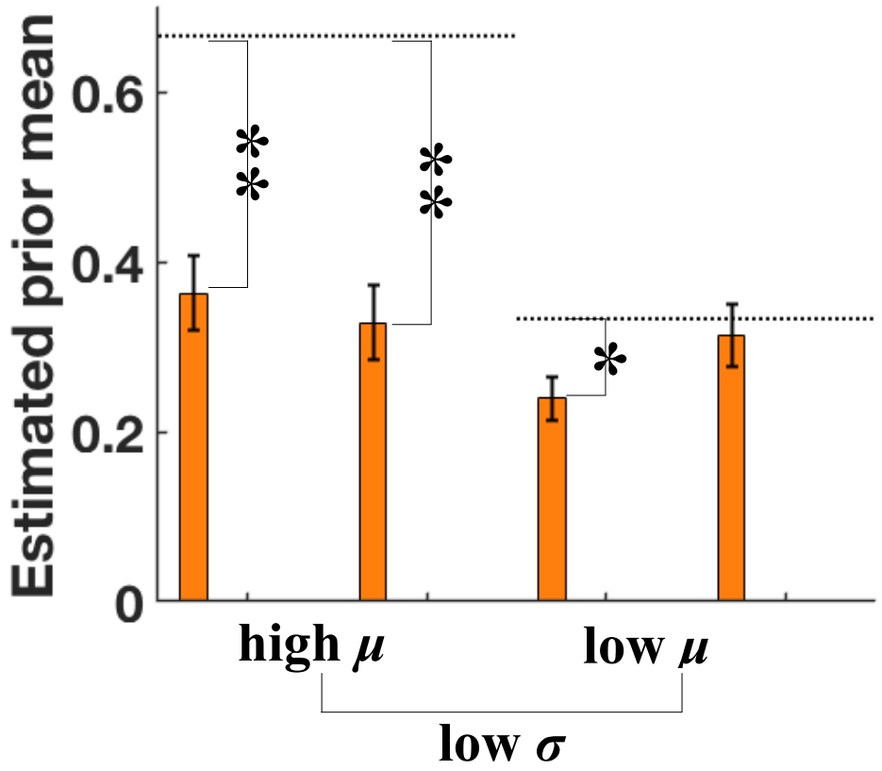


- DBM accounts for subjects' choices much better than RW±
- Estimated prior mean: 0.32 (±0.19), very similar to previous experiment — universal?
- Synthetic data generated from DBM used to fit RW± → $\varepsilon^+ > \varepsilon^-$ (not shown)
- False optimism (chosen arm) ≈ pessimism (unchosen arm), so what's the difference?

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)

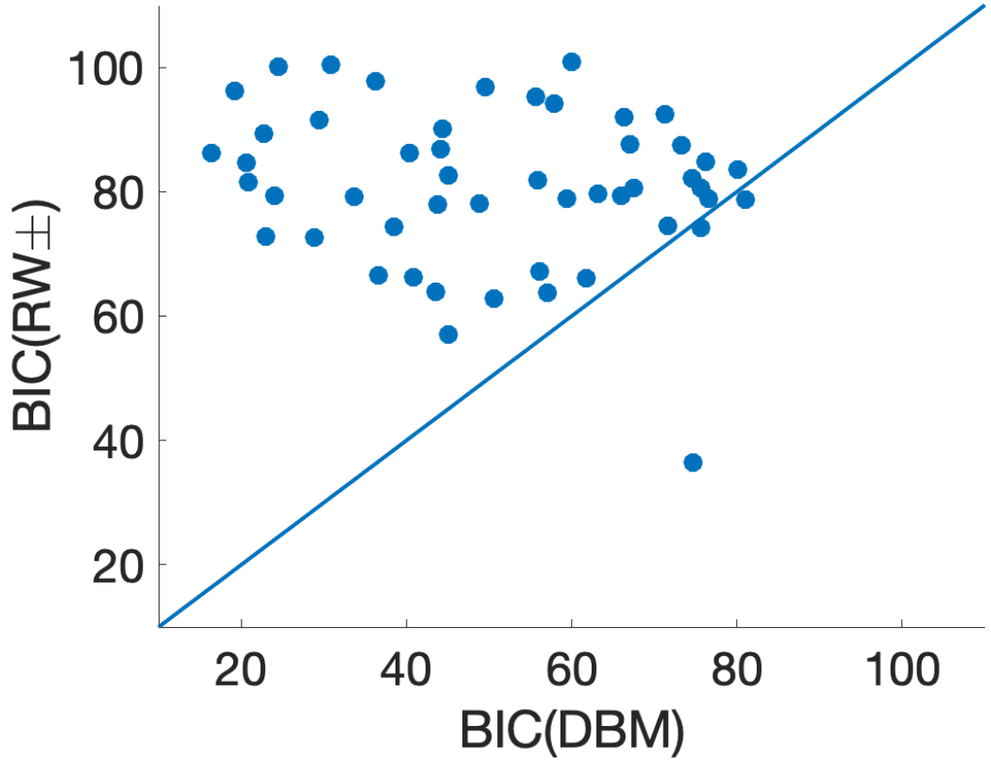


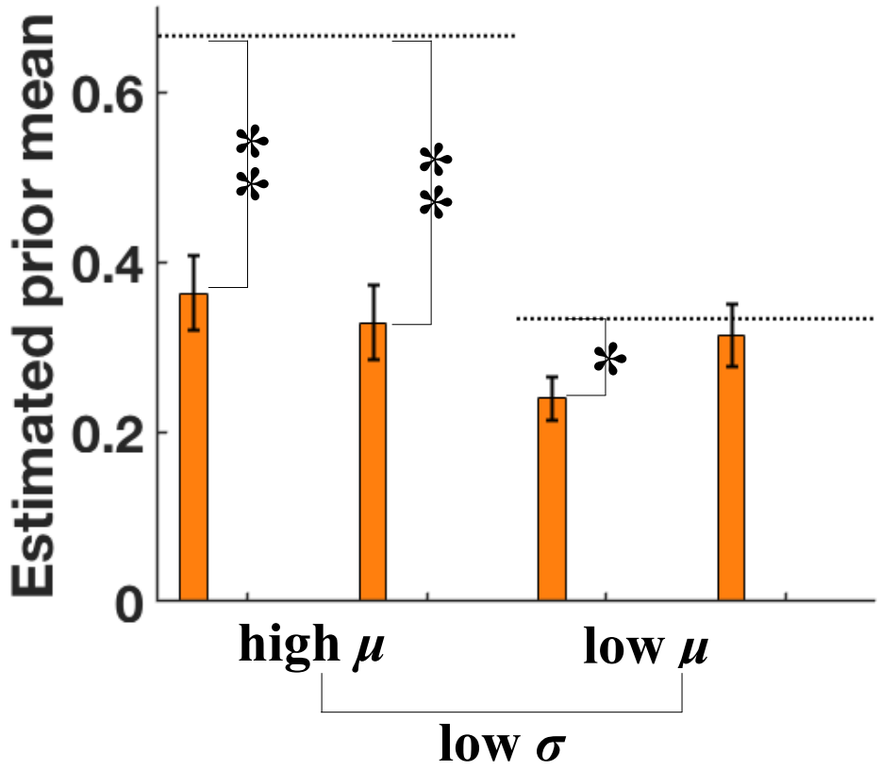


- DBM accounts for subjects' choices much better than RW±
- Estimated prior mean: 0.32 (±0.19), very similar to previous experiment — universal?
- Synthetic data generated from DBM used to fit RW± → $\varepsilon^+ > \varepsilon^-$ (not shown)
- False optimism (chosen arm) ≈ pessimism (unchosen arm), so what's the difference?
  - Increasing devaluation of unchosen arm: choice history (not only reward history) matters

And What of Optimism Bias?

(Zhou, Guo, Yu, *Cogsci*, 2020)





- DBM accounts for subjects' choices much better than RW±
- Estimated prior mean: 0.32 (±0.19), very similar to previous experiment — universal?
- Synthetic data generated from DBM used to fit RW± → $\varepsilon^+ > \varepsilon^-$ (not shown)
- False optimism (chosen arm) ≈ pessimism (unchosen arm), so what's the difference?
  - Increasing devaluation of unchosen arm: choice history (not only reward history) matters
  - Leveraging the property: high choice frequency correlated with high reward rate in individuals who can "find" the good options (bad news for individuals who are stuck in bad choice patterns, e.g. addiction)

Outline

- Computational modeling
- **Applications to psychiatry**
  - ✳ **Addiction**
  - ✳ Depression/anxiety
- Discussion

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

- Task: 3-armed bandit task, real-valued rewards, 3 reward levels, 6 forced-choice followed by 1-6 free choice trials

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

- Task: 3-armed bandit task, real-valued rewards, 3 reward levels, 6 forced-choice followed by 1-6 free choice trials
- **Unequal information condition**: choose 1 option 4 times, 1 option 2 times, 3rd option 0 times

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

- Task: 3-armed bandit task, real-valued rewards, 3 reward levels, 6 forced-choice followed by 1-6 free choice trials
- **Unequal information condition**: choose 1 option 4 times, 1 option 2 times, 3rd option 0 times
- **Equal information condition**: choose each option 2 times

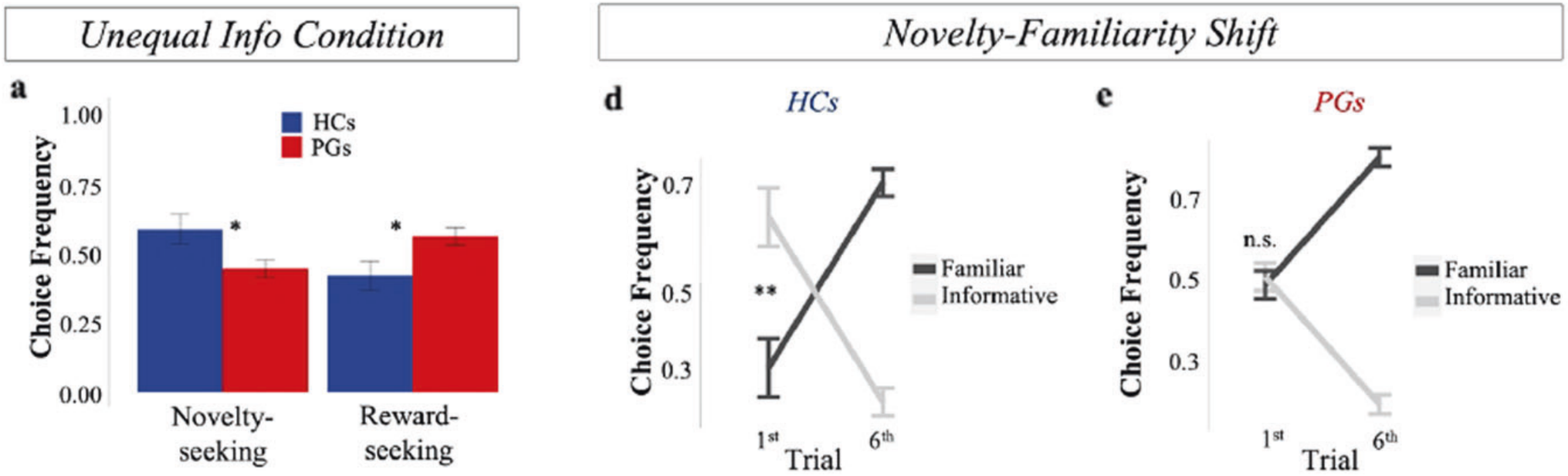
Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

- Task: 3-armed bandit task, real-valued rewards, 3 reward levels, 6 forced-choice followed by 1-6 free choice trials
- **Unequal information condition**: choose 1 option 4 times, 1 option 2 times, 3rd option 0 times
- **Equal information condition**: choose each option 2 times
- Subjects: 40 problem gamblers (PG), 22 healthy controls (HC)

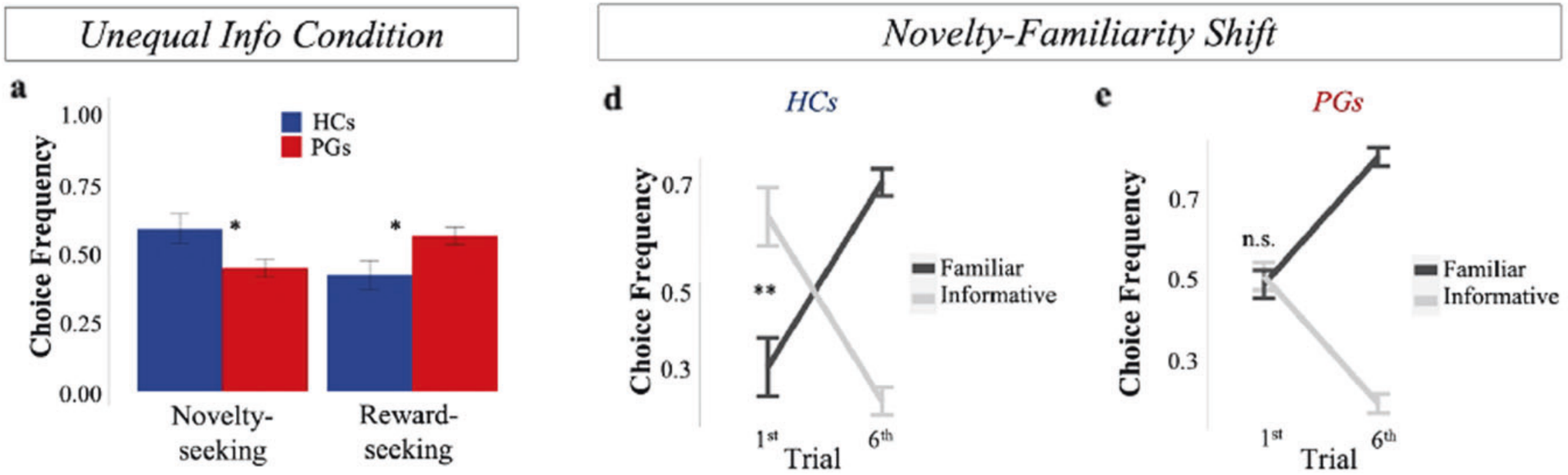
Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)



Gambling Addiction

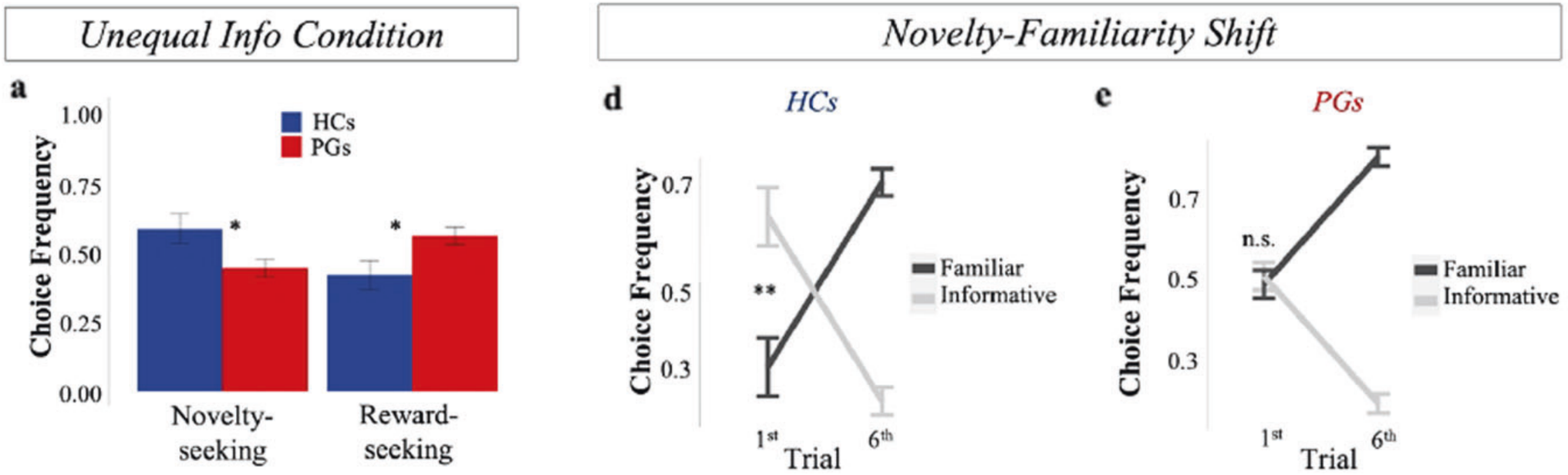
(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)



- HC show stronger novelty seeking (choosing 0-seen option), and lower reward-seeking, tendency than PG

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)



- HC show stronger novelty seeking (choosing 0-seen option), and lower reward-seeking, tendency than PG
- From 1st to 6th forced trial, HC show a significant novelty-familiarity shift, PG show no preference for novelty at the start but do show preference for more familiar options later on

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)
  - $\nu > 0$ (novelty bonus; Gottlieb et al, 2013)

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)
  - $v > 0$ (novelty bonus; Gottlieb et al, 2013)
- HC: large $v > 0$, $k \approx 0$, PG: smaller $v > 0$, larger $k > 0$

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)
  - $v > 0$ (novelty bonus; Gottlieb et al, 2013)
- HC: large $v > 0$, $k \approx 0$, PG: smaller $v > 0$, larger $k > 0$
- HC do not exhibit a general uncertainty bonus, only novelty bonus!

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)
  - $v > 0$ (novelty bonus; Gottlieb et al, 2013)
- HC: large $v > 0$, $k \approx 0$, PG: smaller $v > 0$, larger $k > 0$
- HC do not exhibit a general uncertainty bonus, only novelty bonus!
- PG show smaller novelty bonus, but also an uncertainty penalty/familiarity bonus!

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)
  - $v > 0$ (novelty bonus; Gottlieb et al, 2013)
- HC: large $v > 0$, $k \approx 0$, PG: smaller $v > 0$, larger $k > 0$
- HC do not exhibit a general uncertainty bonus, only novelty bonus!
- PG show smaller novelty bonus, but also an uncertainty penalty/familiarity bonus!
- No group differences in overall performance; simulations: both modes of behavior “optimal”

Gambling Addiction

(Dezza, Noel, Cleeremans, Yu, *Translational Psychiatry*, 2021)

$$V_{t,j}(c) = Q_{t+1,j}(c) + \sum_{1}^{t} i_{t,j}(c) * k + 1_{\text{novel}} * \nu$$

- Decision policy
  - Q-value (E[reward$_{t+1}$|data$_t$]): computed via RW, DBM
  - $k < 0$ (uncertainty bonus; Wilson et al 2014, Dezza et al 2017)
  - $v > 0$ (novelty bonus; Gottlieb et al, 2013)
- HC: large $v > 0$, $k \approx 0$, PG: smaller $v > 0$, larger $k > 0$
- HC do not exhibit a general uncertainty bonus, only novelty bonus!
- PG show smaller novelty bonus, but also an uncertainty penalty/familiarity bonus!
- No group differences in overall performance; simulations: both modes of behavior "optimal"
- But! Reward rates fixed in current design. In non-stationary environment, the PG pattern can be trapped into suboptimal patterns of behavior (addiction ⇒ worse Wisconsin Card Sorting Task, Hosak et al; 2012)

Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

- Task: 2-armed bandit task

Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

- Task: 2-armed bandit task
- Subjects: 16 MDI, 16 HCS

Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

- Task: 2-armed bandit task
- Subjects: 16 MDI, 16 HCS
- Learning model: DBM

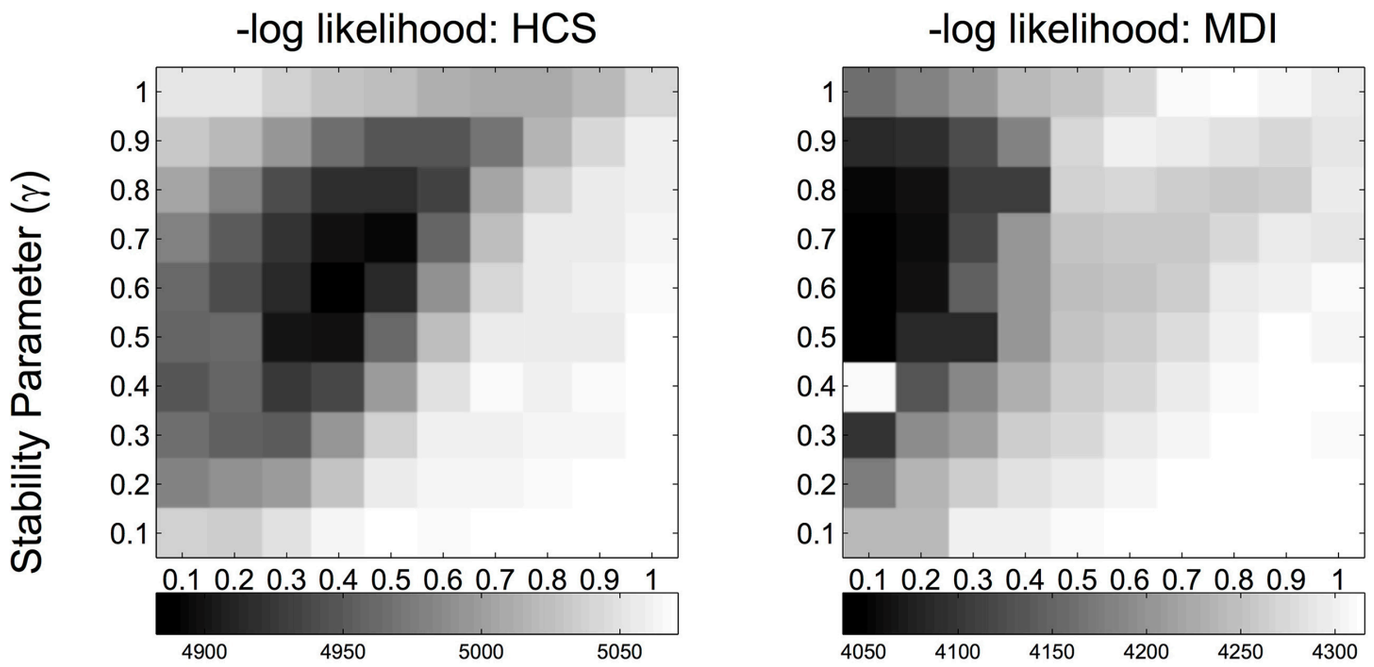
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

- Task: 2-armed bandit task
- Subjects: 16 MDI, 16 HCS
- Learning model: DBM
- Decision policies: softmax, knowledge gradient, ε-greedy, WSLS, τ-switch

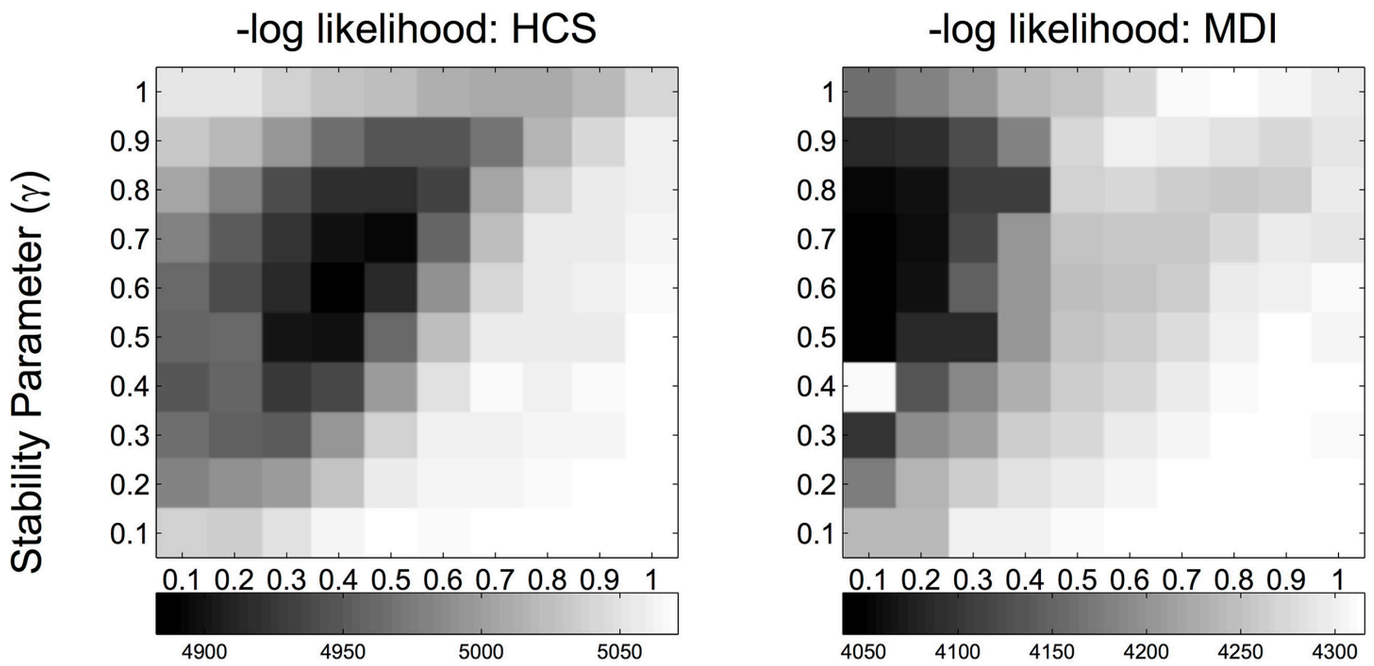
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)



Methamphetamine Addiction

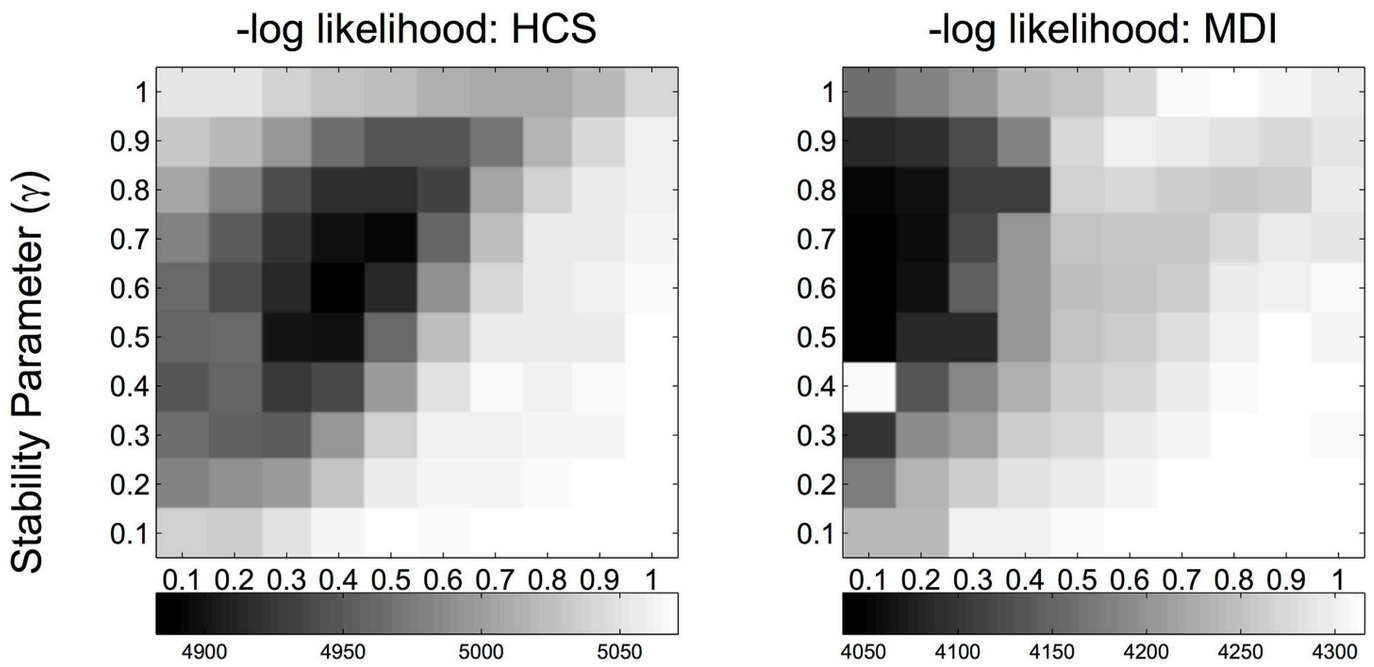
(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)



- MDI expected much lower reward rate in the environment than HCS (similar stability assumptions)

Methamphetamine Addiction

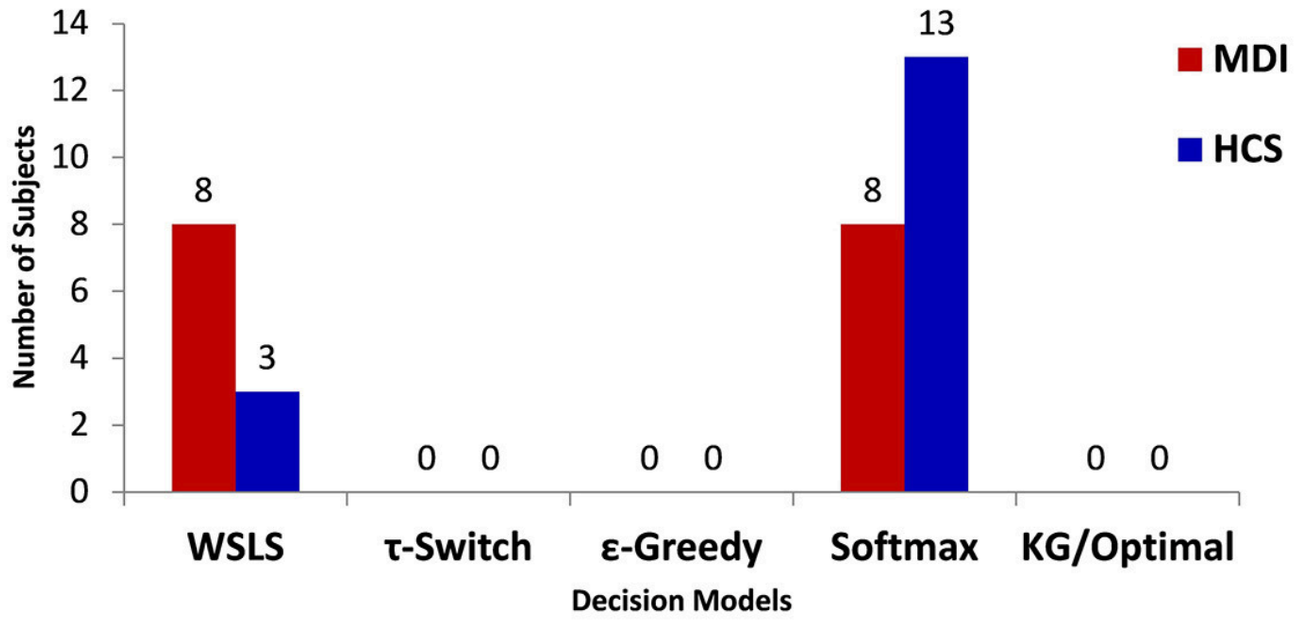
(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)



- MDI expected much lower reward rate in the environment than HCS (similar stability assumptions)
- ⇒ excessive tendency to stick with familiar option (devaluing of unchosen option, not due to assumptions about greater stability)

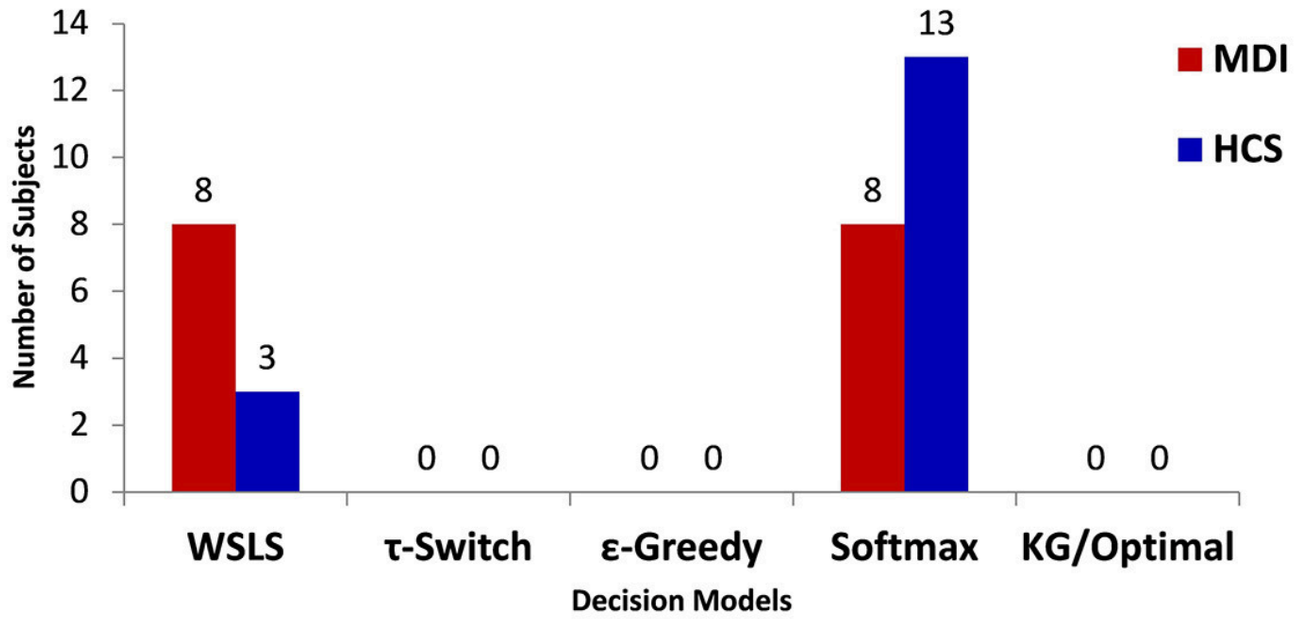
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)



Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

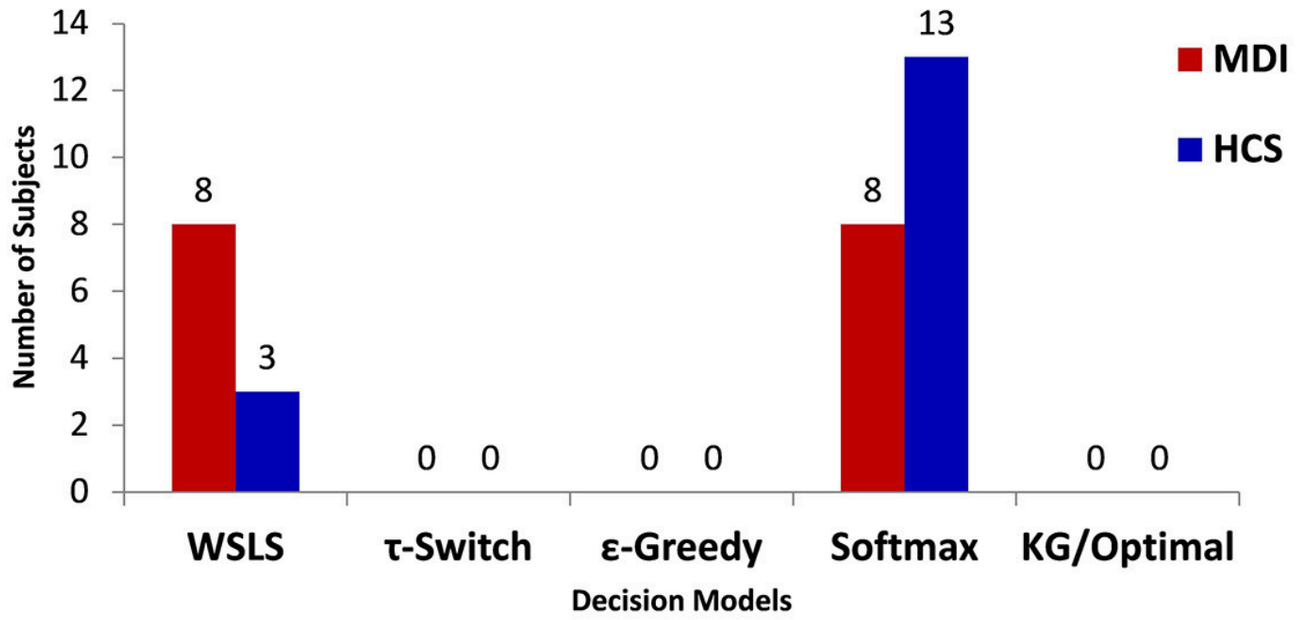


- Identify best model using Bayesian model evidence:

  $p(X|M) = \int p(X|\theta,M)p(\theta,M)\, d\theta$

Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

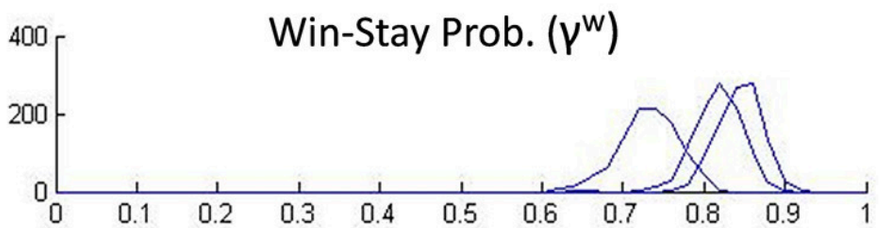


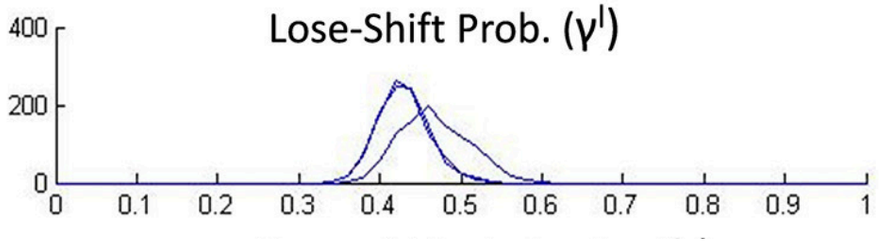
- Identify best model using Bayesian model evidence:
  $p(X|M) = \int p(X|\theta,M)p(\theta,M)\, d\theta$
- MDI more likely to use WSLS (no learning); HCS more likely to use DBM+softmax

Methamphetamine Addiction

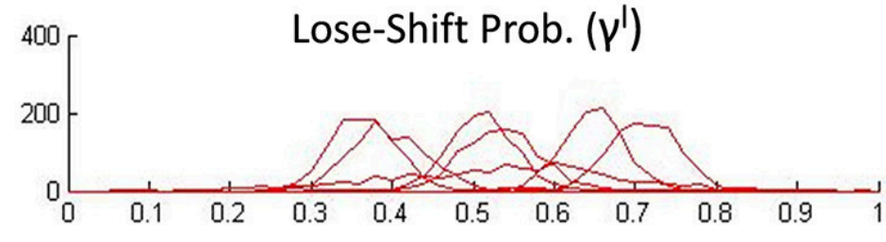
(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS





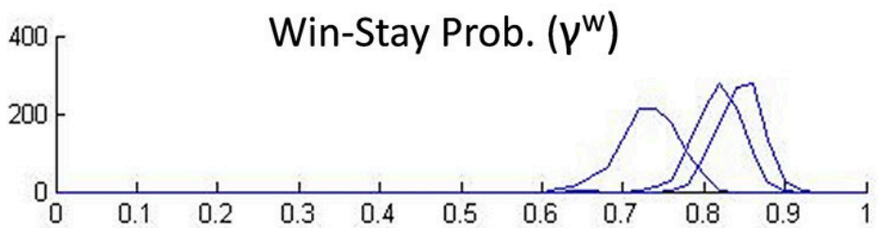
MDI

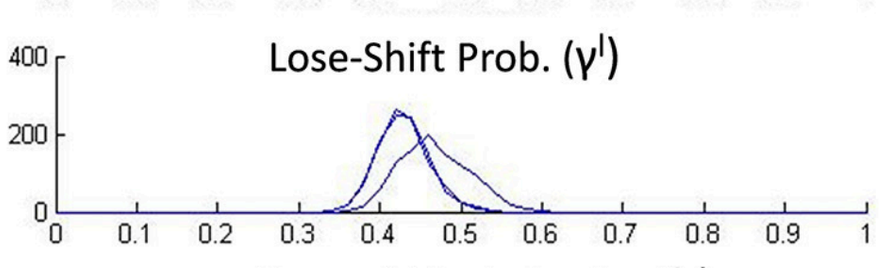


Methamphetamine Addiction

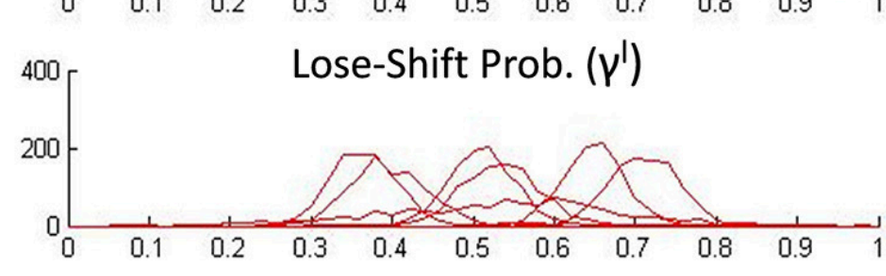
(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS





MDI

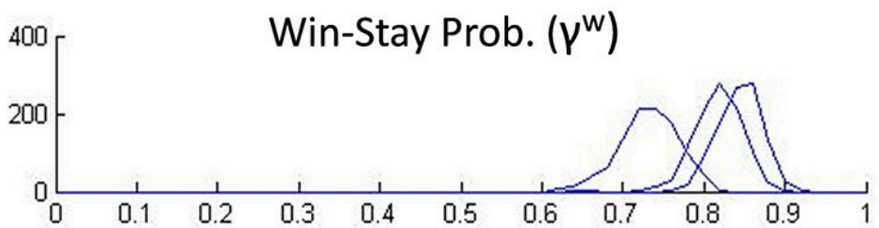


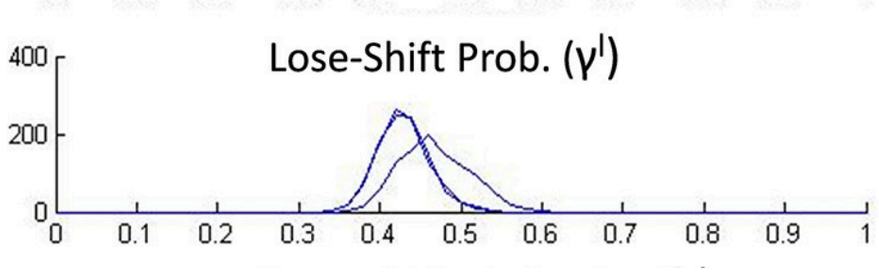
- Of subjects best described by WSLS: MDI have higher P(win-stay)

Methamphetamine Addiction

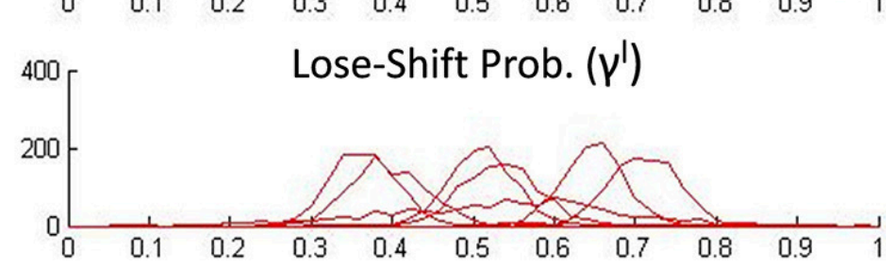
(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS





MDI

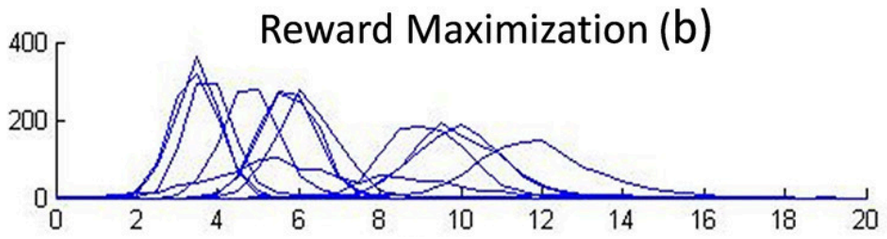


- Of subjects best described by WSLS: MDI have higher P(win-stay)
- Related to greater tendency to persevere

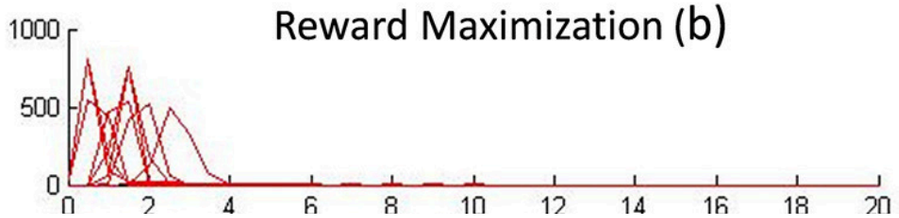
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS



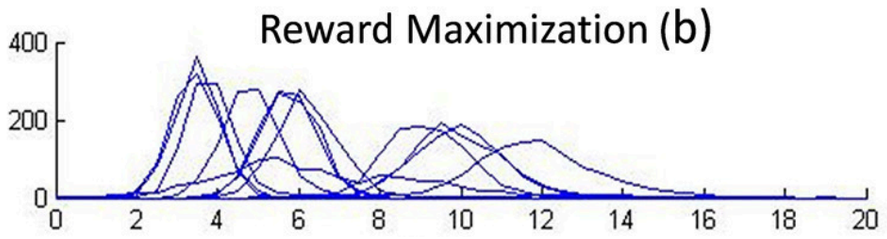
MDI



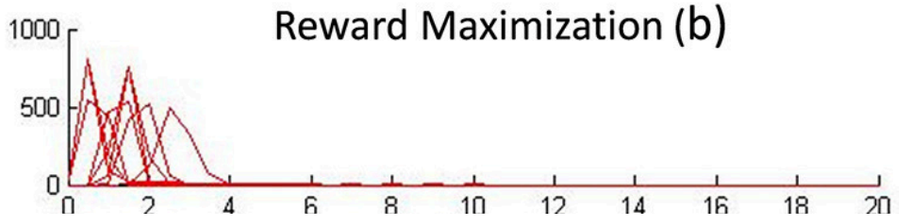
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS



MDI

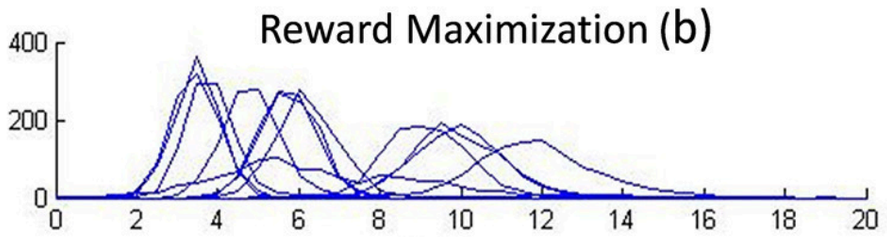


- Of subjects best described by DBM+softmax: MDI maximize expected reward less (lower inverse-temperature parameter b in softmax)

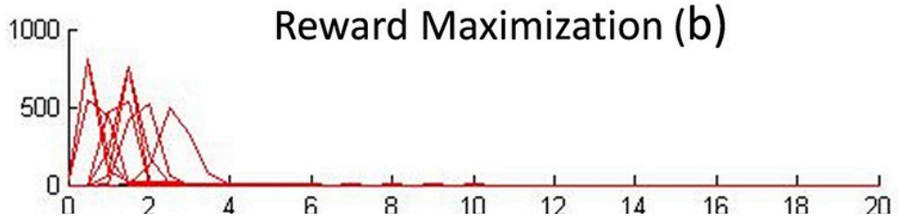
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS



MDI

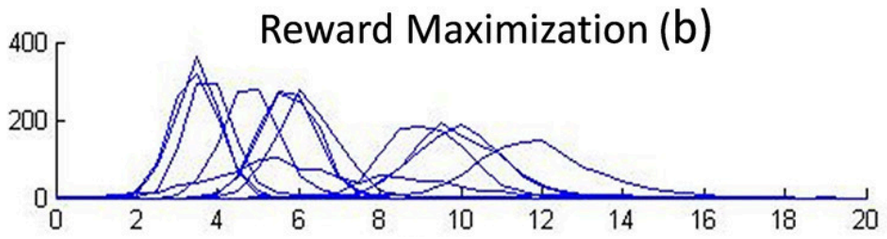


- Of subjects best described by DBM+softmax: MDI maximize expected reward less (lower inverse-temperature parameter *b* in softmax)
- Can be more stochasticity OR sticking with familiar option (need 3-arm bandit task)

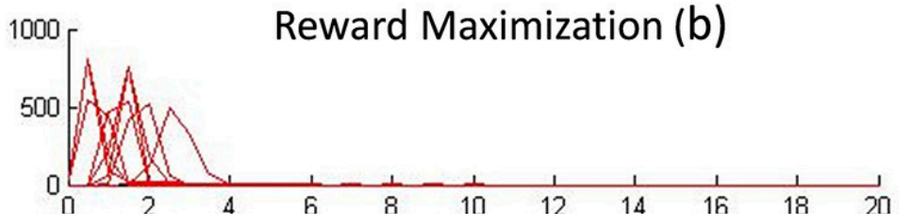
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS



MDI

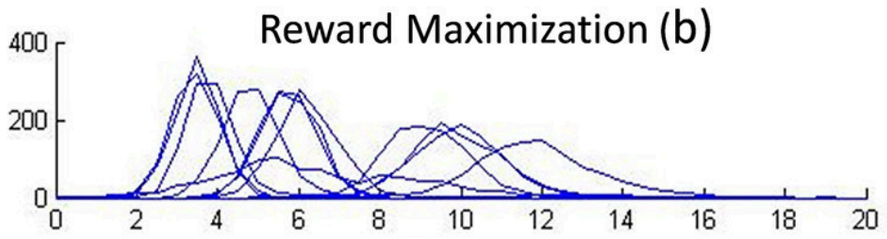


- Of subjects best described by DBM+softmax: MDI maximize expected reward less (lower inverse-temperature parameter *b* in softmax)
- Can be more stochasticity OR sticking with familiar option (need 3-arm bandit task)
- Again: no group difference in overall performance (total points earned)

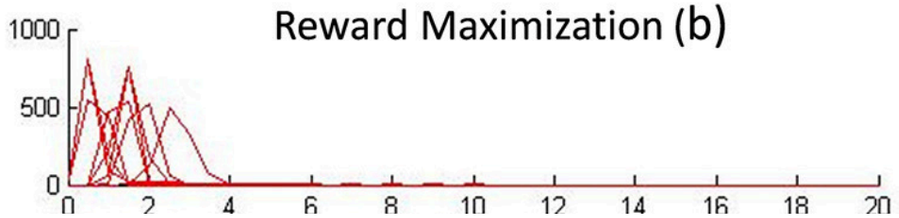
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS



MDI

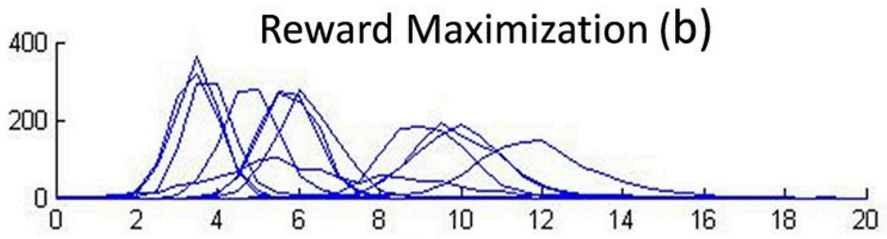


- Of subjects best described by DBM+softmax: MDI maximize expected reward less (lower inverse-temperature parameter *b* in softmax)
- Can be more stochasticity OR sticking with familiar option (need 3-arm bandit task)
- Again: no group difference in overall performance (total points earned)
- VBM analysis: MDI have lower gray matter volume of thalamic dorsal lateral (DL) nucleus, mediated by softmax parameter *b*; DL volume predictive of MDI vs. HCS status

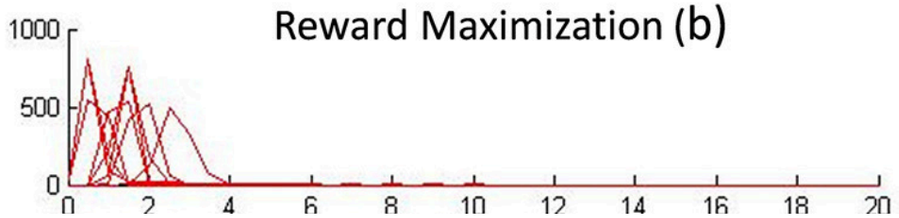
Methamphetamine Addiction

(Harlé, Zhang, Schiff, Mackey, Paulus, Yu, *Frontiers in Psychology*, 2015)

HCS



MDI



- Of subjects best described by DBM+softmax: MDI maximize expected reward less (lower inverse-temperature parameter *b* in softmax)
- Can be more stochasticity OR sticking with familiar option (need 3-arm bandit task)
- Again: no group difference in overall performance (total points earned)
- VBM analysis: MDI have lower gray matter volume of thalamic dorsal lateral (DL) nucleus, mediated by softmax parameter *b*; DL volume predictive of MDI vs. HCS status
- DL nucleus reciprocally connected to ACC (expectancy violation & change detection, Ide et al, 2013)

Addiction: Summary

Addiction: Summary

- Compared to HC's, both addiction groups (PG and MDI) show a "familiarity bonus" (or "uncertainty penalty")

Addiction: Summary

- Compared to HC's, both addiction groups (PG and MDI) show a "familiarity bonus" (or "uncertainty penalty")
  - Apparently mediated by more pessimistic beliefs about unchosen options

Addiction: Summary

- Compared to HC's, both addiction groups (PG and MDI) show a "familiarity bonus" (or "uncertainty penalty")
  - Apparently mediated by more pessimistic beliefs about unchosen options
- No group differences in overall performance in these tasks — BUT fixed reward structure!

Addiction: Summary

- Compared to HC's, both addiction groups (PG and MDI) show a "familiarity bonus" (or "uncertainty penalty")
  - Apparently mediated by more pessimistic beliefs about unchosen options
- No group differences in overall performance in these tasks — BUT fixed reward structure!
- In non-stationary environment, an aversion to novel/unfamiliar options can trap one into suboptimal behavioral patterns

Addiction: Summary

- Compared to HC's, both addiction groups (PG and MDI) show a "familiarity bonus" (or "uncertainty penalty")
  - Apparently mediated by more pessimistic beliefs about unchosen options
- No group differences in overall performance in these tasks — BUT fixed reward structure!
- In non-stationary environment, an aversion to novel/unfamiliar options can trap one into suboptimal behavioral patterns
- Future direction: non-stationary task environments, e.g. change points

Outline

- Computational modeling
- **Applications to psychiatry**
  - Addiction
  - **Depression/anxiety**
- Discussion

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

- Task: 2-armed bandit task

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

- Task: 2-armed bandit task
- Subjects: 53 individuals with a range of anhedonia and state anxiety scores

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

- Task: 2-armed bandit task
- Subjects: 53 individuals with a range of anhedonia and state anxiety scores
- Models: WSLS, DBM+softmax

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

- Task: 2-armed bandit task
- Subjects: 53 individuals with a range of anhedonia and state anxiety scores
- Models: WSLS, DBM+softmax
- Results:

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

- Task: 2-armed bandit task
- Subjects: 53 individuals with a range of anhedonia and state anxiety scores
- Models: WSLS, DBM+softmax
- Results:
  - Higher state anxiety ⇒ WSLS > DBM+softmax, i.e. less learning and more immediate aversion to “failure”

Depression/Anxiety

(Harlé, Guo, Zhang, Paulus, Yu, *PLOS One*, 2017)

- Task: 2-armed bandit task
- Subjects: 53 individuals with a range of anhedonia and state anxiety scores
- Models: WSLS, DBM+softmax
- Results:
  - Higher state anxiety ⇒ WSLS > DBM+softmax, i.e. less learning and more immediate aversion to “failure”
  - Among “softmax subjects", higher anhedonia ⇒ less likely to choose more rewarding option; but since only 2 options, unclear if it is more random stochasticity or some other motivational factor like novelty or uncertainty

Outline

- Computational modeling
- Applications to psychiatry
  - Addiction
  - Depression/anxiety
- **Discussion**

Summary: Three Wrongs Make a Right

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility
    2. Humans utilize **suboptimal** decision policy (Softmax or KG)

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility
    2. Humans utilize **suboptimal** decision policy (Softmax or KG)
    3. Humans **underestimate** prior mean reward rate

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility
    2. Humans utilize **suboptimal** decision policy (Softmax or KG)
    3. Humans **underestimate** prior mean reward rate
- These suboptimalities compensate for one another

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility
    2. Humans utilize **suboptimal** decision policy (Softmax or KG)
    3. Humans **underestimate** prior mean reward rate
- These suboptimalities compensate for one another
    - Lower prior mean mitigates both (false) non-stationarity assumption and simple decision strategy

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility
    2. Humans utilize **suboptimal** decision policy (Softmax or KG)
    3. Humans **underestimate** prior mean reward rate
- These suboptimalities compensate for one another
    - Lower prior mean mitigates both (false) non-stationarity assumption and simple decision strategy
- DBM/pbRL better account of the **computational processes** underlying learning & decision making under uncertainty than previous models (Q-learning, RW±)

Summary: Three Wrongs Make a Right

- Three suboptimalities in human decision-making
    1. Humans **overestimate** environmental volatility
    2. Humans utilize **suboptimal** decision policy (Softmax or KG)
    3. Humans **underestimate** prior mean reward rate
- These suboptimalities compensate for one another
    - Lower prior mean mitigates both (false) non-stationarity assumption and simple decision strategy
- DBM/pbRL better account of the **computational processes** underlying learning & decision making under uncertainty than previous models (Q-learning, RW±)
- May serve as a better model for identifying neural correlates and understanding the computational roles played by different brain areas

Discussion

Discussion

Basic Science ⇒ Psychiatry

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and “into the brain”

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks
- Better latent variables for identifying brain correlates, and thus possible targets of future clinical interventions

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks
- Better latent variables for identifying brain correlates, and thus possible targets of future clinical interventions
- Better differentiation of different psychiatric populations

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks
- Better latent variables for identifying brain correlates, and thus possible targets of future clinical interventions
- Better differentiation of different psychiatric populations

Psychiatry ⇒ Basic Science

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks
- Better latent variables for identifying brain correlates, and thus possible targets of future clinical interventions
- Better differentiation of different psychiatric populations

Psychiatry ⇒ Basic Science

- New dimensions to explore in behavioral experiments, e.g. # arms, non-stationarity

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks
- Better latent variables for identifying brain correlates, and thus possible targets of future clinical interventions
- Better differentiation of different psychiatric populations

Psychiatry ⇒ Basic Science

- New dimensions to explore in behavioral experiments, e.g. # arms, non-stationarity
- Internal "knobs" that can be tuned too far one way or the other, e.g. prior belief about unchosen options

Discussion

Basic Science ⇒ Psychiatry

- Going beyond scales and behavior, and "into the brain"
- Cognitive representations and internal states, e.g. prior beliefs, uncertainty, novelty
- Better integration of results/insights from different tasks
- Better latent variables for identifying brain correlates, and thus possible targets of future clinical interventions
- Better differentiation of different psychiatric populations

Psychiatry ⇒ Basic Science

- New dimensions to explore in behavioral experiments, e.g. # arms, non-stationarity
- Internal "knobs" that can be tuned too far one way or the other, e.g. prior belief about unchosen options
- New brain areas that deserve a closer look, e.g. what does the thalamic LD nucleus do?!

Acknowledgment

Acknowledgment

- **Yu Lab**
  - Experimental design/data collection: Henry Qiu, Alvita Tran, Shunan Zhang
  - Modeling/data analysis: Dalin Guo, Shunan Zhang, Chaitanya Ryali, Corey Zhou, Zoe He

Acknowledgment

- **Yu Lab**
  - Experimental design/data collection: Henry Qiu, Alvita Tran, Shunan Zhang
  - Modeling/data analysis: Dalin Guo, Shunan Zhang, Chaitanya Ryali, Corey Zhou, Zoe He
- **Collaborators**
  - Katia Harlé, Martin Paulus
  - Florent Meyniel, Stefano Palminteri (RW± data)

Acknowledgment

- **Yu Lab**
  - Experimental design/data collection: Henry Qiu, Alvita Tran, Shunan Zhang
  - Modeling/data analysis: Dalin Guo, Shunan Zhang, Chaitanya Ryali, Corey Zhou, Zoe He
- **Collaborators**
  - Katia Harlé, Martin Paulus
  - Florent Meyniel, Stefano Palminteri (RW± data)
- **Funding**
  - CRCNS (NSF BCS, NIH NIDA), ARL, MURI, UCSD

Openings: Postdoc, PhD

AJYU@UCSD.EDU

Questions?

Learning & DM under Uncertainty

Learning & DM under Uncertainty

- Exact reward distribution of options unknown, but can learn from experience, and may have general info about "environment"

Learning & DM under Uncertainty

- Exact reward distribution of options unknown, but can learn from experience, and may have general info about "environment"
- How does such partial knowledge affect human decision-making?

Learning & DM under Uncertainty

- Exact reward distribution of options unknown, but can learn from experience, and may have general info about "environment"
- How does such partial knowledge affect human decision-making?
- Computational modeling:

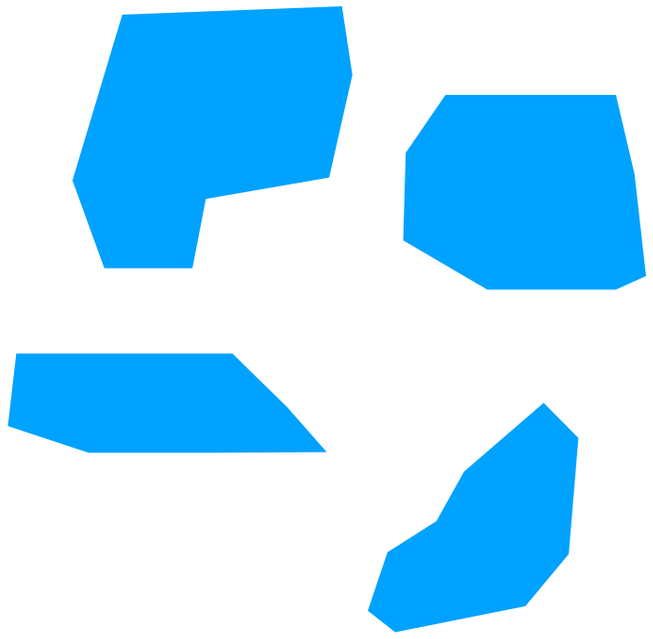
Learning & DM under Uncertainty

- Exact reward distribution of options unknown, but can learn from experience, and may have general info about "environment"
- How does such partial knowledge affect human decision-making?
- Computational modeling:
  - characterize human learning & decision processes

Learning & DM under Uncertainty

- Exact reward distribution of options unknown, but can learn from experience, and may have general info about "environment"
- How does such partial knowledge affect human decision-making?
- Computational modeling:
  - characterize human learning & decision processes
  - understand provenance & consequence of these processes

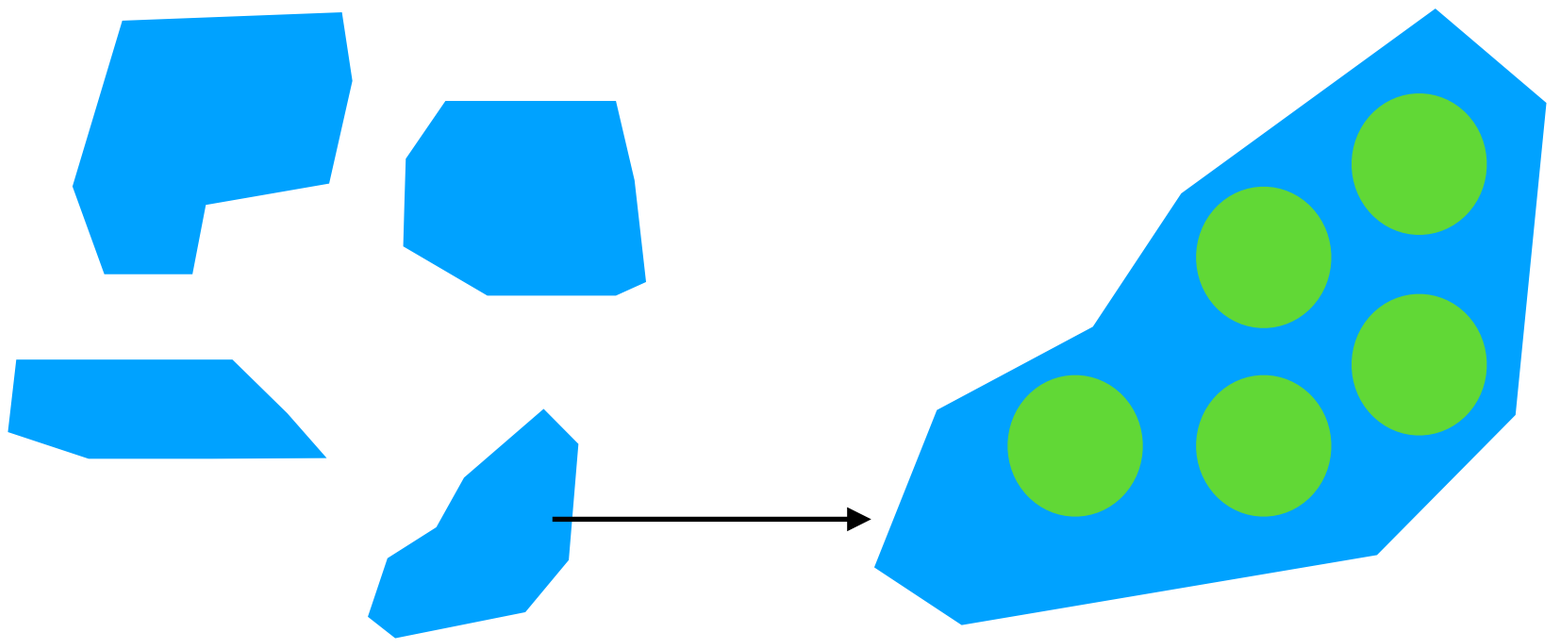
Cover Story: Ice fishing Game



4 environments

(4 lakes)

Cover Story: Ice fishing Game



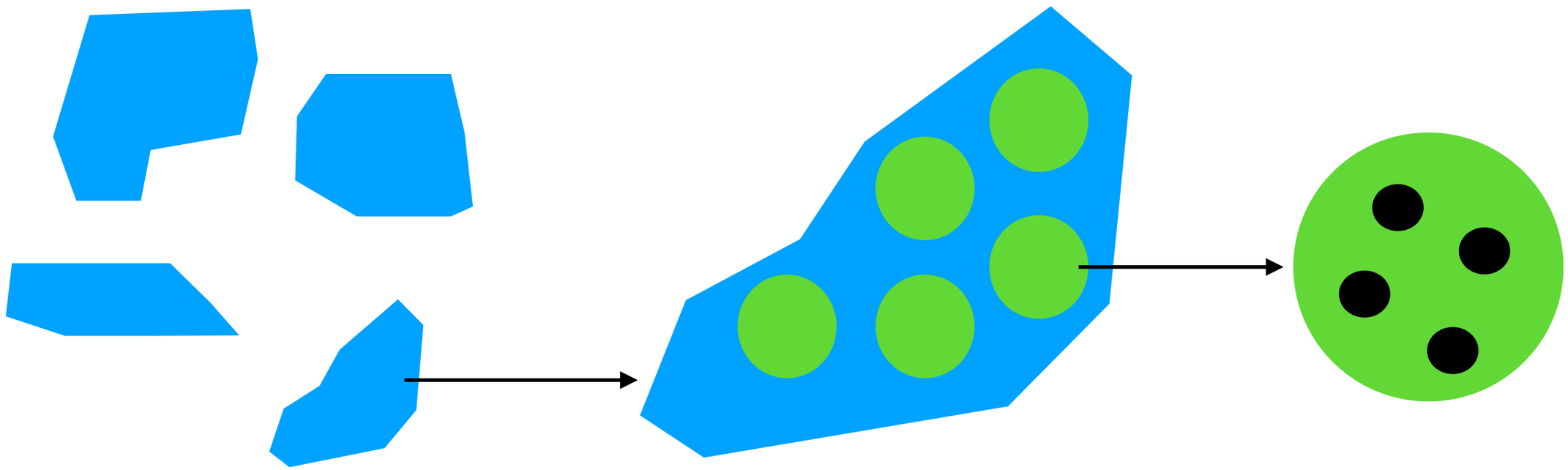
4 environments

(4 lakes)

50 games/environment

(50 camps/lake)

Cover Story: Ice fishing Game



4 environments

(4 lakes)

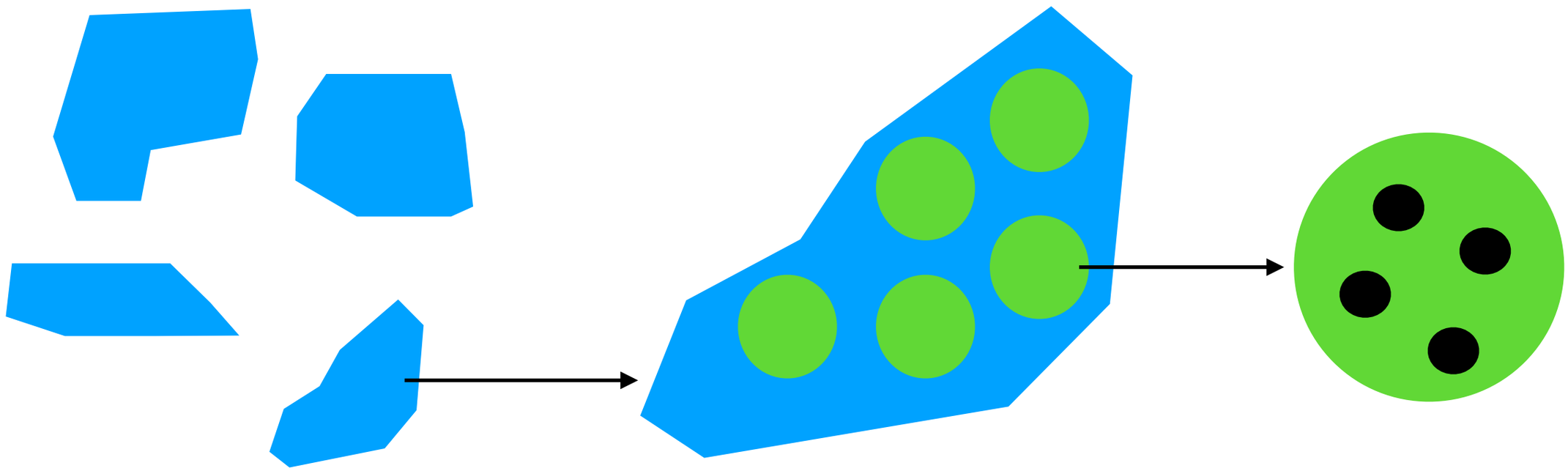
50 games/environment

(50 camps/lake)

4 arms/game

(4 holes, 15 attempts)

Cover Story: Ice fishing Game



4 environments
(4 lakes)

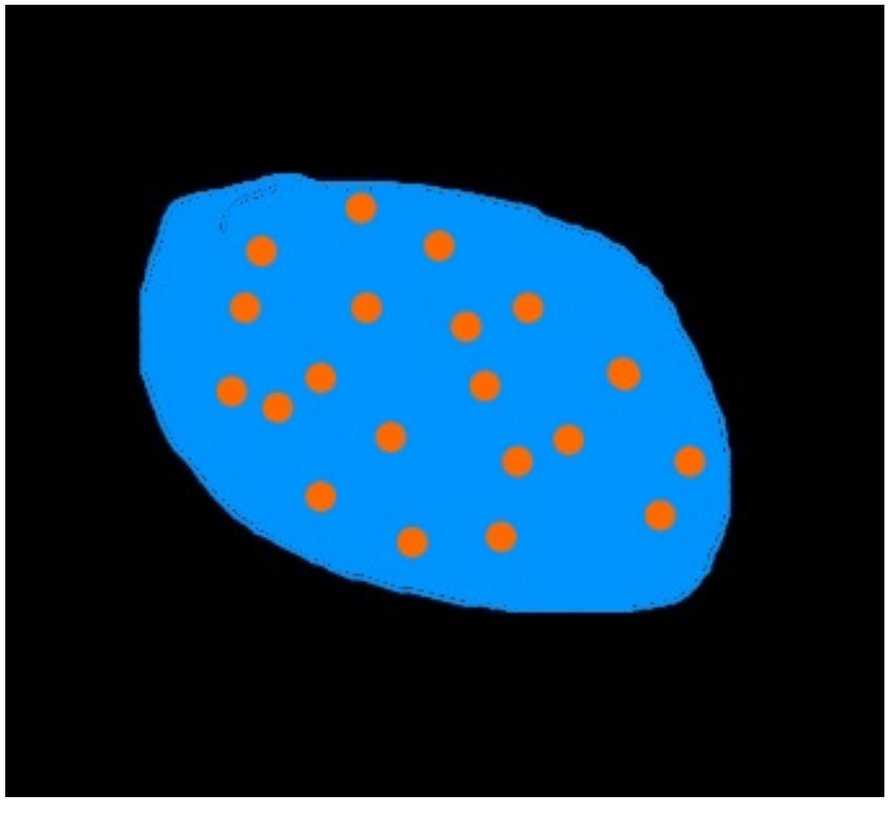
50 games/environment
(50 camps/lake)

4 arms/game
(4 holes, 15 attempts)

The 4 lakes differ in

- Overall abundance of fish
- Variability of abundance across holes within each lake

Prior Info: Fishing Report

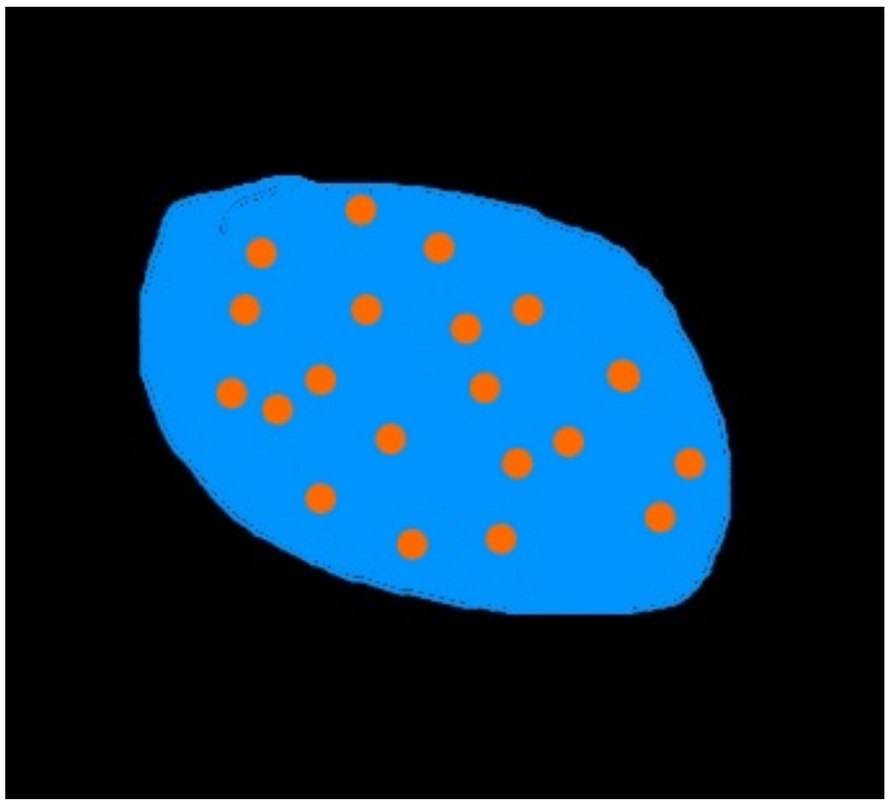


Exploratory fishing report of 20 exploratory holes				
5	7	5	7	7
7	6	7	7	8
6	8	7	6	7
7	5	7	7	6

High abundance, low variance

Prior Info: Fishing Report

General information about reward distribution of environment



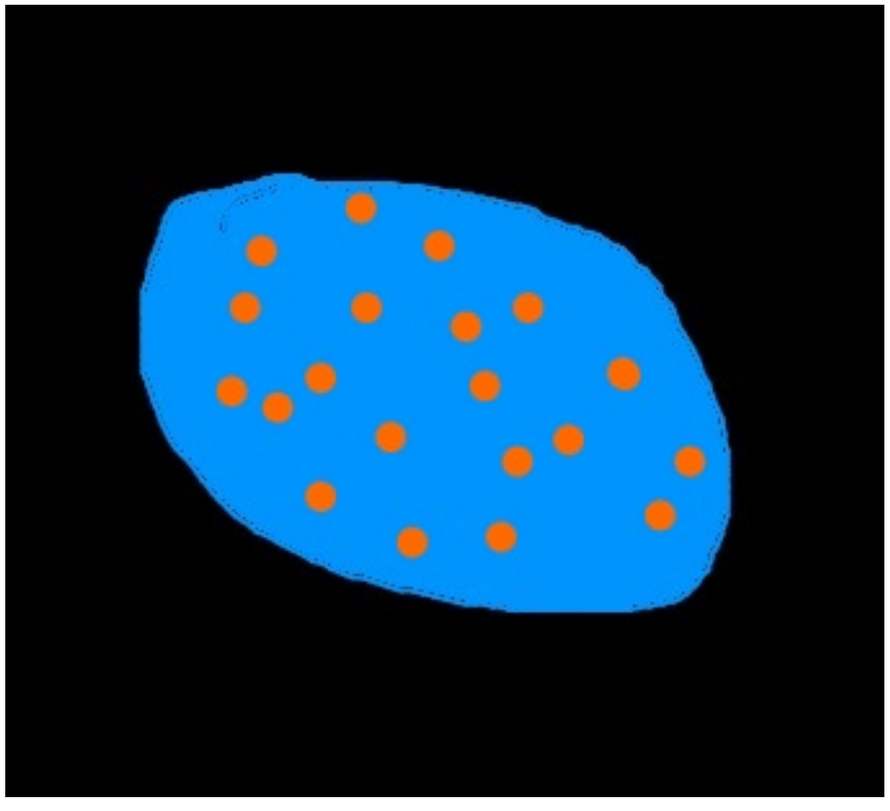
Exploratory fishing report of 20 exploratory holes				
5	7	5	7	7
7	6	7	7	8
6	8	7	6	7
7	5	7	7	6

High abundance, low variance

Prior Info: Fishing Report

General information about reward distribution of environment

- **20** random samples from the true Beta distribution



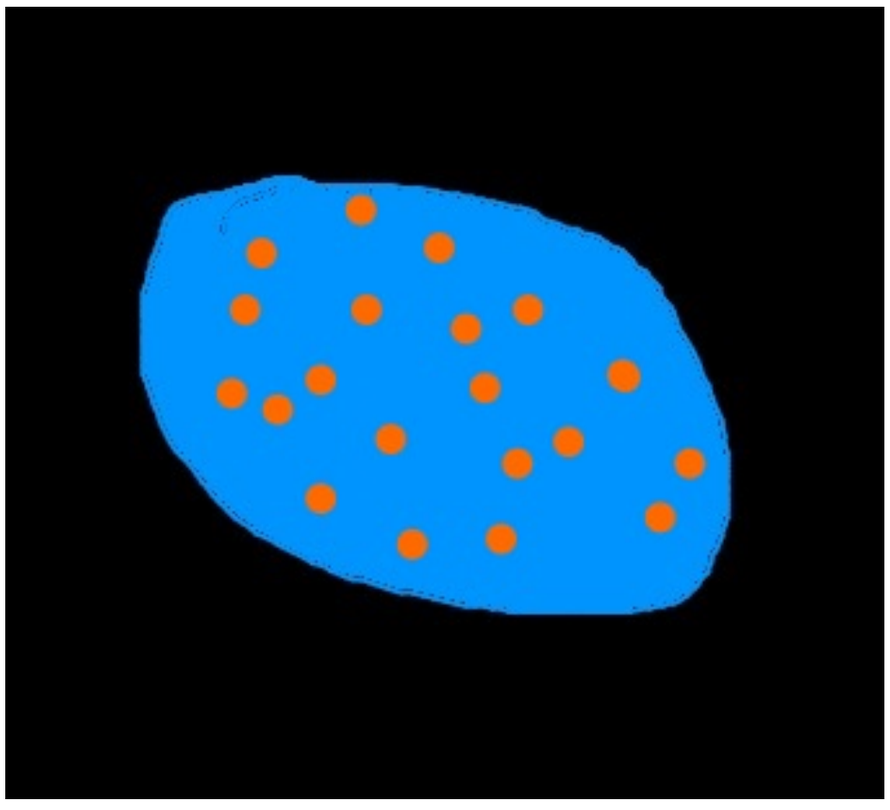
Exploratory fishing report of 20 exploratory holes				
5	7	5	7	7
7	6	7	7	8
6	8	7	6	7
7	5	7	7	6

High abundance, low variance

Prior Info: Fishing Report

General information about reward distribution of environment

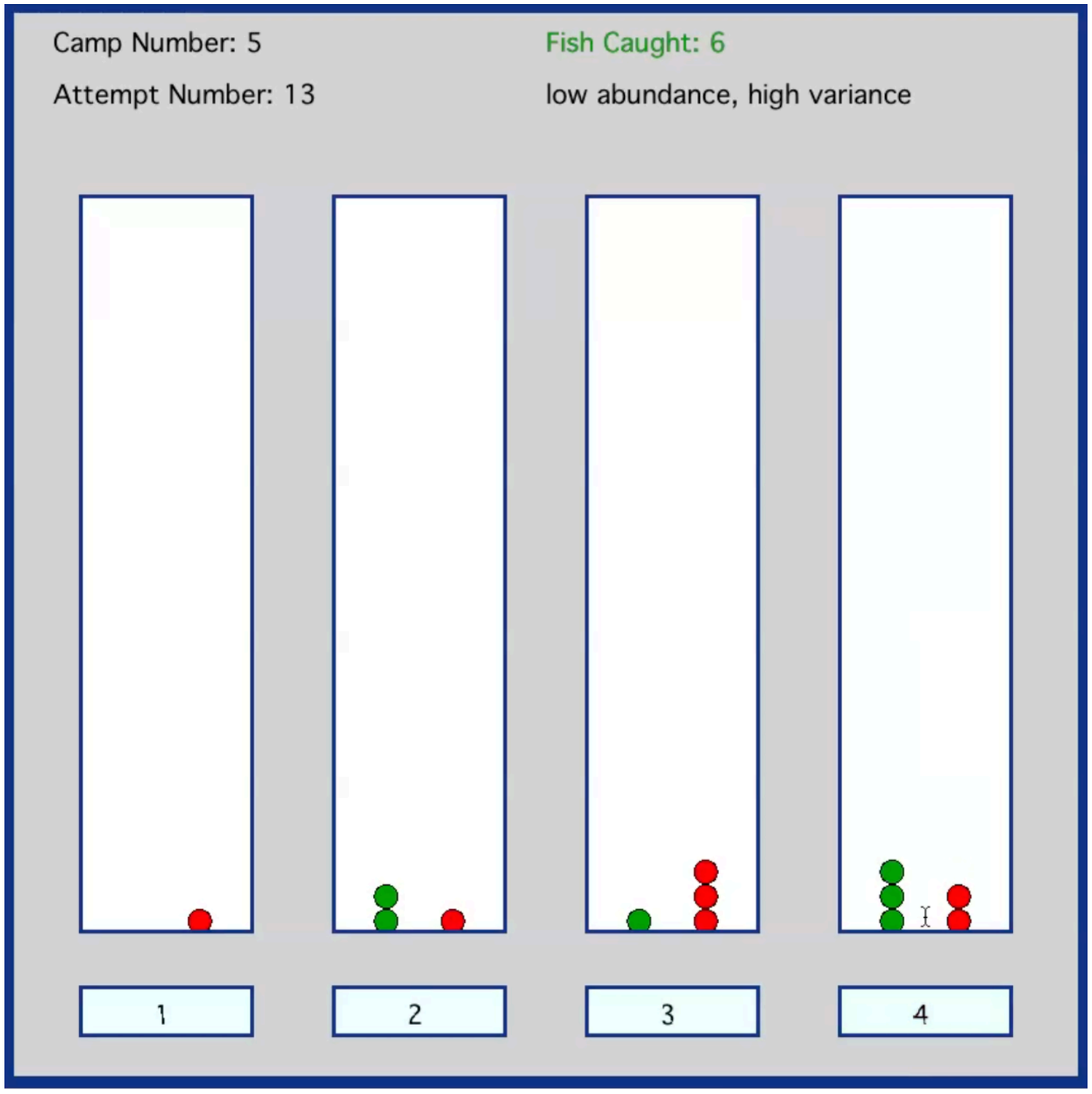
- **20** random samples from the true Beta distribution
- Number: **# fish caught out of 10 attempts**



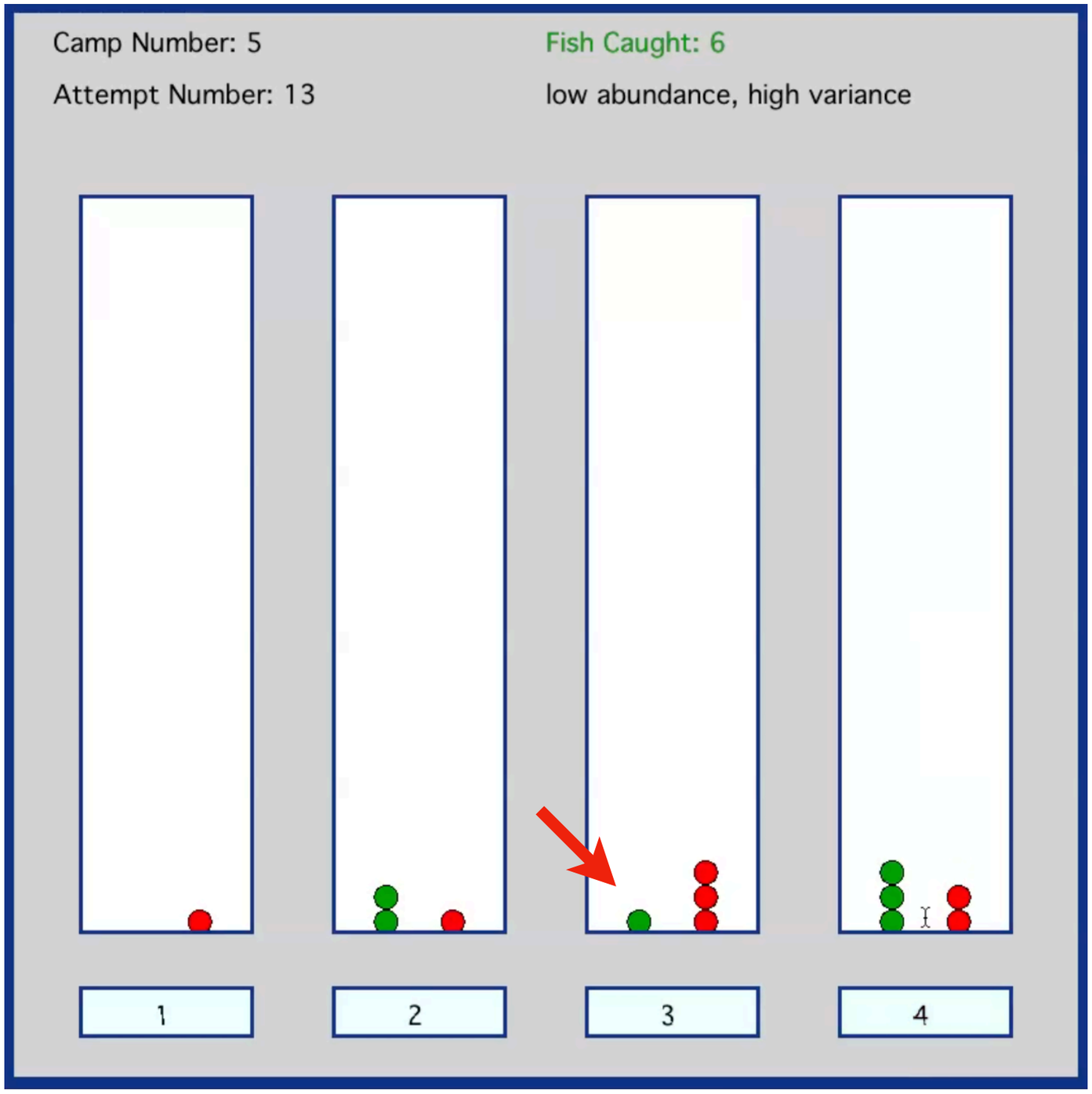
Exploratory fishing report of 20 exploratory holes				
5	7	5	7	7
7	6	7	7	8
6	8	7	6	7
7	5	7	7	6

High abundance, low variance

Experimental Interface

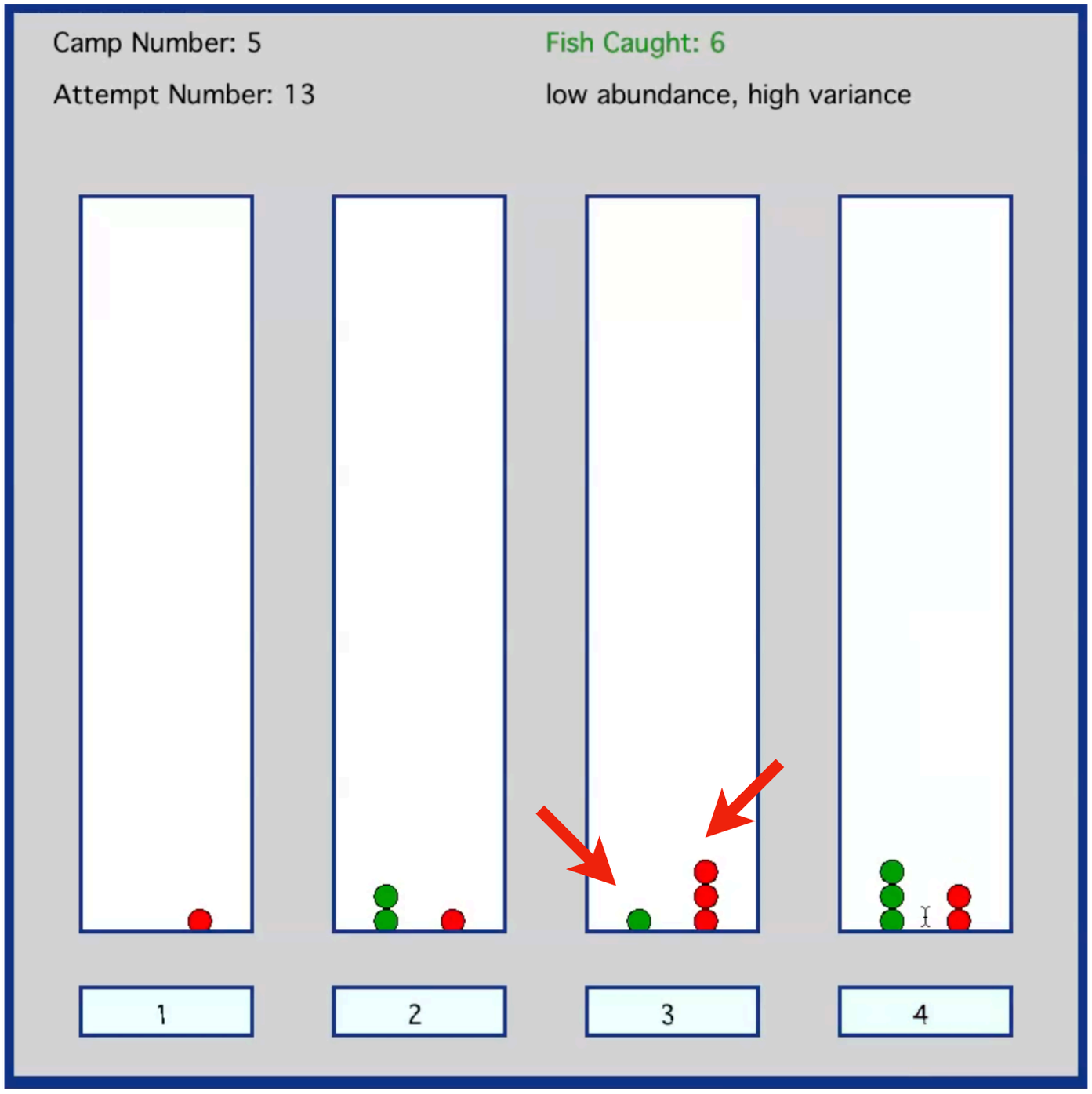


Experimental Interface



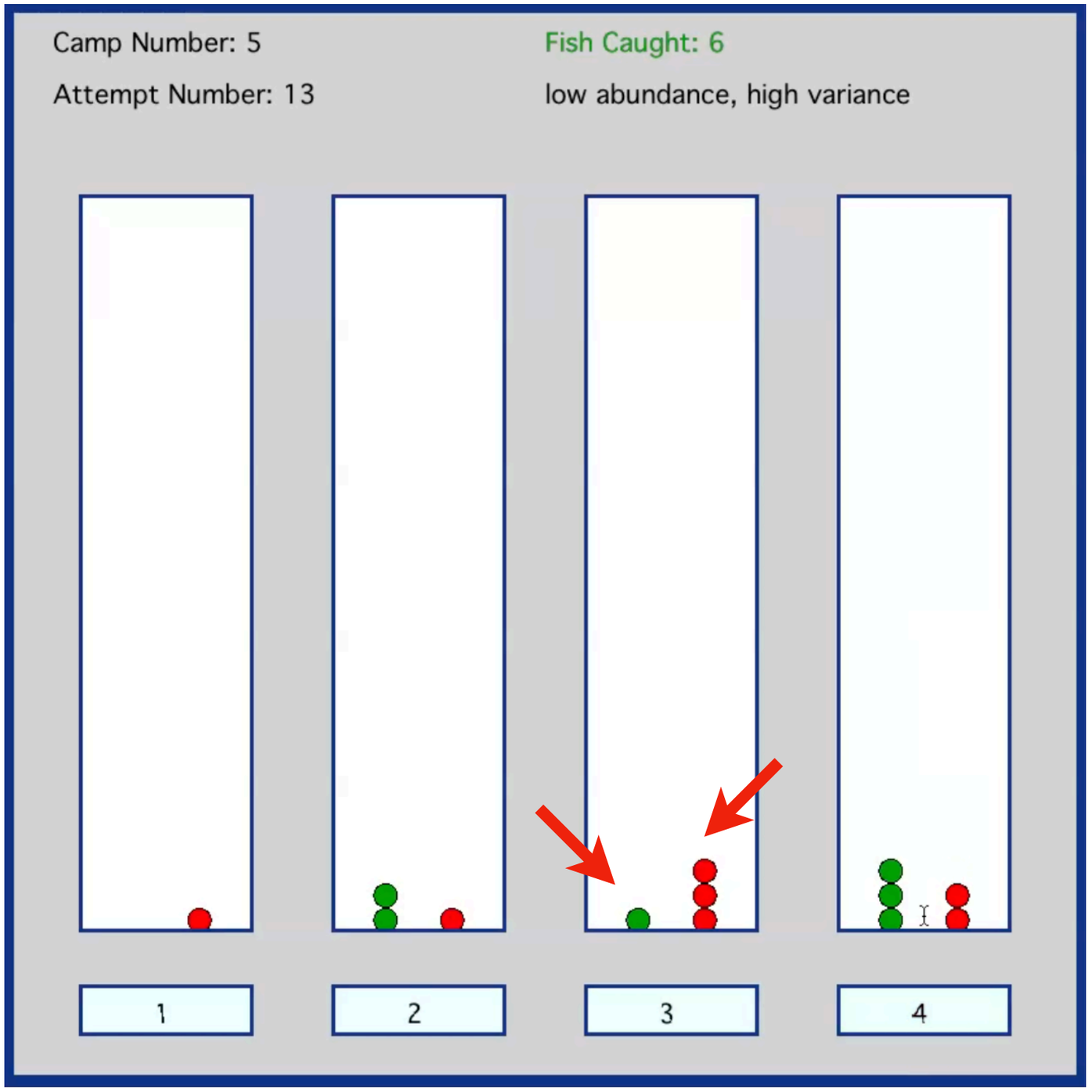
- green: reward

Experimental Interface



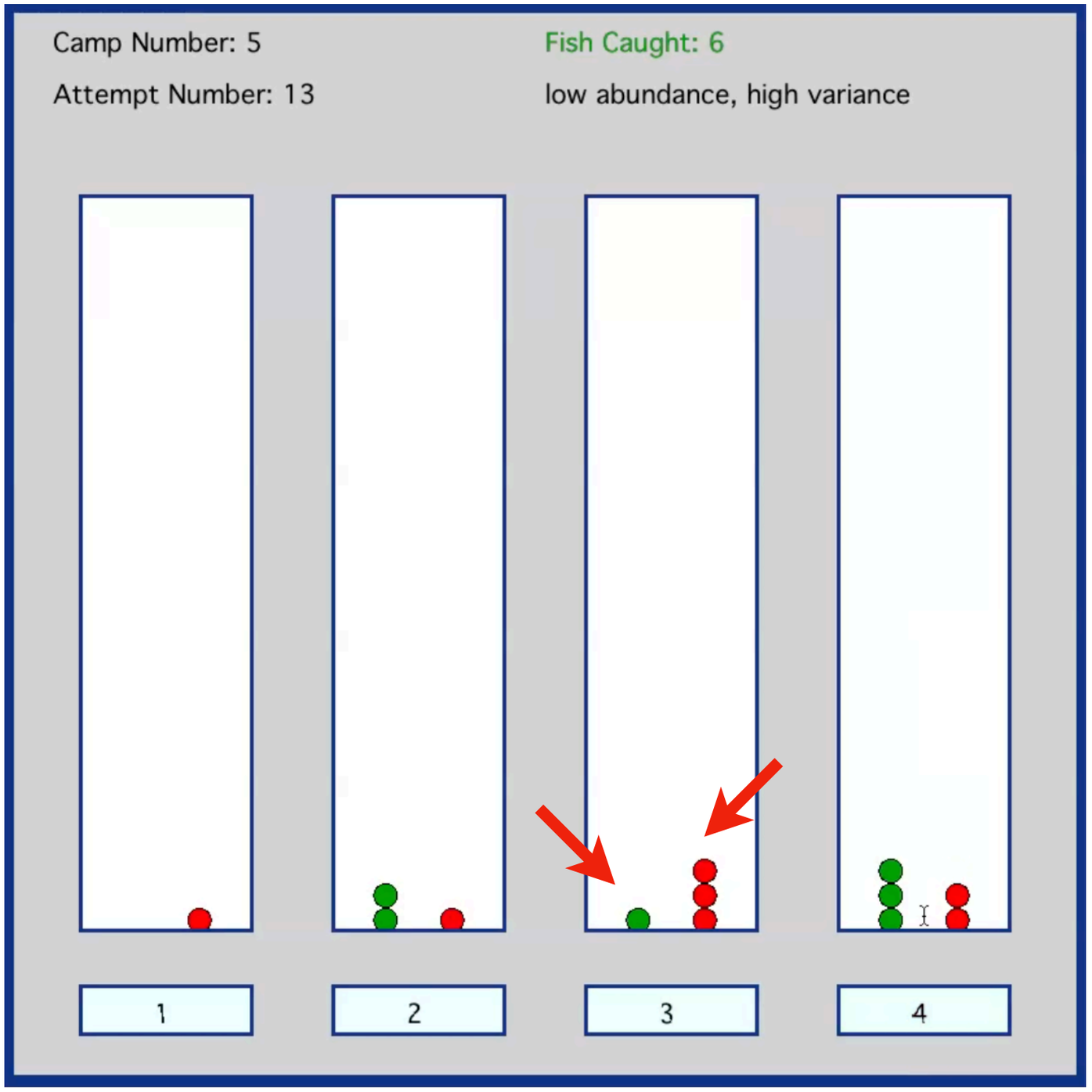
- green: reward
- red: no reward

Experimental Interface



- green: reward
- red: no reward
- also info about environment & attempt #

Experimental Interface



- green: reward
- red: no reward
- also info about environment & attempt #
- visual display minimizes working memory demands/variability

Baseline Comparisons

Baseline Comparisons

Optimal policy

Baseline Comparisons

Optimal policy

- maximizes expected total future reward

Baseline Comparisons

Optimal policy

- maximizes expected total future reward
- computationally expensive, inconsistent with human behavior (Zhang & Yu, 2013)

Baseline Comparisons

Optimal policy

- maximizes expected total future reward
- computationally expensive, inconsistent with human behavior (Zhang & Yu, 2013)
- approximate with Knowledge Gradient (KG), approximately optimal (Rhyzov et al, 2012)

Baseline Comparisons

Optimal policy

- maximizes expected total future reward
- computationally expensive, inconsistent with human behavior (Zhang & Yu, 2013)
- approximate with Knowledge Gradient (KG), approximately optimal (Rhyzov et al, 2012)
- provides “upper-bound”

Baseline Comparisons

Optimal policy

- maximizes expected total future reward
- computationally expensive, inconsistent with human behavior (Zhang & Yu, 2013)
- approximate with Knowledge Gradient (KG), approximately optimal (Rhyzov et al, 2012)
- provides “upper-bound”

Random policy

Baseline Comparisons

Optimal policy

- maximizes expected total future reward
- computationally expensive, inconsistent with human behavior (Zhang & Yu, 2013)
- approximate with Knowledge Gradient (KG), approximately optimal (Rhyzov et al, 2012)
- provides “upper-bound”

Random policy

- chooses randomly & equally among the arms

Baseline Comparisons

Optimal policy

- maximizes expected total future reward
- computationally expensive, inconsistent with human behavior (Zhang & Yu, 2013)
- approximate with Knowledge Gradient (KG), approximately optimal (Rhyzov et al, 2012)
- provides "upper-bound"

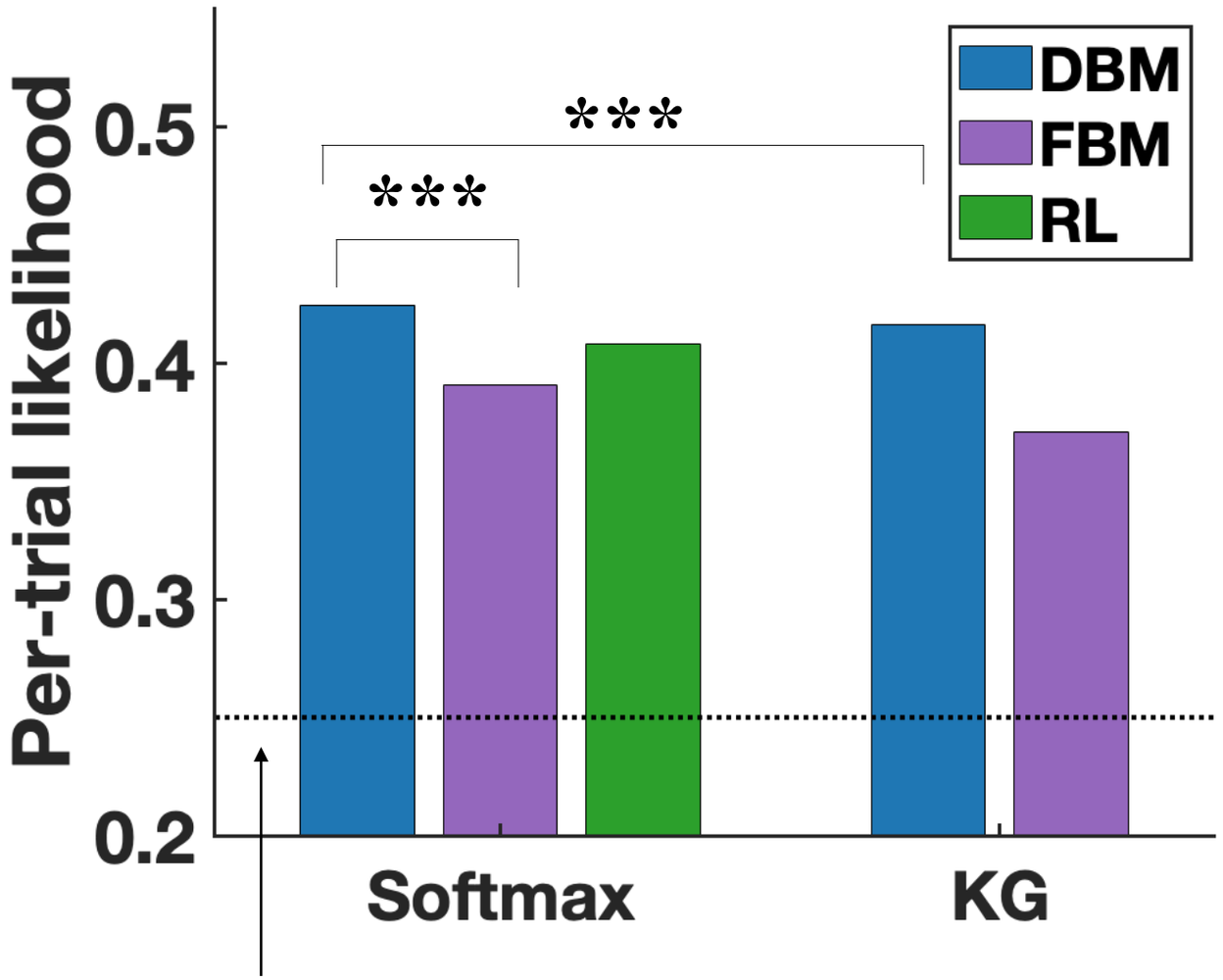
Random policy

- chooses randomly & equally among the arms
- expected reward = prior mean, e.g. 2/3, 1/3

Decision Model: Knowledge Gradient (KG)

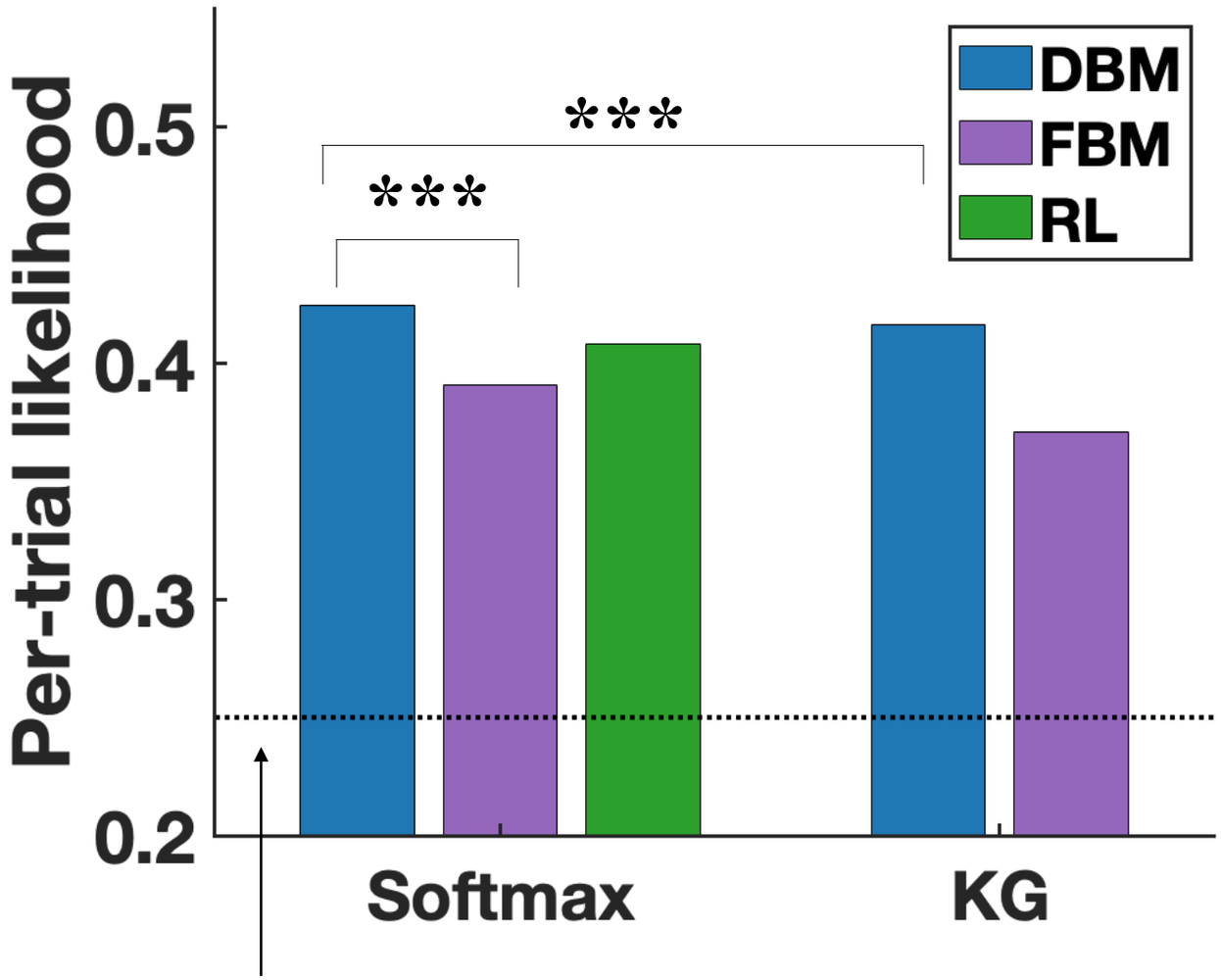
- An **approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy
- **Myopic**: when computing the knowledge gain, it commits to one more exploratory decision, and assumes to exploit in all remaining trials
- Decision rule: **current estimated reward rate** + **knowledge gain * horizon** (how many trials left)

Model Comparison



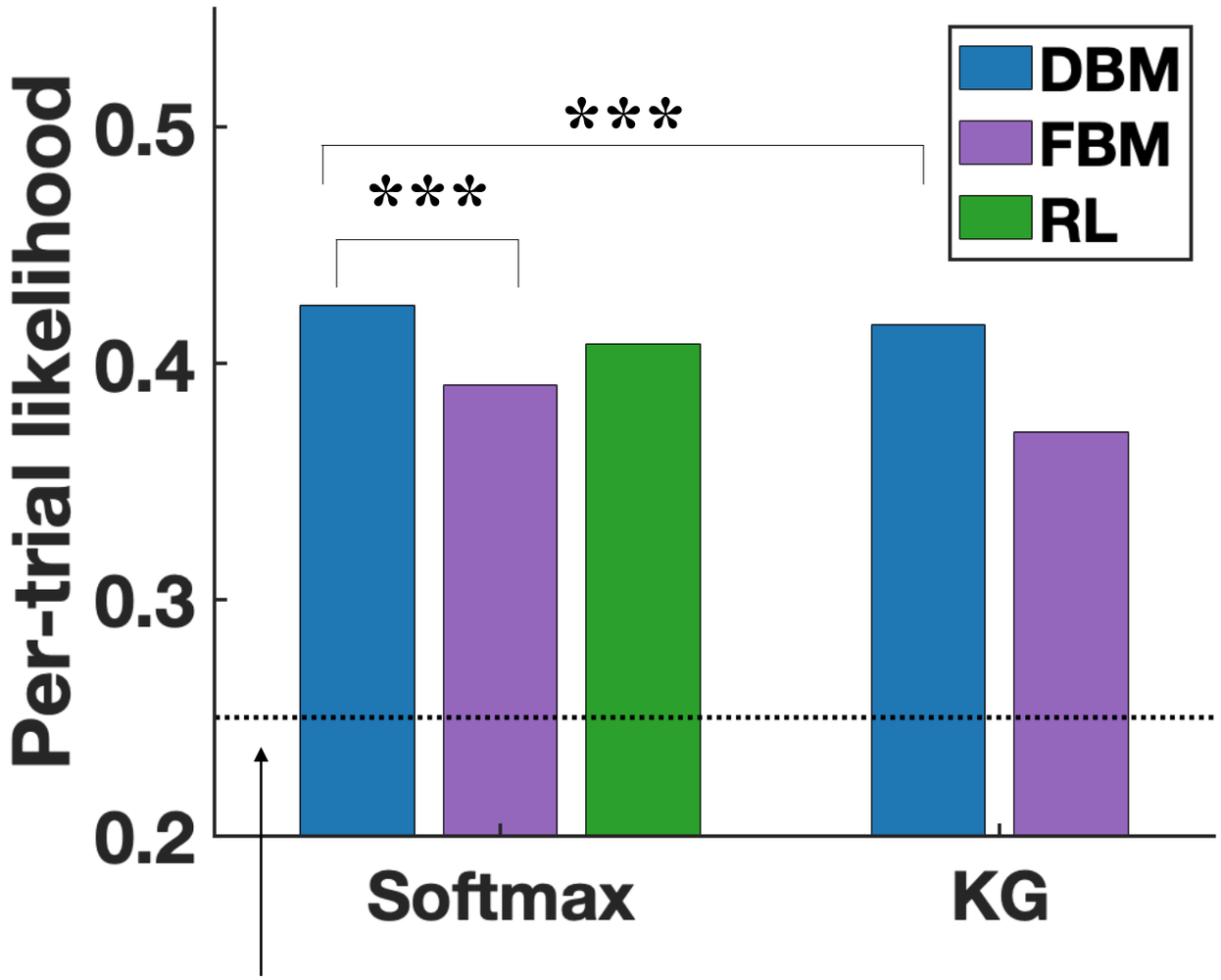
chance level (0.25)

Model Comparison



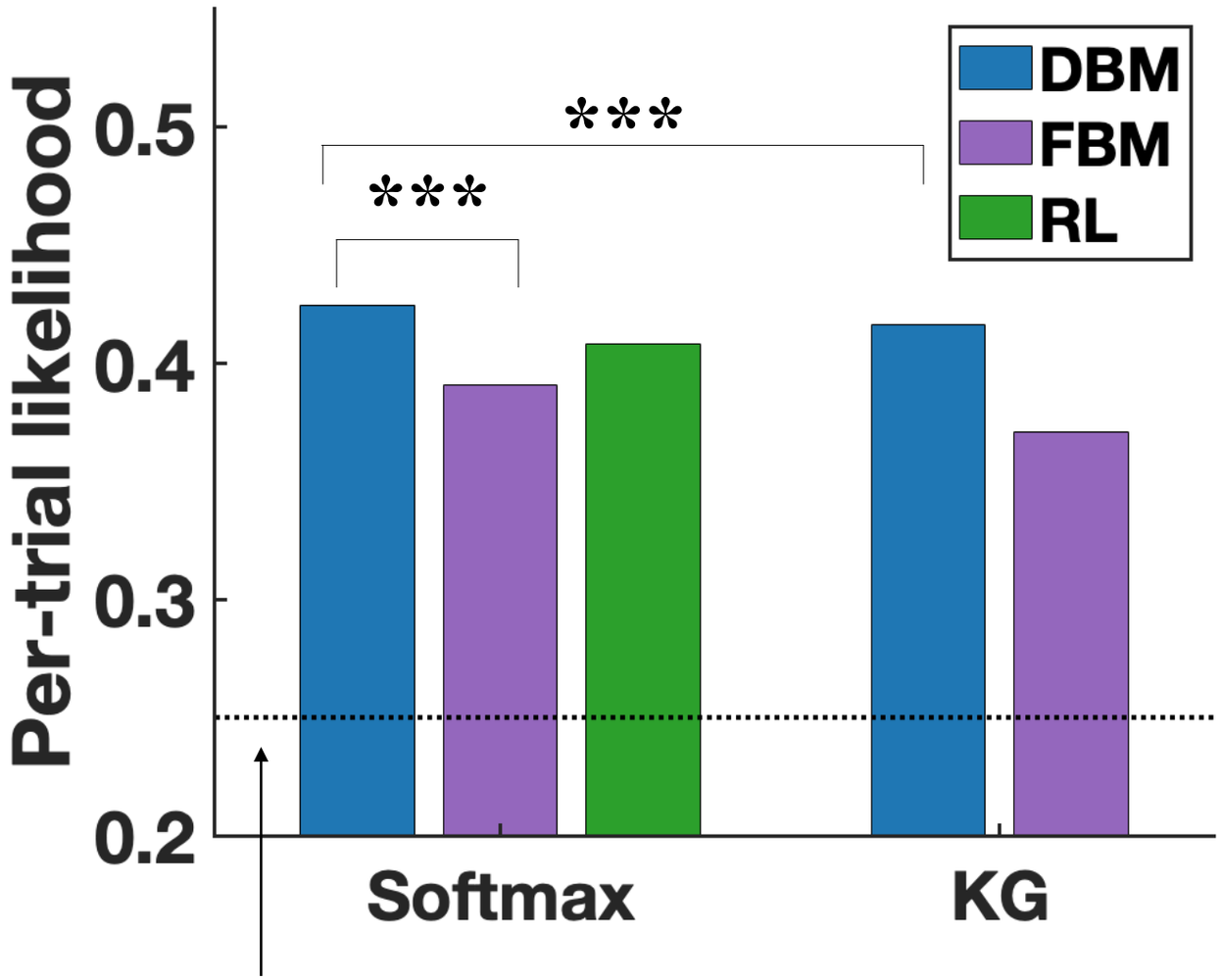
- 10-fold cross-validation (5 games held out)

Model Comparison



- 10-fold cross-validation (5 games held out)
- DBM+softmax predicts per-trial choice best

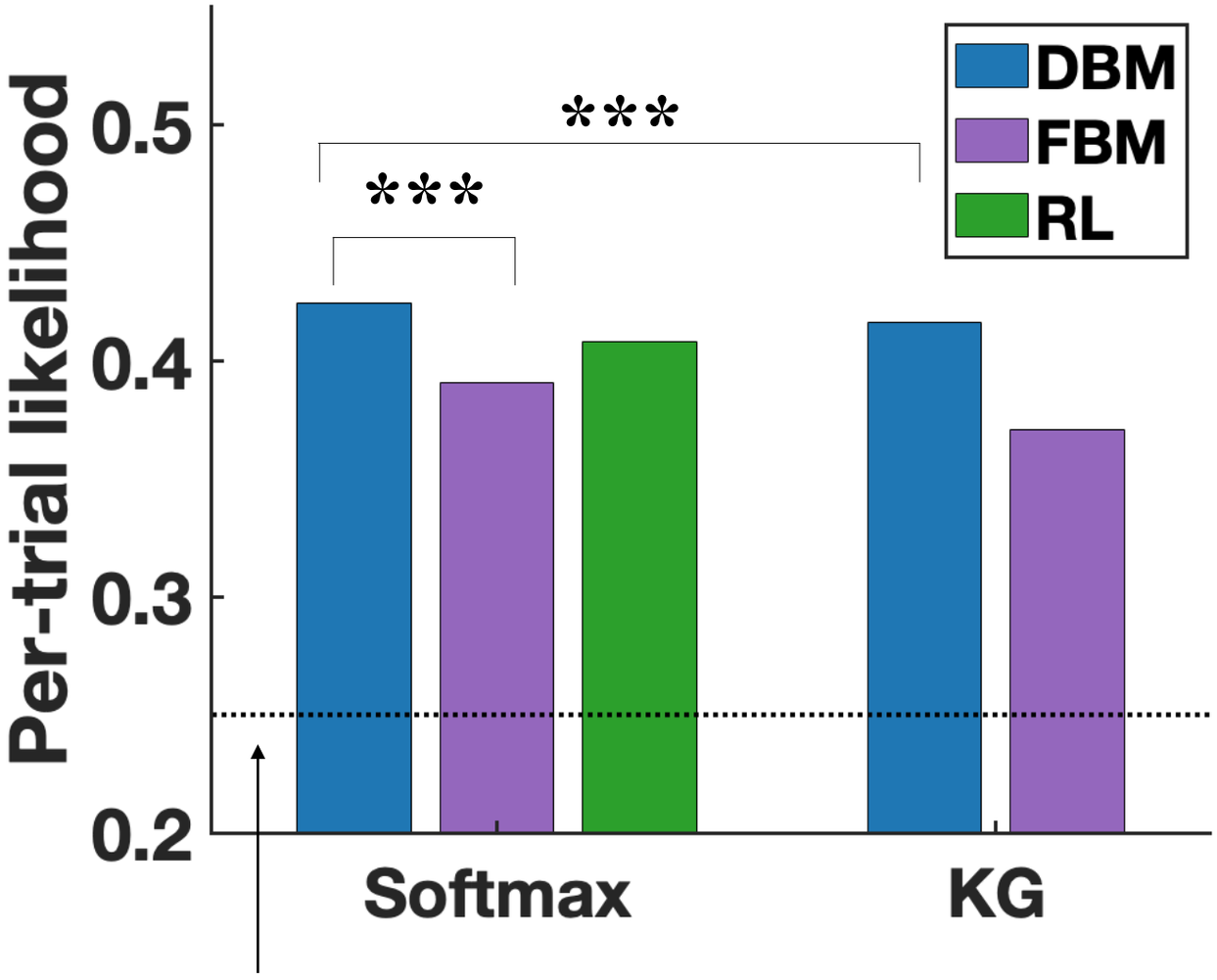
Model Comparison



chance level (0.25)

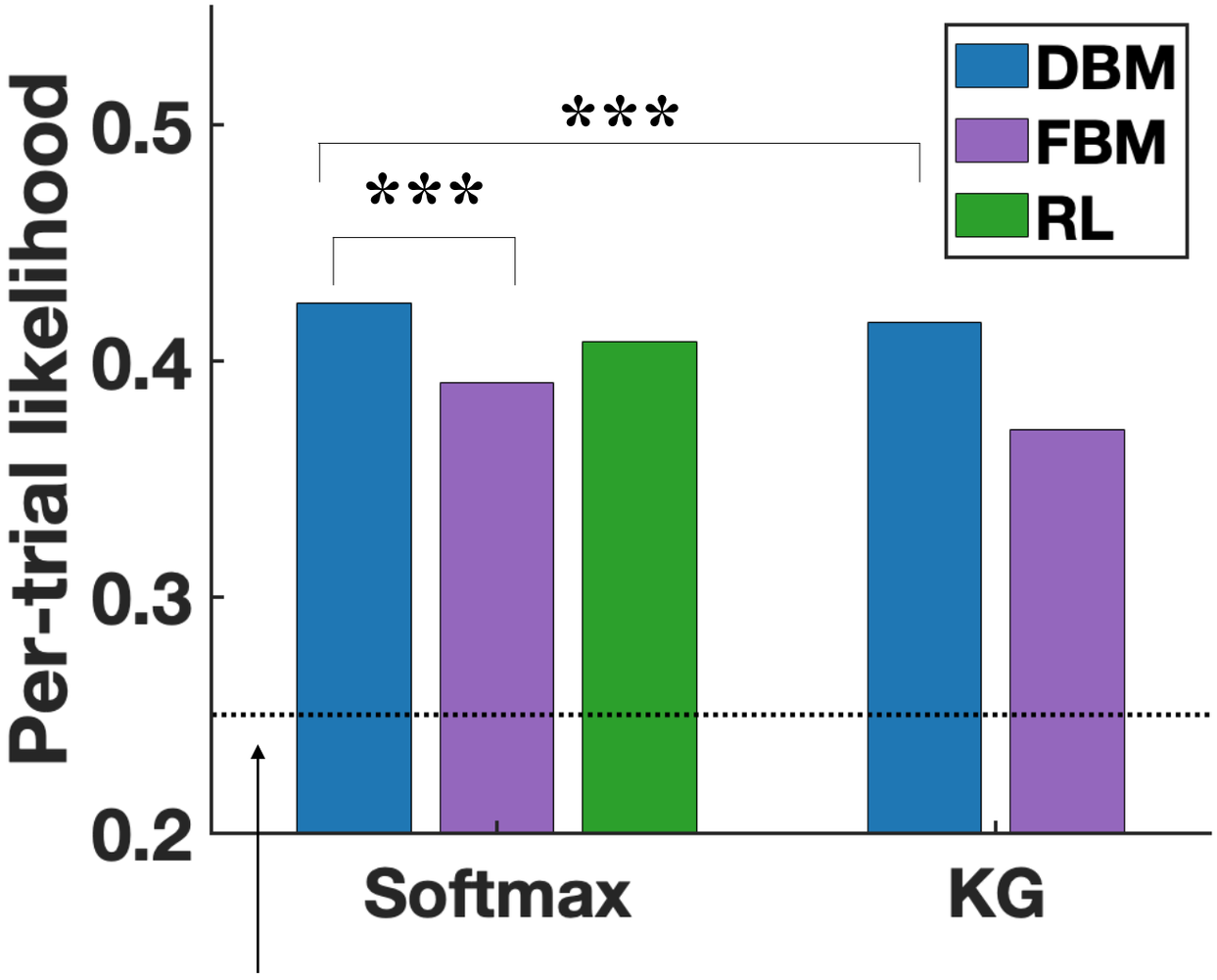
- 10-fold cross-validation (5 games held out)
- DBM+softmax predicts per-trial choice best
- volatility overestimation

Model Comparison



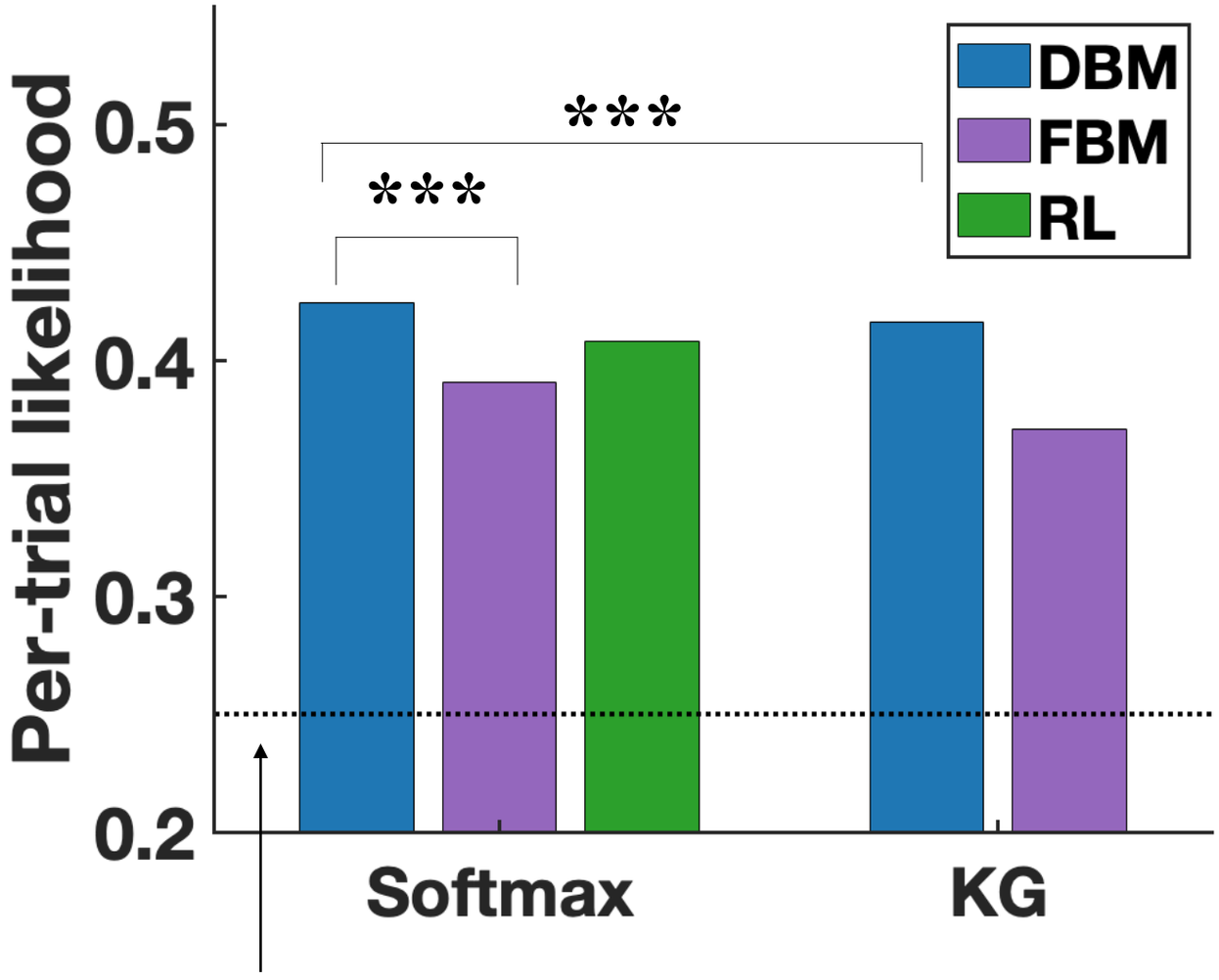
- 10-fold cross-validation (5 games held out)
- DBM+softmax predicts per-trial choice best
- volatility overestimation
- persistent prior bias

Model Comparison



- 10-fold cross-validation (5 games held out)
- DBM+softmax predicts per-trial choice best
- volatility overestimation
- persistent prior bias
- simplistic decision policy

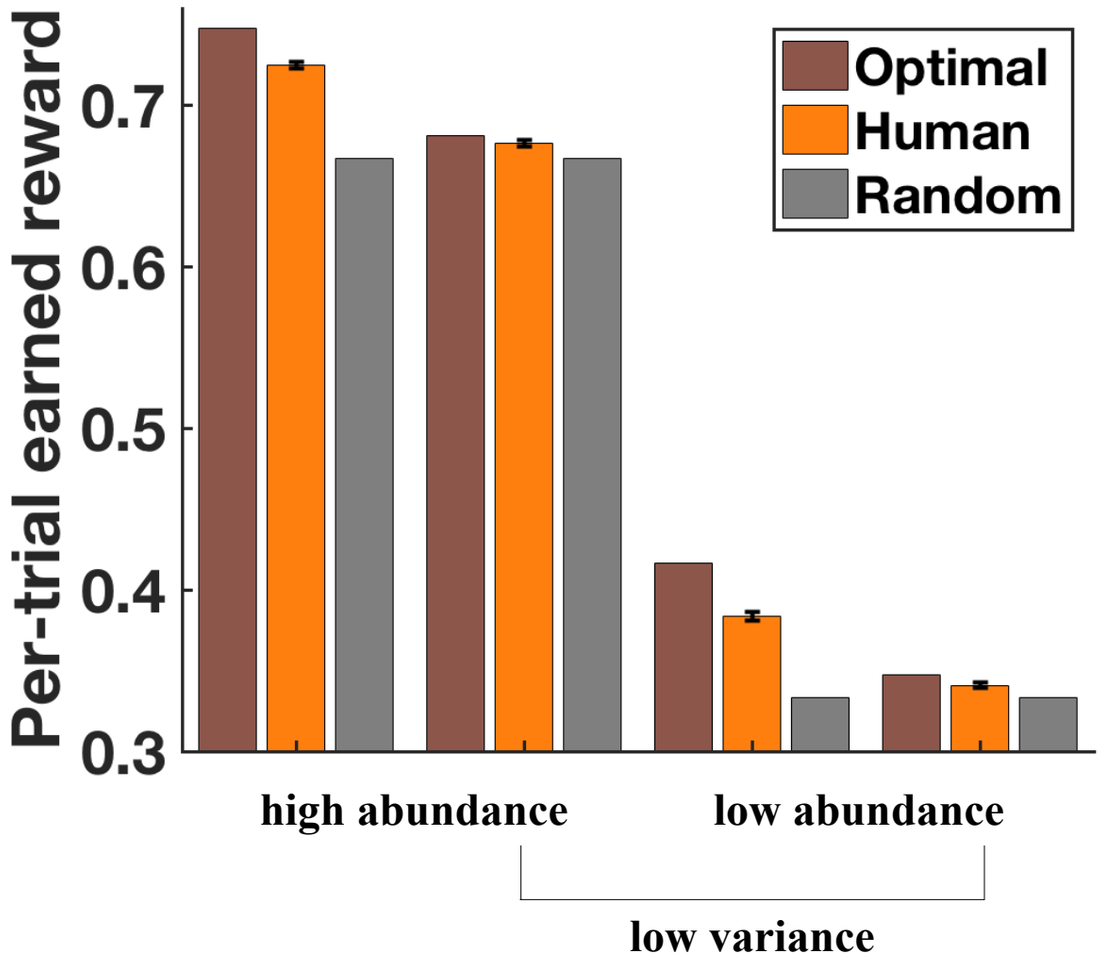
Model Comparison



- 10-fold cross-validation (5 games held out)
- DBM+softmax predicts per-trial choice best
- volatility overestimation
- persistent prior bias
- simplistic decision policy
- Hallmark of DBM: lose-shift rate sensitive to frequency of reward in recent history

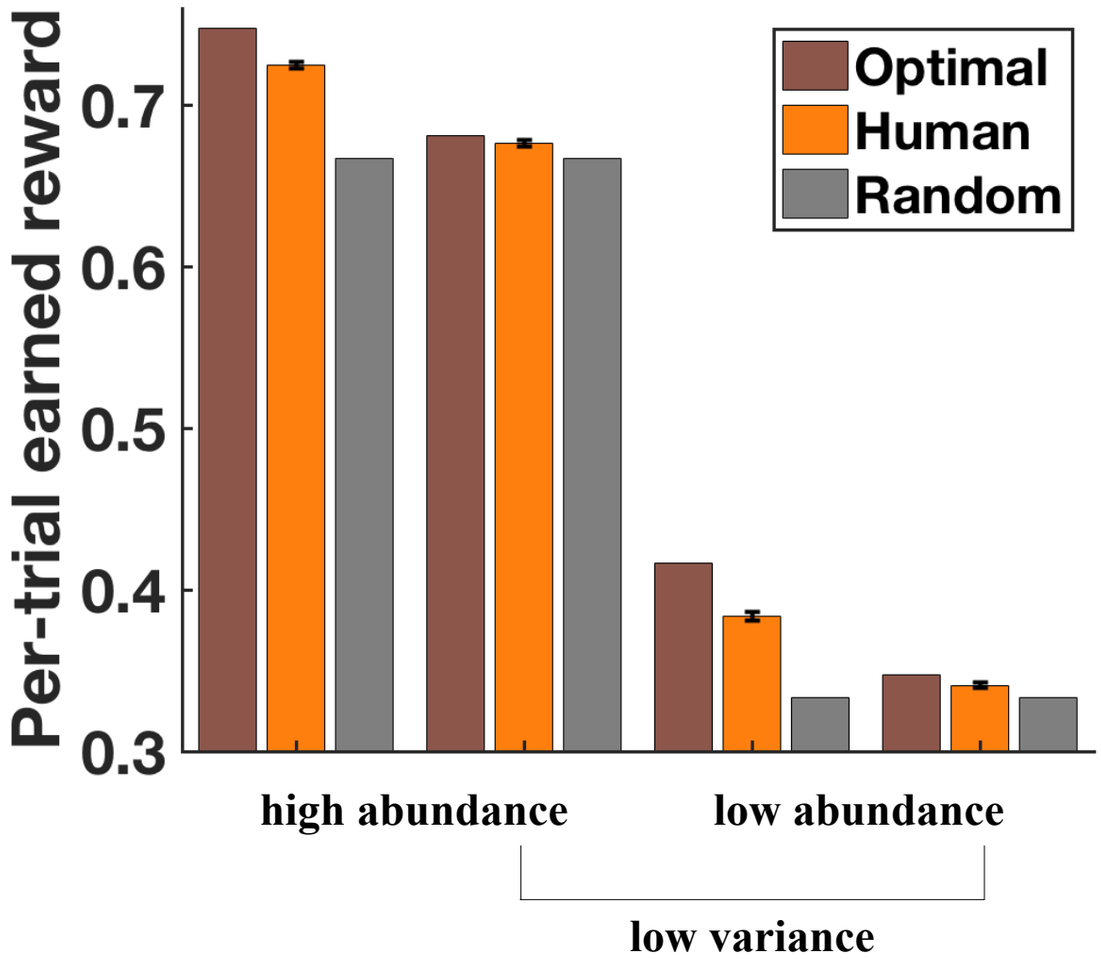
Human Comparison to Baselines

Performance: average reward per trial



Human Comparison to Baselines

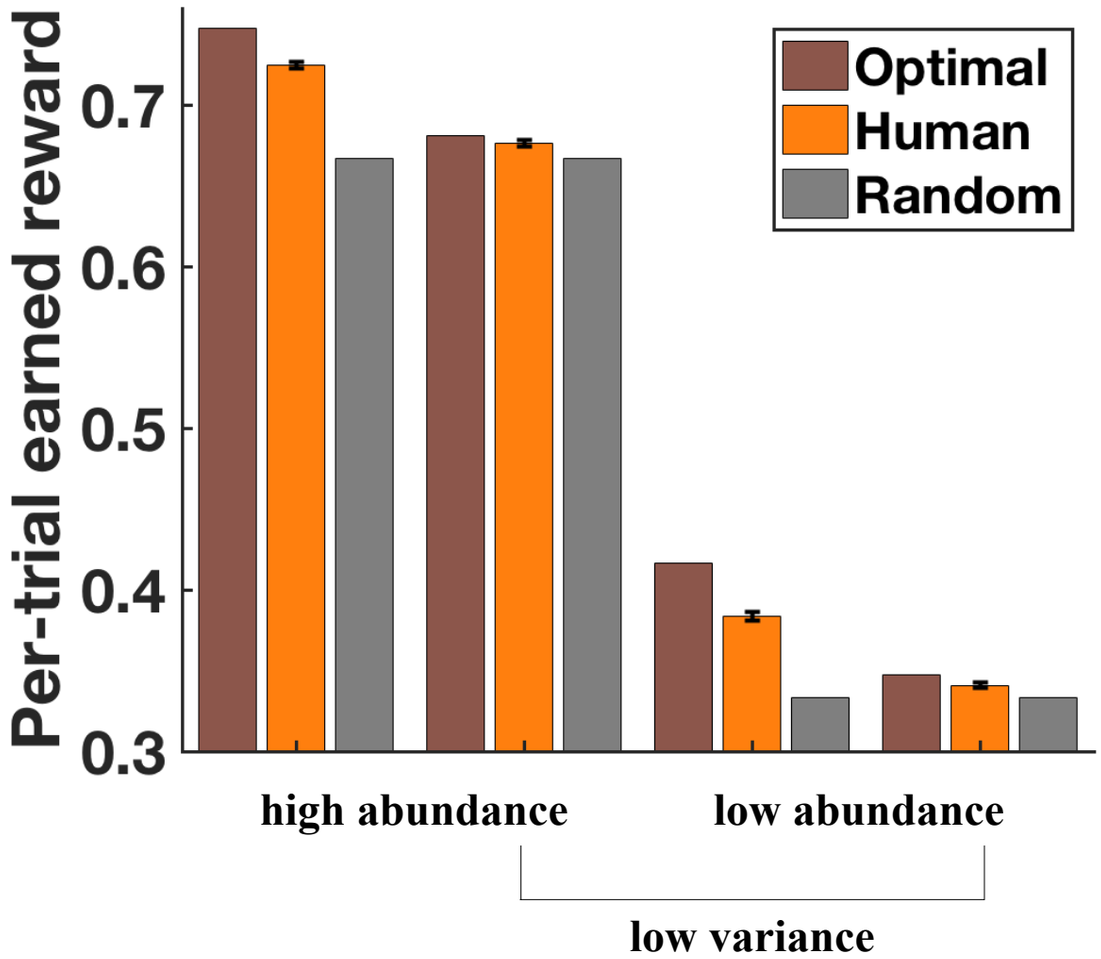
Performance: average reward per trial



- All 4 environments: subjects close to optimal, and significantly better than random policy ($p<0.01$)

Human Comparison to Baselines

Performance: average reward per trial



- All 4 environments: subjects close to optimal, and significantly better than random policy ($p<0.01$)
- Biased E[reward] even more mysterious: experienced average reward > true prior mean

Nearly Optimal: Knowledge Gradient (KG)

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy
- Myopically approximate value of exploration (assume one more exploratory decision, followed by exploitation)

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy
- Myopically approximate value of exploration (assume one more exploratory decision, followed by exploitation)

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy
- Myopically approximate value of exploration (assume one more exploratory decision, followed by exploitation)

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

- Decision rule:

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy
- Myopically approximate value of exploration (assume one more exploratory decision, followed by exploitation)

$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

- Decision rule:

$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

Nearly Optimal: Knowledge Gradient (KG)

- **Approximation** to the optimal policy *(Ryzhov et al., 2012)*
- Computationally cheaper than the optimal policy
- Myopically approximate value of exploration (assume one more exploratory decision, followed by exploitation)

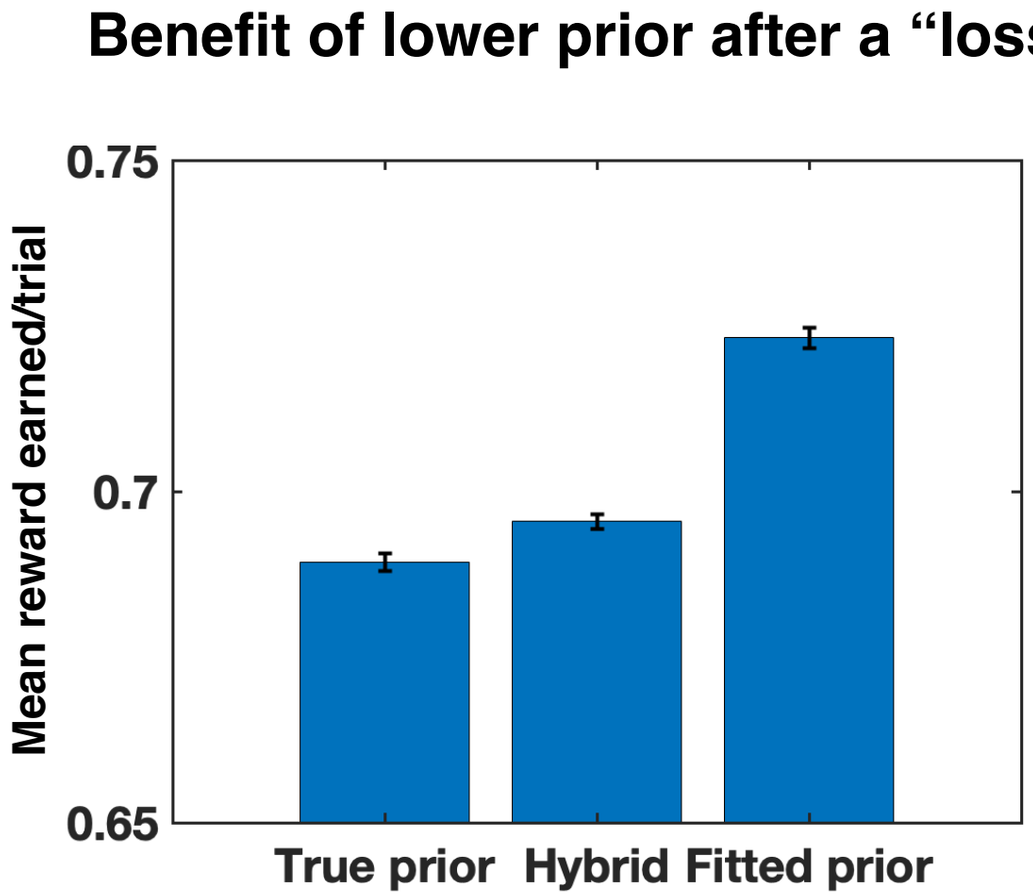
$$v_k^{\mathrm{KG},t} = \mathbb{E}\left[\max_{k'} \theta_{k'}^{t+1} \,|\, D^t = k, \mathbf{q}^t\right] - \max_{k'} \theta_{k'}^t$$

- Decision rule:

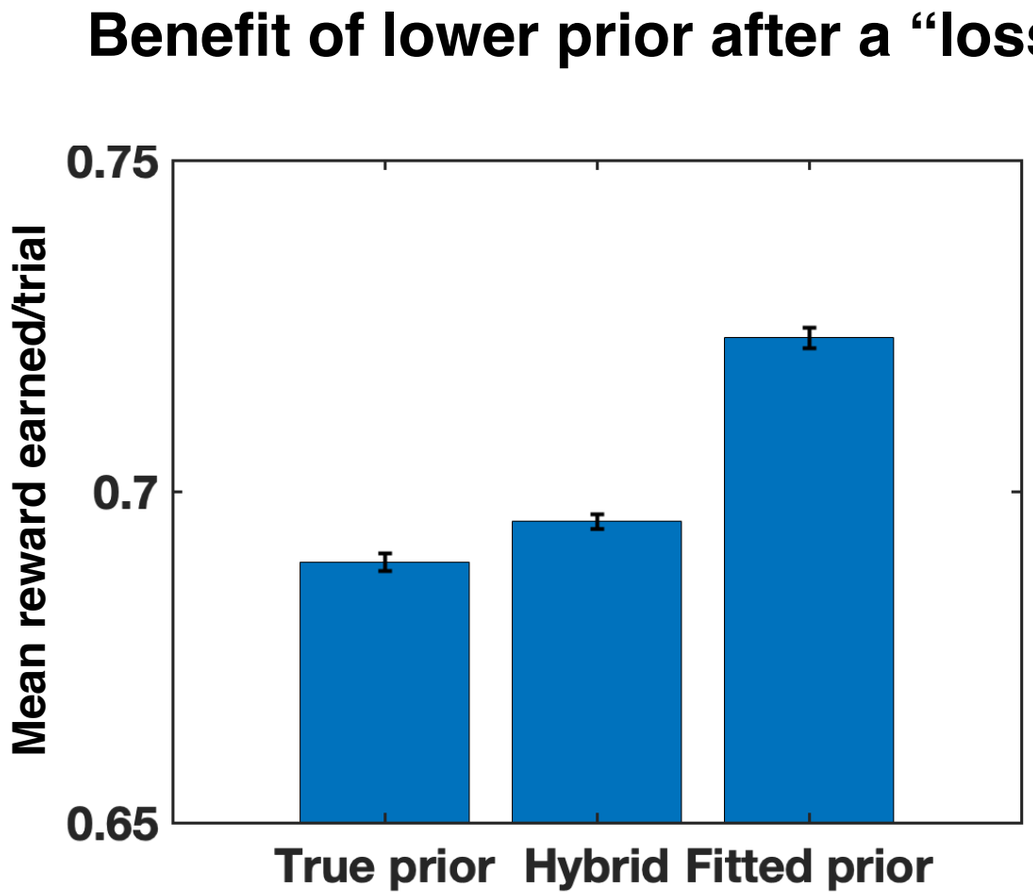
$$D^{\mathrm{KG},t} = \arg\max_k \theta_k^t + (T - t - 1)\, v_k^{\mathrm{KG},t}$$

- Captures different aspects of human behavior than softmax

Simulation: Lower Prior Post-Loss Helps a Bit

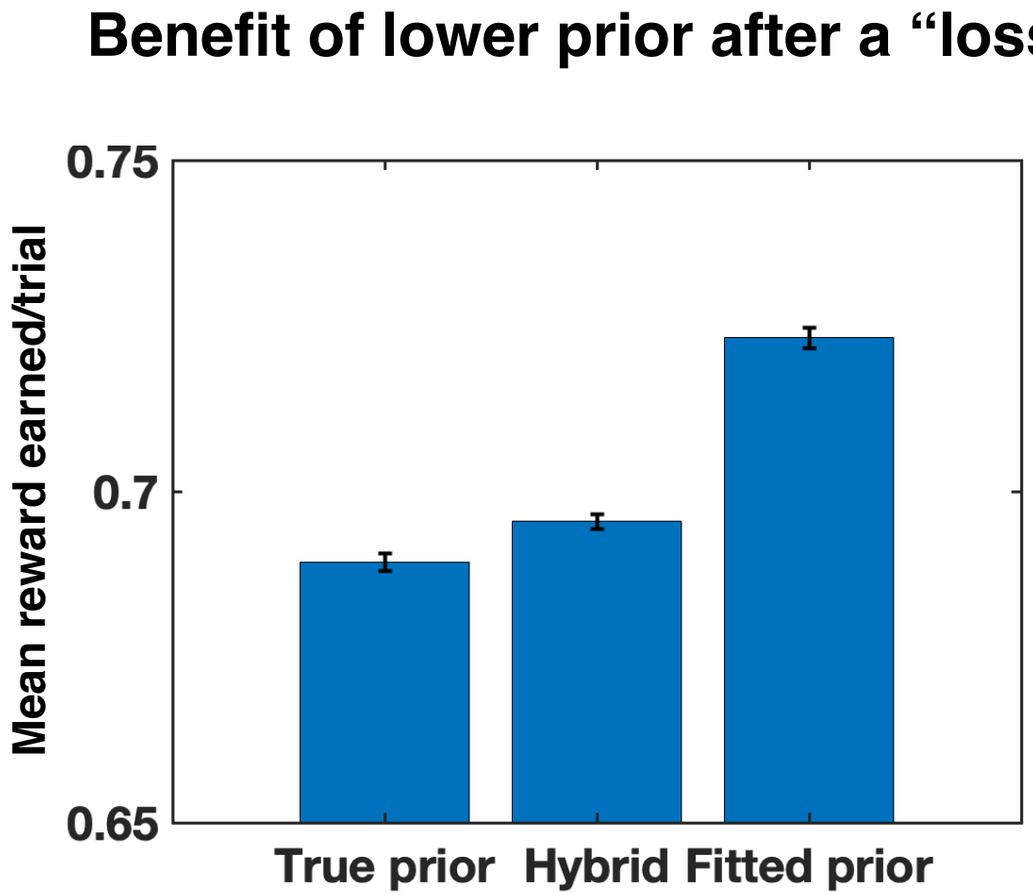


Simulation: Lower Prior Post-Loss Helps a Bit



- Fitted prior (θ_0= .35) earns more reward than true prior (θ_0=.67)

Simulation: Lower Prior Post-Loss Helps a Bit



- Fitted prior (θ_0= .35) earns more reward than true prior (θ_0=.67)
- Hybrid model (replacing true prior with fitted prior after each **loss**) captures **some** of the benefit

DBM = Persistently Biased RL (pbRL)

DBM = Persistently Biased RL (pbRL)

$$\hat{\theta}_k^t = a + b\hat{\theta}_k^{t-1} + cR_t$$

DBM = Persistently Biased RL (pbRL)

$$\hat{\theta}_k^t = a + b\hat{\theta}_k^{t-1} + cR_t$$

- pbRL has simply leaky integrating dynamics (neurally plausible)

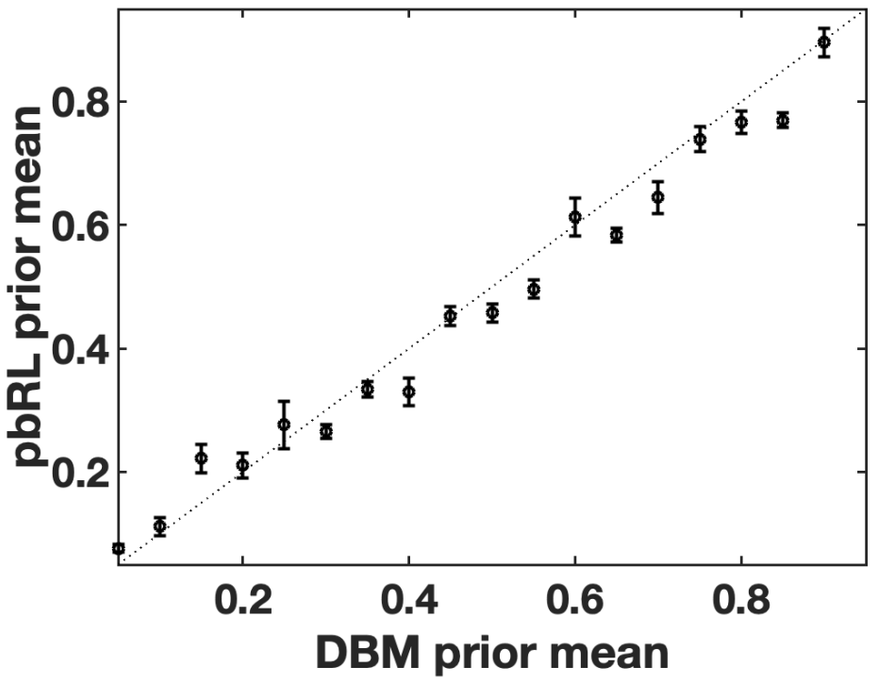
DBM = Persistently Biased RL (pbRL)

$$\hat{\theta}_k^t = a + b\hat{\theta}_k^{t-1} + cR_t$$

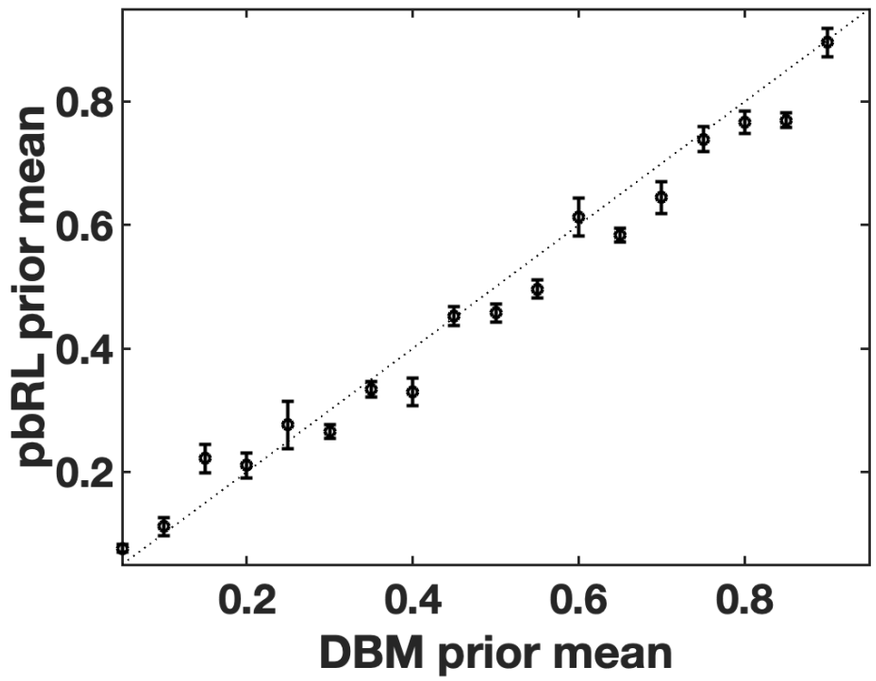
- pbRL has simply leaky integrating dynamics (neurally plausible)
- Proven equivalence between DBM and pbRL makes it possible to fit pbRL directly and statistically interpret parameters

DBM = Persistently Biased RL (pbRL)

Estimation of *α*



Estimation of prior mean



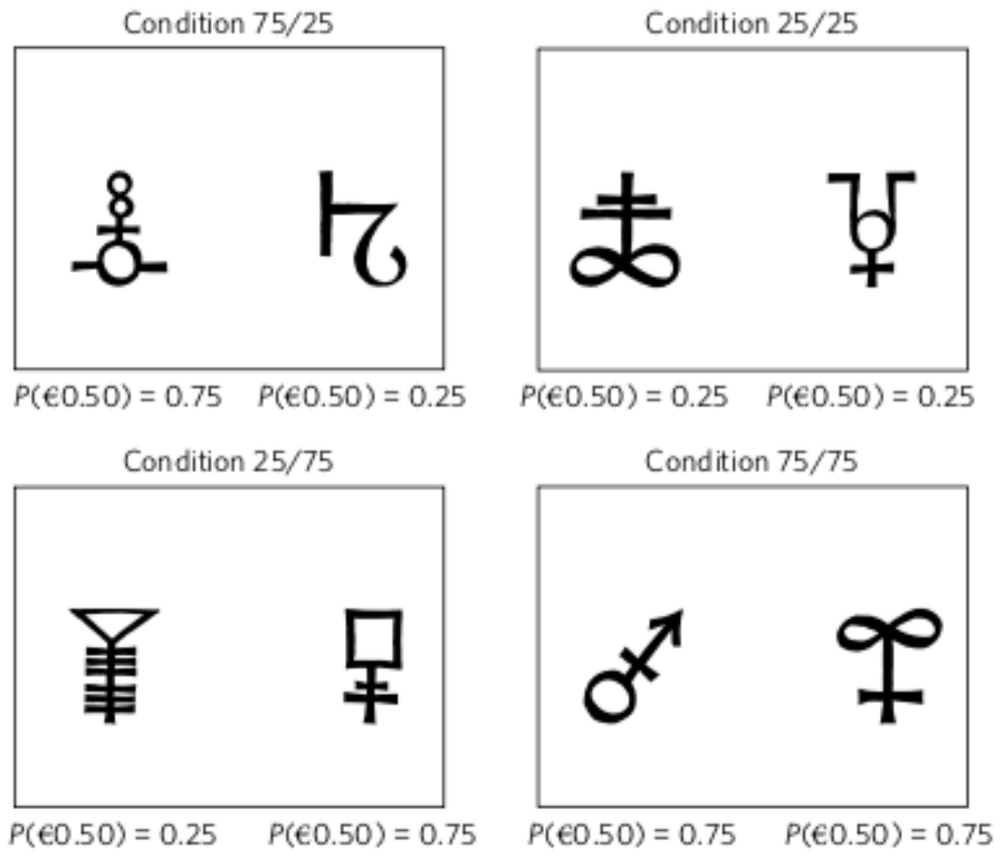
$$\hat{\theta}_k^t = a + b\hat{\theta}_k^{t-1} + cR_t$$

- pbRL has simply leaky integrating dynamics (neurally plausible)
- Proven equivalence between DBM and pbRL makes it possible to fit pbRL directly and statistically interpret parameters
- Validated on our dataset ⇒ normative (statistical) grounding of RL

And What of Optimism Bias?

(Lefebvre et al, 2017)

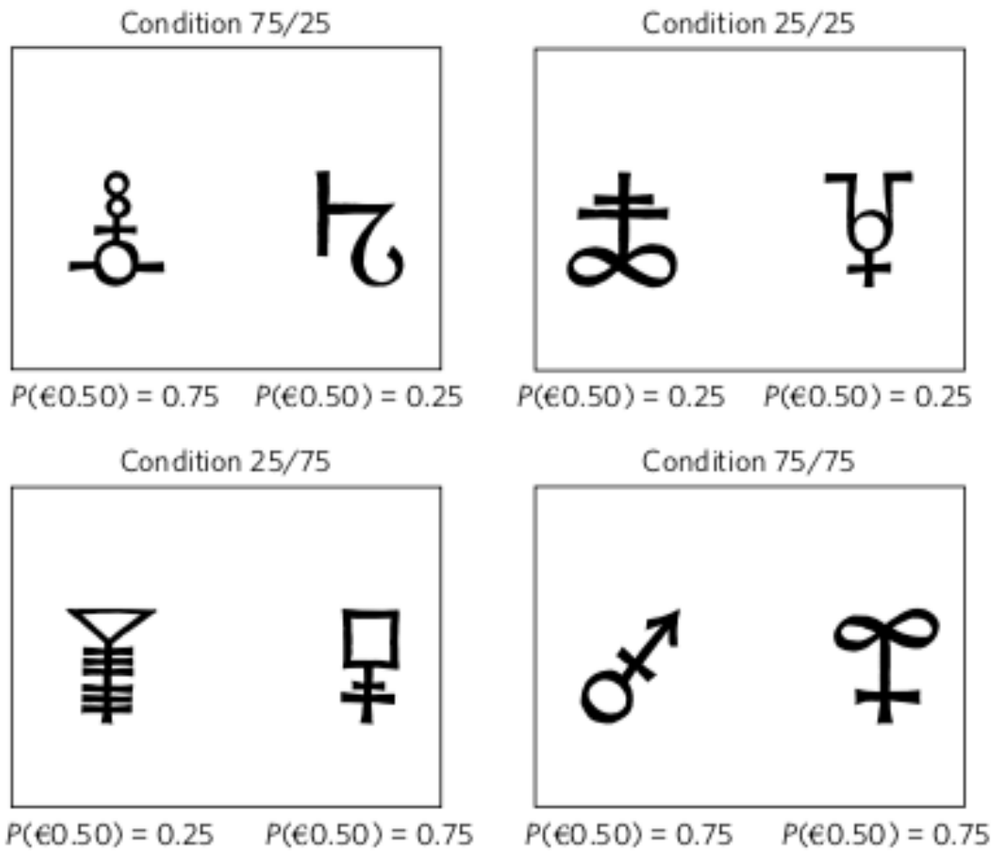
Experimental Design



And What of Optimism Bias?

(Lefebvre et al, 2017)

Experimental Design

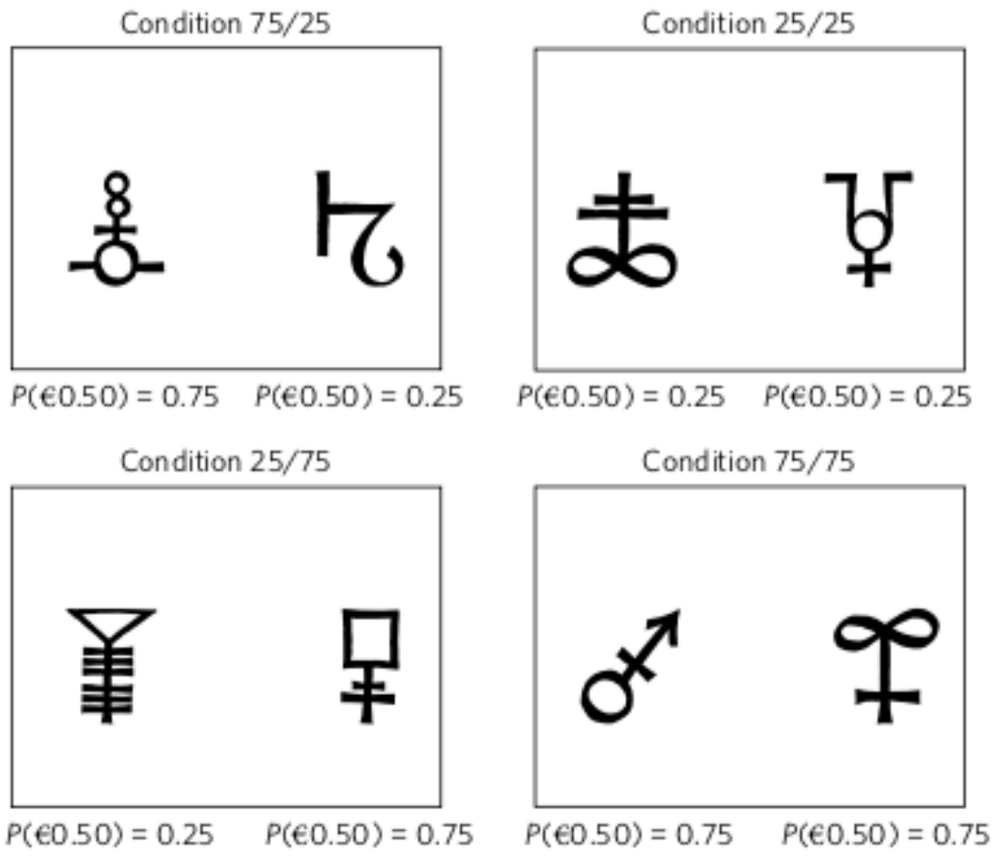


- 2-armed bandit task

And What of Optimism Bias?

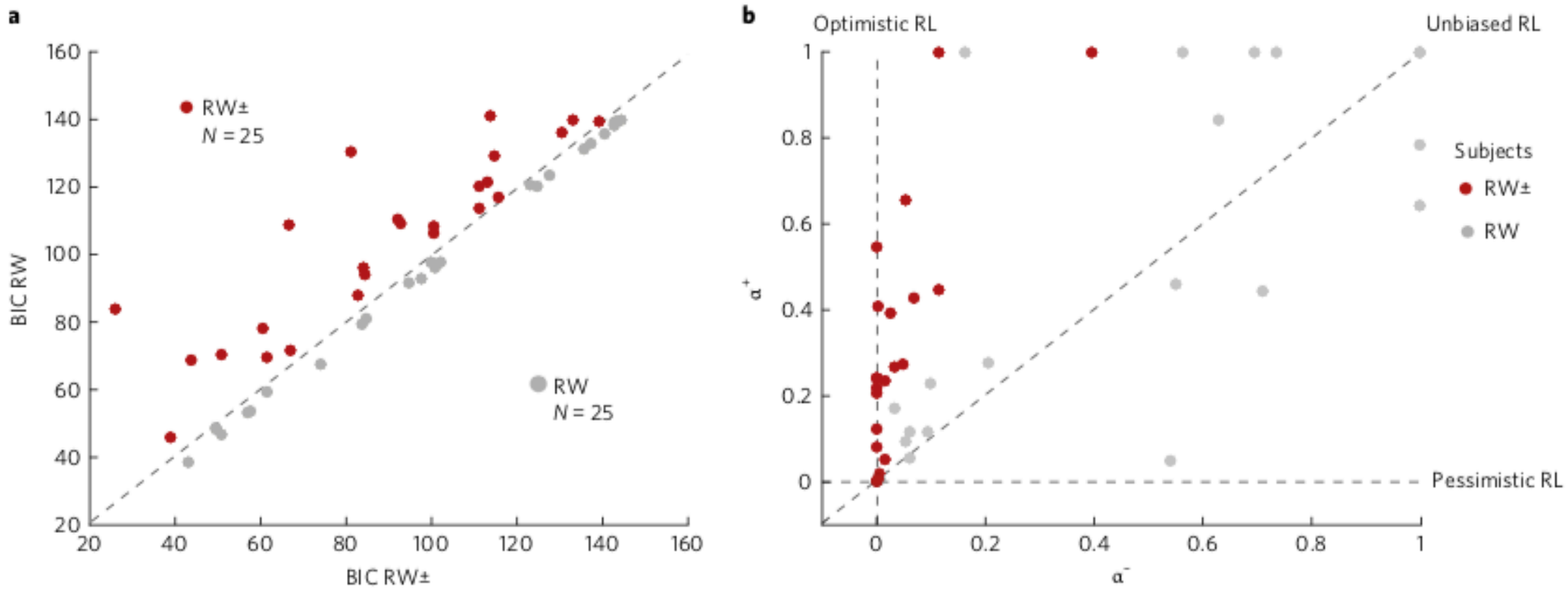
(Lefebvre et al, 2017)

Experimental Design



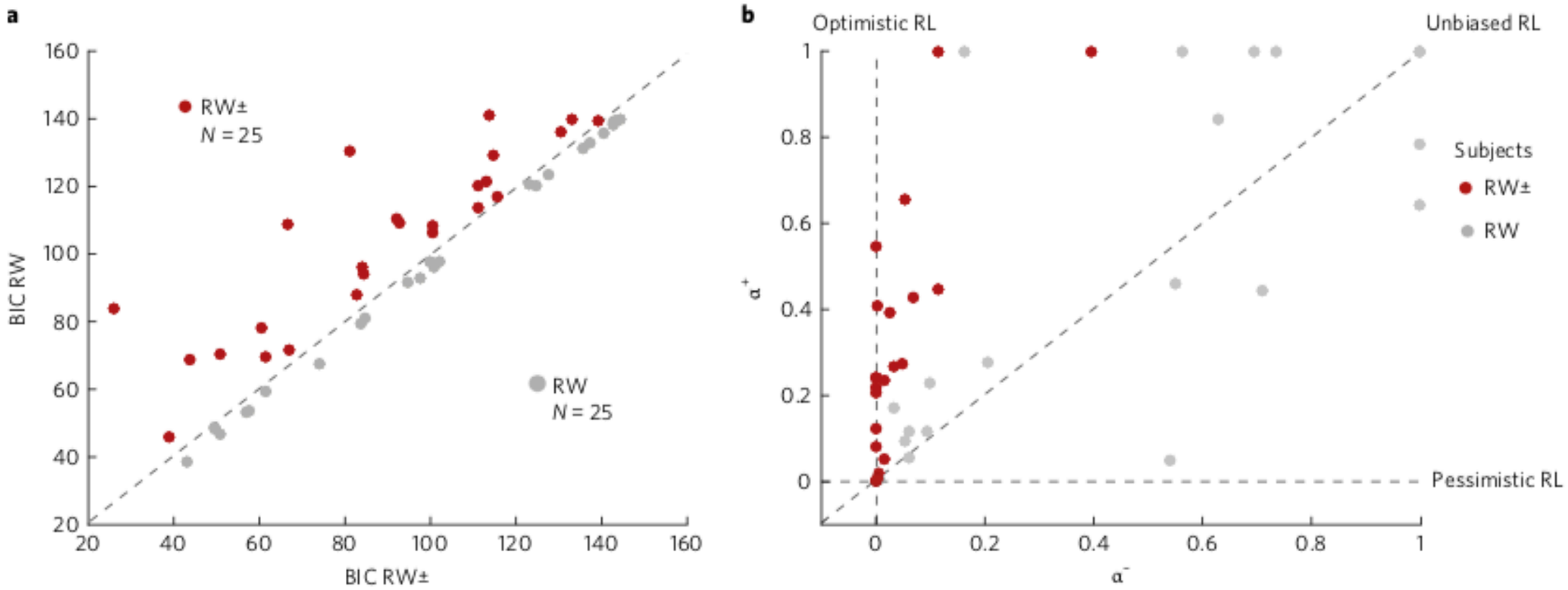
- 2-armed bandit task
- 2 x 2 design: mean x variance

And What of Optimism Bias?



$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

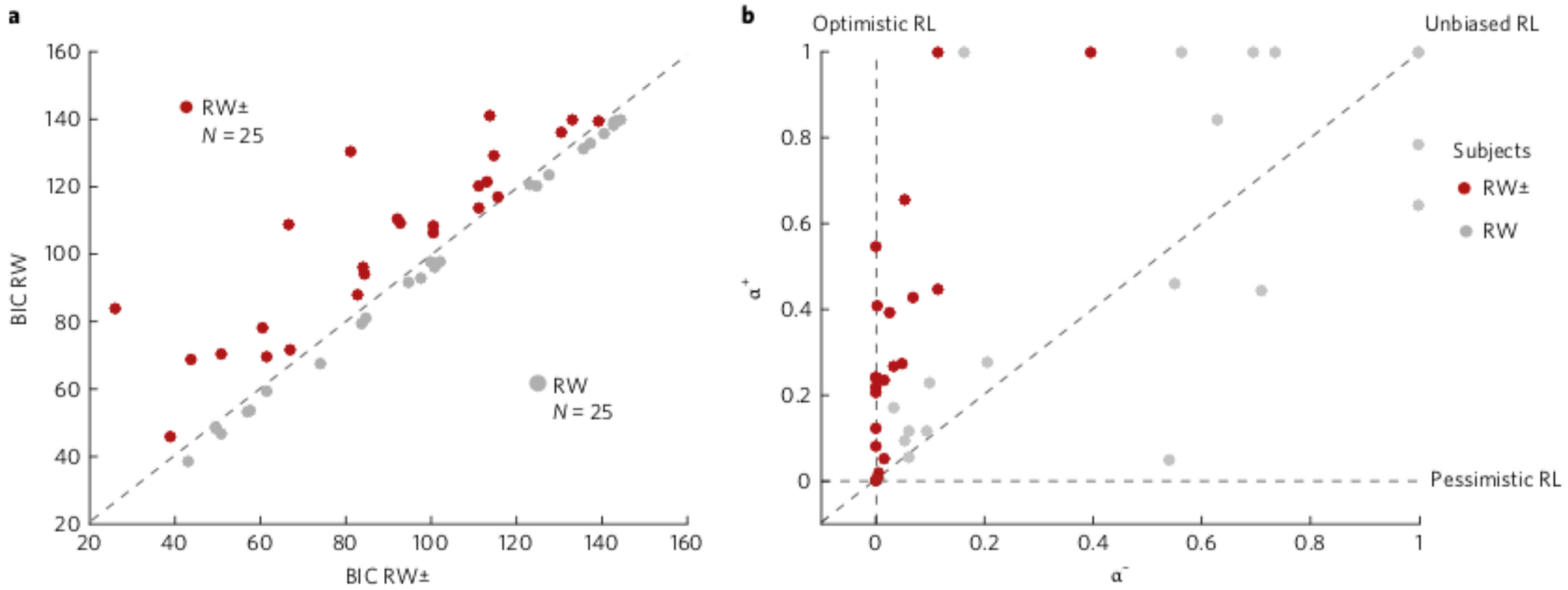
And What of Optimism Bias?



$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017): 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$

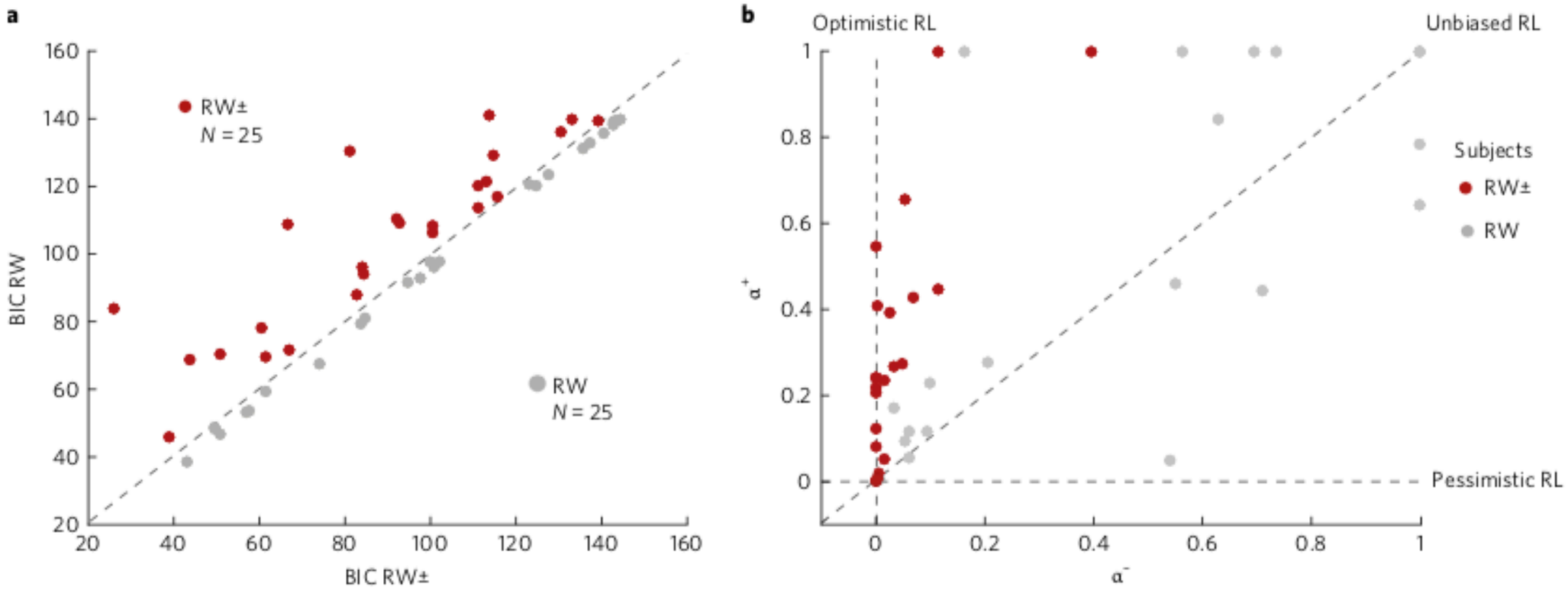
And What of Optimism Bias?



$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017): 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$
- Could this actually be explained by reward rate under-estimation?

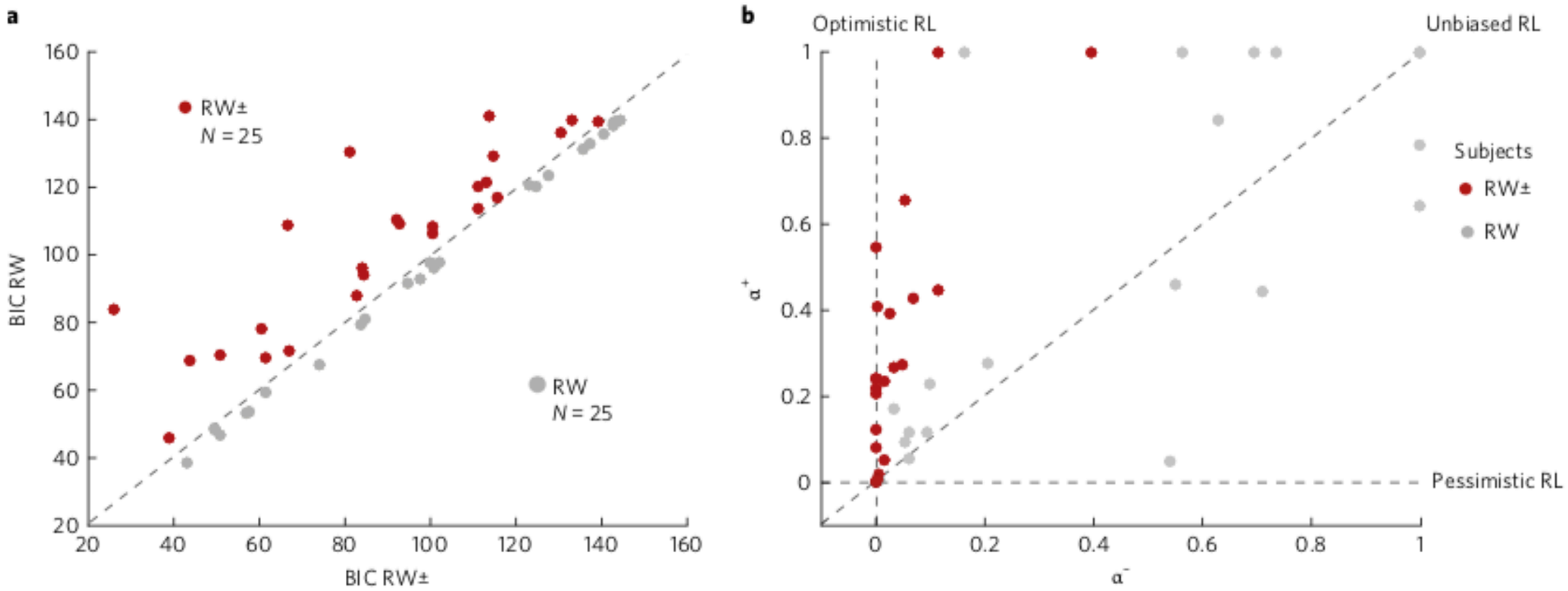
And What of Optimism Bias?



$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017): 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$
- Could this actually be explained by reward rate under-estimation?
  - apparent contradiction: devaluation of chosen arm = pessimism

And What of Optimism Bias?



$$\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \begin{cases} \epsilon^+(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} > 0 \\ \epsilon^-(R_t - \hat{\theta}_k^{t-1}), & \textbf{if } R_t - \hat{\theta}_k^{t-1} < 0 \end{cases}$$

- Lefebvre et al (2017): 1/2 subjects better fit by RW± than RW; $\varepsilon^+ > \varepsilon^-$
- Could this actually be explained by reward rate under-estimation?
  - apparent contradiction: devaluation of chosen arm = pessimism
  - however: devaluation of unchosen arm = optimism