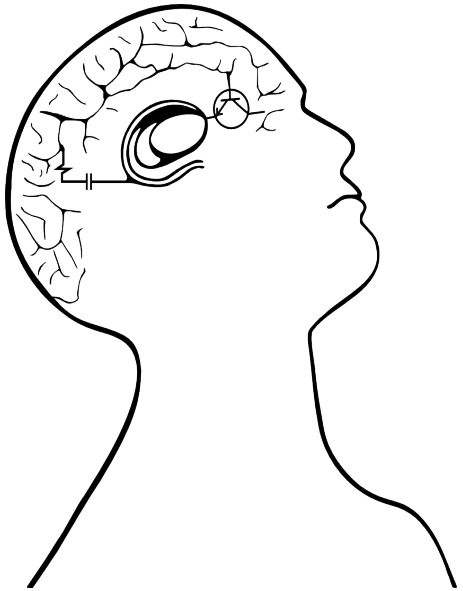
Probing for latent states during reinforcement learning

(what do you do when you don't know what to do)

Michael J Frank
Lab for Neural Computation and Cognition
Brown University

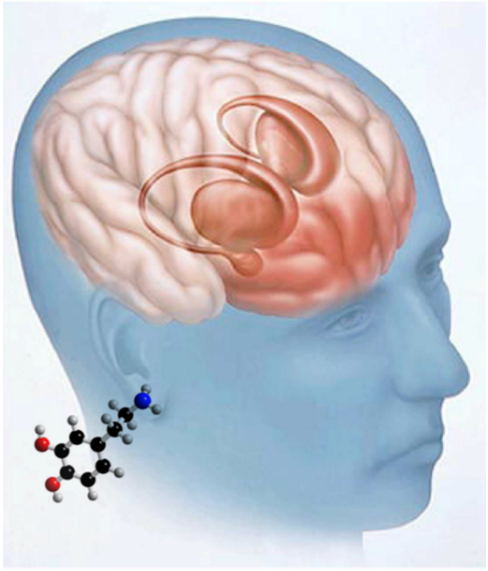


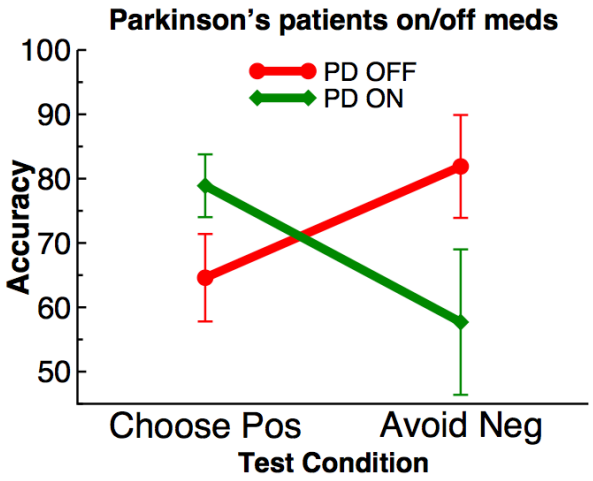
LNCC

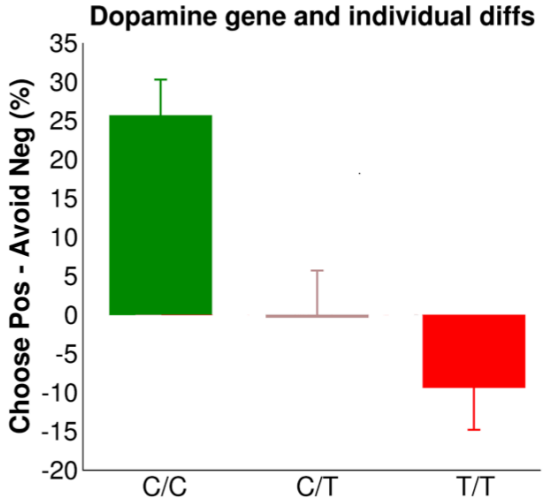
BROWN

Jeff Cockburn

Striatal dopamine and reward-based learning / choice



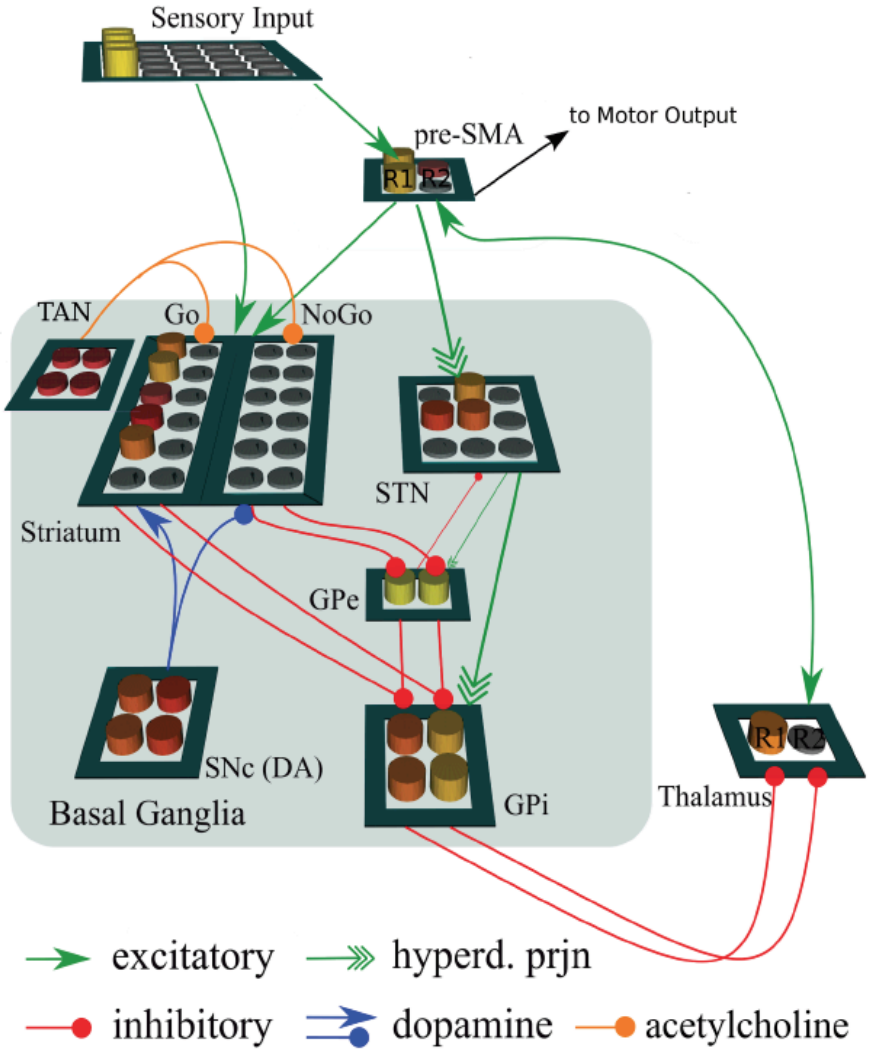


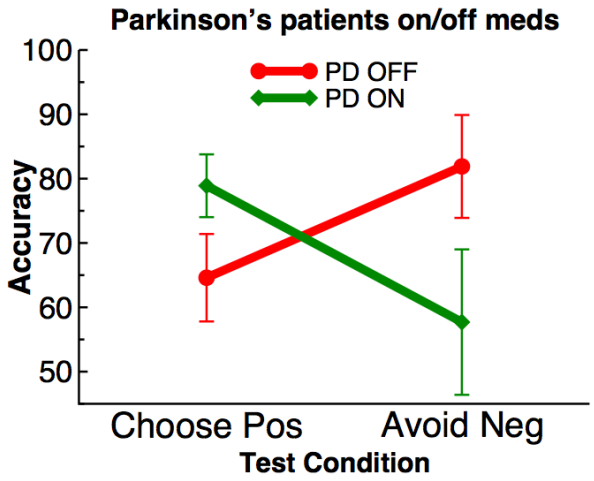


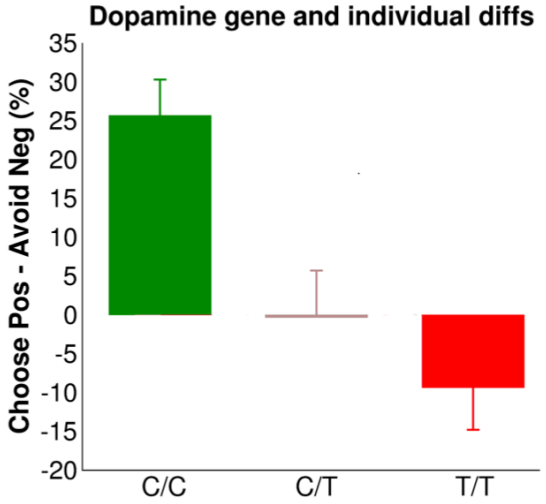
Frank et al., 2004; 2007; Doll et al 2011; Collins & Frank 2014; Cockburn et al 2014; Cox et al 2015

Pessiglione et al 2006; Palminteri et al 2009; Cools et al 2006; Jocham et al 2011….

Striatal dopamine and reward-based learning / choice



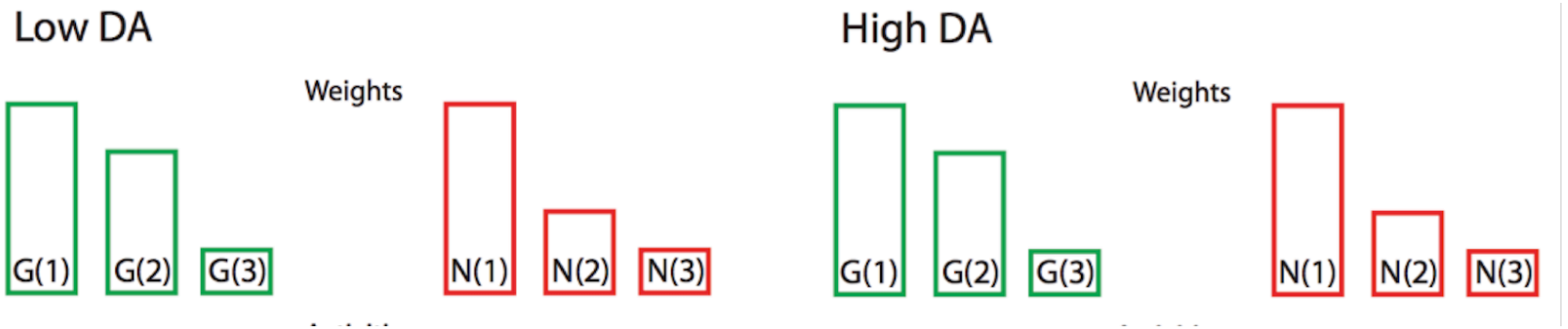




Frank et al., 2004; 2007; Doll et al 2011; Collins & Frank 2014; Cockburn et al 2014; Cox et al 2015

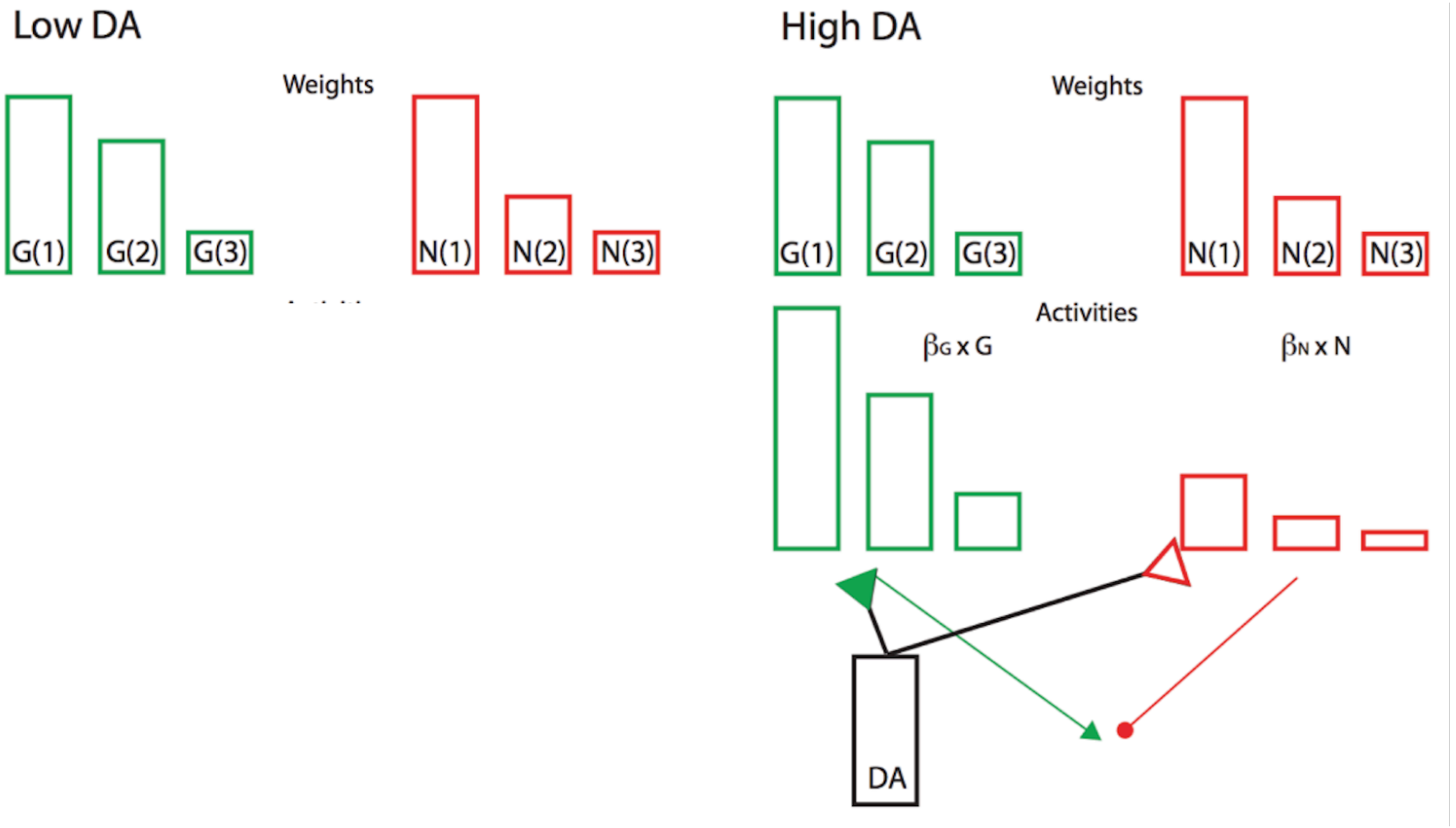
Pessiglione et al 2006; Palminteri et al 2009; Cools et al 2006; Jocham et al 2011….

Striatal dopamine and reward-based learning / choice



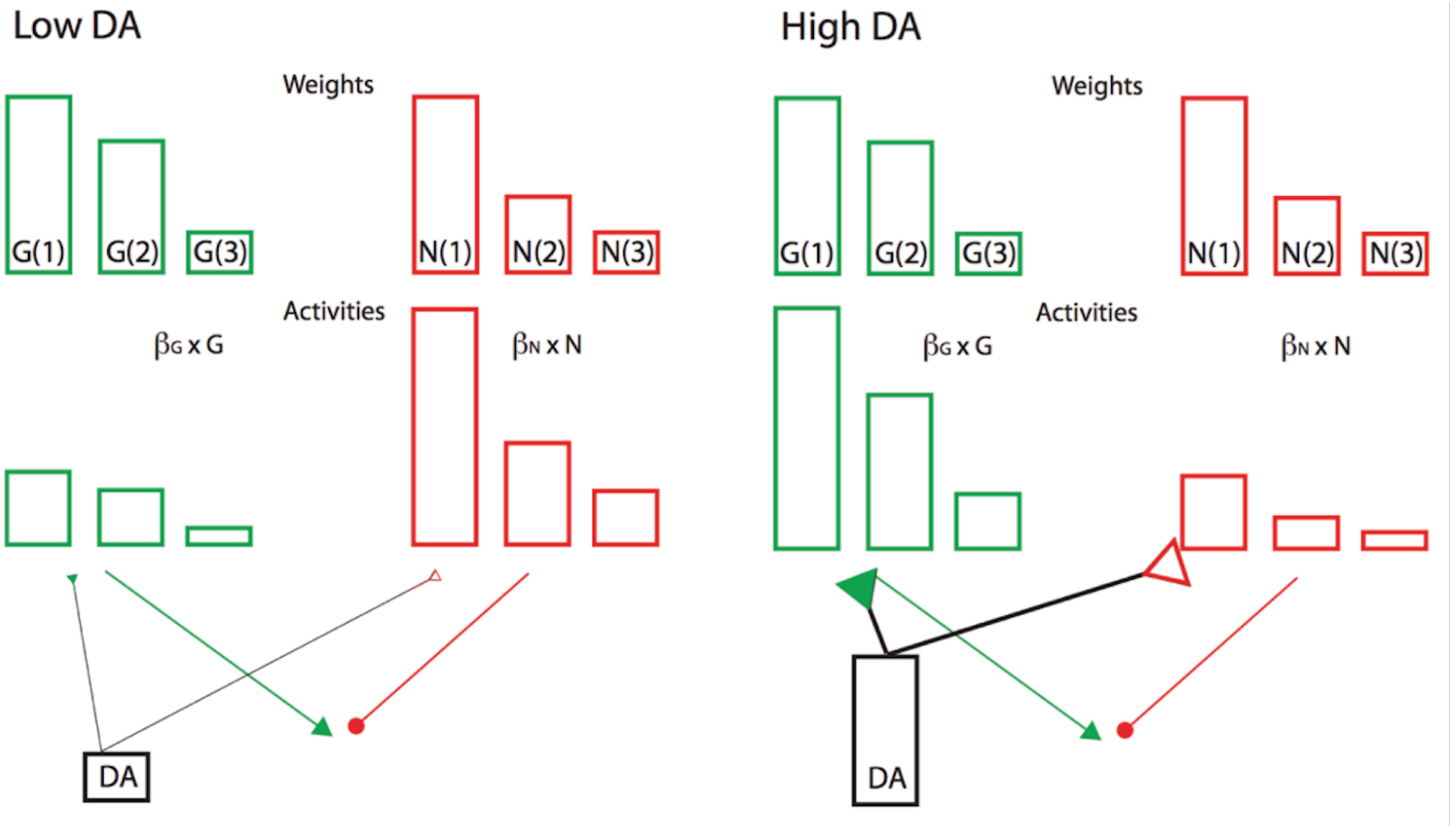
Collins & Frank 2014 *Psychological Review*

Striatal dopamine and reward-based learning / choice



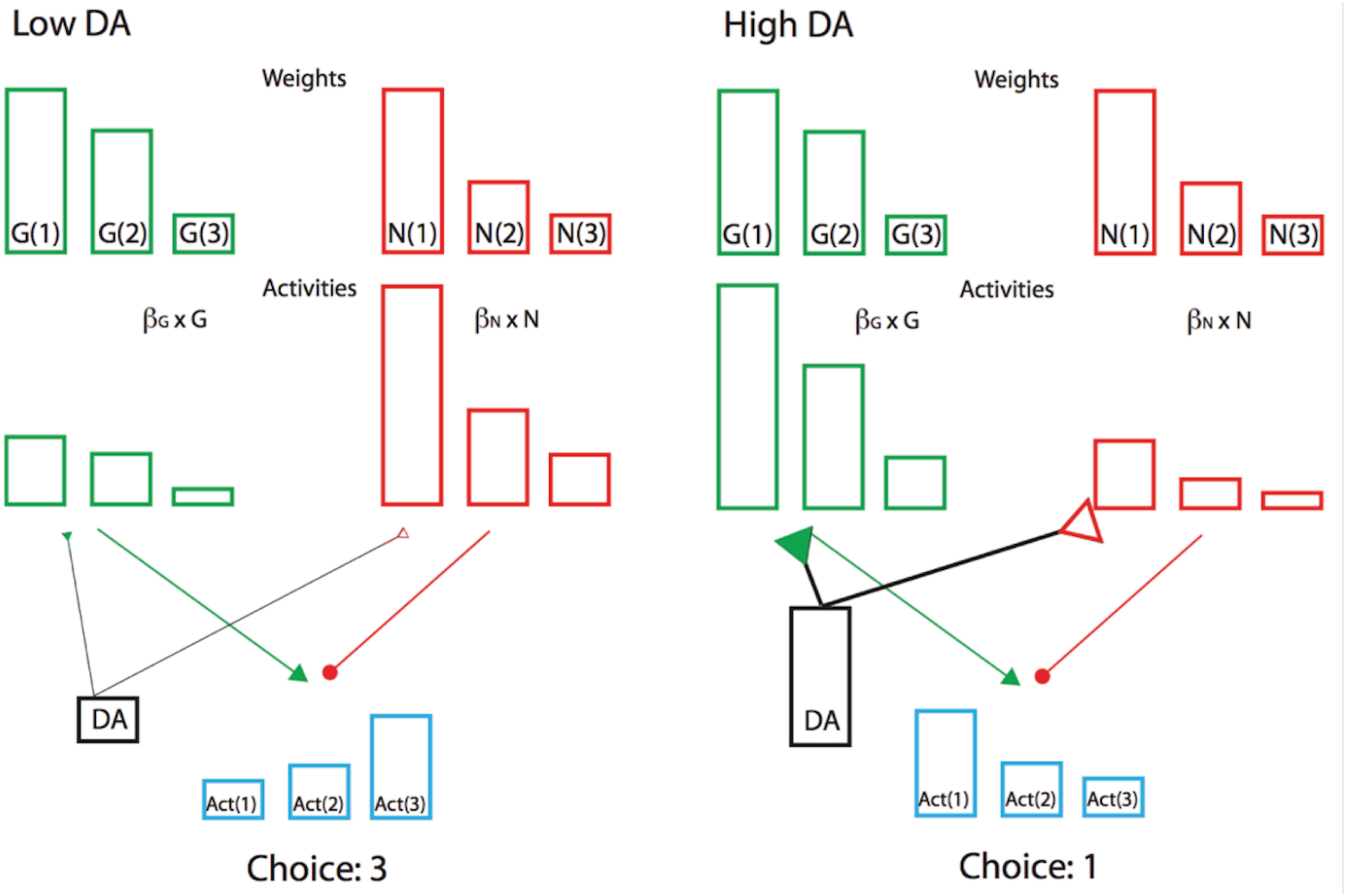
Collins & Frank 2014 *Psychological Review*

Striatal dopamine and reward-based learning / choice



Collins & Frank 2014 *Psychological Review*

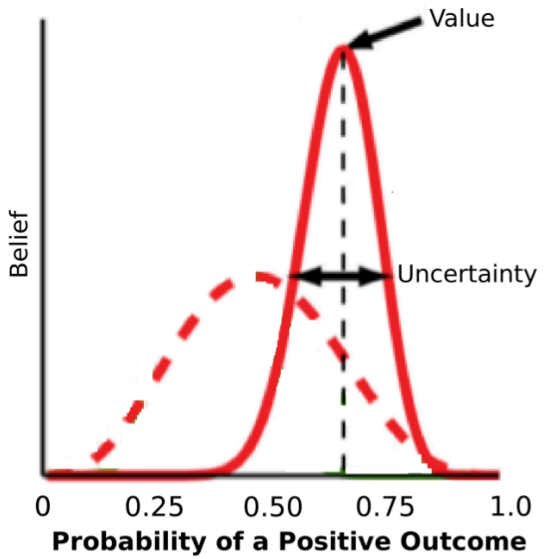
Striatal dopamine and reward-based learning / choice



Collins & Frank 2014 *Psychological Review*

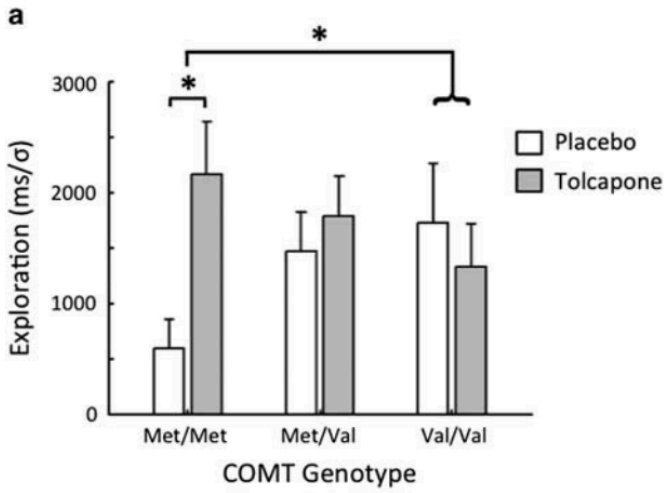
Exploration

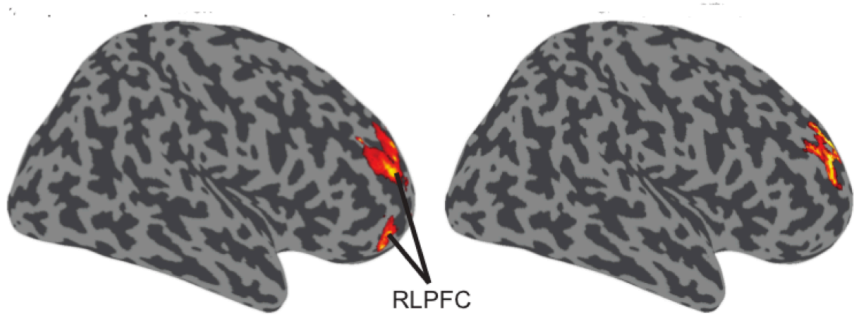
- By exploiting learned strategies, we can get a certain amount of reward
- But when to explore?
- Theory: Explore based on relative *uncertainty* about whether other actions might yield better outcomes than status quo



Exploration

- By exploiting learned strategies, we can get a certain amount of reward
- But when to explore?
- Theory: Explore based on relative *uncertainty* about whether other actions might yield better outcomes than status quo



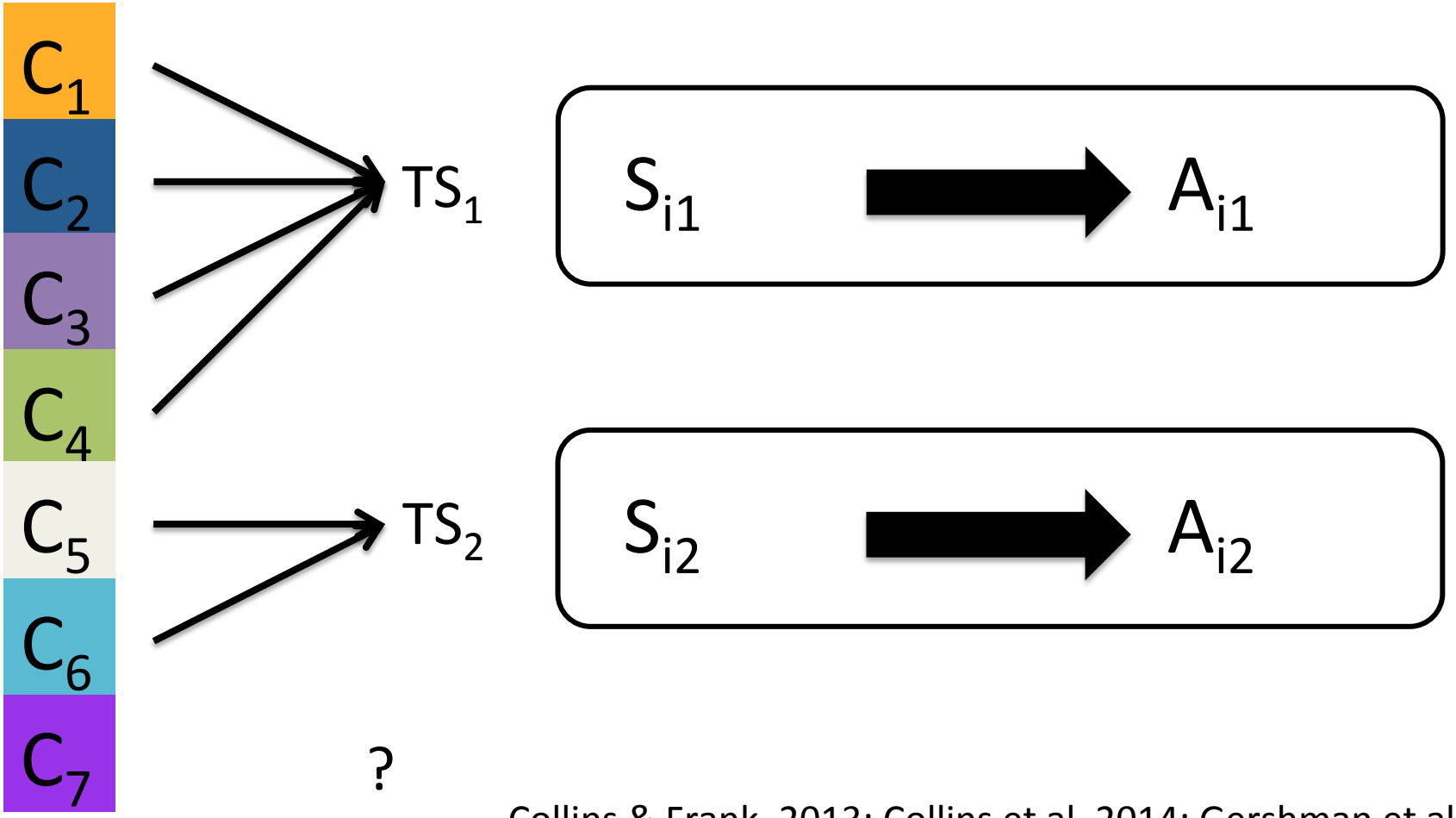


$\epsilon > 0$ *('explorers')* *explorers > non-explorers*

Frank et al., 2009; Badre et al., 2012; Cavanagh et al, 2012; Kayser et al 2015

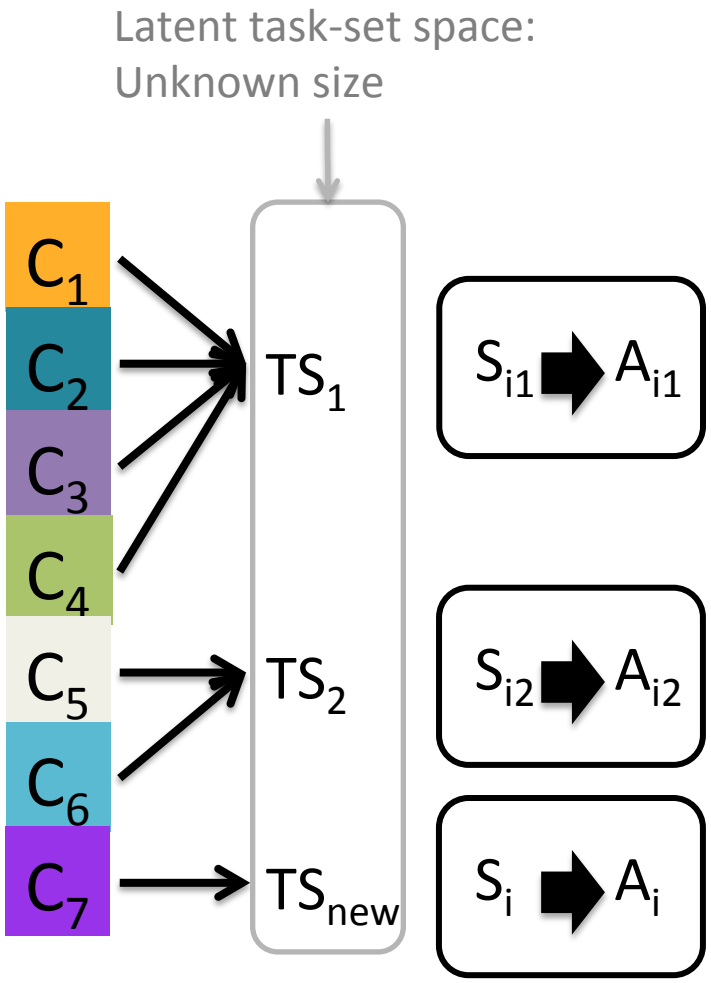
Wilson et al, 2014; Yu & Dayan, 2005 etc

Inferring latent task-set states



Collins & Frank, 2013; Collins et al, 2014; Gershman et al 2010
Hampton & O'Doherty, 2006; Wilson et al 2014

Inferring latent task-set states



- Model online learning:
  - build TS space:
    - C-TS; TS_i: S-A.
  - infer TS(t):
    - Approximate Bayesian inference

Prior on TS in a new context:

$P_0(TS = TS_j \mid C_{new}) \sim N(TS_j \mid C^*)$

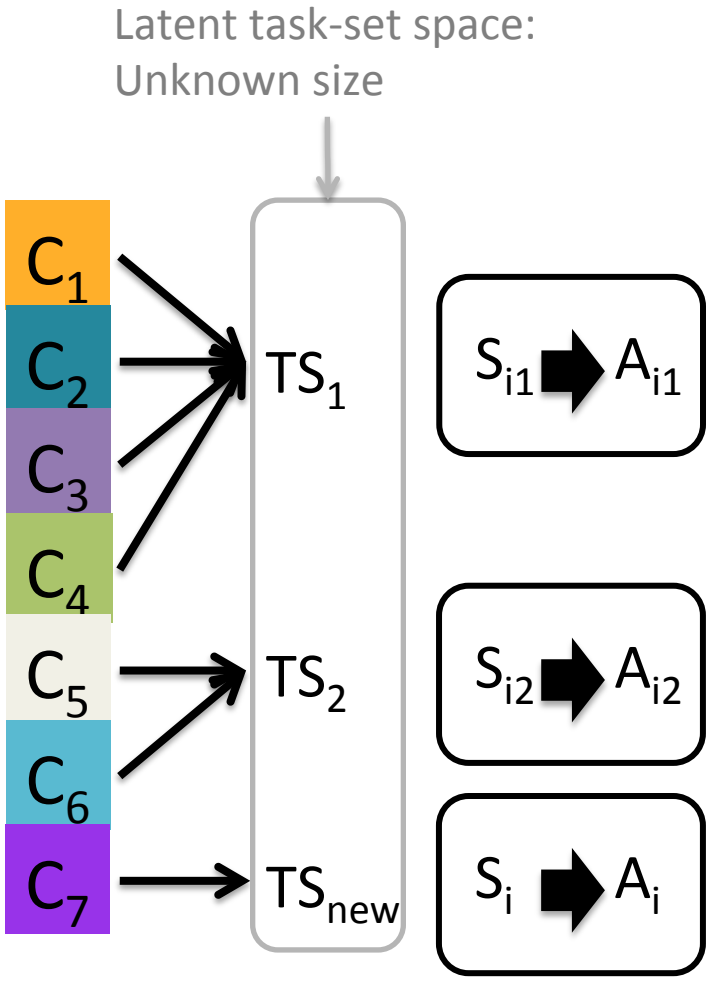
$P_0(TS = new \mid C_{new}) \sim \alpha$

$\alpha > 0$: Clustering parameter.

Collins & Frank, Psych Review, 2013; Collins et al, 2014, etc

Inferring latent task-set states

- Model online learning:
  - build TS space:
    - C-TS; TS_i: S-A.
  - infer TS(t):
    - Approximate Bayesian



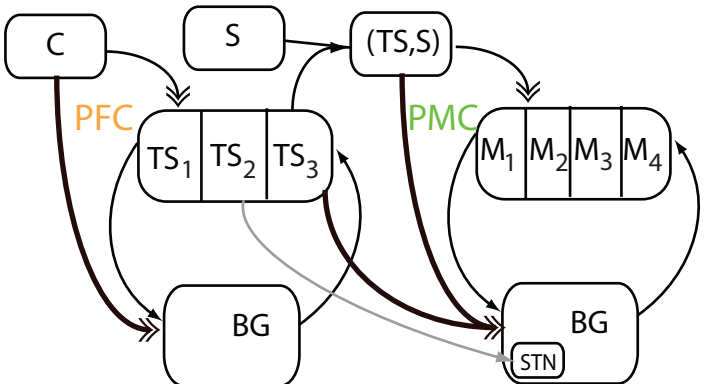
Collins & Frank, Psych Review, 2013

Prior on TS in a new context:

$P_0(TS = TS_j \mid C_{new}) \sim N(TS_j \mid C^*)$

$P_0(TS = new \mid C_{new}) \sim \alpha$

$\alpha > 0$: Clustering parameter.



What's in an outcome?

- reward & information -

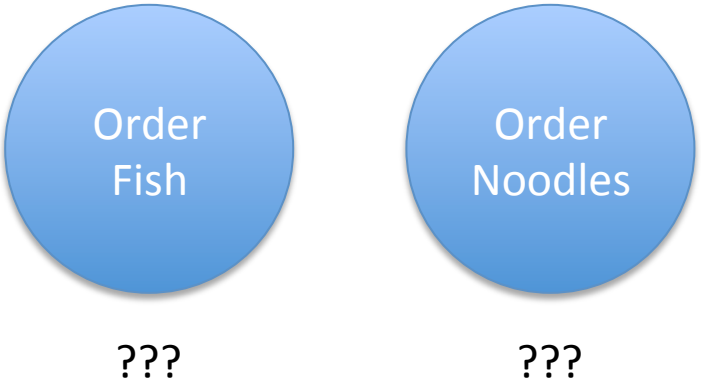
How can we select actions to **inform** us about relevant task states –
even if the actions themselves have little direct reward value**?**

Jeff Cockburn

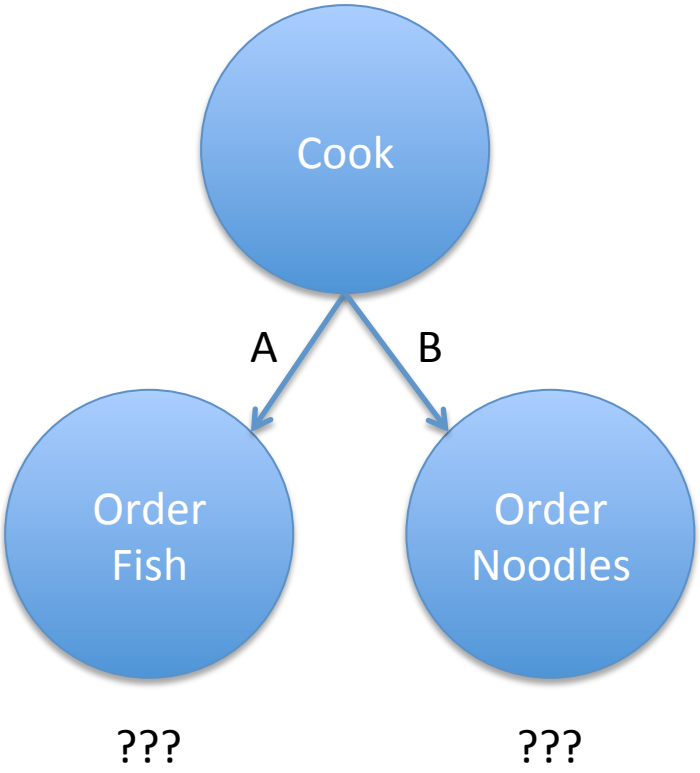
Dayan & Daw, 2008; Friston et al 2015

The sushi restaurant process

The sushi restaurant process



The sushi restaurant process



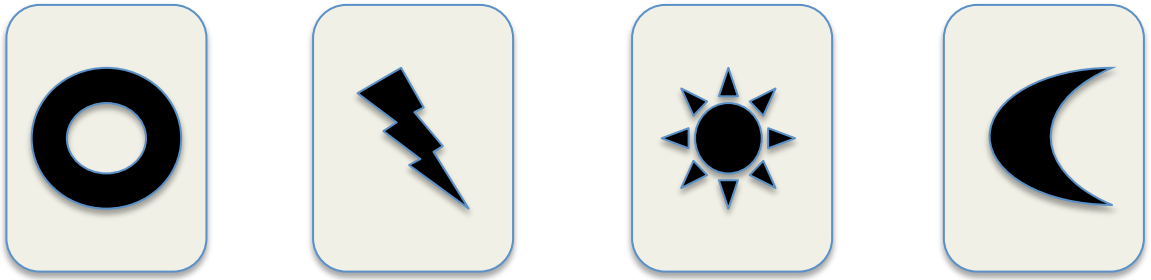
Experiment 1

Do people seek information?

When and what for?

Experiment 1

- Design -

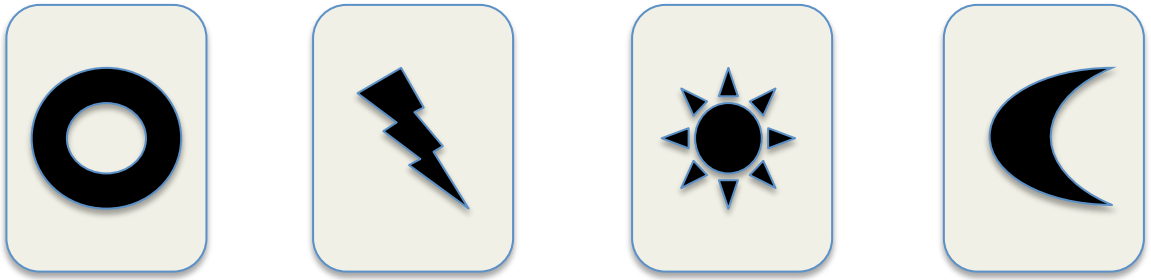


Experiment 1

- Design -

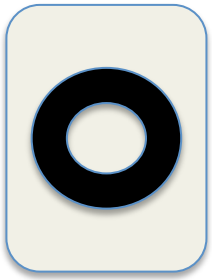
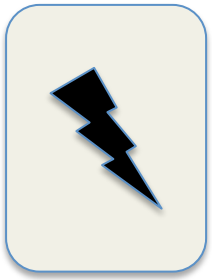
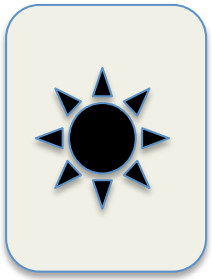
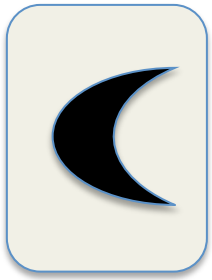
Deck A

Deck B



Experiment 1

- Design -

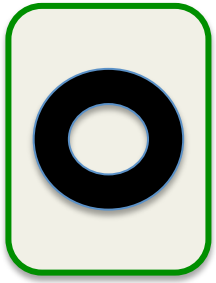
Deck A	15% 50 pnt	5% 50 pnt	100% 0.25 pnt	100% 1 pnt
				
Deck B	5% 50 pnt	15% 50 pnt	100% 0.75 pnt	100% 1 pnt

Experiment 1

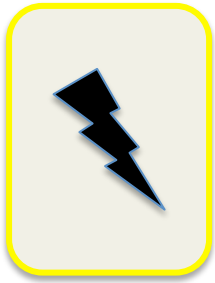
- Design: reward structure -

Deck A

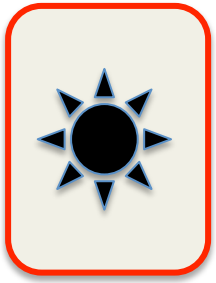
15%
50 pnt
EV=7.5



5%
50 pnt
EV=5.0



100%
0.25 pnt
EV=0.25



100%
1 pnt
EV=1

Deck B

5%
50 pnt
EV=5.0

15%
50 pnt
EV=7.5

100%
0.75 pnt
EV=0.75

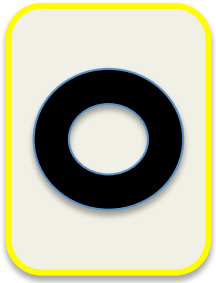
100%
1 pnt
EV=1

Experiment 1

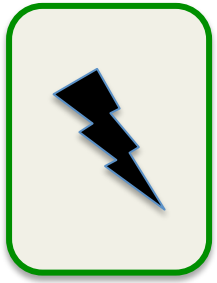
- Design: reward structure -

Deck A

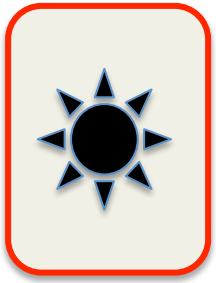
15%
50 pnt
EV=7.5



5%
50 pnt
EV=5.0



100%
0.25 pnt
EV=0.25



100%
1 pnt
EV=1

Deck B

5%
50 pnt
EV=5.0

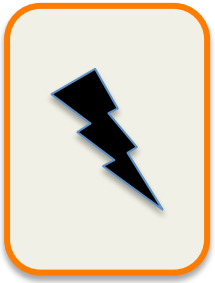
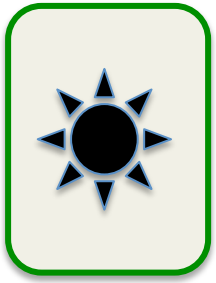
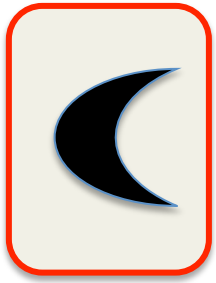
15%
50 pnt
EV=7.5

100%
0.75 pnt
EV=0.75

100%
1 pnt
EV=1

Experiment 1

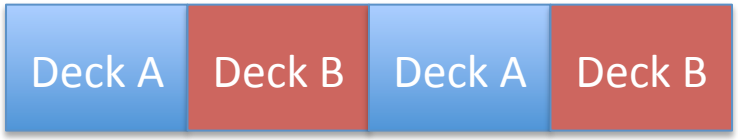
- Design: information structure -

Deck A	15% 50 pnt EV=7.5	5% 50 pnt EV=5.0	100% 0.25 pnt EV=0.25	100% 1 pnt EV=1
				
Deck B	5% 50 pnt EV=5.0 **I(O;D)<0.1**	15% 50 pnt EV=7.5 **I(O;D)<0.1**	100% 0.75 pnt EV=0.75 **I(O;D)=1**	100% 1 pnt EV=1 **I(O;D)=0**

Experiment 1

- Design: Protocol -

Training: Blocks of 10 trials – deck known

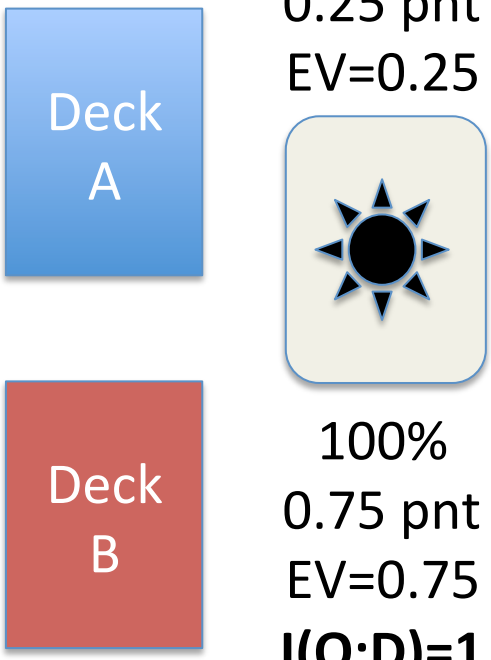


Testing: 5% deck switch – deck unknown

Experiment 1

- Design: Conditions -

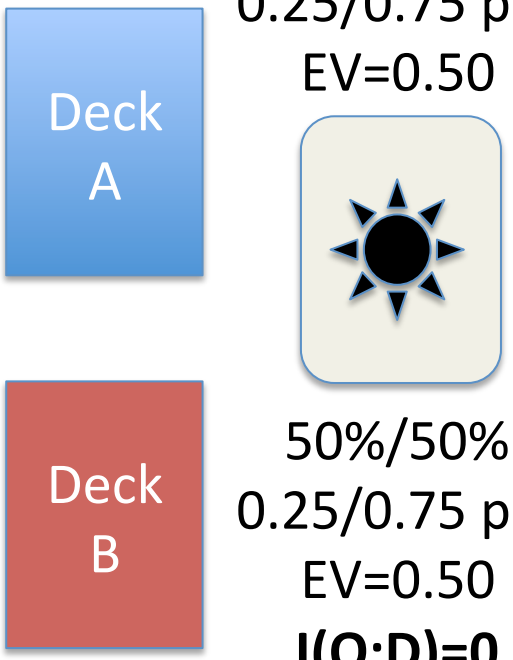
Informative Condition



100%
0.25 pnt
EV=0.25

100%
0.75 pnt
EV=0.75
I(O;D)=1

Uninformative Condition



50%/50%
0.25/0.75 pnt
EV=0.50

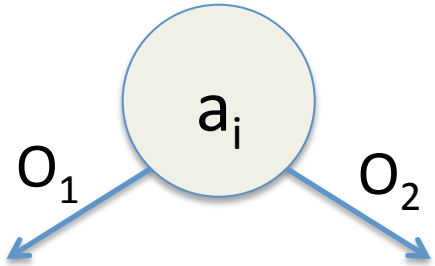
50%/50%
0.25/0.75 pnt
EV=0.50
I(O;D)=0

What you should do when you don't know what to do

A touch of theoretical work

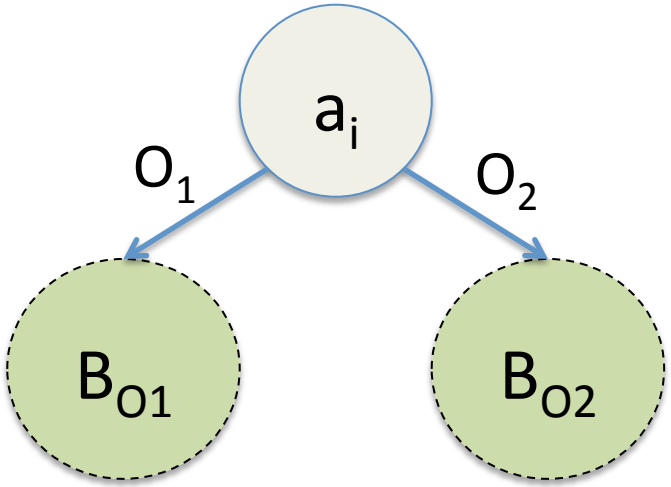
A Bayesian forward model

- *N*-step look-ahead -



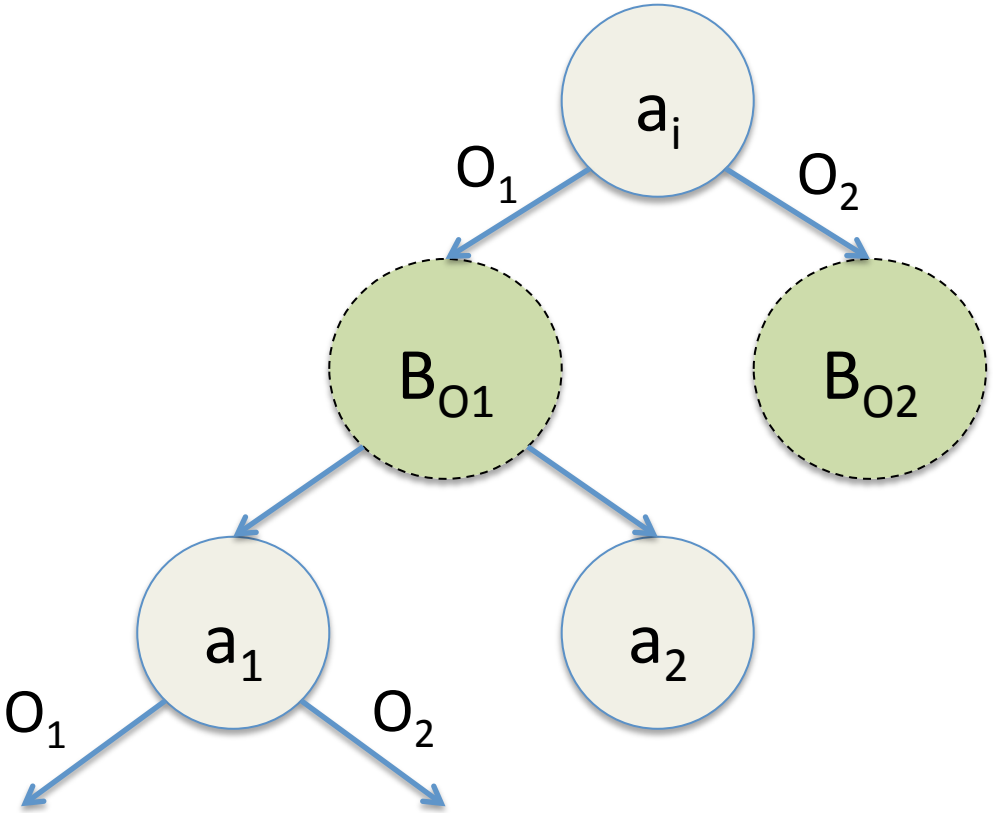
A Bayesian forward model

- *N*-step look-ahead -



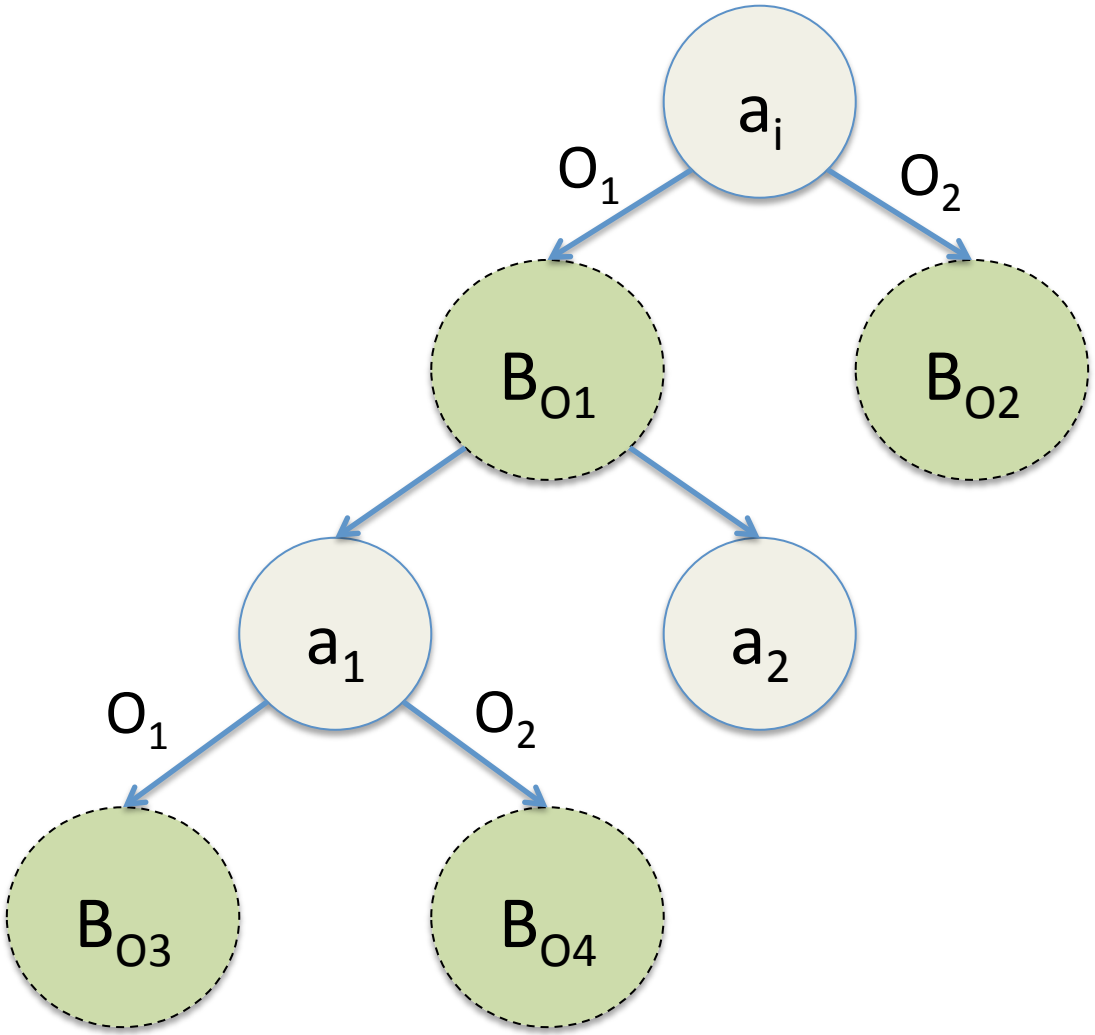
A Bayesian forward model

- *N*-step look-ahead -



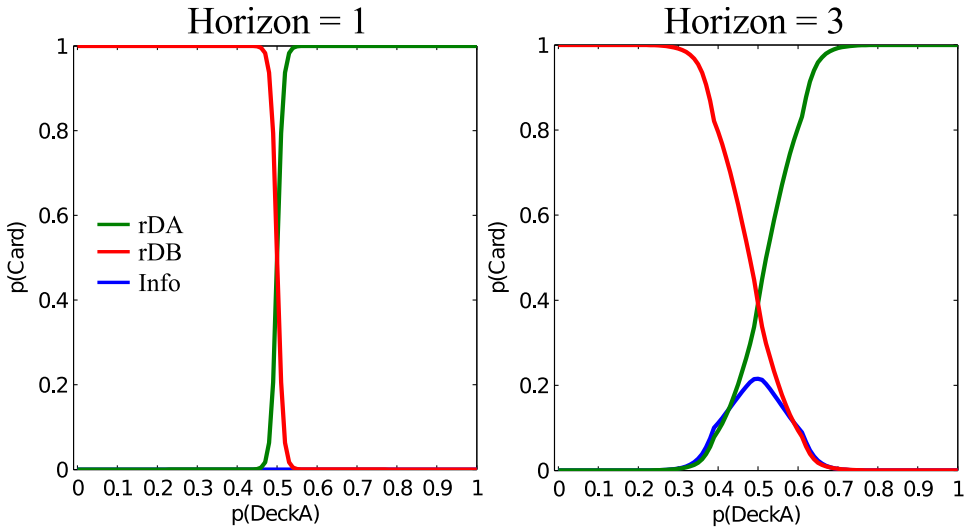
A Bayesian forward model

- *N*-step look-ahead -



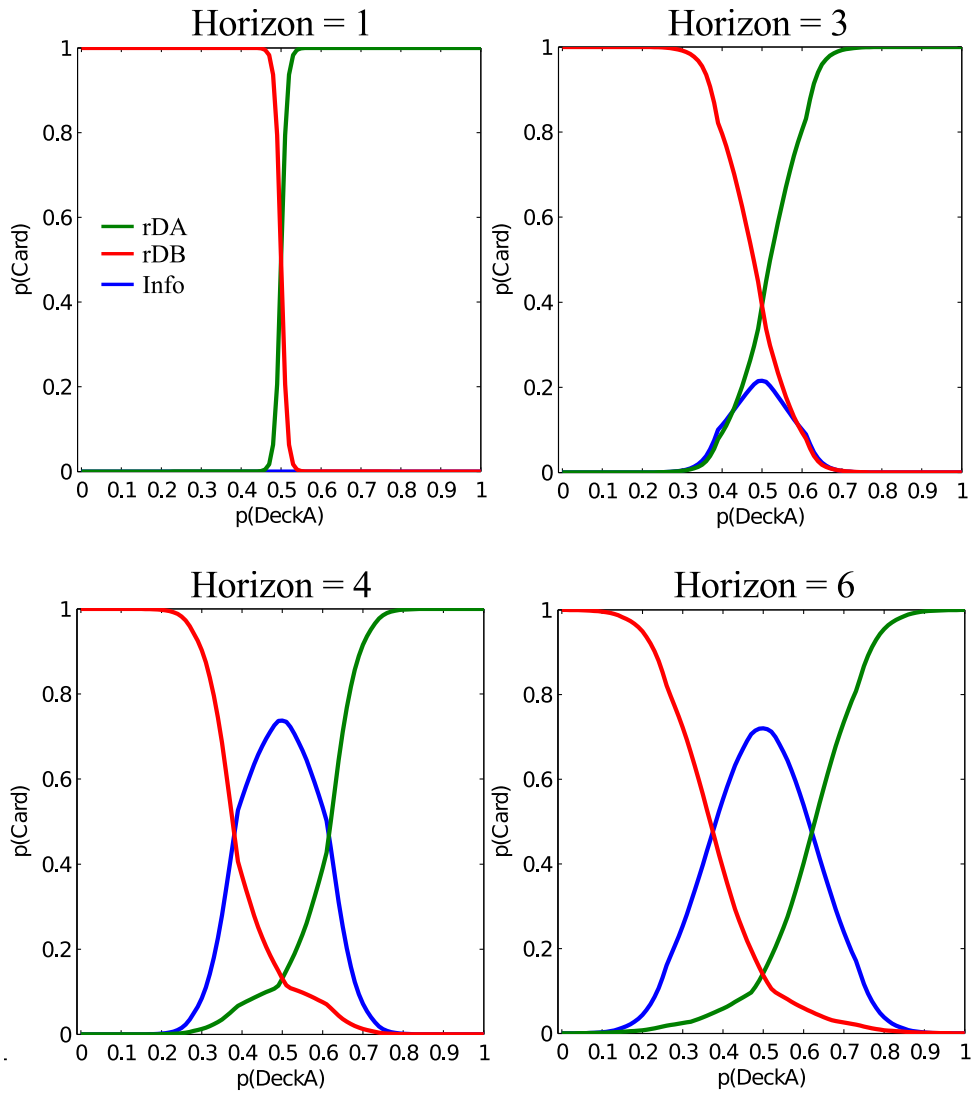
A Bayesian forward model

- Sensitivity to depth of look-ahead -



A Bayesian forward model

- Sensitivity to depth of look-ahead -



A heuristic alternative

- expected reward & information gain -

$$SV(a_i) = EV[a_i] + H(D_B) * I(O_i; D)$$

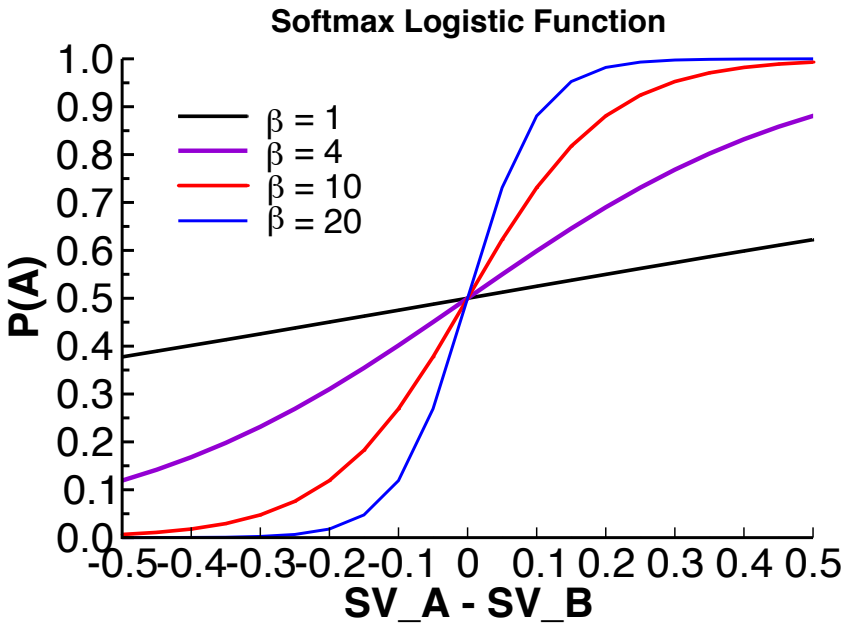
- Keep track of probability each deck is in play based on outcomes observed.
- Probe when:
  - most uncertain about latent Deck AND:
  - probing action is more informative

A heuristic alternative

- expected reward & information gain -

$$SV(a_i) = EV[a_i] + H(D_B) * I(O_i; D)$$

- Keep track of probability each deck is in play based on outcomes observed.
- Probe when:
  - most uncertain about latent Deck AND:
  - probing action is more informative



A heuristic alternative

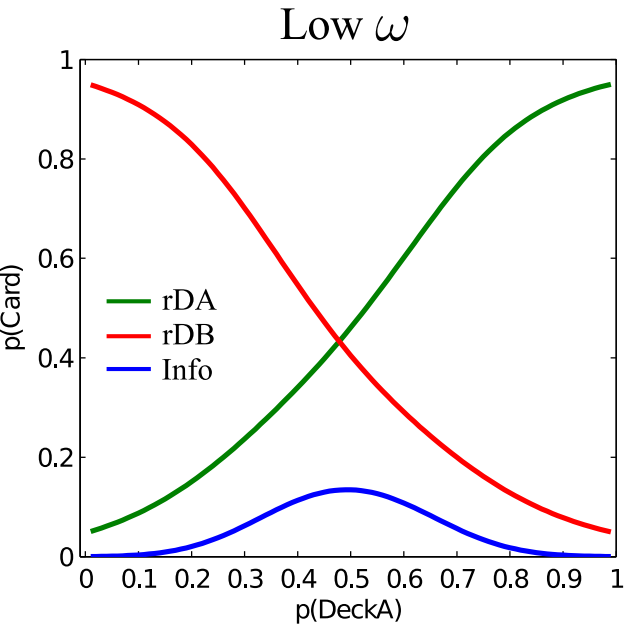
- Coping with individual differences -

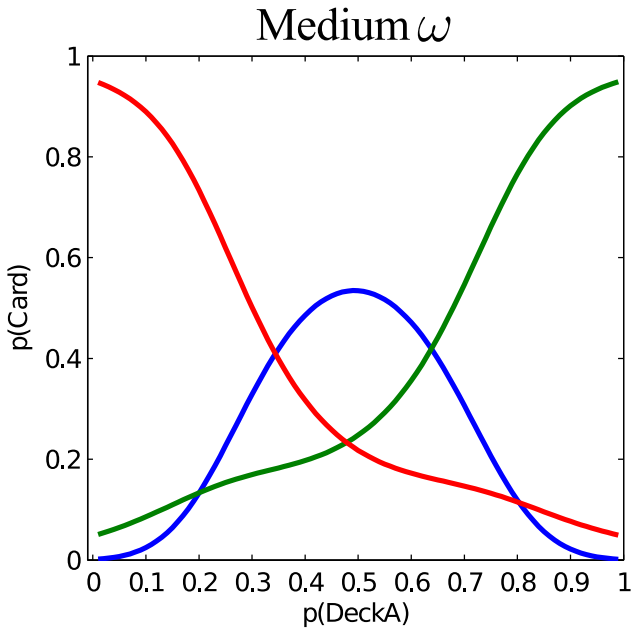
$$SV(a_i) = EV[a_i] + \boldsymbol{\omega} * H(D_B) * I(O_i; D)$$

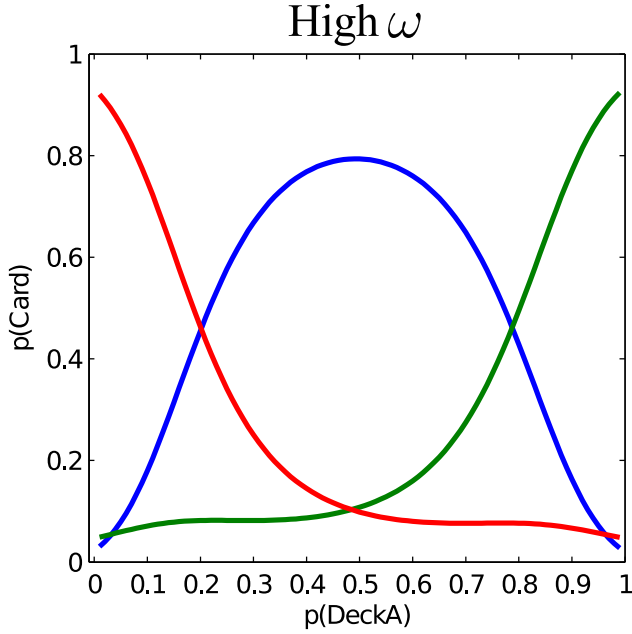
A heuristic alternative

- variable weighting factor-

$$SV(a_i) = EV[a_i] + \boldsymbol{\omega} * H(D_B) * I(O_i; D)$$

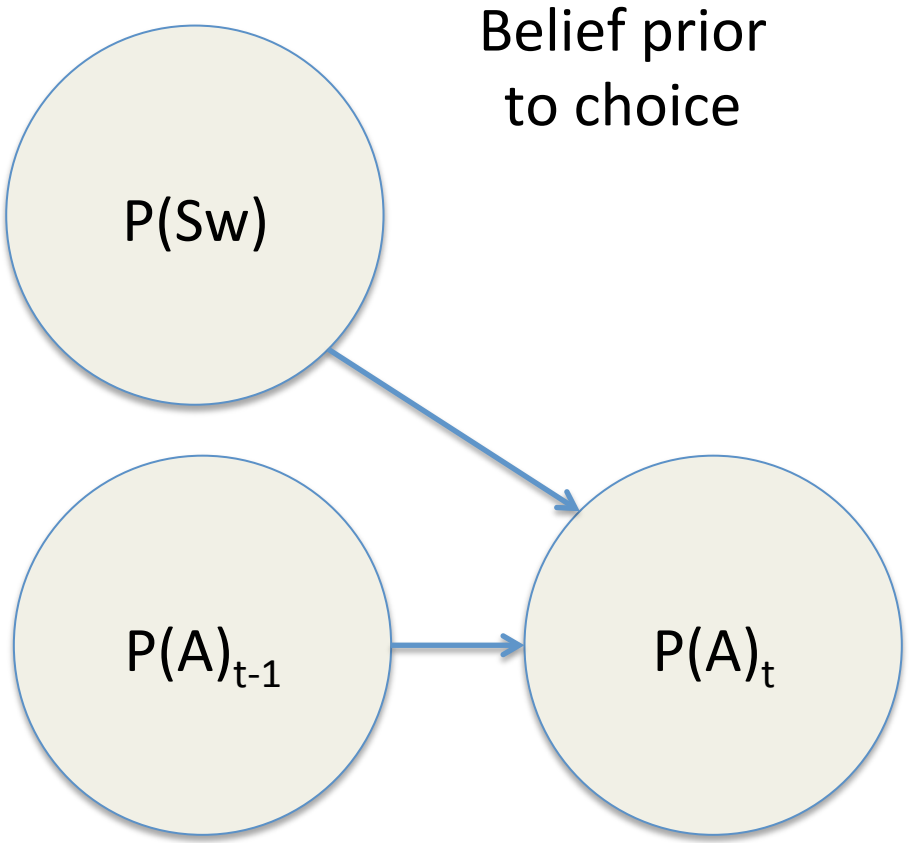






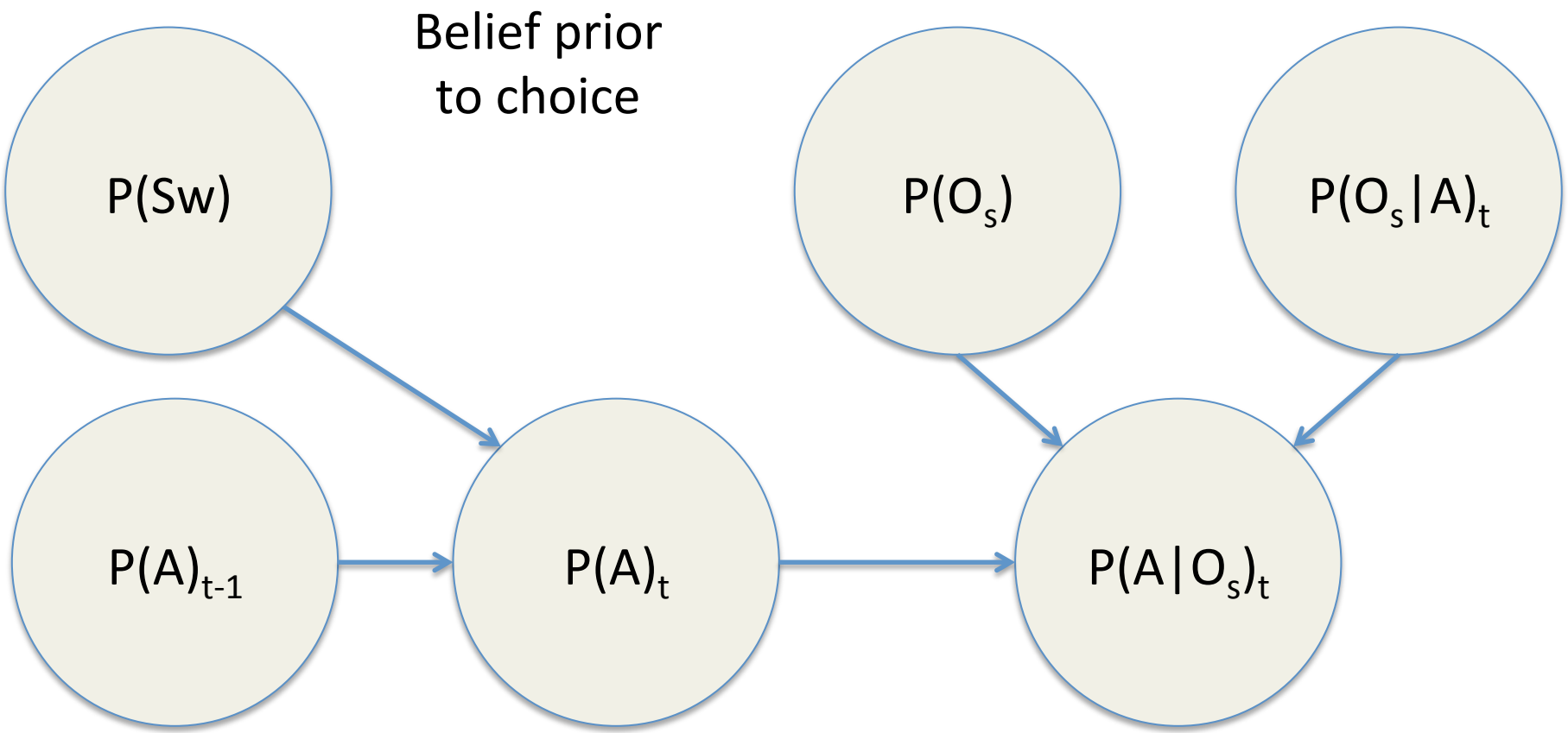
The optimal observer

- Inferred belief that Deck A is in play -



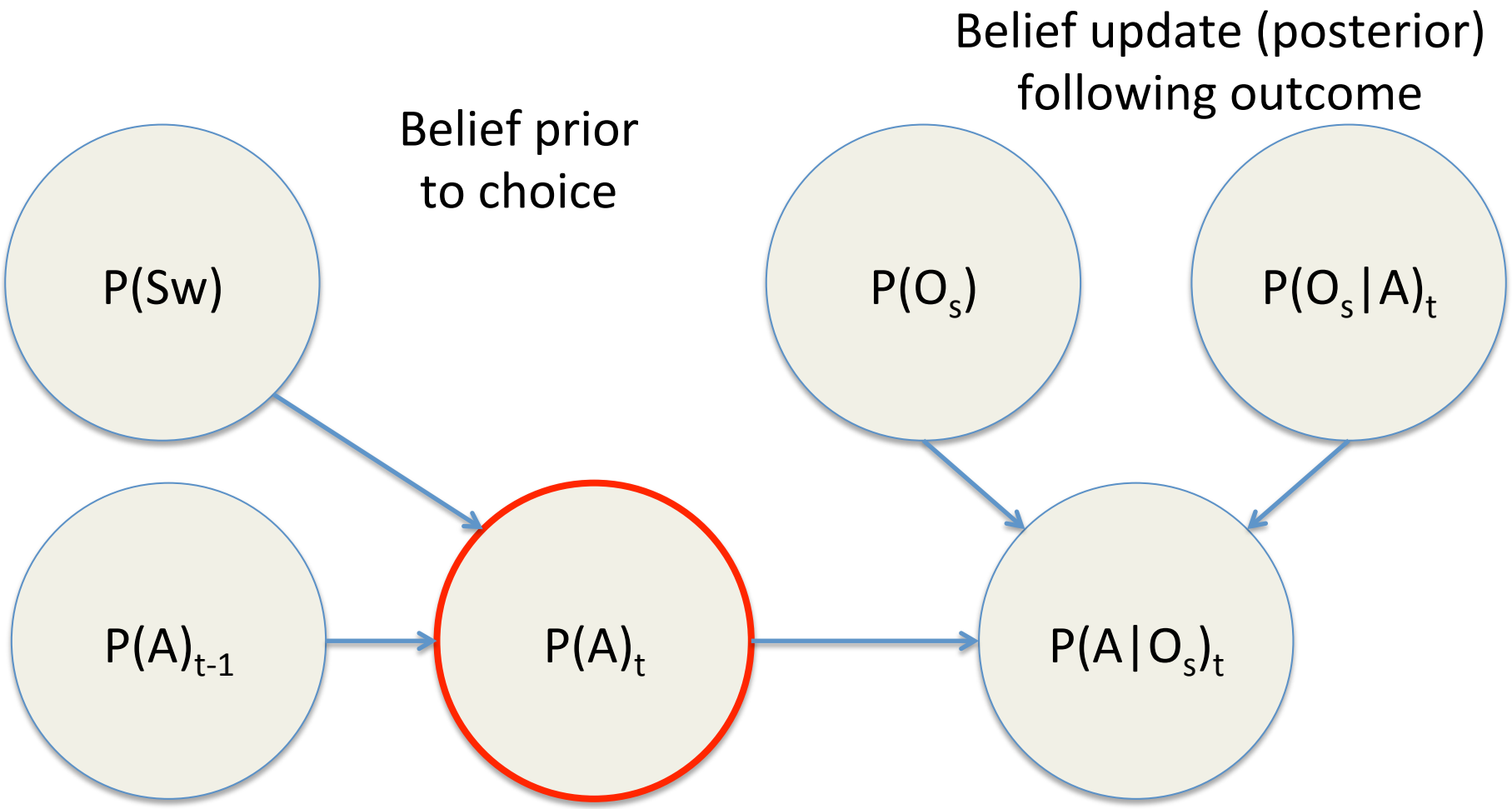
The optimal observer

- Inferred belief that Deck A is in play -



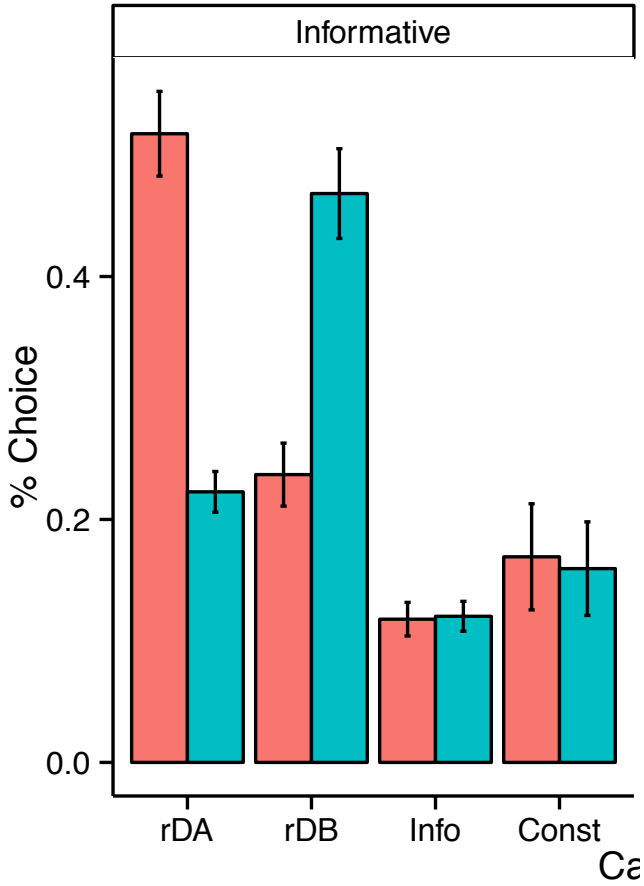
The optimal observer

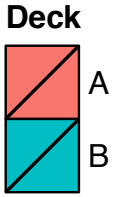
- Inferred belief that Deck A is in play -



Empirical Data: Response patterns

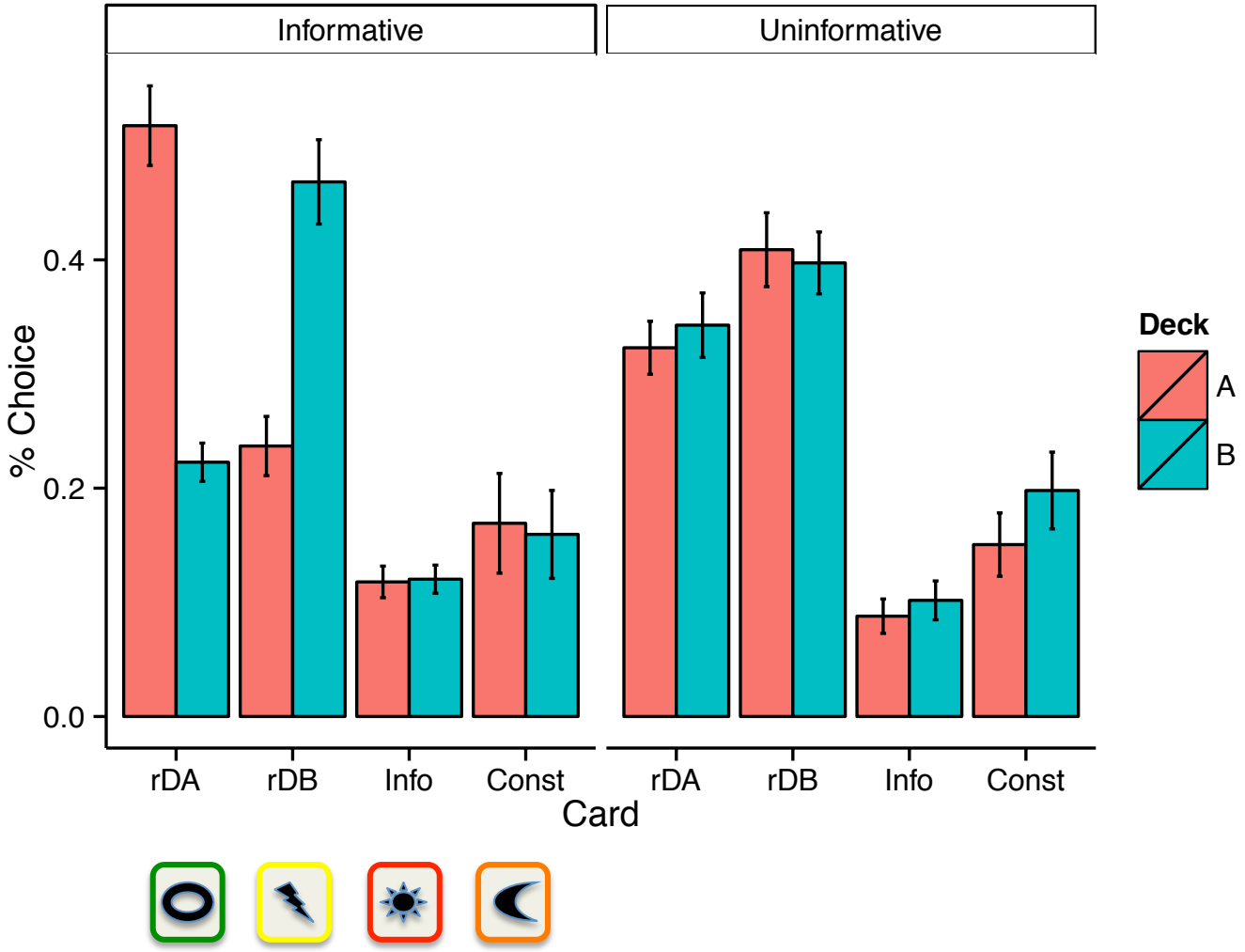
- Deck in play is unknown -





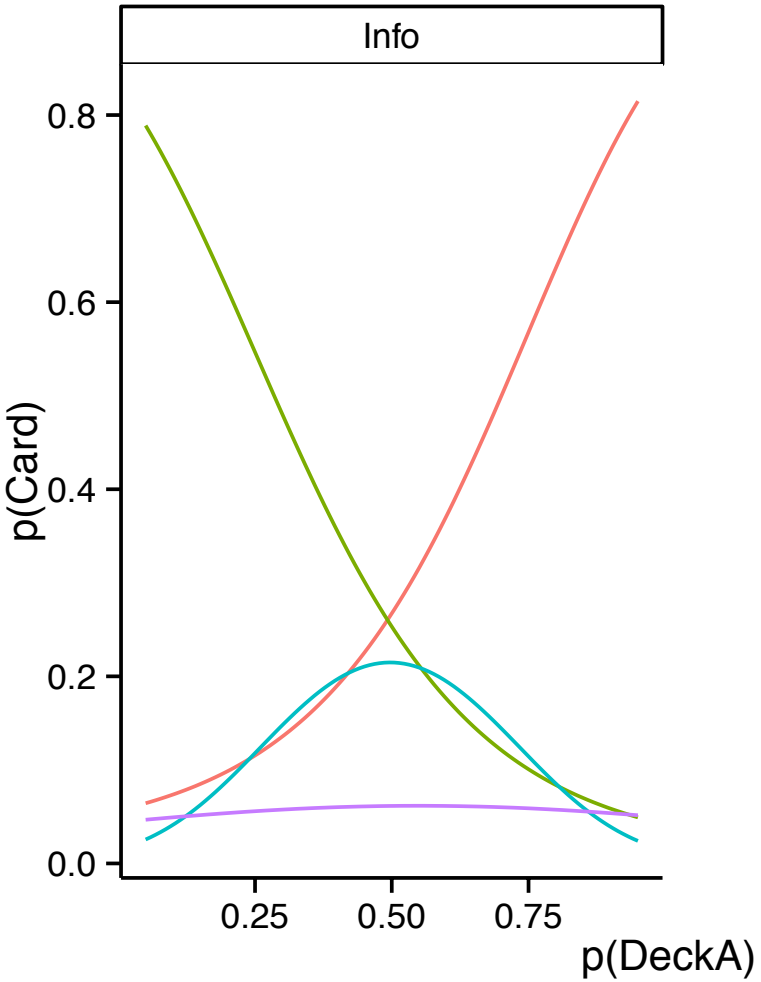
Empirical Data: Response patterns

- Deck in play is unknown -



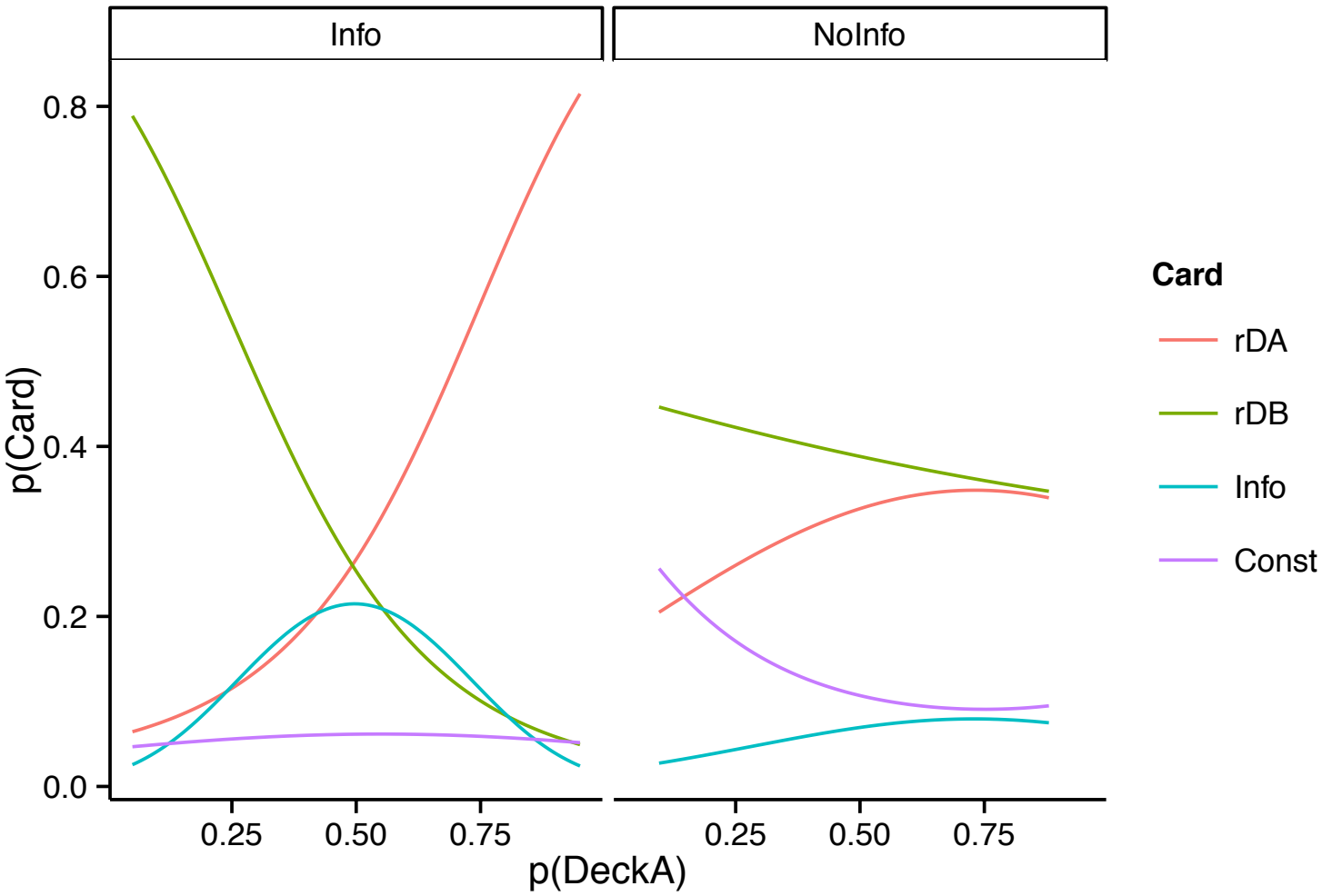
Logistic Regression Fits

- Effect of deck belief -



Logistic Regression Fits

- Effect of deck belief -



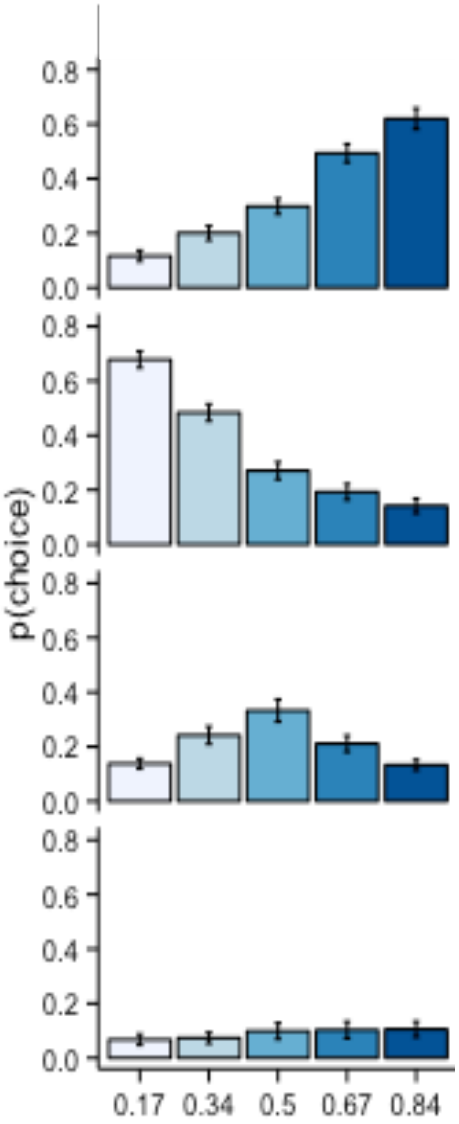
Raw data – informative condition

rDA

rDB

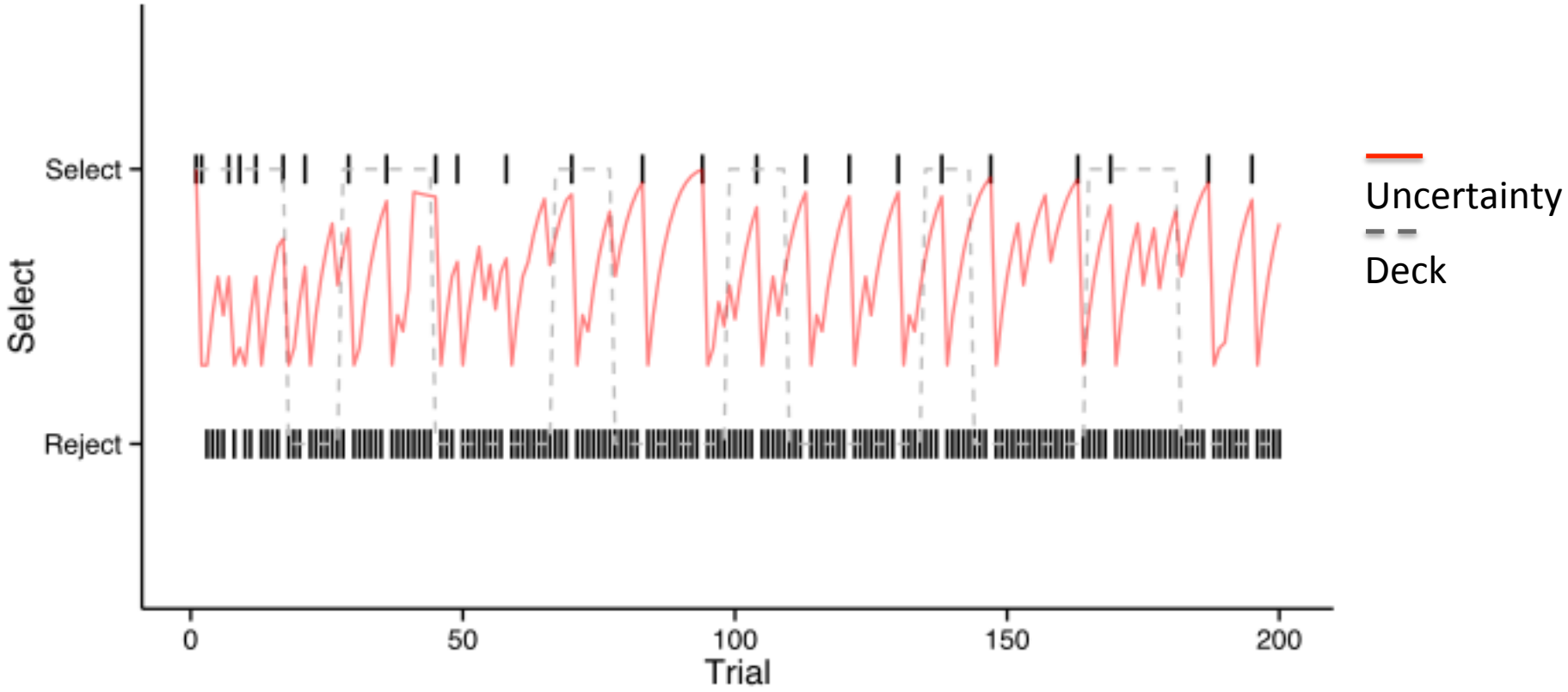
Info

Const



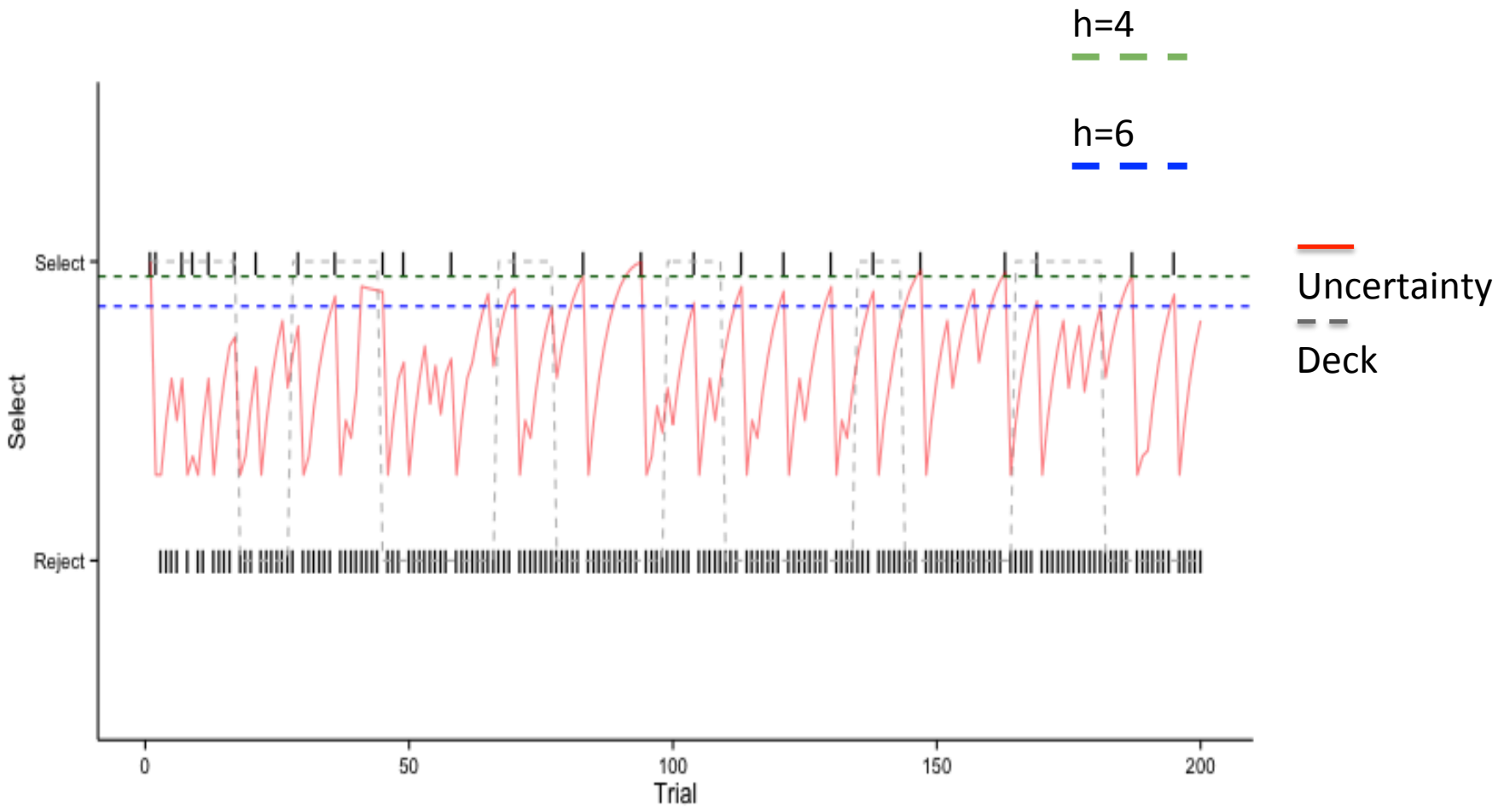
Trial-by-trial response patterns

- Select vs Reject Info card -



Trial-by-trial response patterns

- Select vs Reject Info card -



Experiment 2

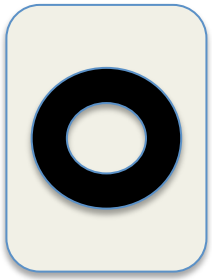
Sensitivity to the cost of information (opportunity costs)

Experiment 2

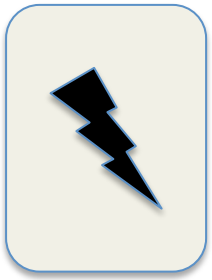
- Design: Structure -

Deck A

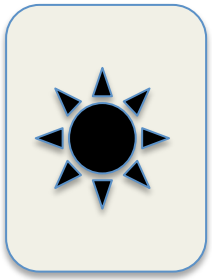
40%
100 pnt
EV=40



10%
100 pnt
EV=10



100%
1 pnt
EV=1



Deck B

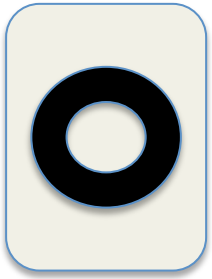
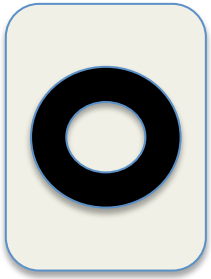
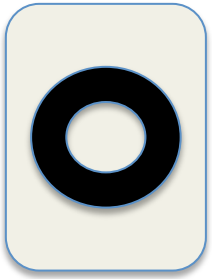
10%
100 pnt
EV=10

40%
100 pnt
EV=40

100%
2 pnt
EV=2

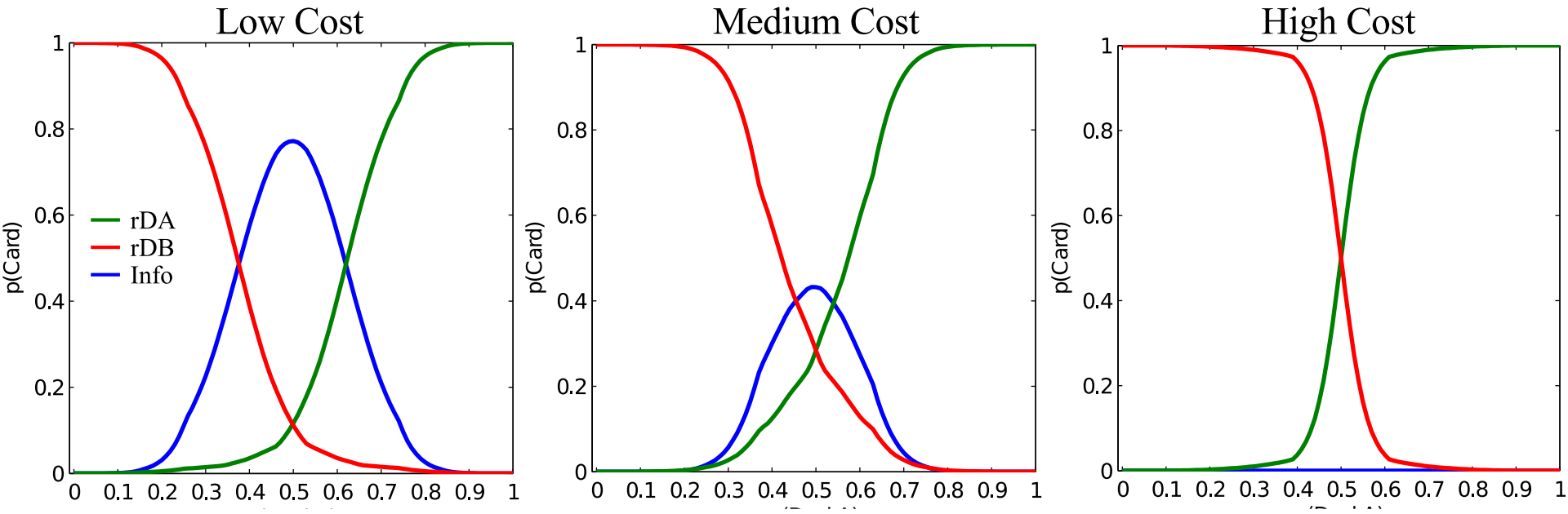
Experiment 2

- Cost Conditions -

Low cost	Med cost	High cost
40% 100 pnt **EV=40**	70% 100 pnt **EV=70**	90% 100 pnt **EV=90**
		
10% 100 pnt **EV=10**	40% 100 pnt **EV=40**	60% 100 pnt **EV=60**

Sensitivity to the cost of information

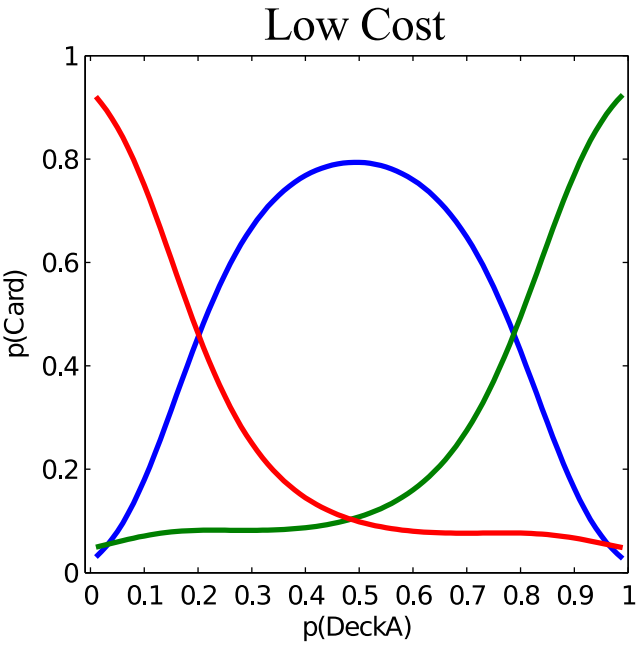
- **optimally** tracking missed reward opportunity -

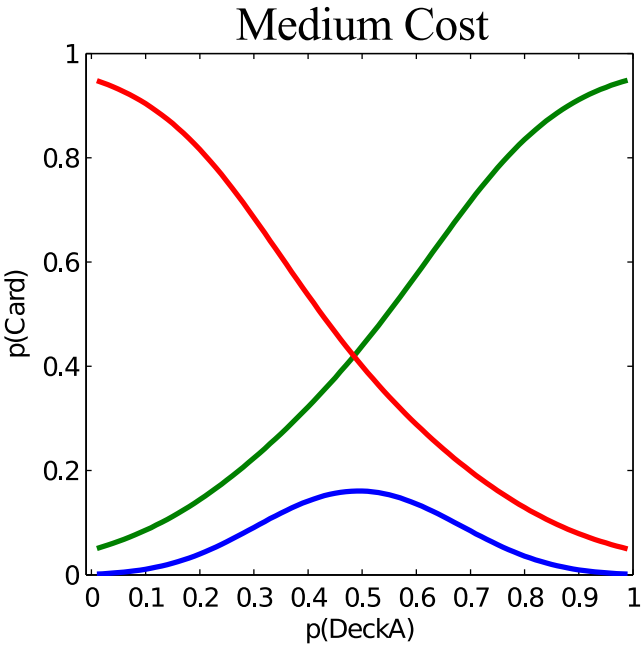


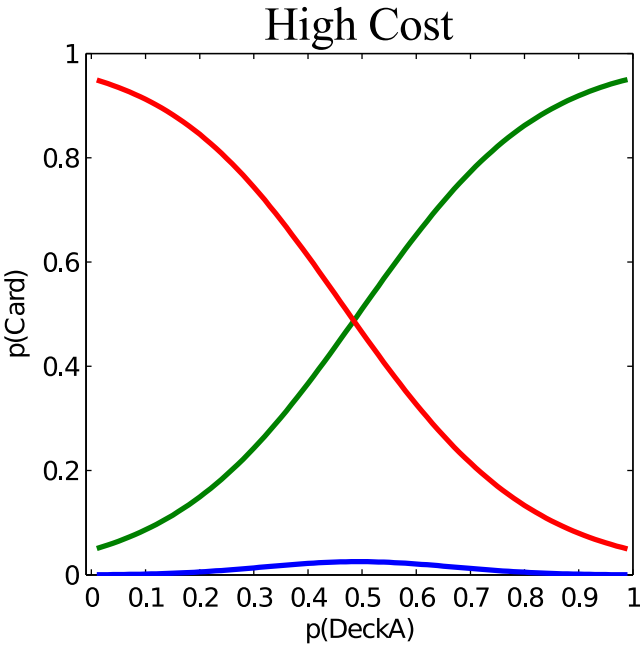
Sensitivity to the cost of information

- **heuristically** tracking missed reward opportunity -

$$SV(a_i) = EV[a_i] + \boldsymbol{\omega} * H(D) * I(O_i;D)$$

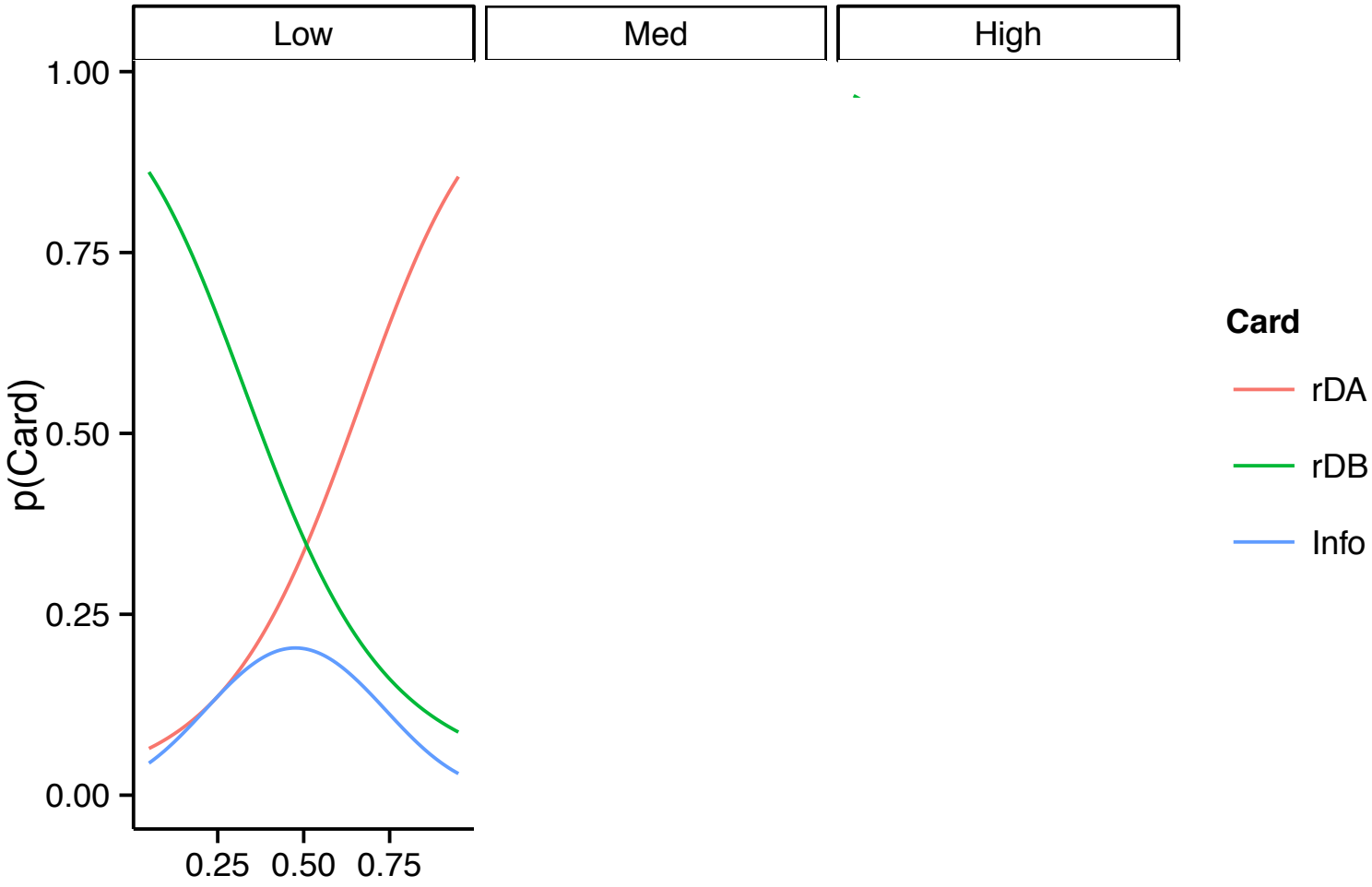






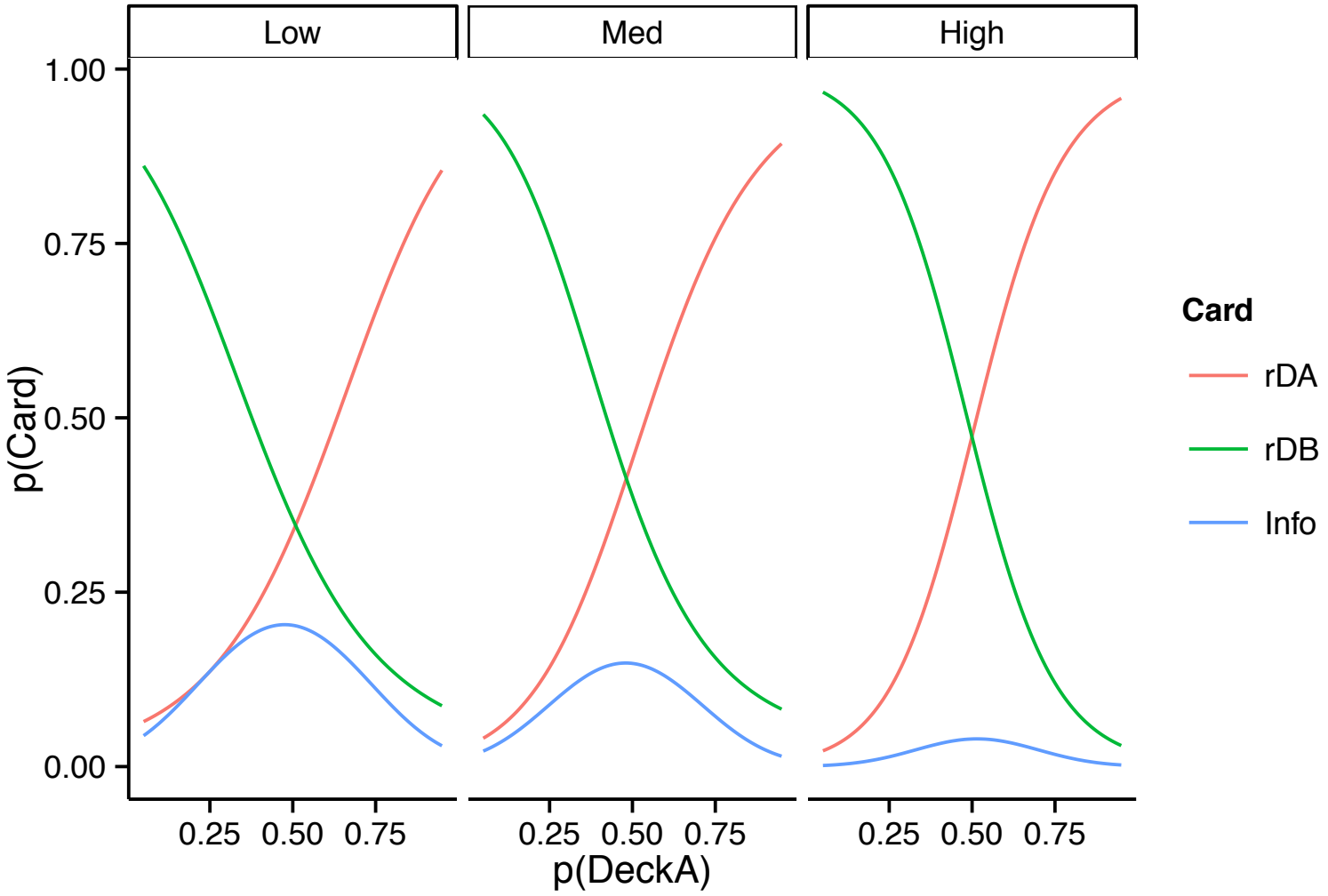
Data: Response patterns

- effect of belief & cost on choice -



Data: Response patterns

- effect of belief & cost on choice -



Experiment 3
Non-pecuniary value of information

Is information valued when it shouldn't influence choice?

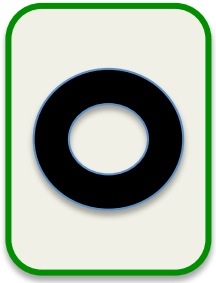
cf. Bromberg-Martin et al., 2009

Experiment 3

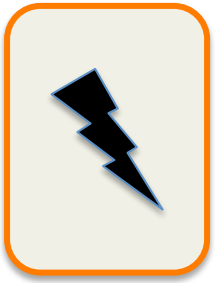
- Design: Reward structure -

Deck A

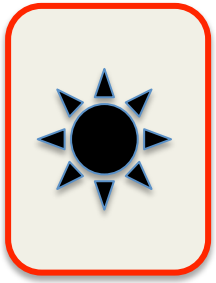
30%
100 pnt
EV=30



10%
100 pnt
EV=10



100%
1 pnt
EV=1



Deck B

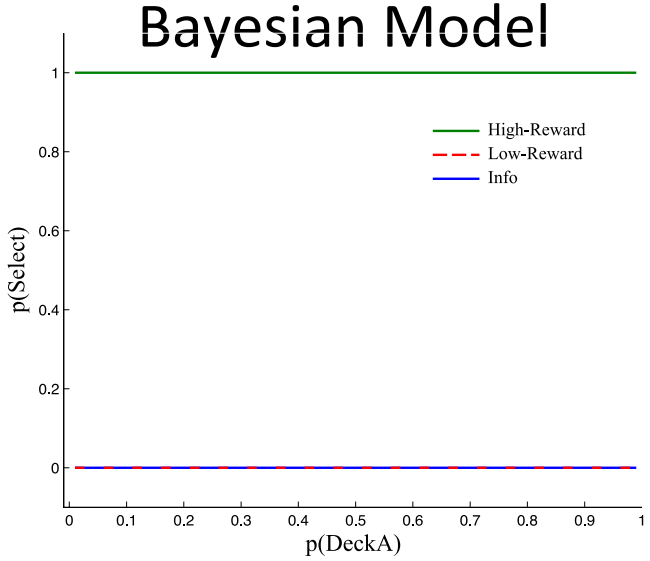
60%
100 pnt
EV=60

30%
100 pnt
EV=30

100%
2 pnt
EV=2

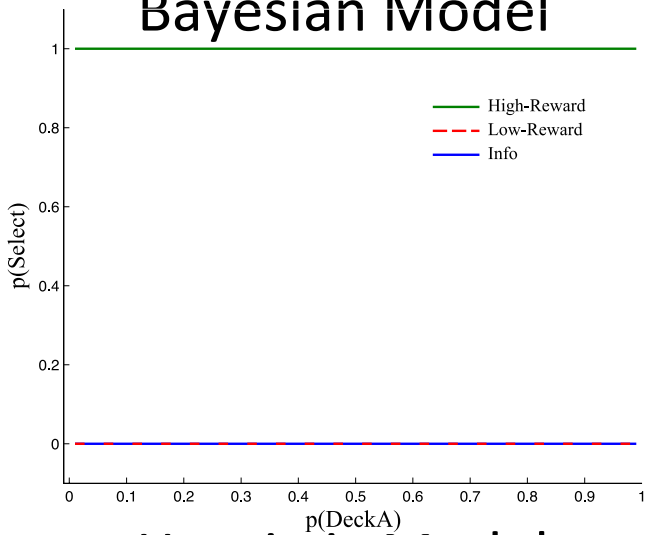
‘Non-instrumental’ value of information

- Model and human response patterns -

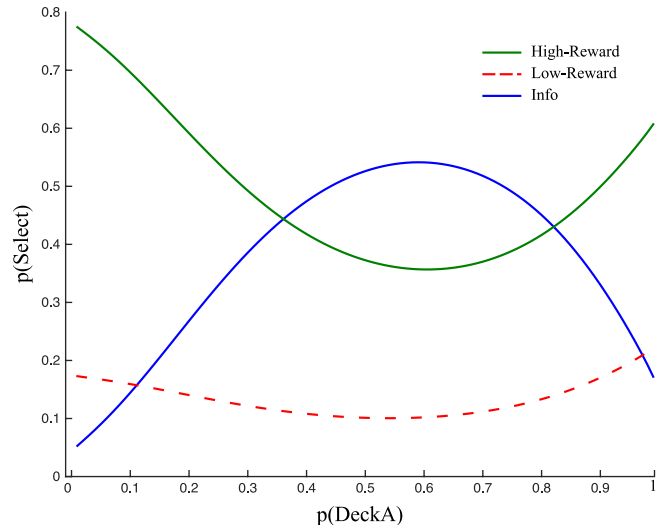


‘Non-instrumental’ value of information
- Model and human response patterns -

Bayesian Model



Heuristic Model

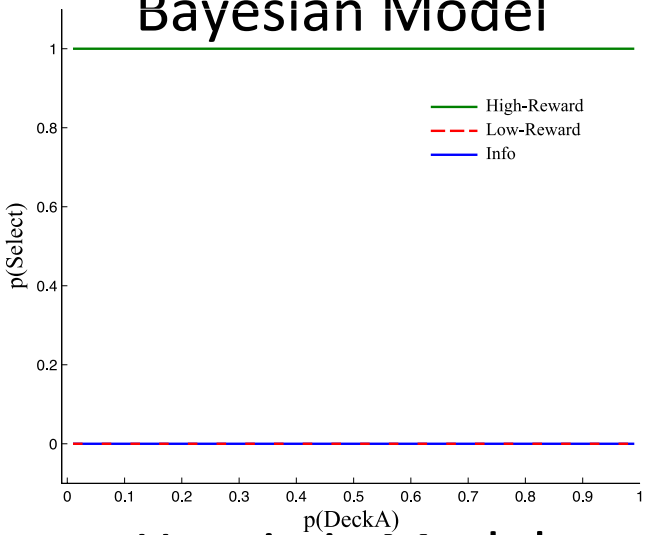


$$SV(a_i) = EV[a_i] + \boldsymbol{\omega} * H(D) * I(O_i;D)$$

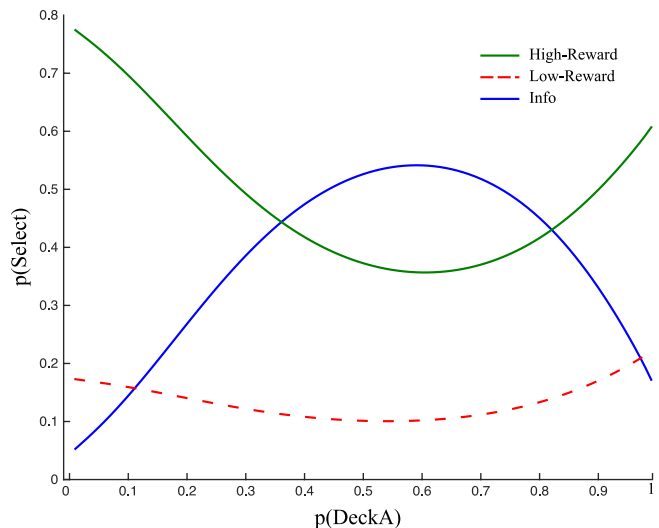
‘Non-instrumental’ value of information

- Model and human response patterns -

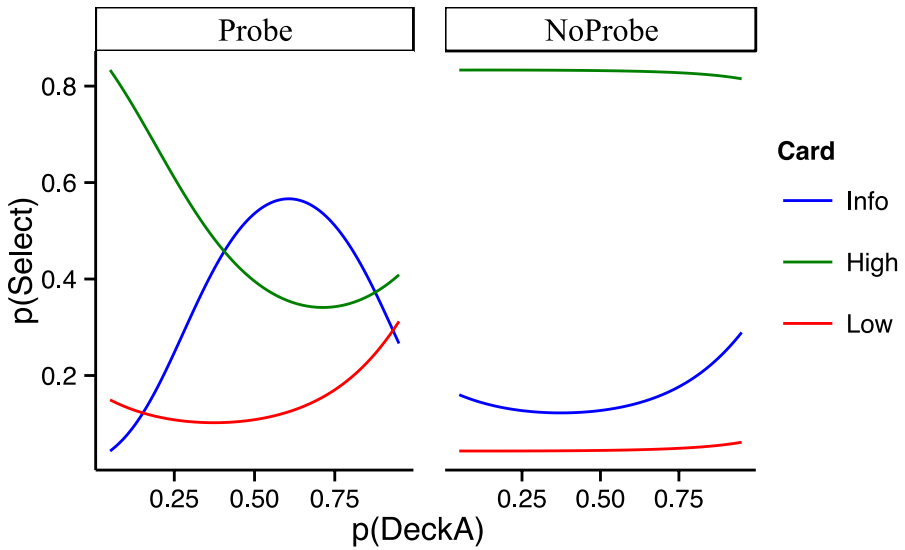
Bayesian Model



Heuristic Model



Participants



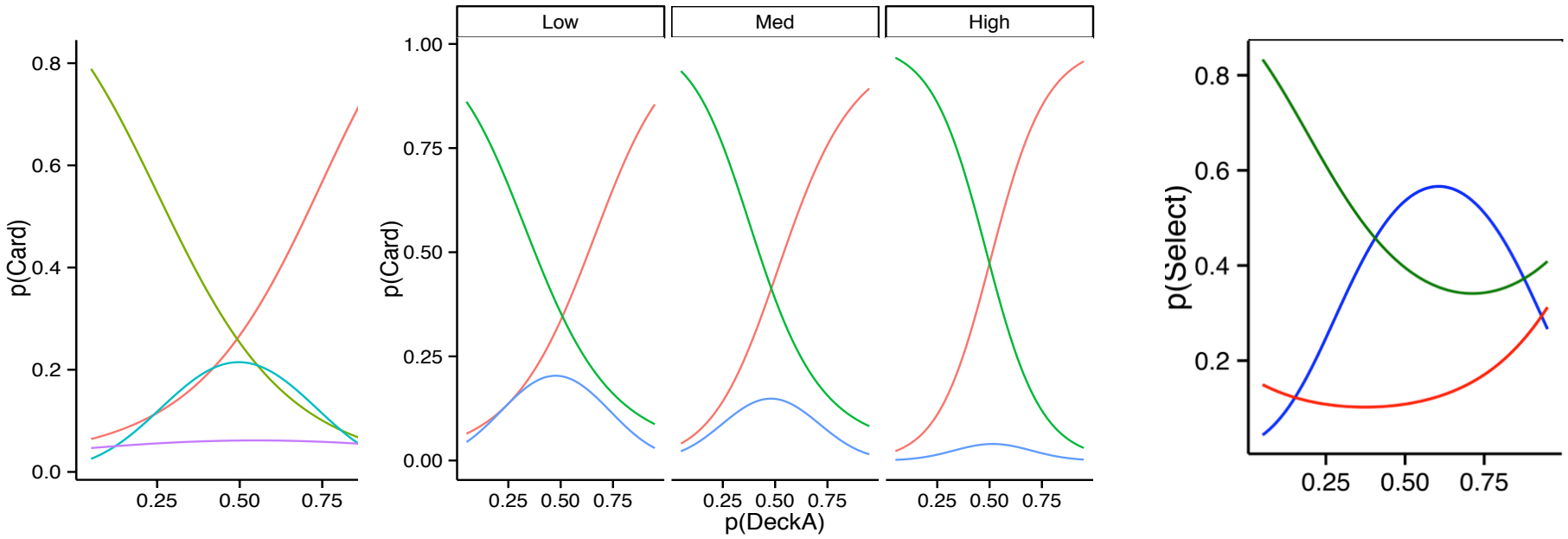
Patterns of information valuation

- behavioral summary -

Uncertainty-driven Information seeking

Sensitivity to information cost

Non-instrumental information seeking



$$SV(a_i) = EV[a_i] + \omega * H(D) * I(O_i;D)$$

Experiment 4

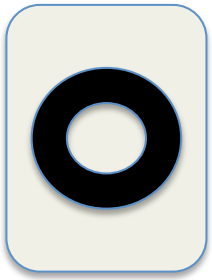
EEG correlates: is information ‘rolled’ into value or represented separately?

Experiment 4: EEG

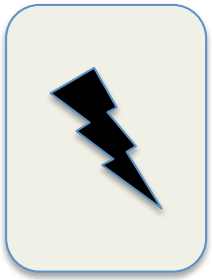
- Design -

Deck A

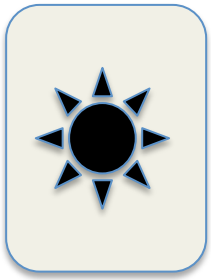
40% 40 pnt
Or 5 pnt
EV=19



25% 25pnt
Or 5 pnt
EV=10



100%
2 pnt
EV=2



Deck B

15% 40 pnt
Or 5 pnt
EV=10

75% 25 pnt
Or 5 pnt
EV=20

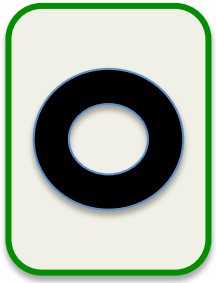
100%
3 pnt
EV=3

Experiment 4: EEG

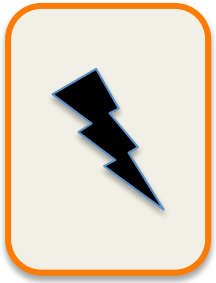
- Design: reward structure -

Deck A

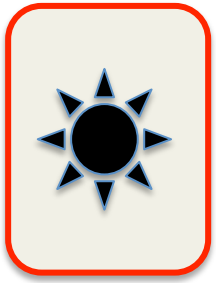
40% 40 pnt
Or 5 pnt
EV=19



25% 25pnt
Or 5 pnt
EV=10



100%
2 pnt
EV=2



Deck B

15% 40 pnt
Or 5 pnt
EV=10

75% 25 pnt
Or 5 pnt
EV=20

100%
3 pnt
EV=3

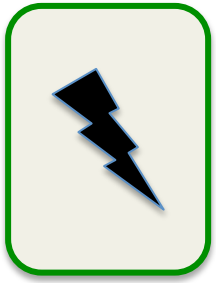
Experiment 4: EEG

- Design: reward structure -

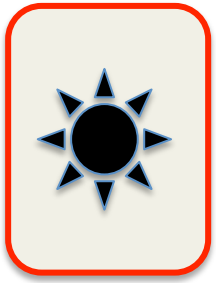
Deck A

40% 40 pnt
Or 5 pnt
EV=19

25% 25pnt
Or 5 pnt
EV=10



100%
2 pnt
EV=2



Deck B

15% 40 pnt
Or 5 pnt
EV=10

75% 25 pnt
Or 5 pnt
EV=20

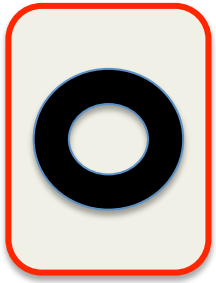
100%
3 pnt
EV=3

Experiment 4: EEG

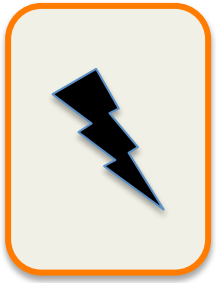
- Design: Information structure -

Deck A

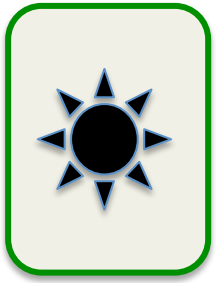
40% 40 pnt
Or 5 pnt
EV=19



25% 25pnt
Or 5 pnt
EV=10



100%
2 pnt
EV=2



Deck B

15% 40 pnt
Or 5 pnt
EV=10
I(O;D)<0.1

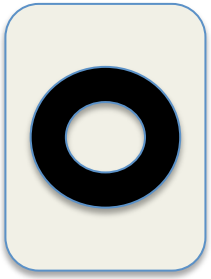
75% 25 pnt
Or 5 pnt
EV=20
I(O;D)=0.2

100%
3 pnt
EV=3
I(O;D)=1

Experiment 4: EEG

- Design: Choice -

40% 40 pnt
Or 5 pnt
EV=19



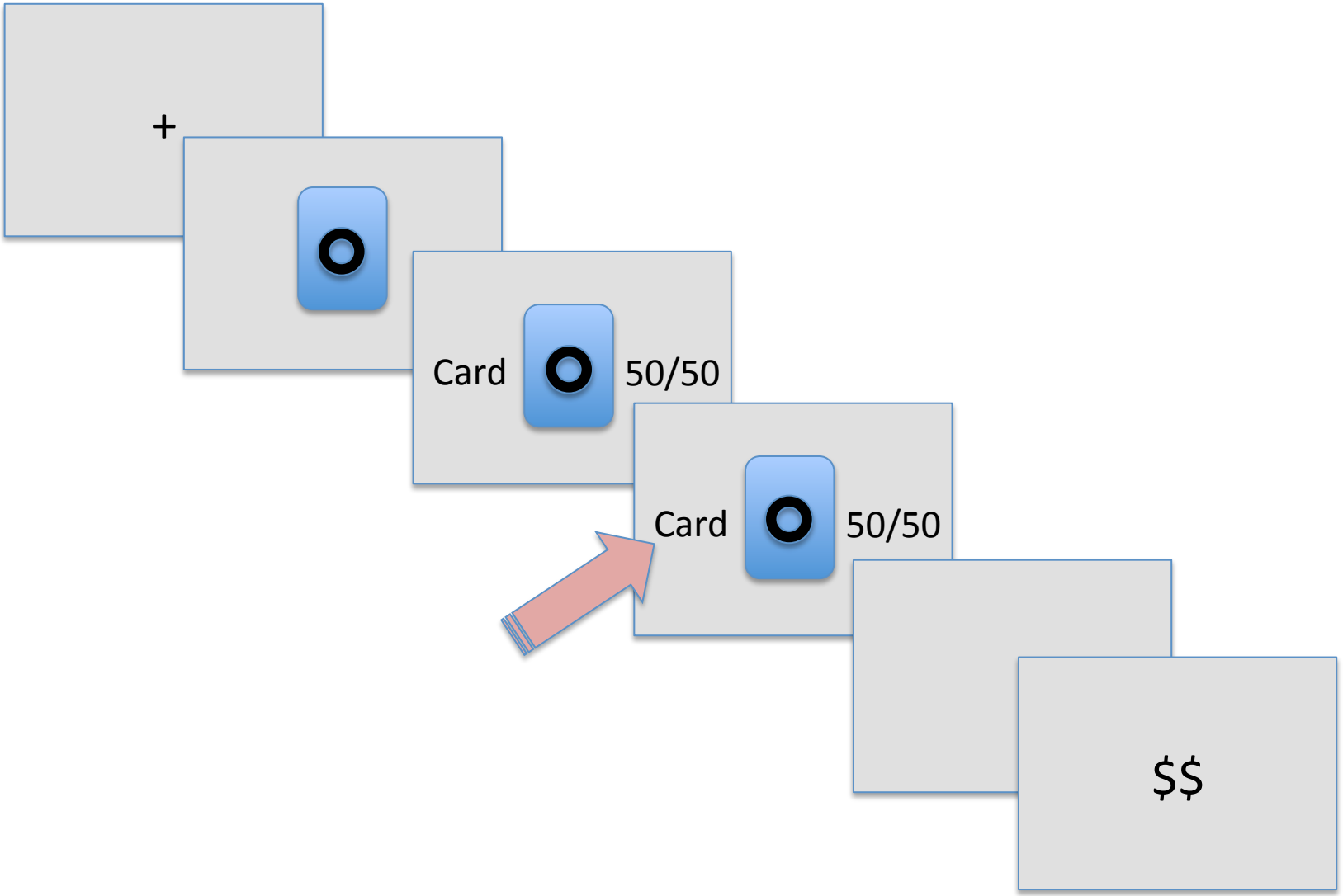
15% 40 pnt
Or 5 pnt
EV=10
I(O;D)<0.1

50/50 Draw

50% 22 pnts
50% 8 pnts
EV=15

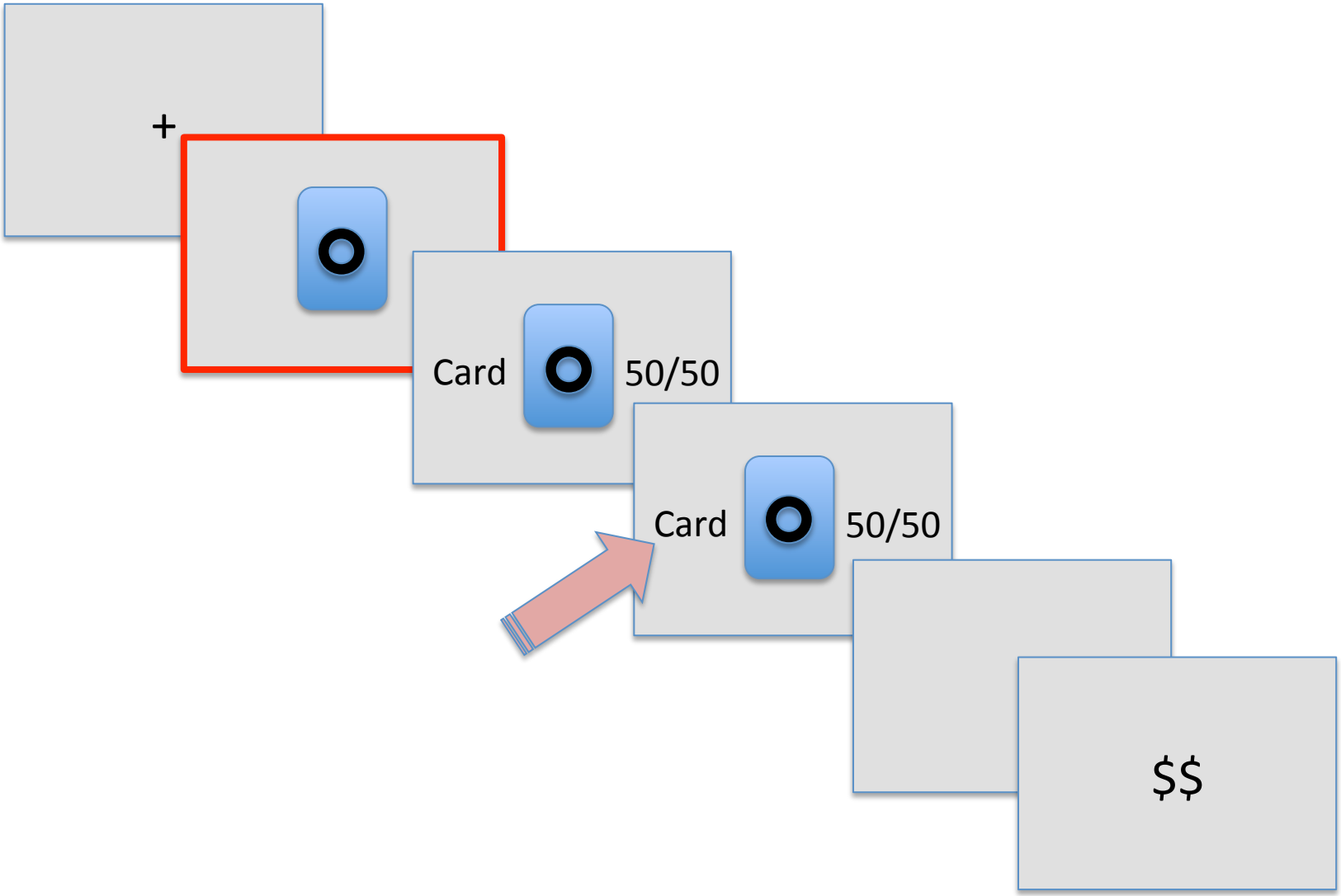
Experiment 4: EEG

- Design: Dynamics -



Experiment 4: EEG

- Design: Dynamics -

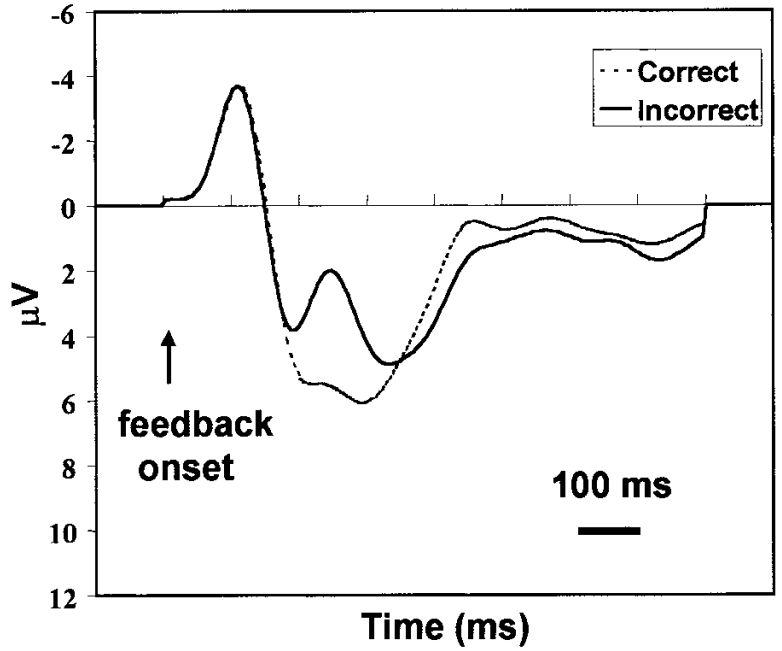


EEG predictions

- components associated with value -

N2 modulation

- expected reward -

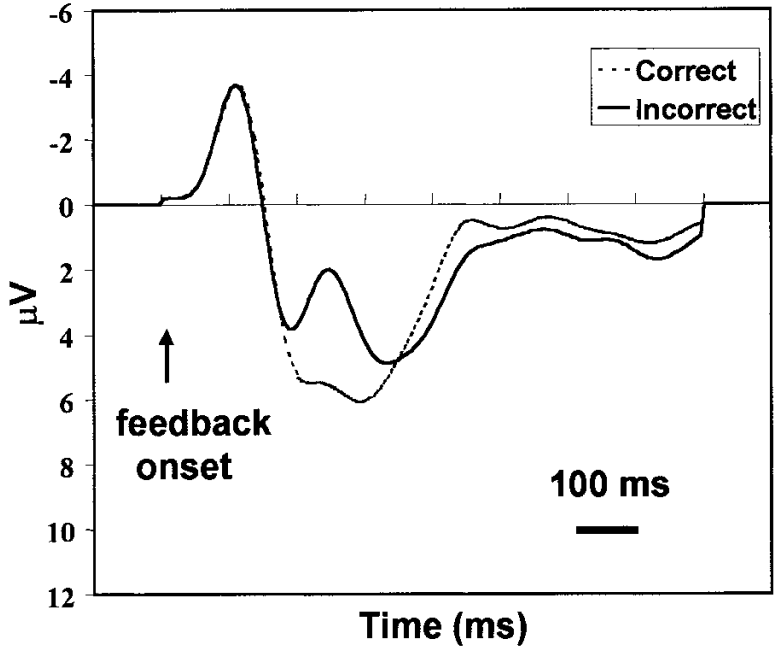


Holroyd et at., 2002

EEG predictions

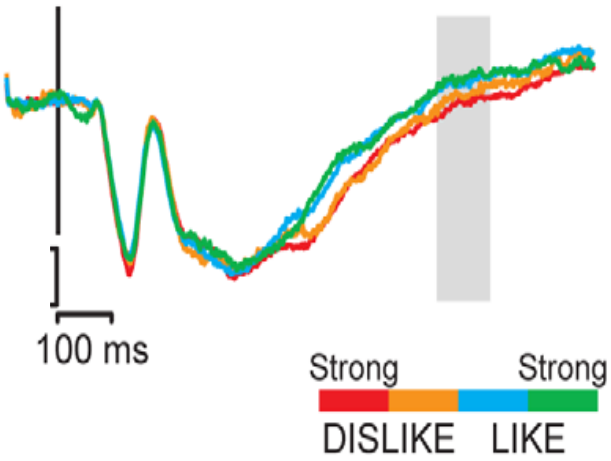
- components associated with value -

N2 modulation
- expected reward -



Holroyd et at., 2002

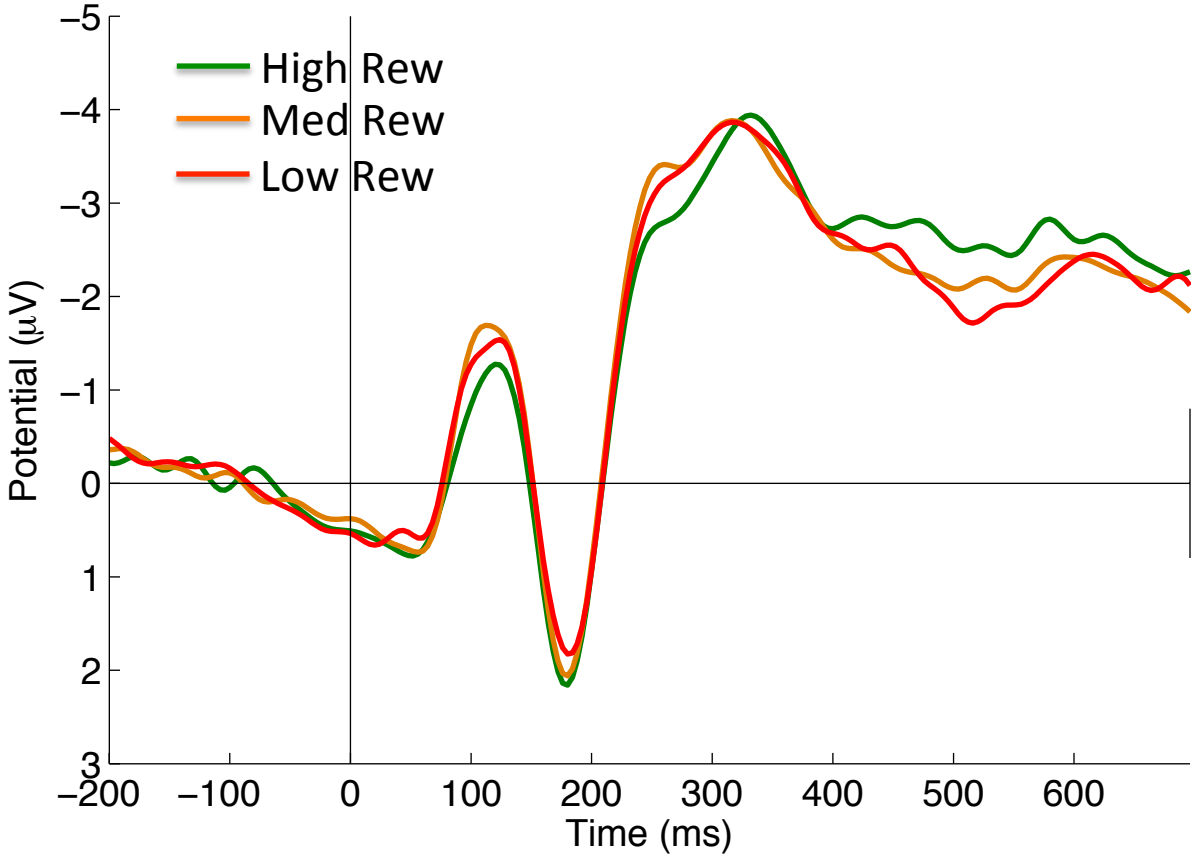
Late-wave modulation
- subjective value -



Harris et at., 2011

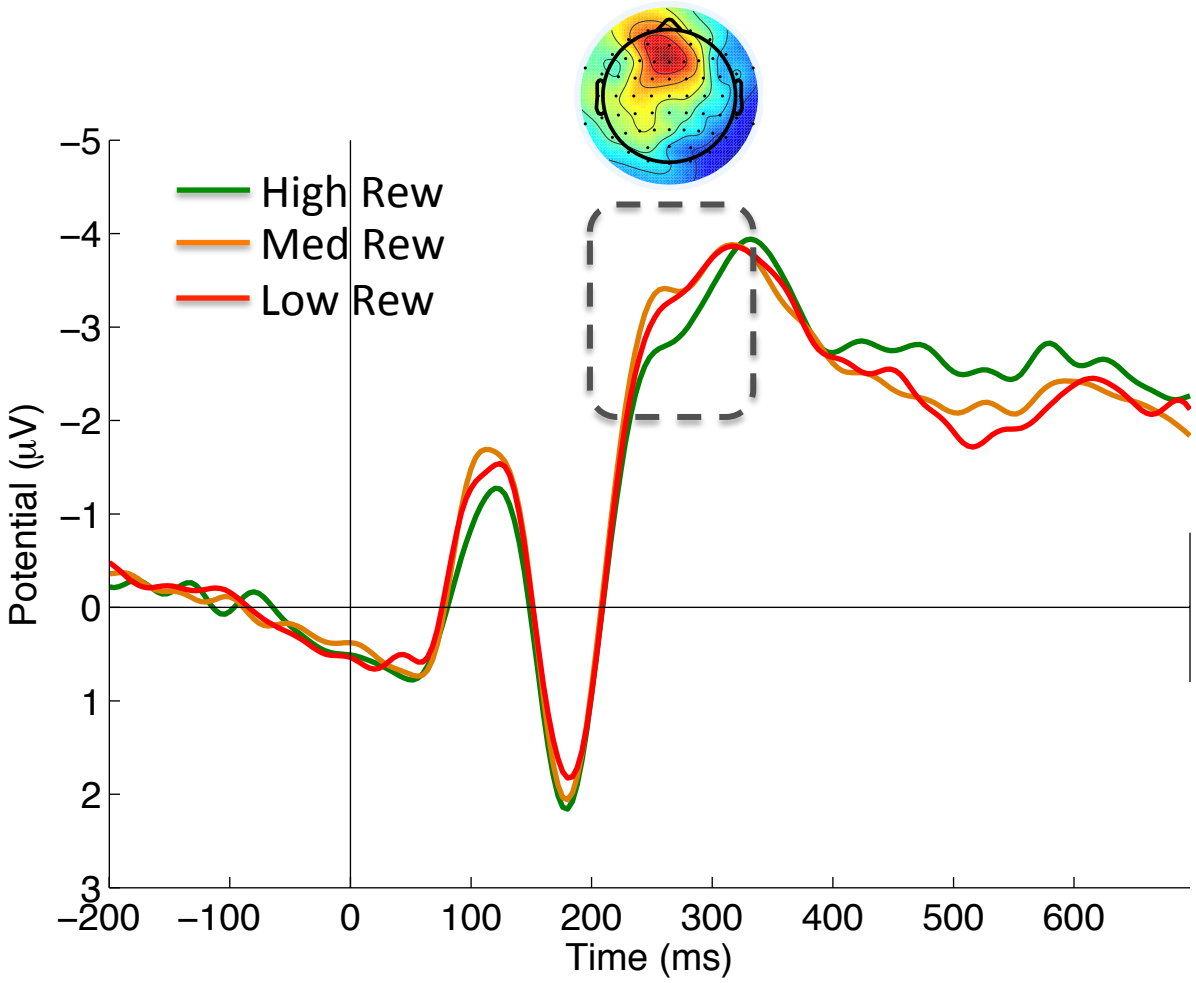
Low-Uncertainty waveforms

- Anterior sites: By expected reward gain -



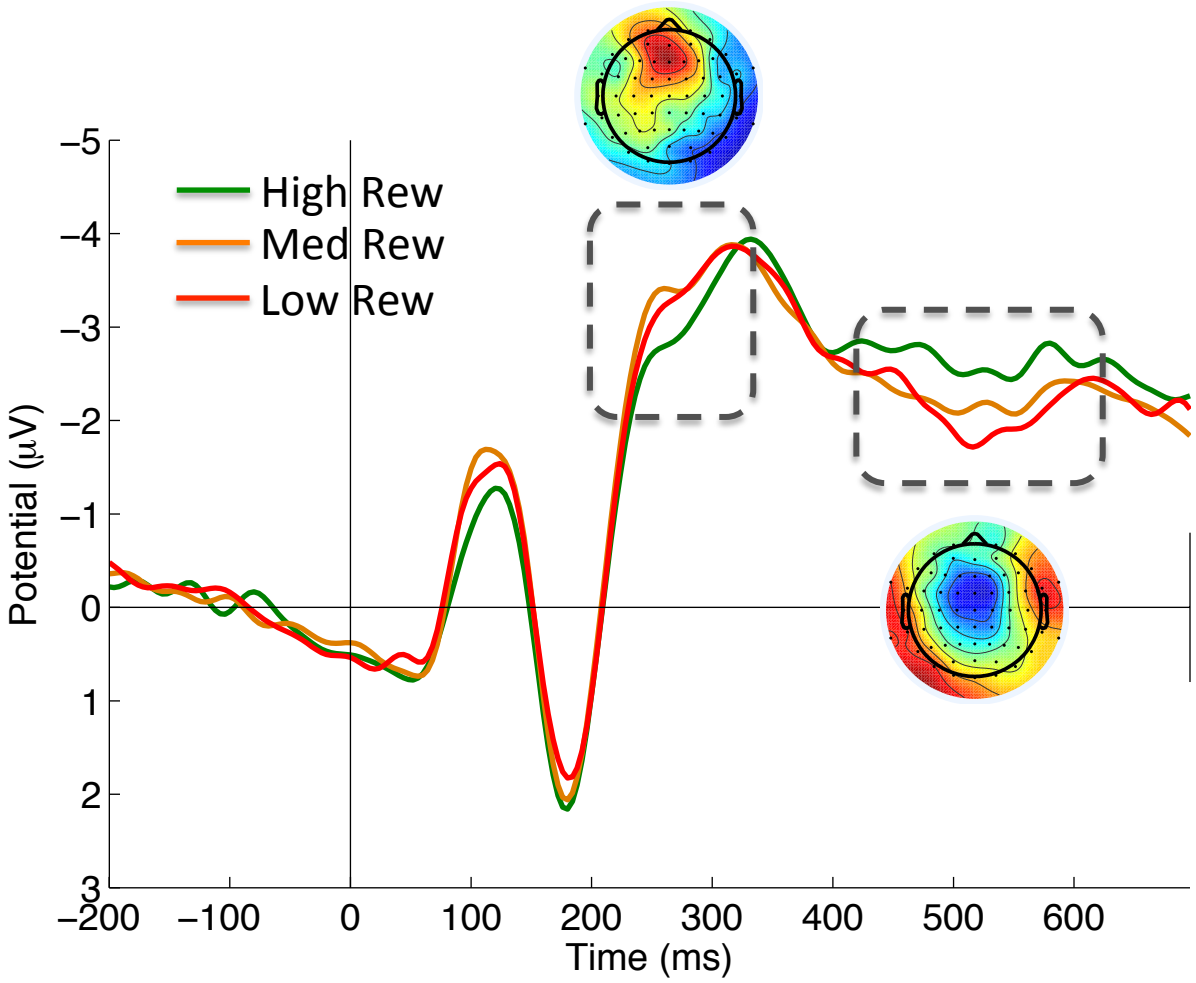
Low-Uncertainty waveforms

- Anterior sites: By expected reward gain -



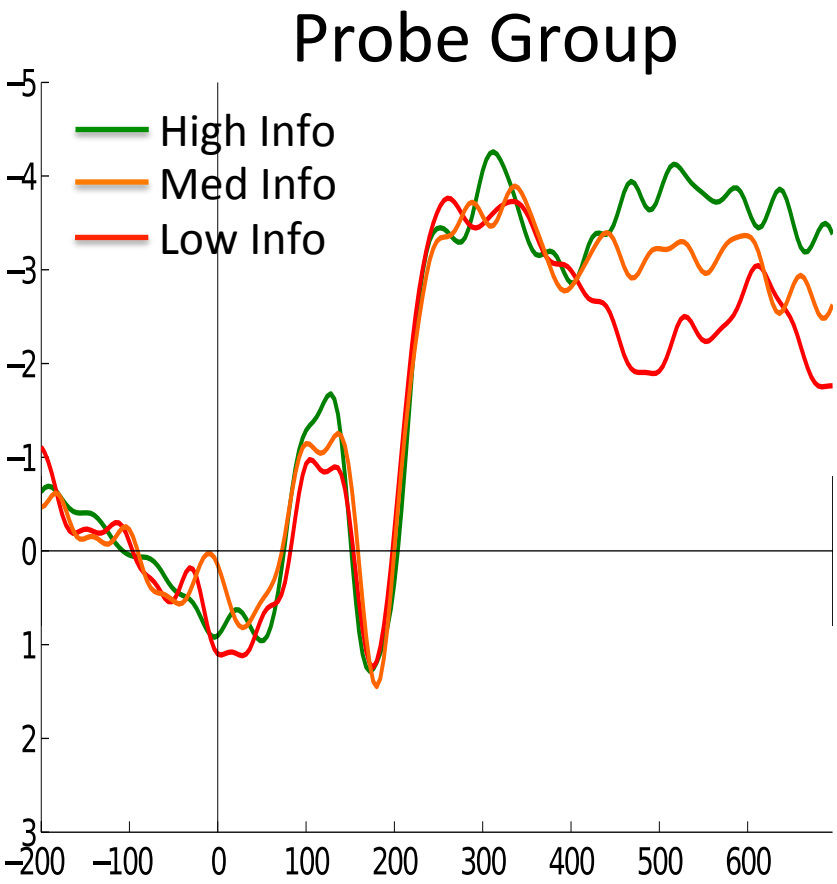
Low-Uncertainty waveforms

- Anterior sites: By expected reward gain -



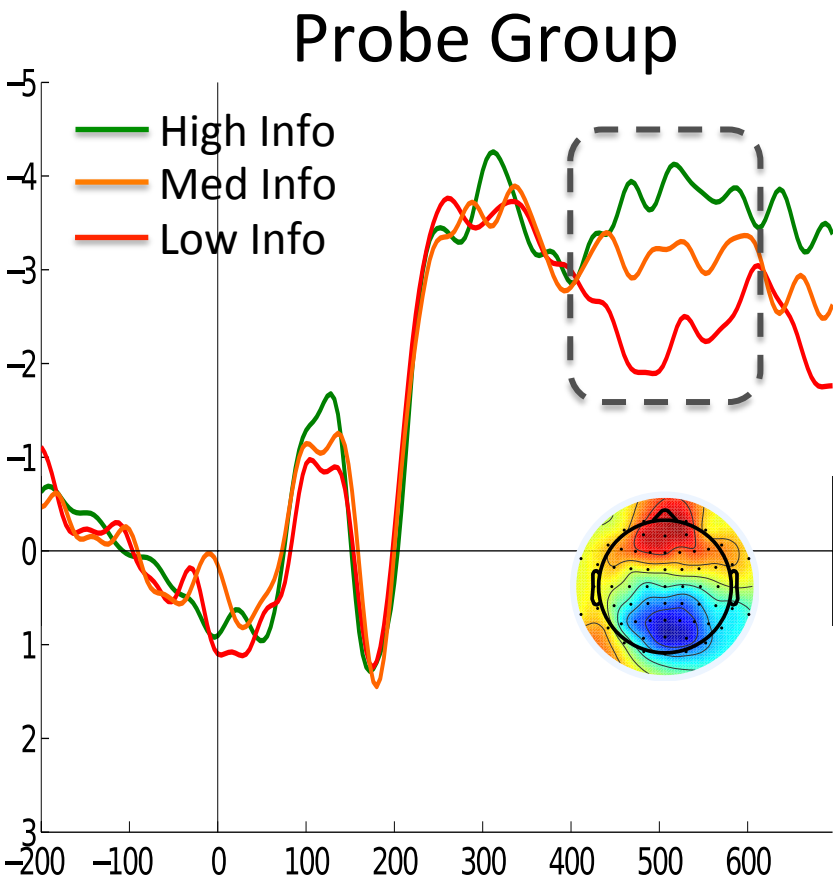
High-Uncertainty waveforms

- Anterior sites: By expected information gain -



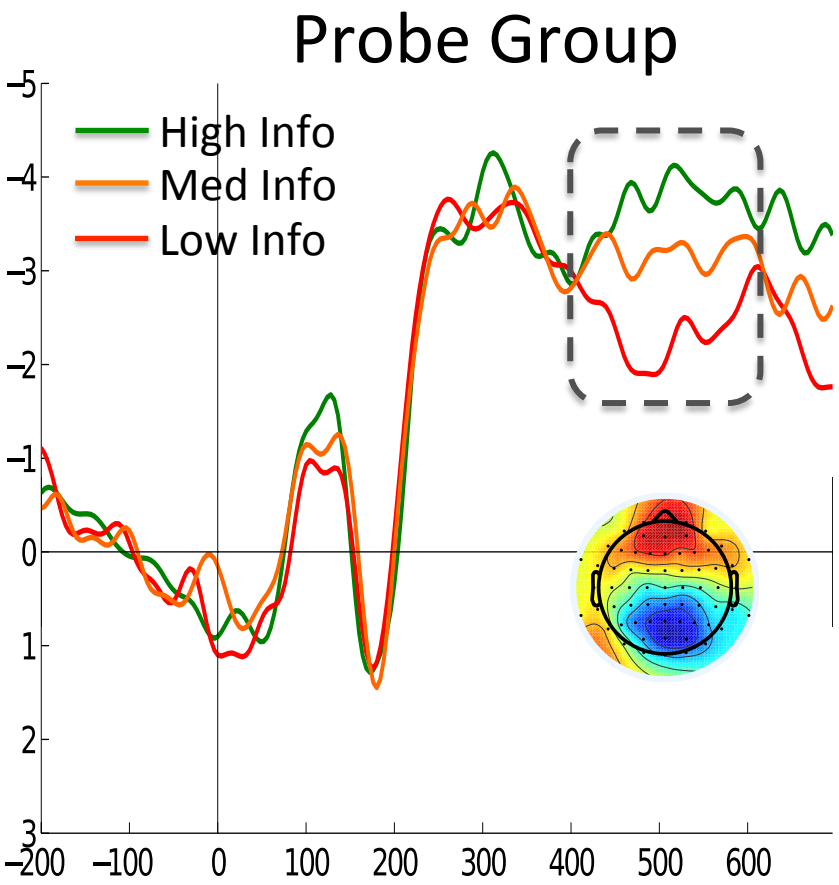
High-Uncertainty waveforms

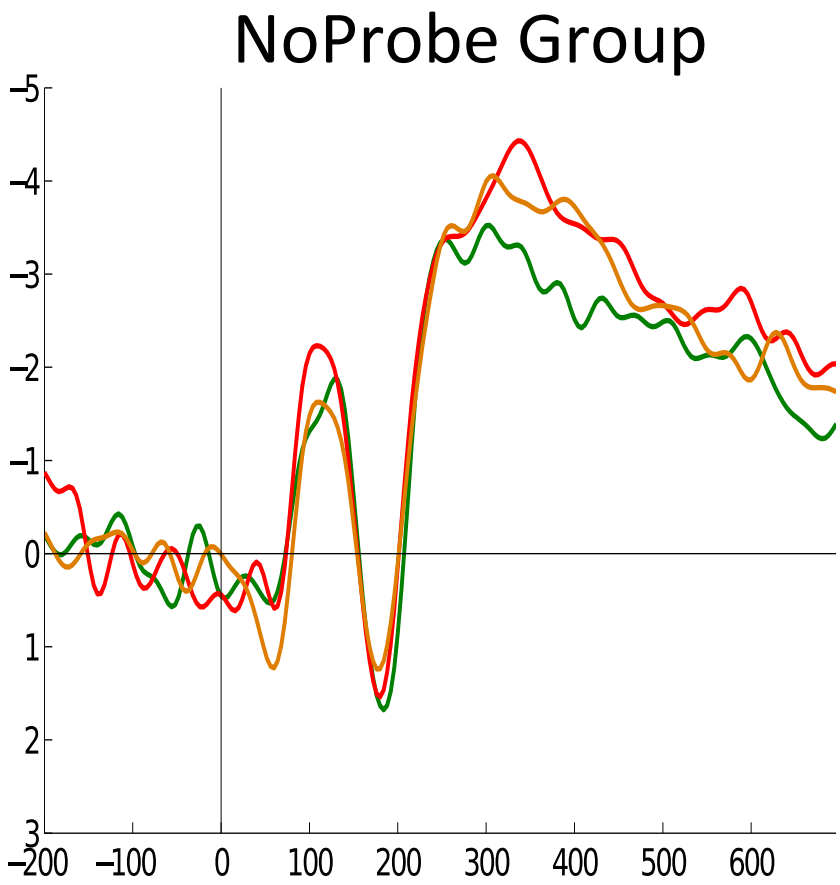
- Anterior sites: By expected information gain -



High-Uncertainty waveforms

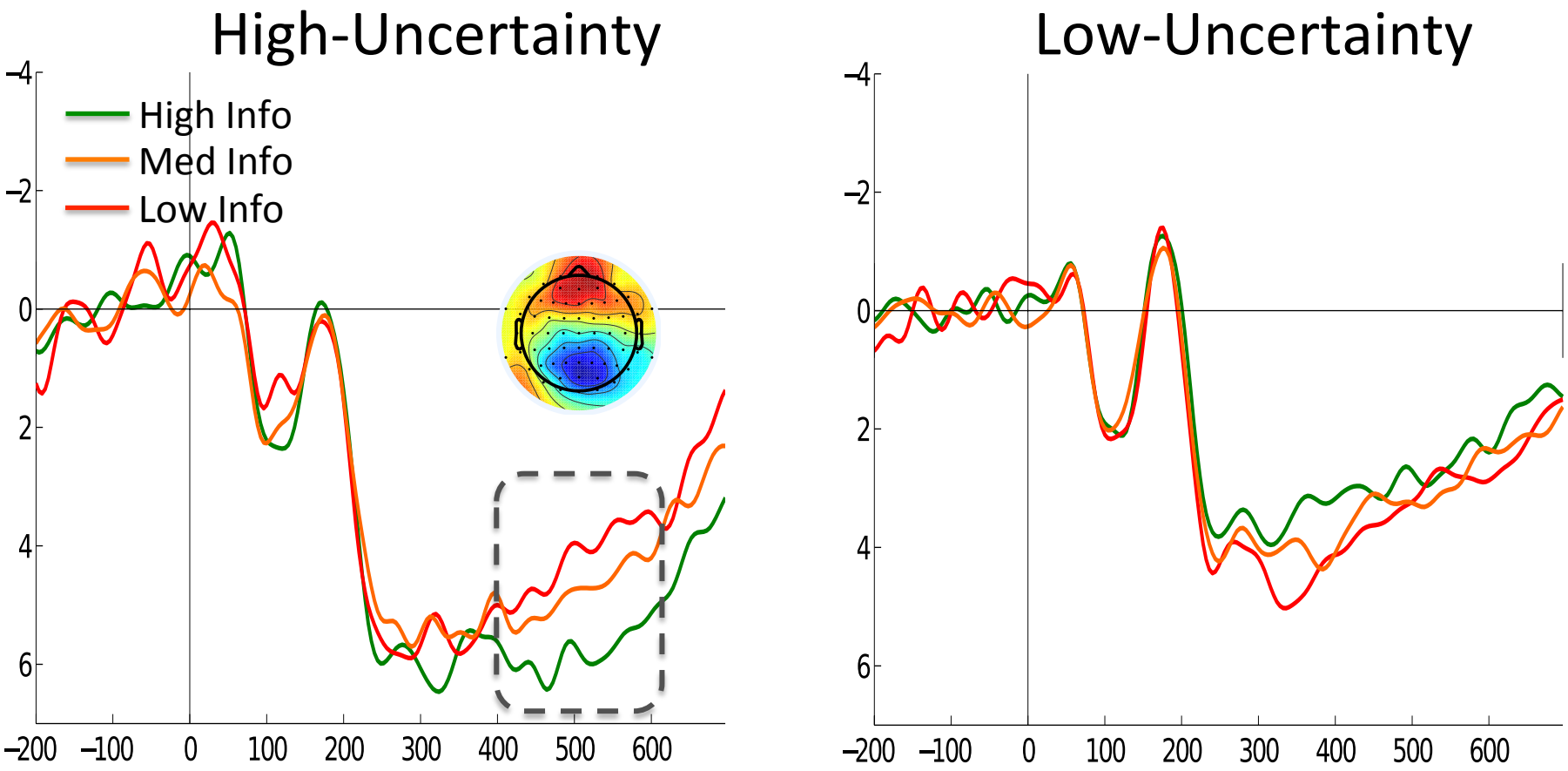
- Anterior sites: By expected information gain -





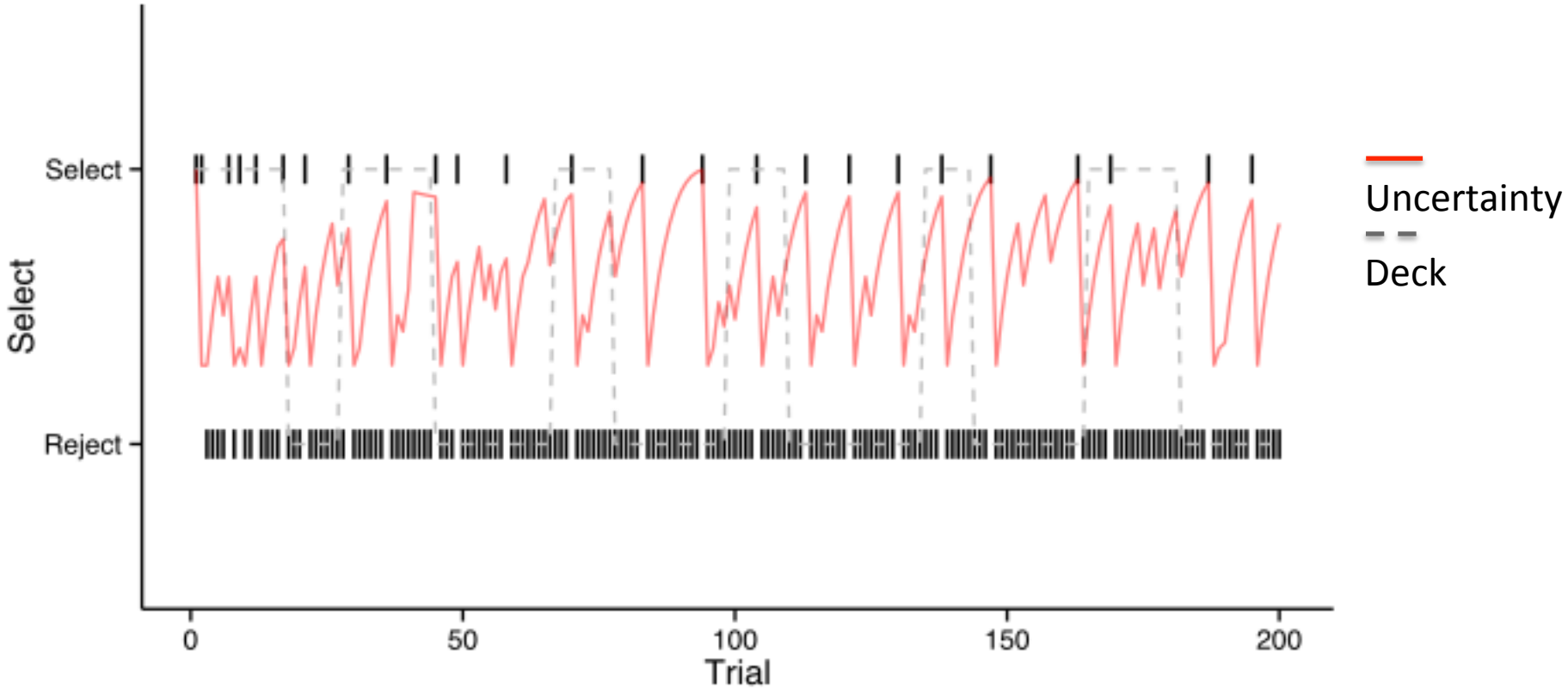
Probe group waveforms

- Posterior sites: By expected information gain -



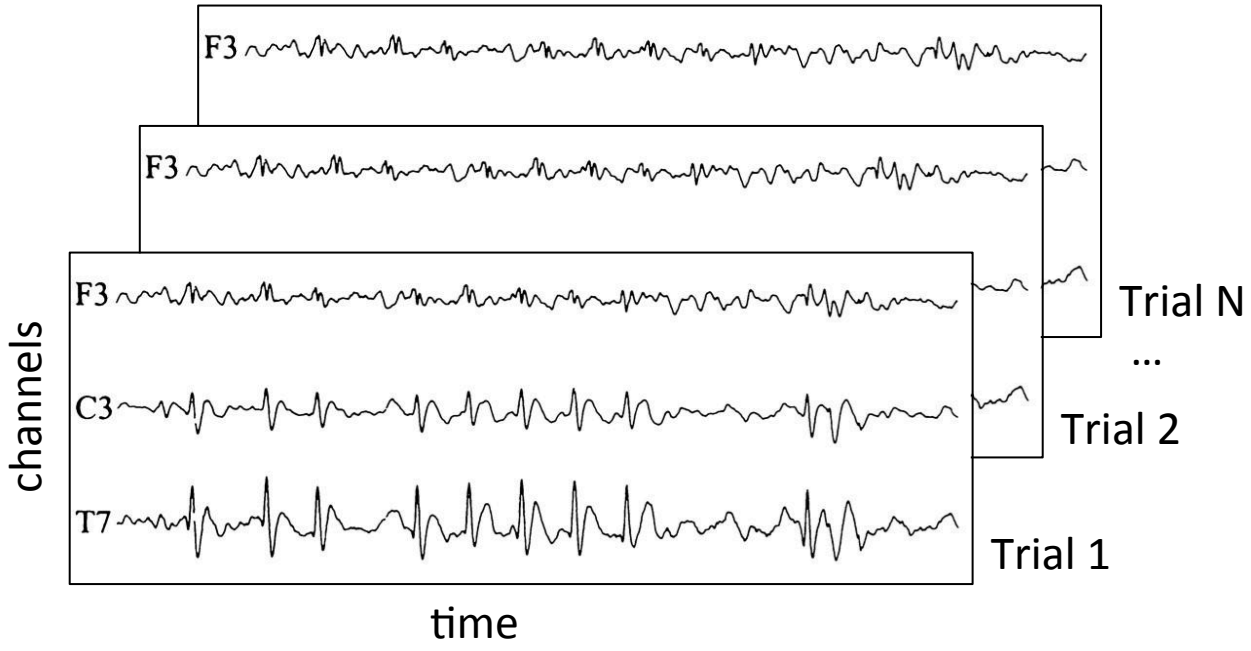
Model-based analysis

- applying GLMs to single-trial data -



Model-based analysis

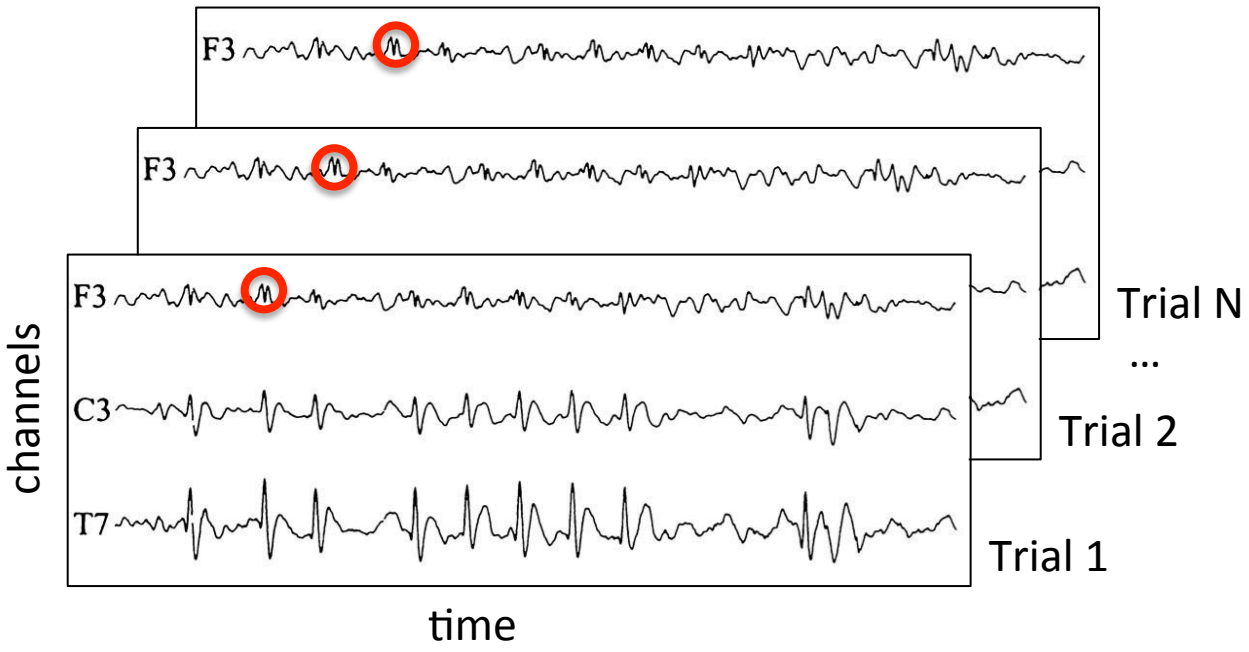
- applying GLMs to single-trial data -



Model-based analysis

- applying GLMs to single-trial data -

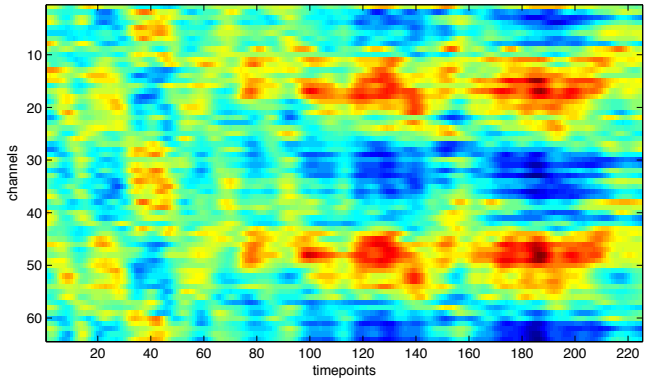
$$V_{c:i,tp:j,tr:k} = \beta_0 + \beta_1 X_{tr:k}$$

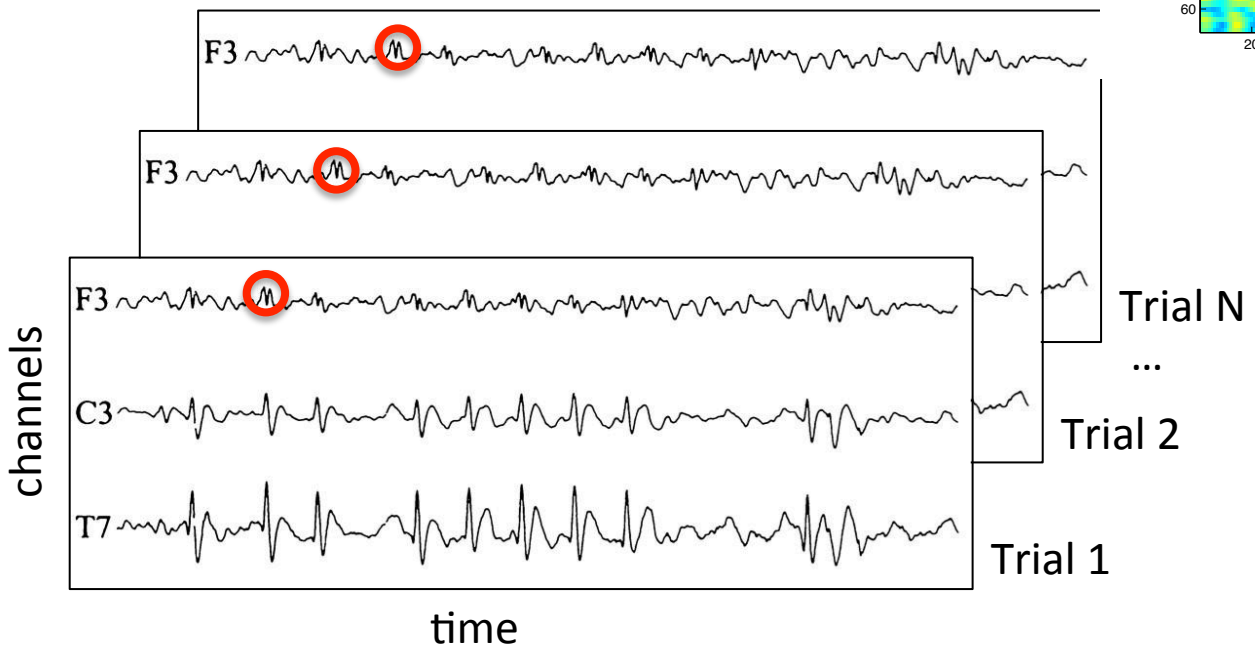


Model-based analysis

- applying GLMs to single-trial data -

$$V_{c:i,tp:j,tr:k} = \beta_0 + \beta_1 X_{tr:k}$$





A model-based analysis

- sites associated with value -

$$SV(a_i) = EV[ai] + \omega * H(D) * I(O_i;D)$$

A model-based analysis

- sites associated with value -

$$SV(a_i) = EV[ai] + \omega * H(D) * I(O_i;D)$$

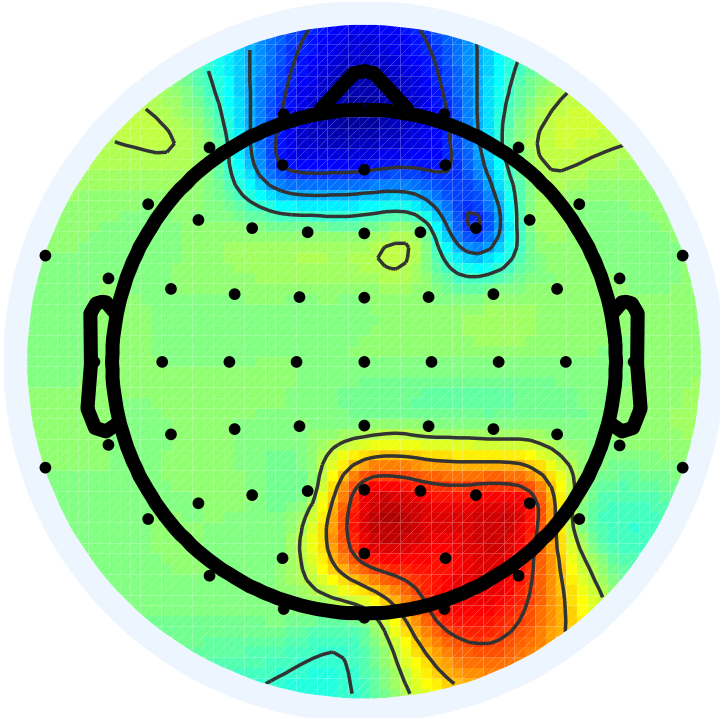
$$V_{tr:k} = \beta_0 + \beta_1(Info + EV)_{tr:k} + \beta_2(Info - EV)_{tr:k}$$

A model-based analysis

- sites associated with value -

$$SV(a_i) = EV[ai] + \omega * H(D) * I(O_i;D)$$

$$V_{tr:k} = \beta_0 + \beta_1(Info + EV)_{tr:k} + \beta_2(Info - EV)_{tr:k}$$

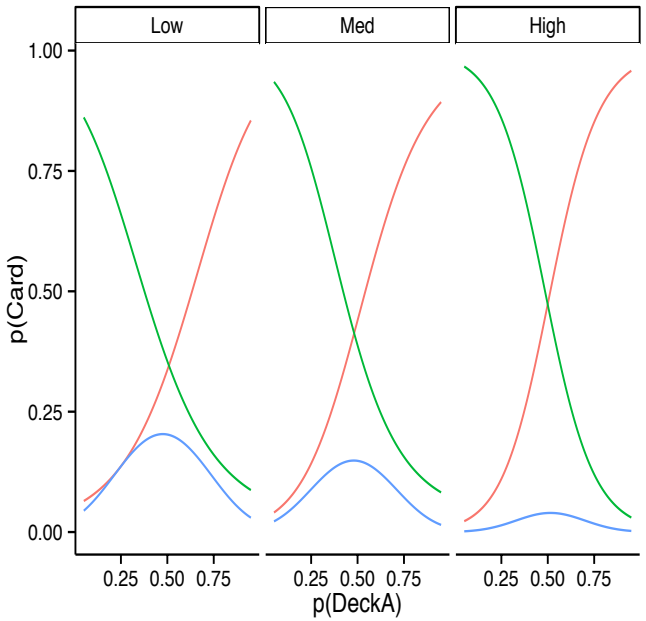


Conclusions

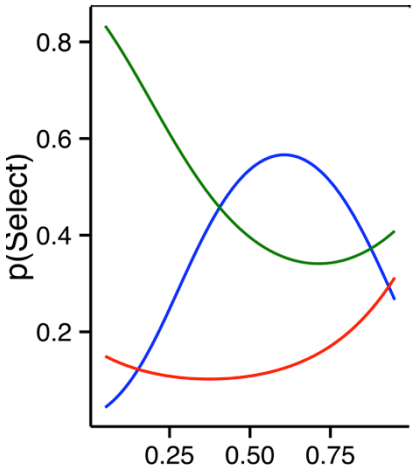
- a case for the value of information -

$$SV(a_i) = EV[ai] + \omega * H(D) * I(O_i;D)$$

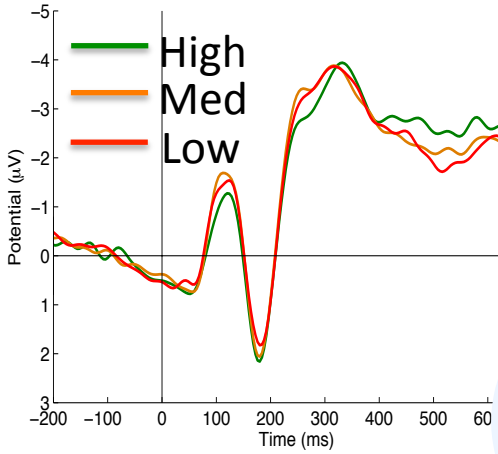
Sensitivity to information cost

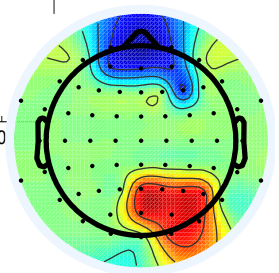


Non-instrumental information seeking

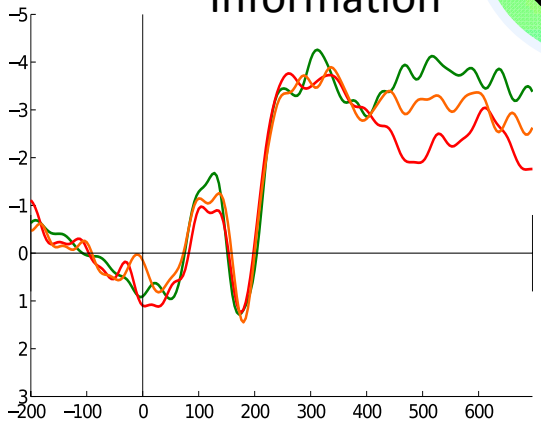


Reward





Information

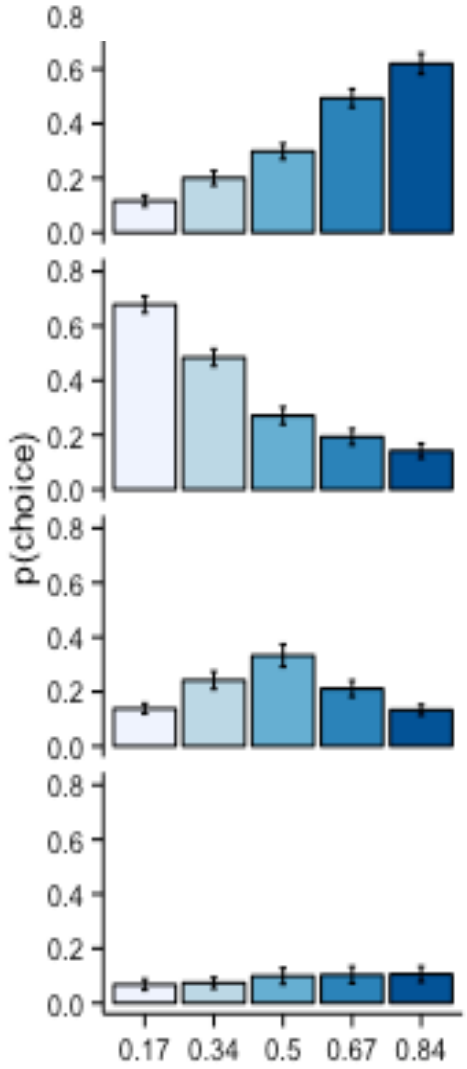


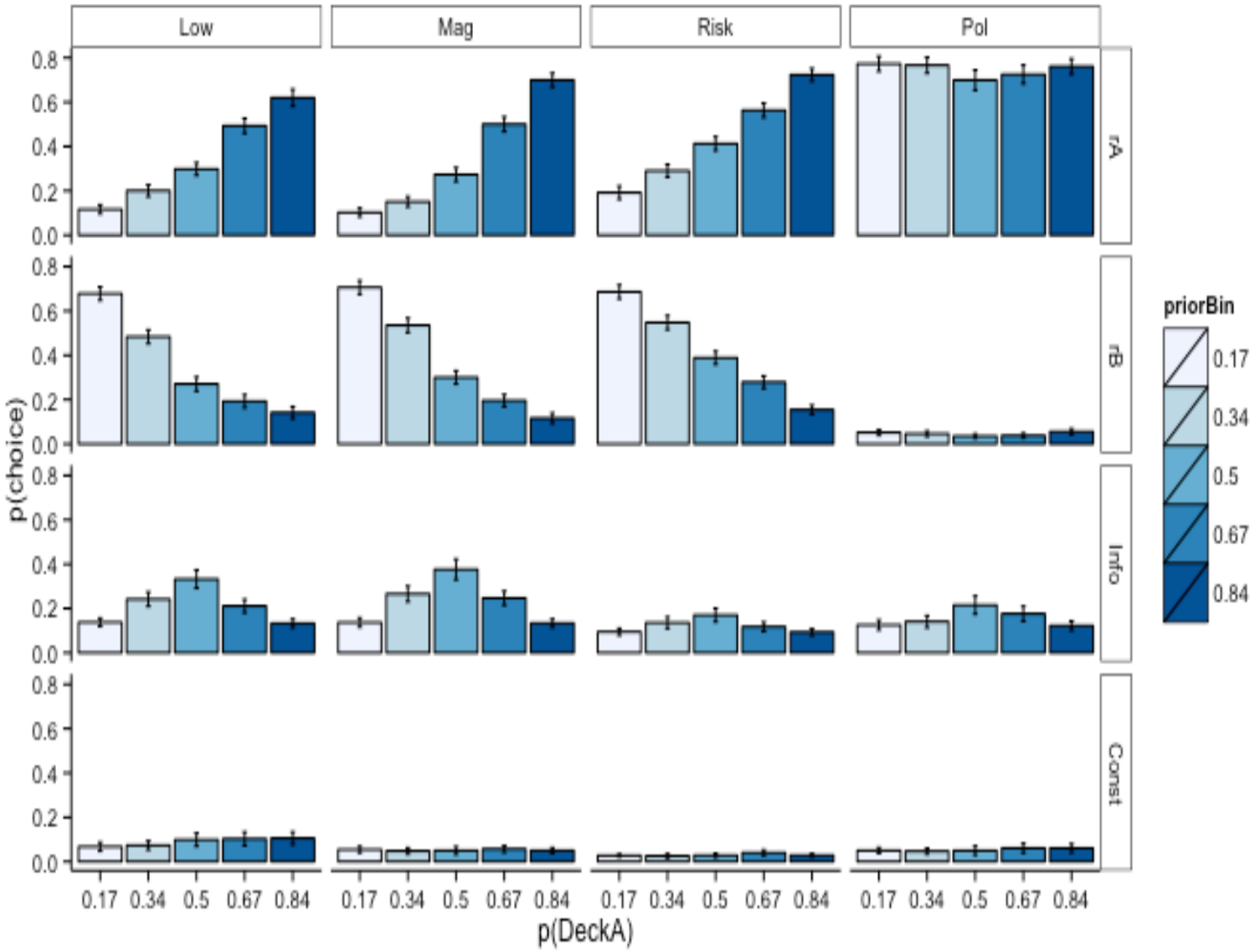
Conclusions

- Information is valuable
  - Probe actions that are most informative about latent state, especially when uncertain
  - Heuristic model approximates optimal solution but makes distinct predictions for single policy versions
  - EEG signals dissociate value and information as predicted

Thanks!

Jeff Cockburn


p(choice)
0.0
0.2
0.4
0.6
0.8
0.17
0.34
0.5
0.67
0.84
P(deckA)


Low
Mag
Risk
Pol
rA
rB
Info
Const
p(choice)
p(DeckA)
0.0
0.2
0.4
0.6
0.8
0.17
0.34
0.5
0.67
0.84
priorBin

Experiment 4

- Design: Protocol -

Training: 30 trial blocks – deck **known**

Testing: 5% deck switch – deck **unknown**

Experiment 4

- Design: Protocol -

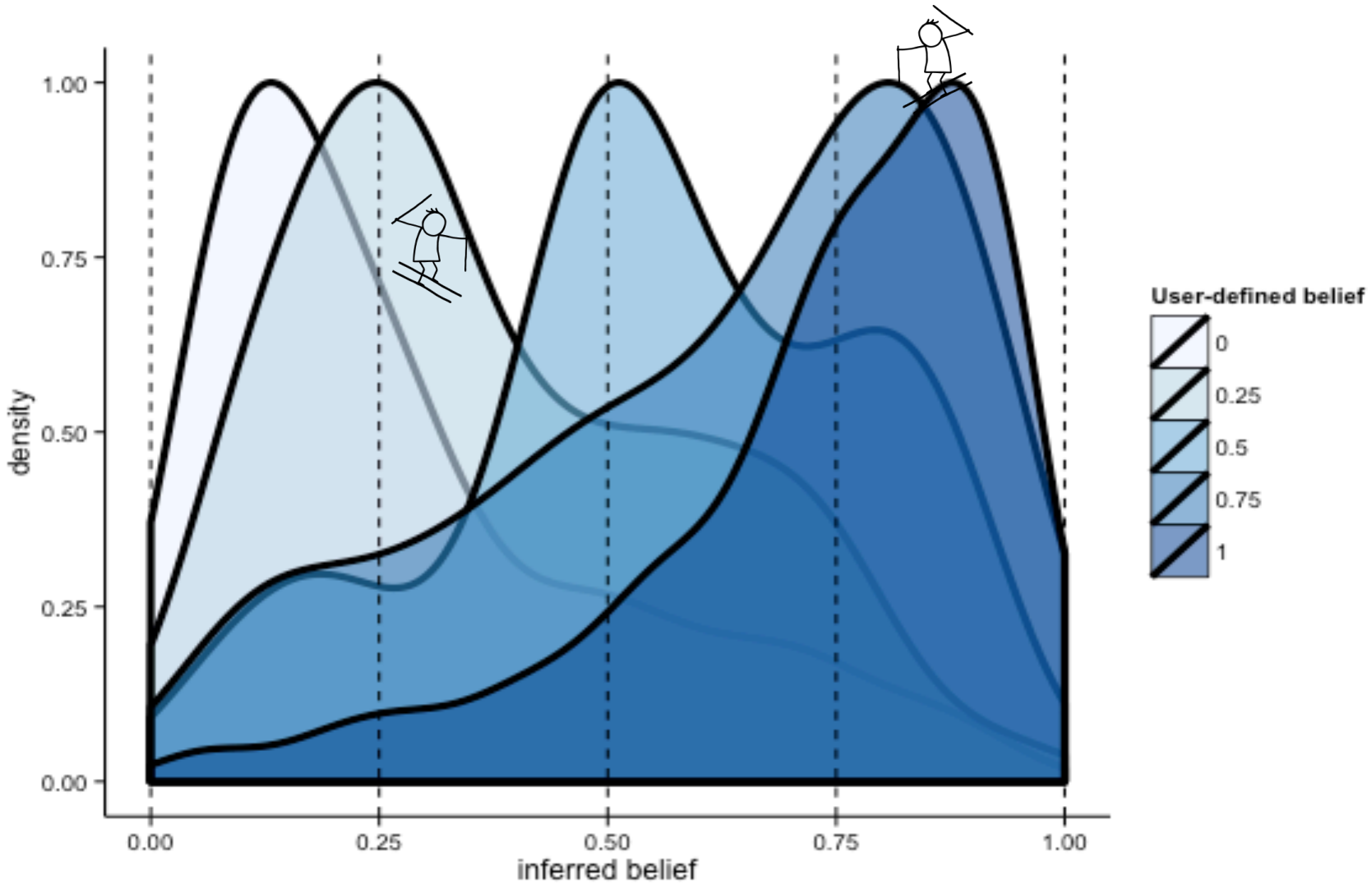
Training: 30 trial blocks – deck **known**

Testing: 5% deck switch – deck **unknown**

Belief Probe: after 10% of trials

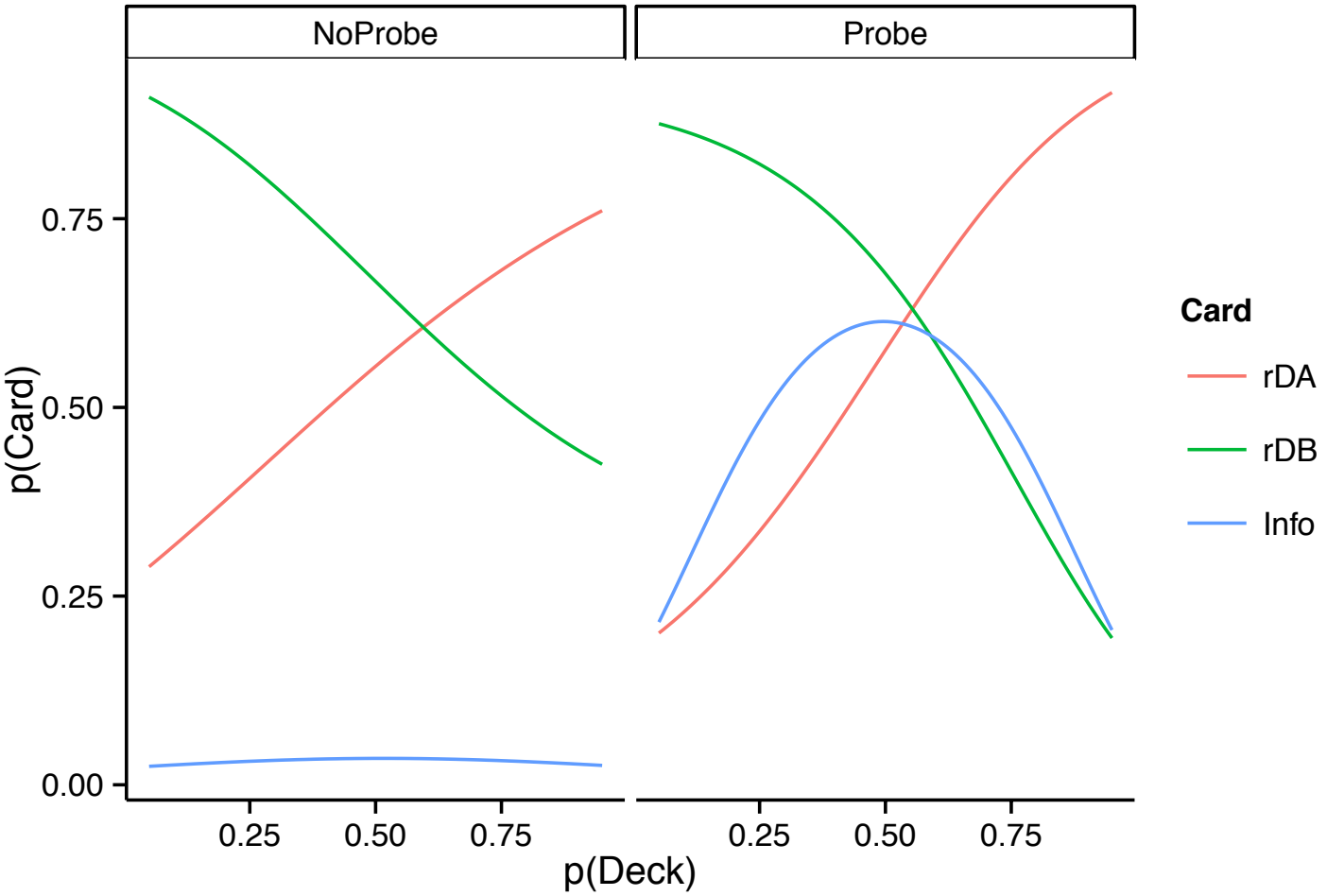
Optimal observer's validity

- match between participant's & inferred belief -

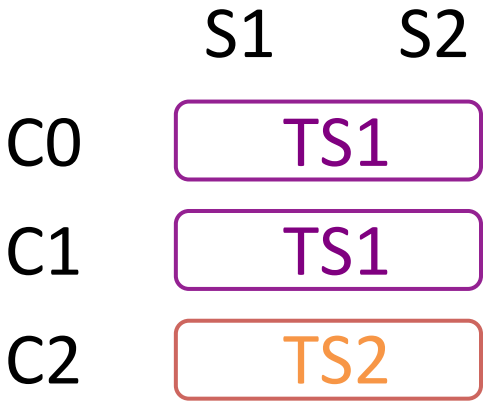


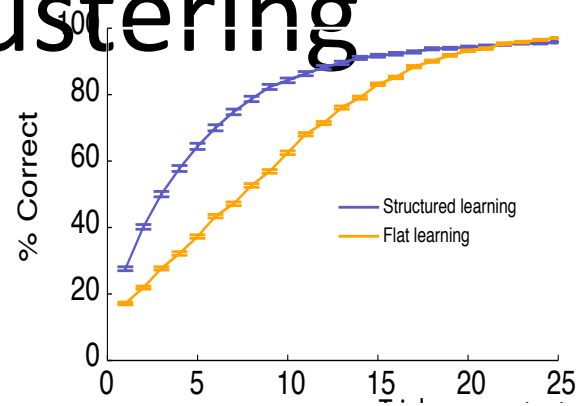
EEG study: Response patterns

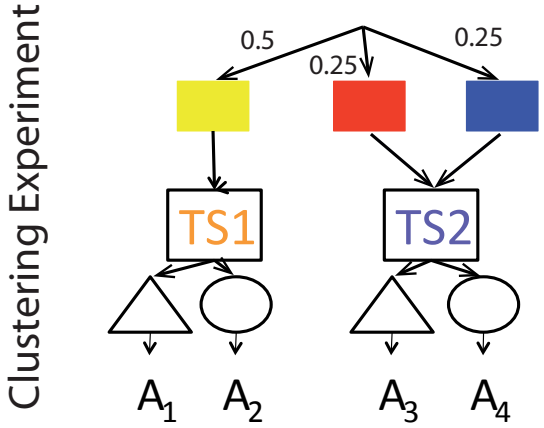
- Probe vs. NoProbe groups -

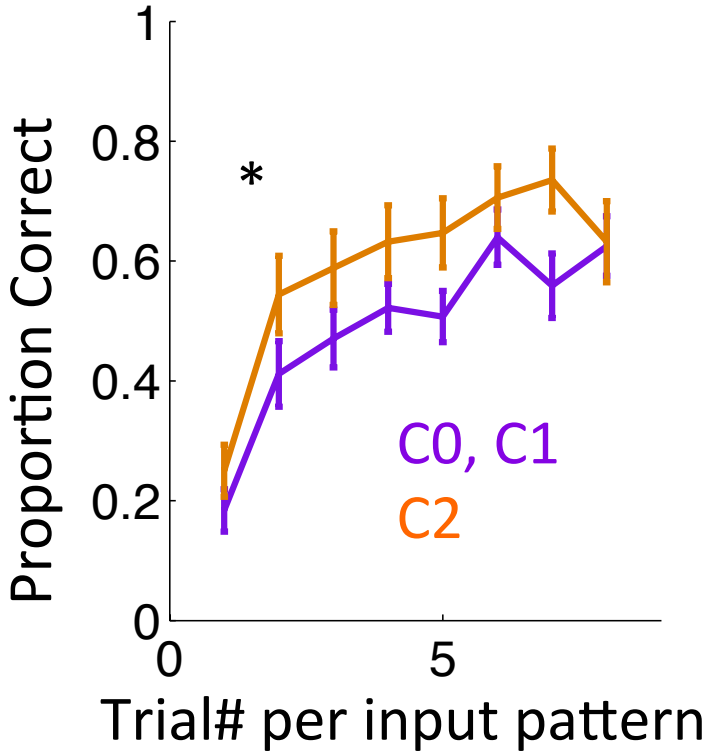


Initial clustering



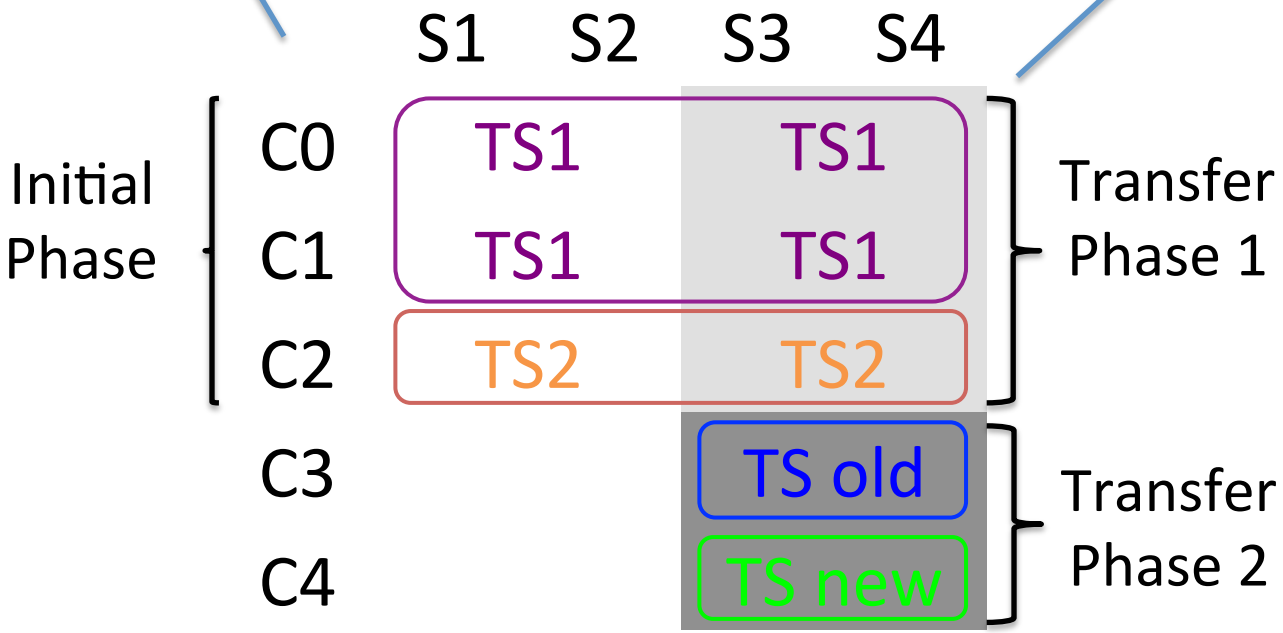






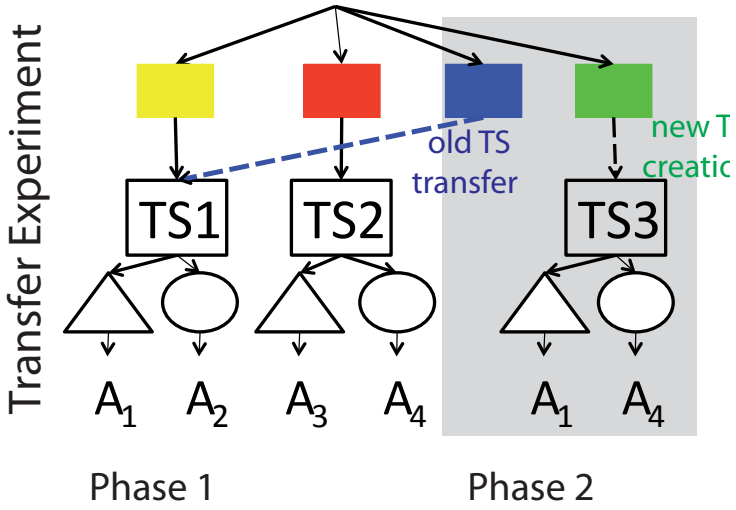
No early clustering benefit - early structure learning is costly

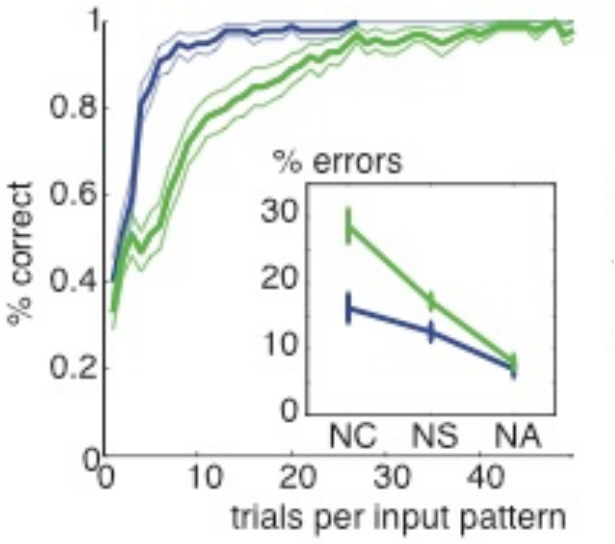
Structure learning affords transfer of new information within learned clusters



Structure learning affords transfer of known rules to new contexts – with CRP-like prior

Neural Network – Results



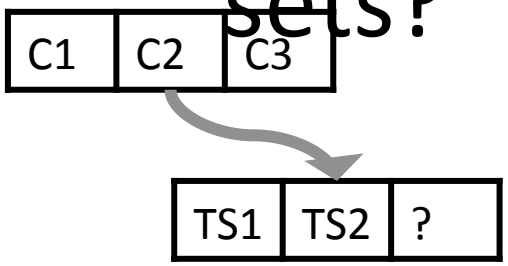


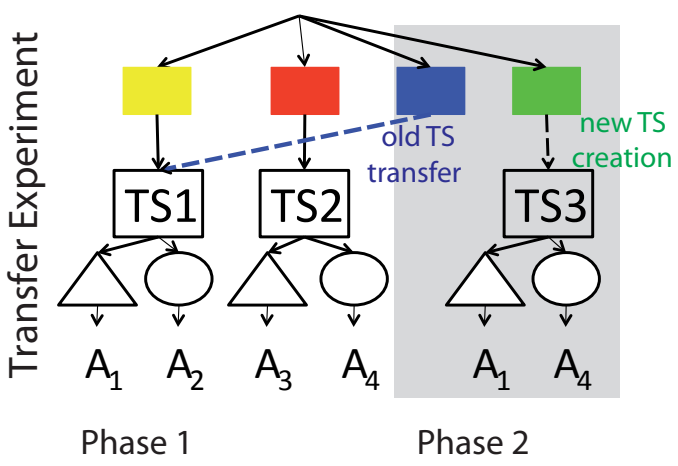
Test phase
new contexts:
Old TS
New TS

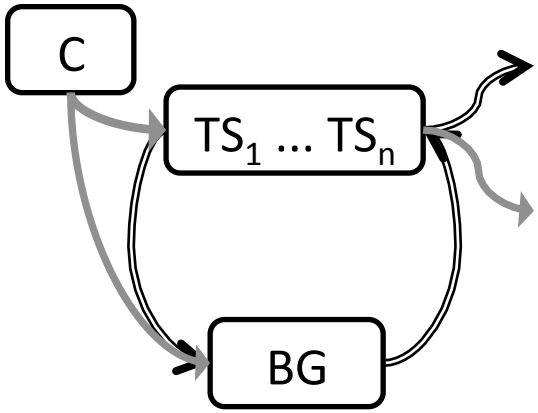
Collins & Frank, 2013

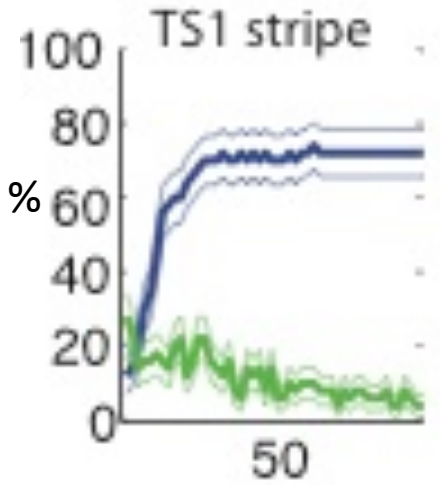
The network learns efficiently unsupervised
Clusters Contexts
Predicts positive, negative transfer

How does the network transfer task sets?





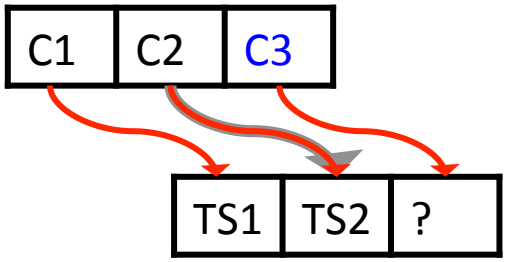


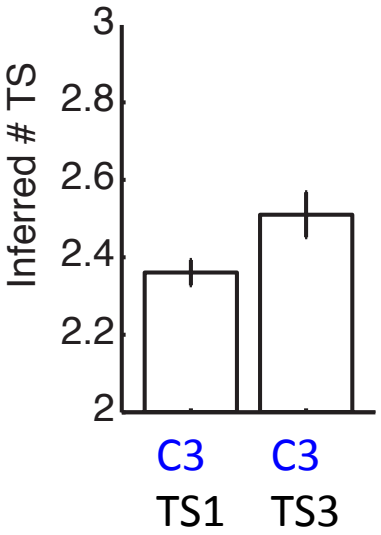


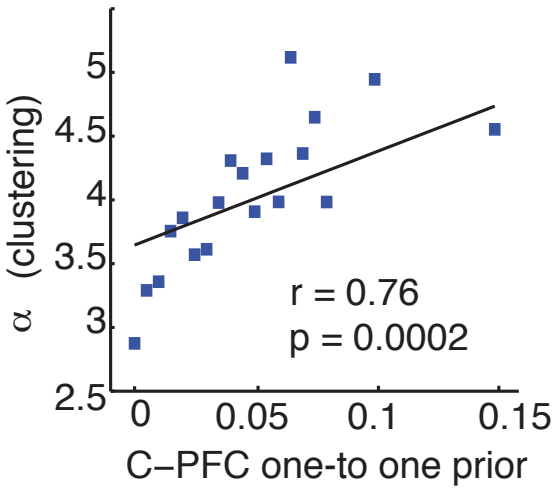
Collins & Frank, 2013

Transfer occurs through reselection of learned PFC states (TS) in new contexts

Algorithmic model vs. Neural network







Collins & Frank, 2013

Both models are approximations of the same process: TS space building