Where am I and What Should I do Next? Overcoming perceptual aliasing in sequential tasks

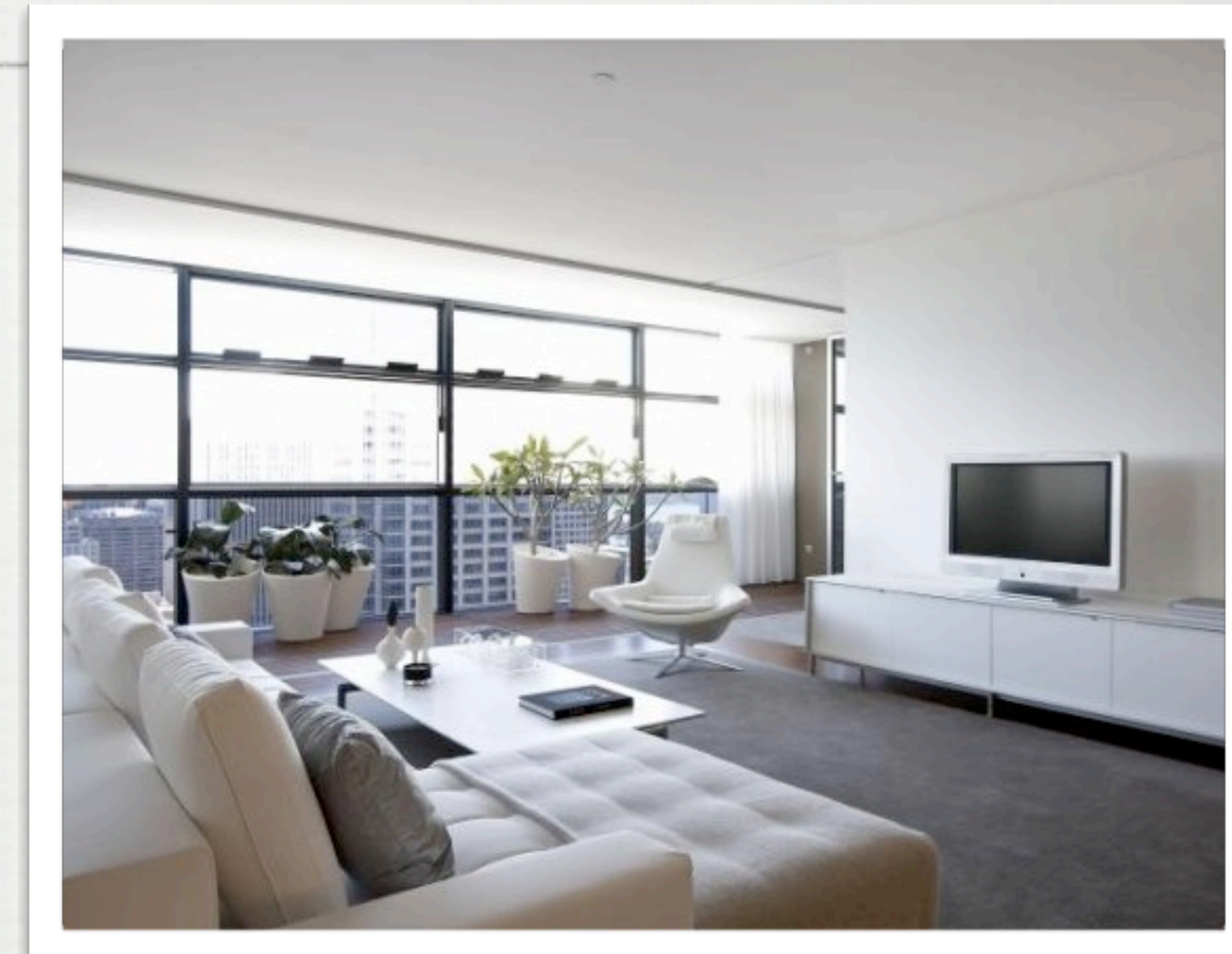
Todd M. Gureckis - New York University

Machine Learning Meets Human Learning

Machine Learning Meets Human Learning

Already pretty "Serious"

Human Learning



Machine Learning

Dynamic Decision Making

Understanding human learning in dynamic/active tasks environments

- Interactions between the behavior of the individual and the 'state' of the system such that each new decision alters future decisions
- Past Examples:
  - Sugar production Factory or the person control task (Berry & Broadbent, 1984, 1987; Stanley, et al. 1989; Sun, Slusarz, Terry, 2005)
  - Micro-world tasks: Fungus eater on Mars task (Toda, 1962), fire-fighters task (Brehmer & Allard, 1991), control of Predator-Prey systems (Dorner & Preubler, 1990; Jensen & Brehmer, 2003)
  - Dynamic motor control tasks (Baddeley, Ingram, & Miall, 2003; Chhabra & Jacobs, 2006)

Reinforcement Learning

- A general computational framework for learning through interacting with the environment
- Extended and developed in the machine learning literature into a full fledged framework for making sequences of actions and discovering optimal behavioral strategies in an unknown environment (i.e., Sutton & Barto and many others)
- Basic principal: start with the normative solution (Bellman equations) and to then implement principled approximations that can be computed online by a learning agent

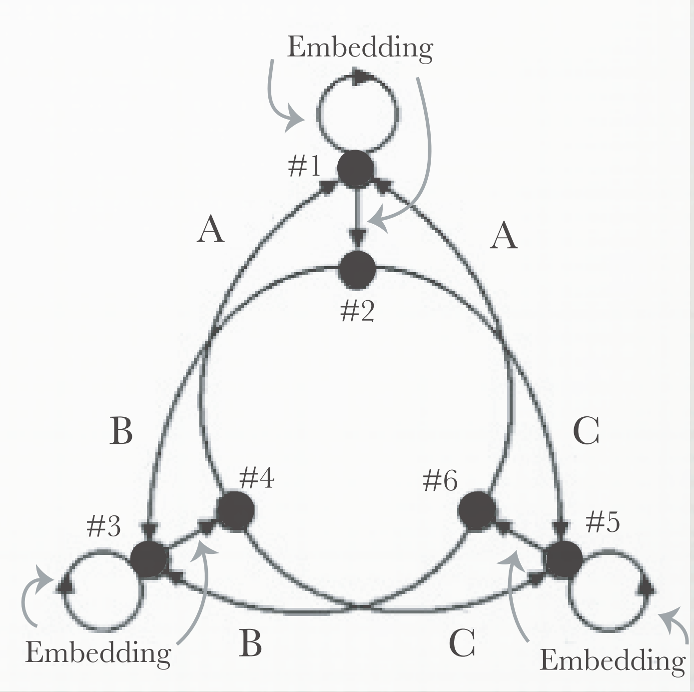
The Environment

- First useful approximation is to assume that the environment behaves according to a Markov Decision Process

- World is completely specified by the one-step dynamics

$$Pr\{r_{t+1} = r | a_t\}$$

$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t\}$$



The target

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

How to behave in an unknown world

- **Keep an ongoing estimate of the value of various "states"/"situations" under a certain policy (i.e., maintain an estimate of the value of taking action *a_t* in each state *s_t)***

- **Once you know how to evaluate a policy, there are a number of ways to actually arrive at (near)-optimal policies**

- **In many of the most interesting cases, it essentially reduces to something like:** ***start out exploring a lot, then slowly become more and more biased by the values you've experienced.***

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

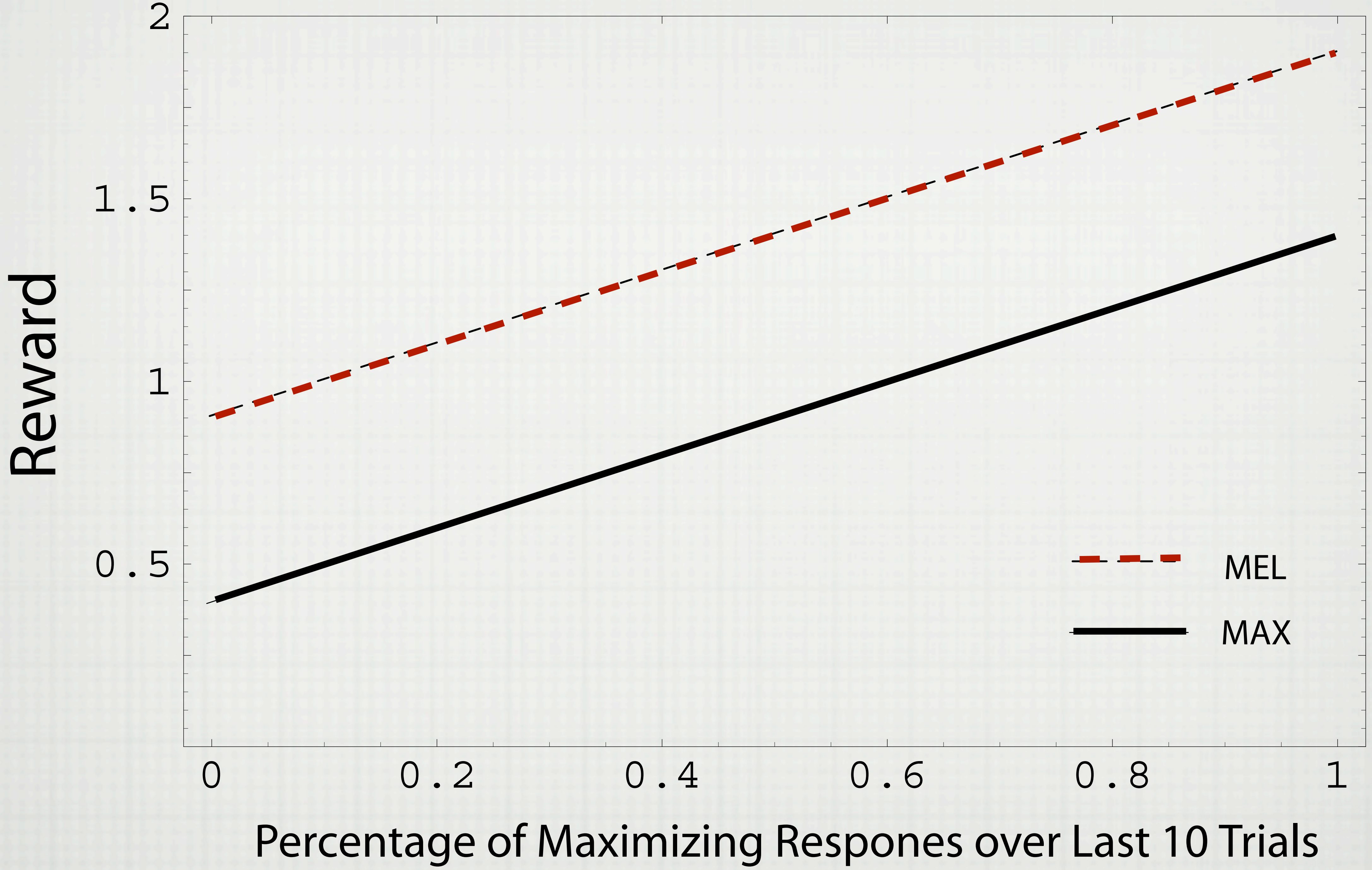
What are the states? How does the human known when actions have changed the state?

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

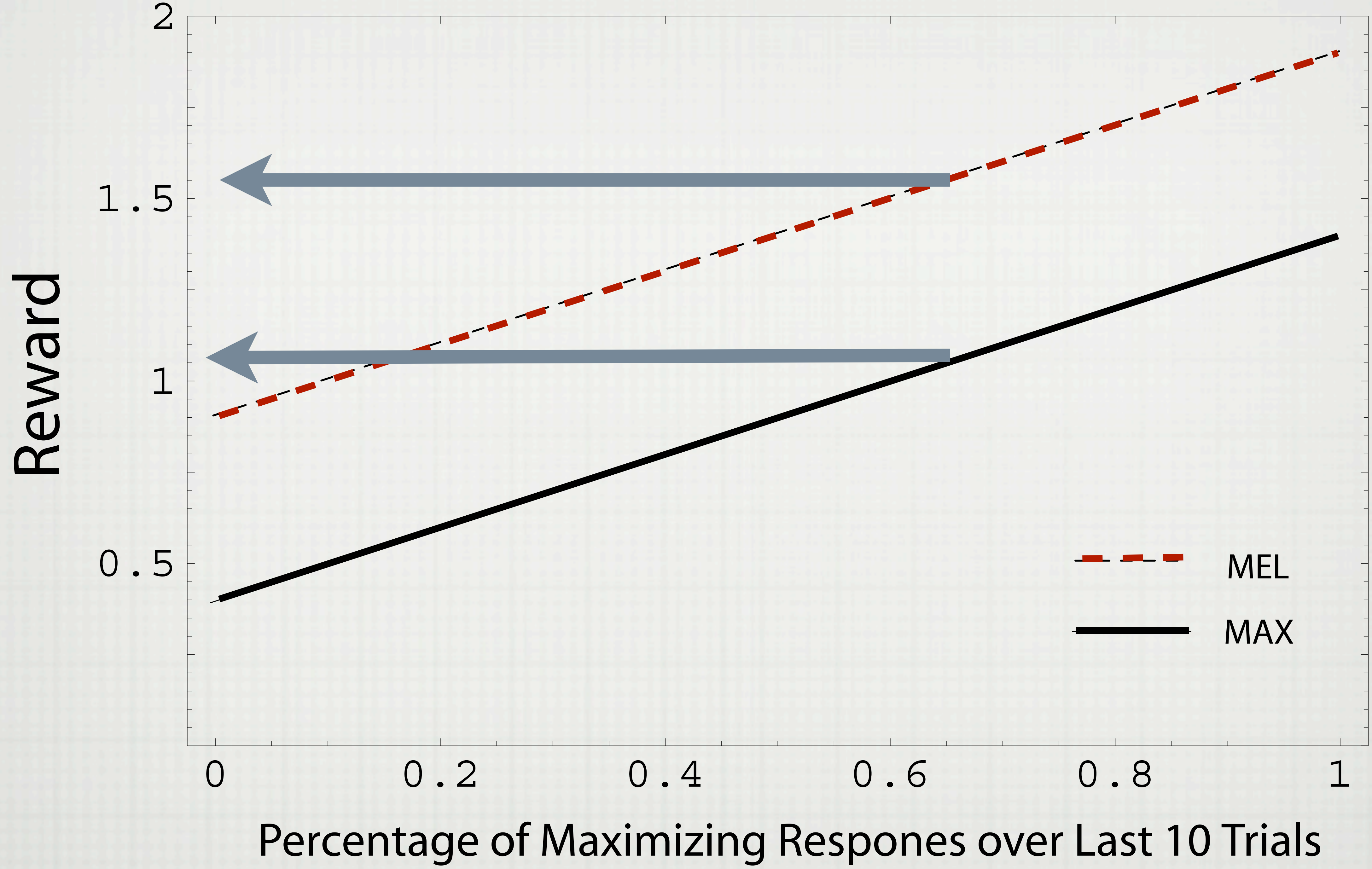
State Representations and the Problem of Perceptual Aliasing

(Whitehead & Ballard, 1991)

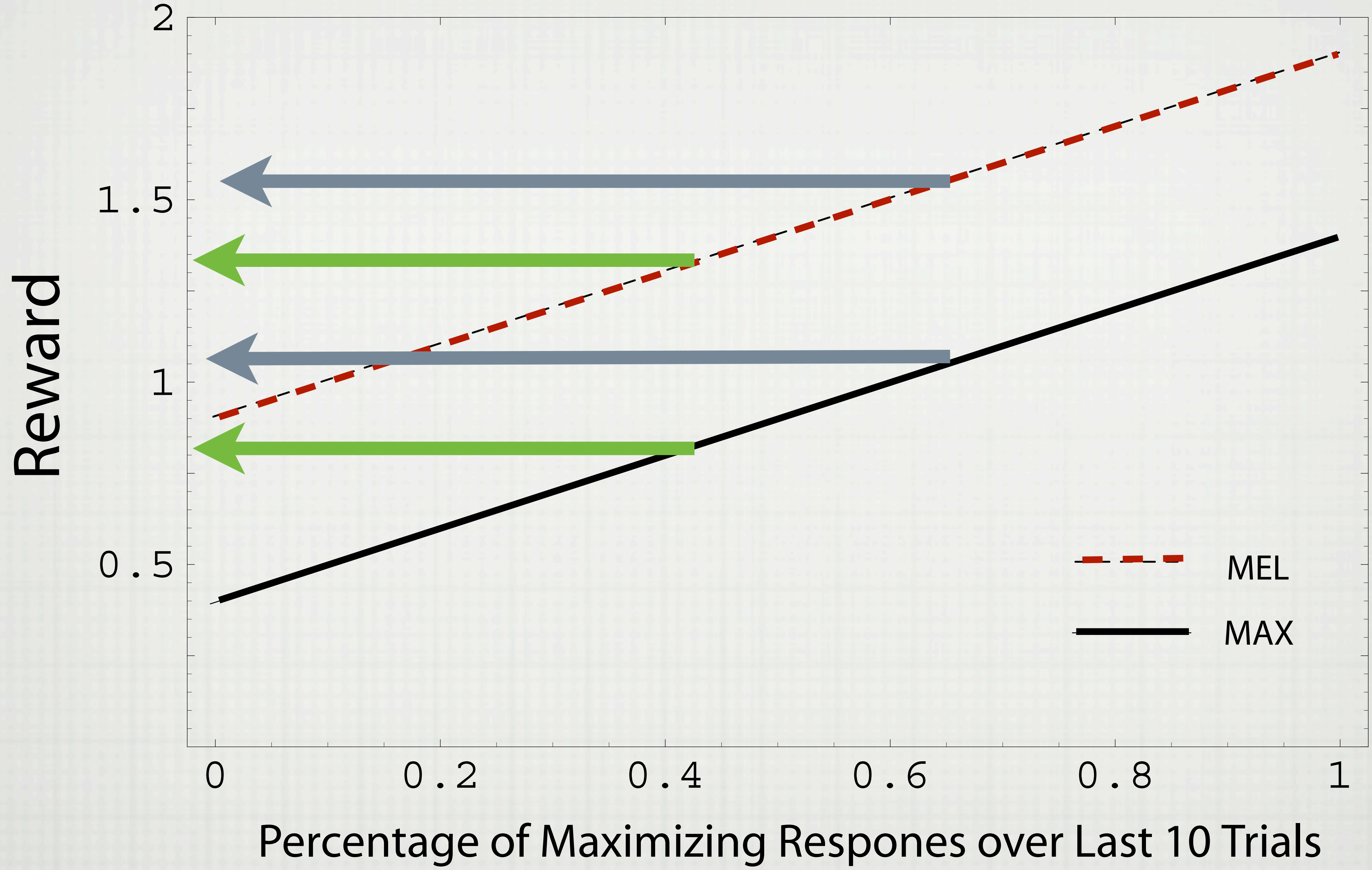
- The well known “secret” of most real-world artificial RL system is that considerable care has to go into constructing the state structure of the task
- When the state structure doesn’t match the world, or the agent adopts a bad representation performance likely suffers... all convergence bets are off
- However, how does this issue play out in human learning?



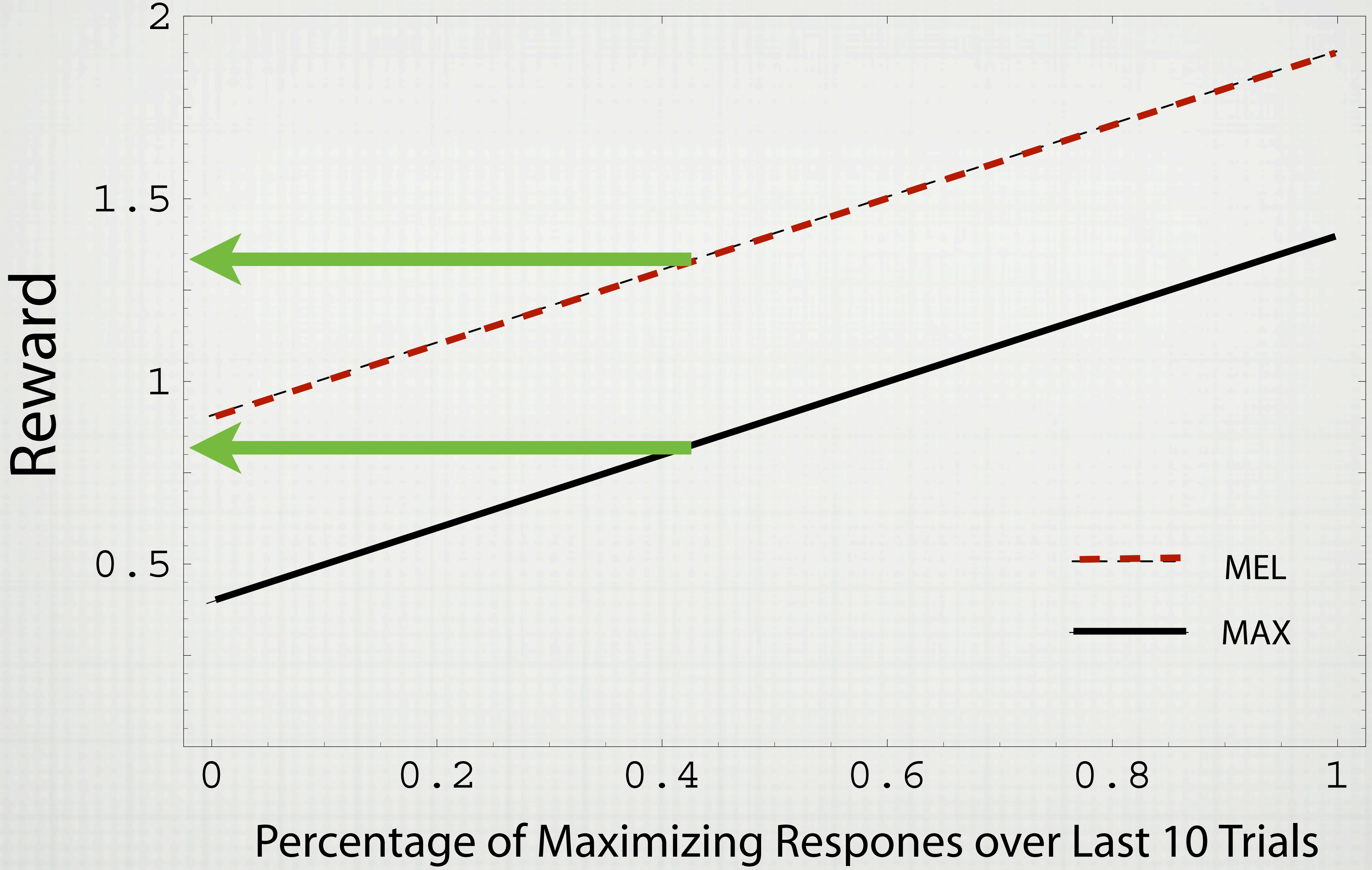
2
1.5
1
0.5
Reward
0
0.2
0.4
0.6
0.8
1
MEL
MAX
Percentage of Maximizing Respones over Last 10 Trials



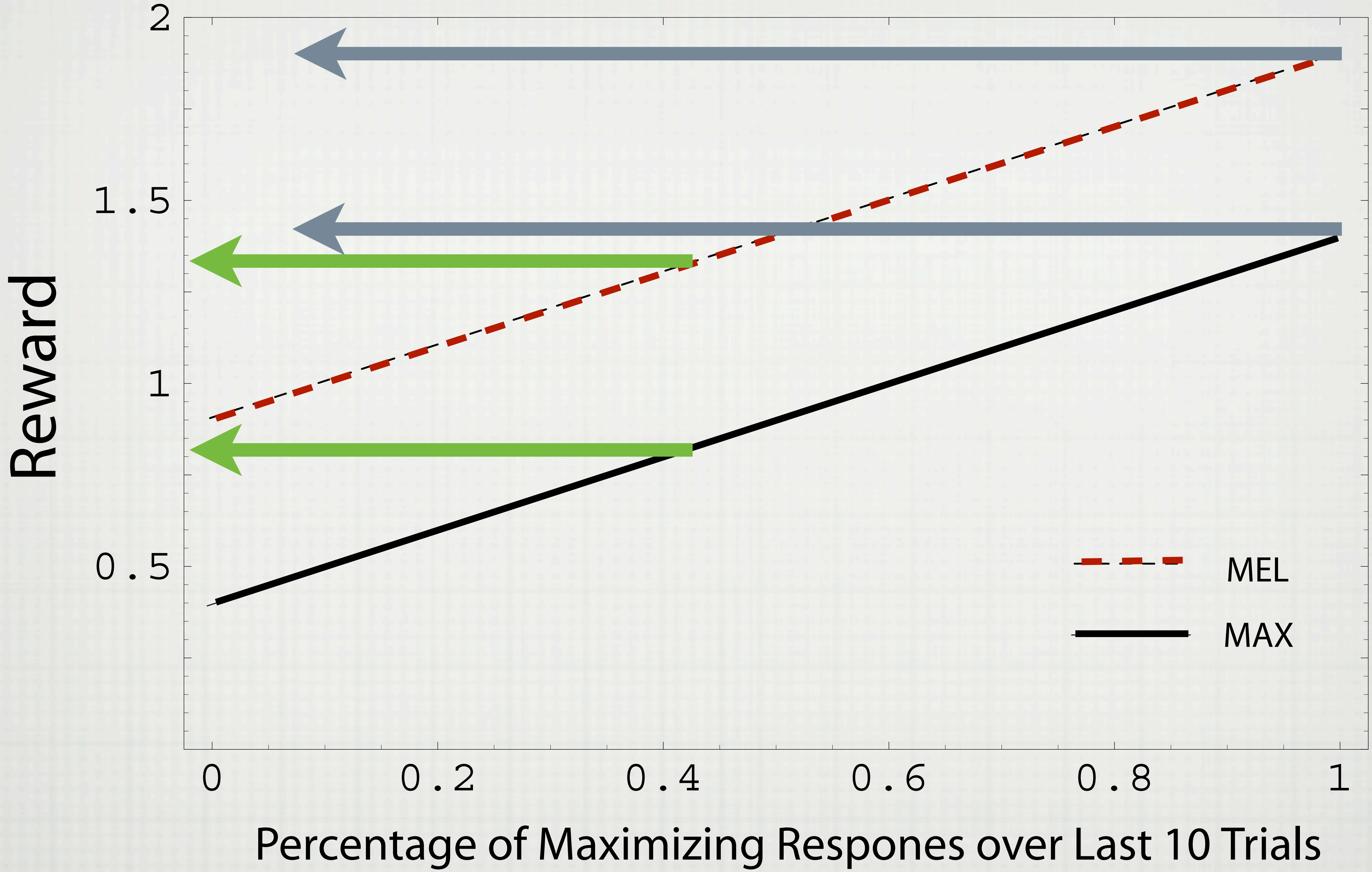
2
1.5
1
0.5
Reward
0
0.2
0.4
0.6
0.8
1
MEL
MAX
Percentage of Maximizing Respones over Last 10 Trials



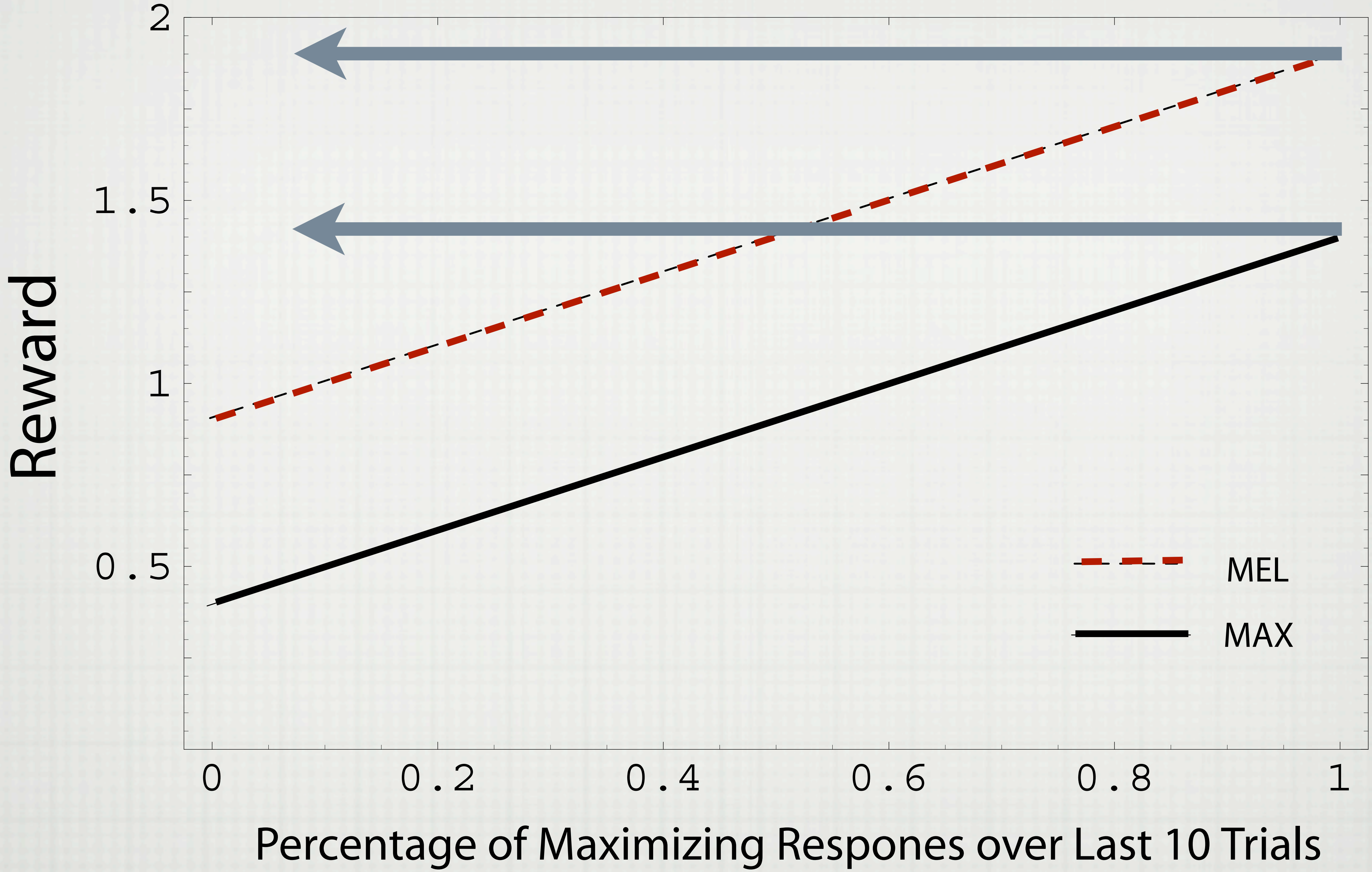
Reward
2
1.5
1
0.5
0
0.2
0.4
0.6
0.8
1
Percentage of Maximizing Respones over Last 10 Trials
MEL
MAX



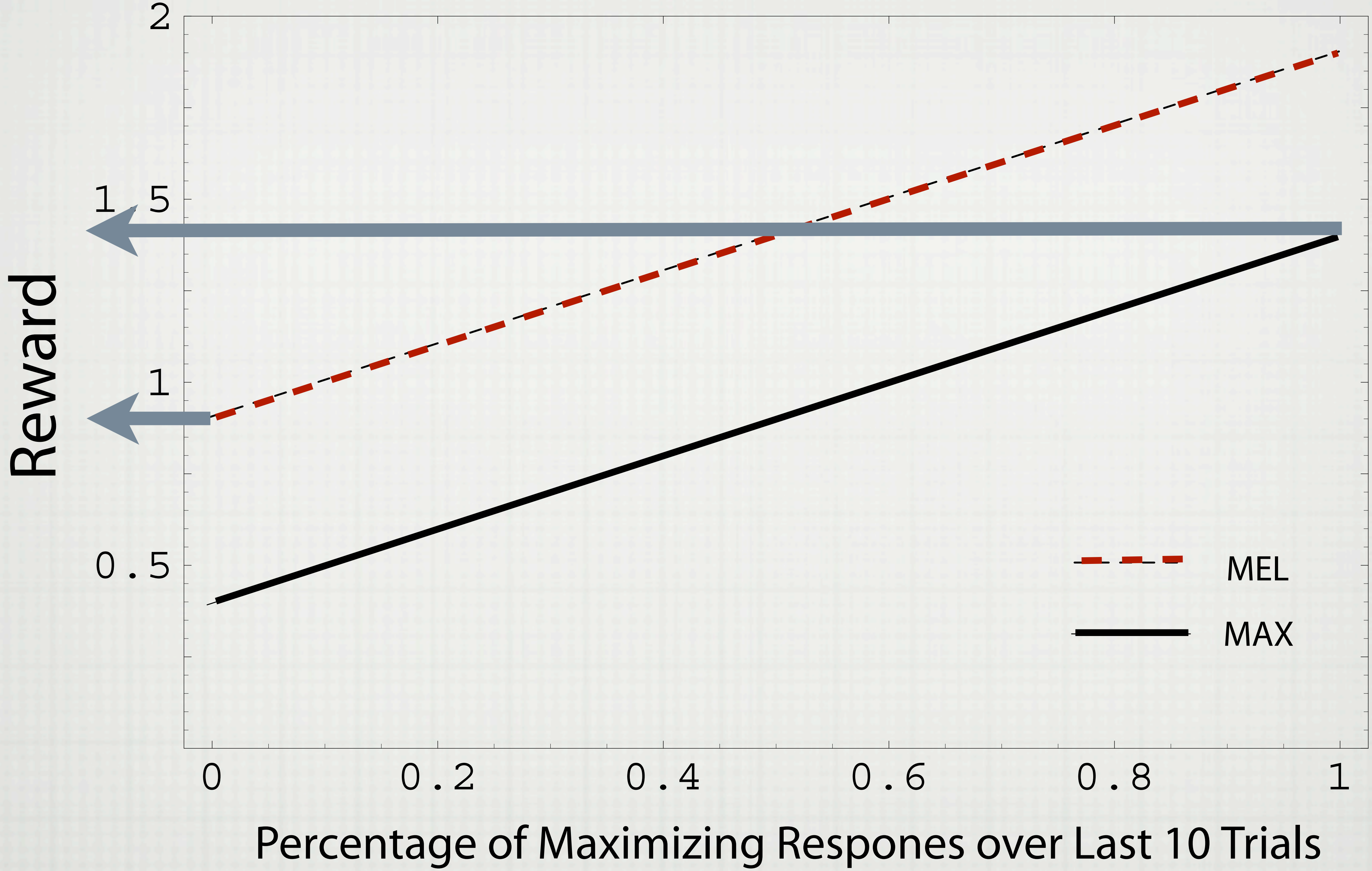
2
1.5
1
0.5
Reward
0
0.2
0.4
0.6
0.8
1
Percentage of Maximizing Respones over Last 10 Trials
MEL
MAX



Reward
2
1.5
1
0.5
0
0.2
0.4
0.6
0.8
1
Percentage of Maximizing Respones over Last 10 Trials
MEL
MAX



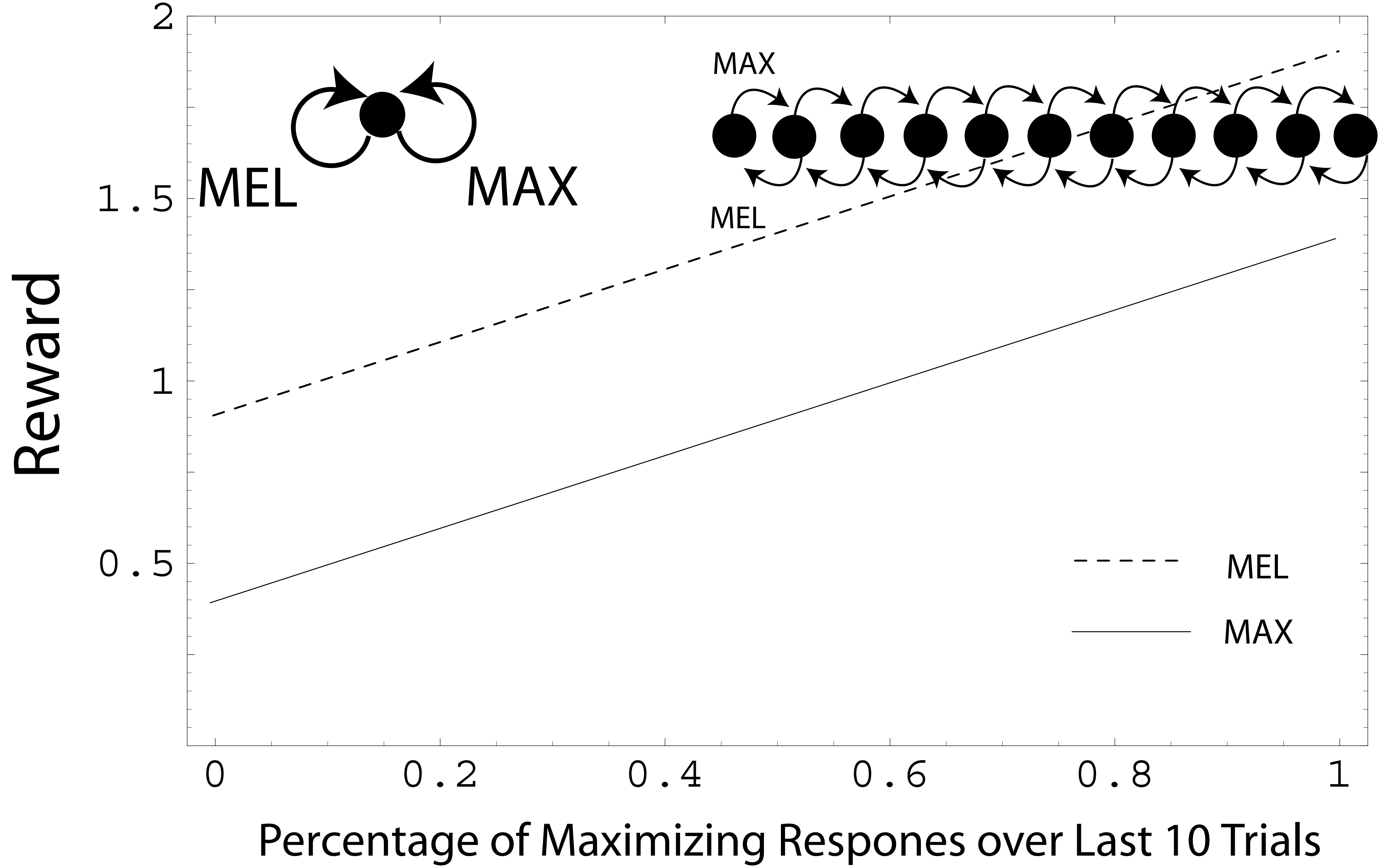
2
1.5
1
0.5
Reward
0
0.2
0.4
0.6
0.8
1
MEL
MAX
Percentage of Maximizing Respones over Last 10 Trials



2
1.5
1
0.5
Reward
0
0.2
0.4
0.6
0.8
1
Percentage of Maximizing Respones over Last 10 Trials
MEL
MAX

What makes this task difficult?

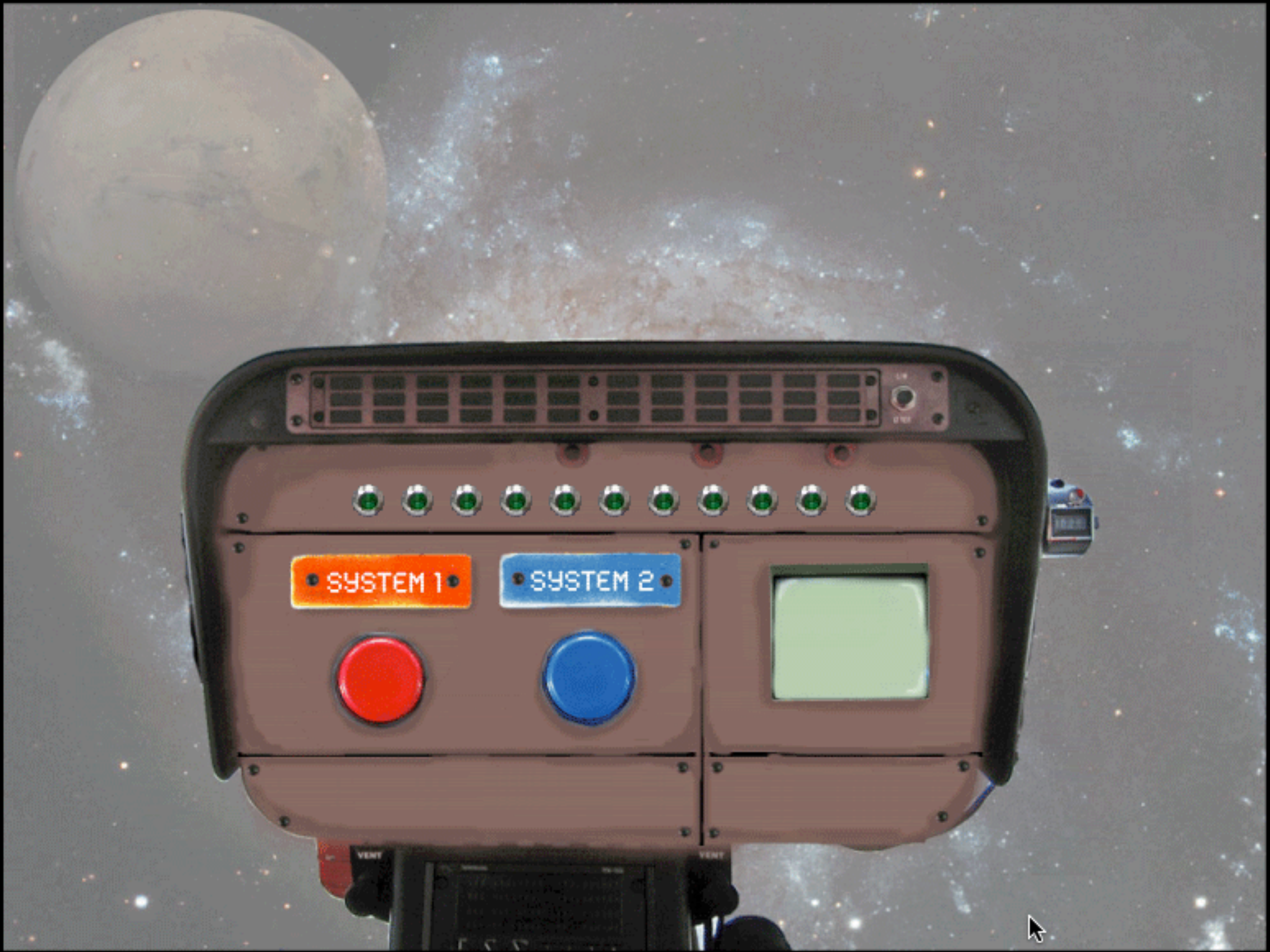
- The task is inherently non-stationary in that the values of particular actions drift over time in response to what you do.
- Difficulty may lie not only in the continuous memory demands or even appreciating future outcomes, but the fact that the relevant task states are "perceptually aliased"
- Prediction: By giving subjects "landmark" cue about the "states" the system transitions through, it expands out the problem in a way that allows people to associate experienced rewards with particular states of the system
- Not to be confused with memory aid (although maybe related): state information just allows you to associate the value of being in a particular place with the reward you get while you are there



2
1.5
1
0.5
0
0.2
0.4
0.6
0.8
1
Reward
Percentage of Maximizing Respones over Last 10 Trials
MEL
MAX
MAX
MEL
MEL
MAX


WELCOME TO THE N.A.S.A. MARS FARMING PROJECT

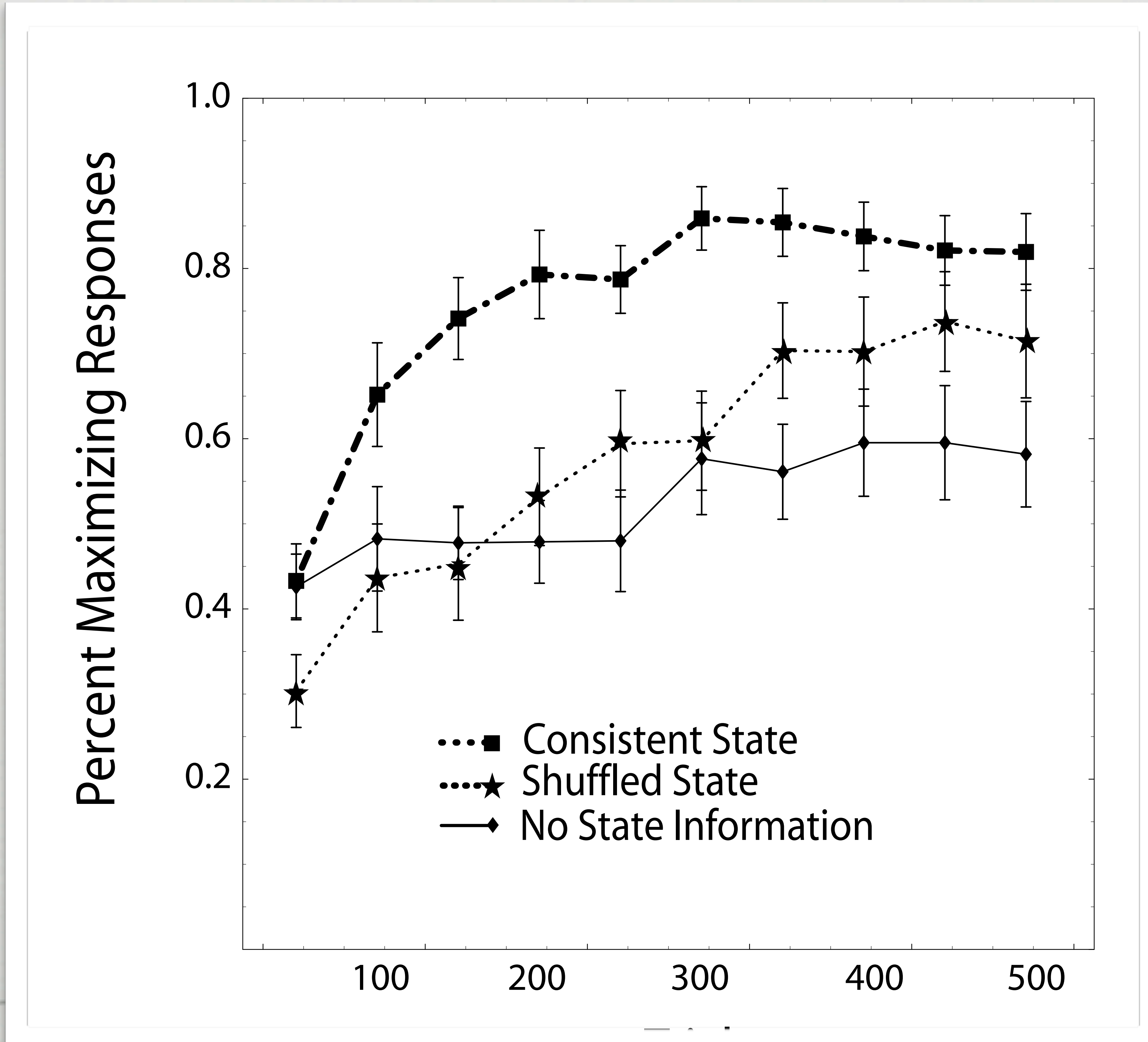

SYSTEM 1
SYSTEM 2
Choose


SYSTEM 1
SYSTEM 2

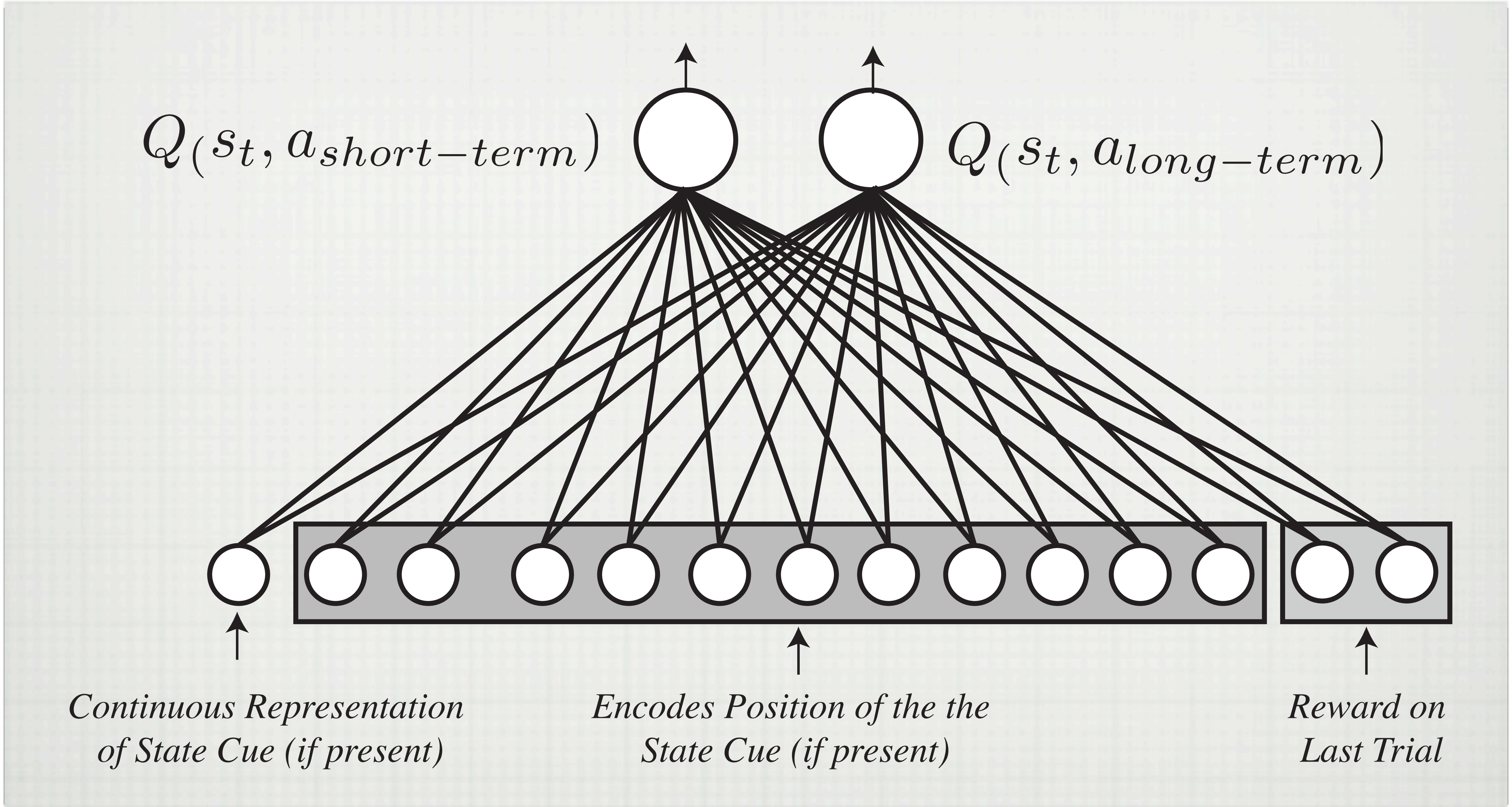
Experiment 1

Gureckis & Love, in press

No State	Shuffled Cue	Consistent Cue
No additional information provided besides rewards on the screen	Mapping between “latent” task states and display was unique but transitions were randomize per subject	Mapping between “latent” task states and display unique, but also moved in orderly direction


Percent Maximizing Responses
1.0
0.8
0.6
0.4
0.2
100
200
300
400
500
Consistent State
Shuffled State
No State Information

State Representation



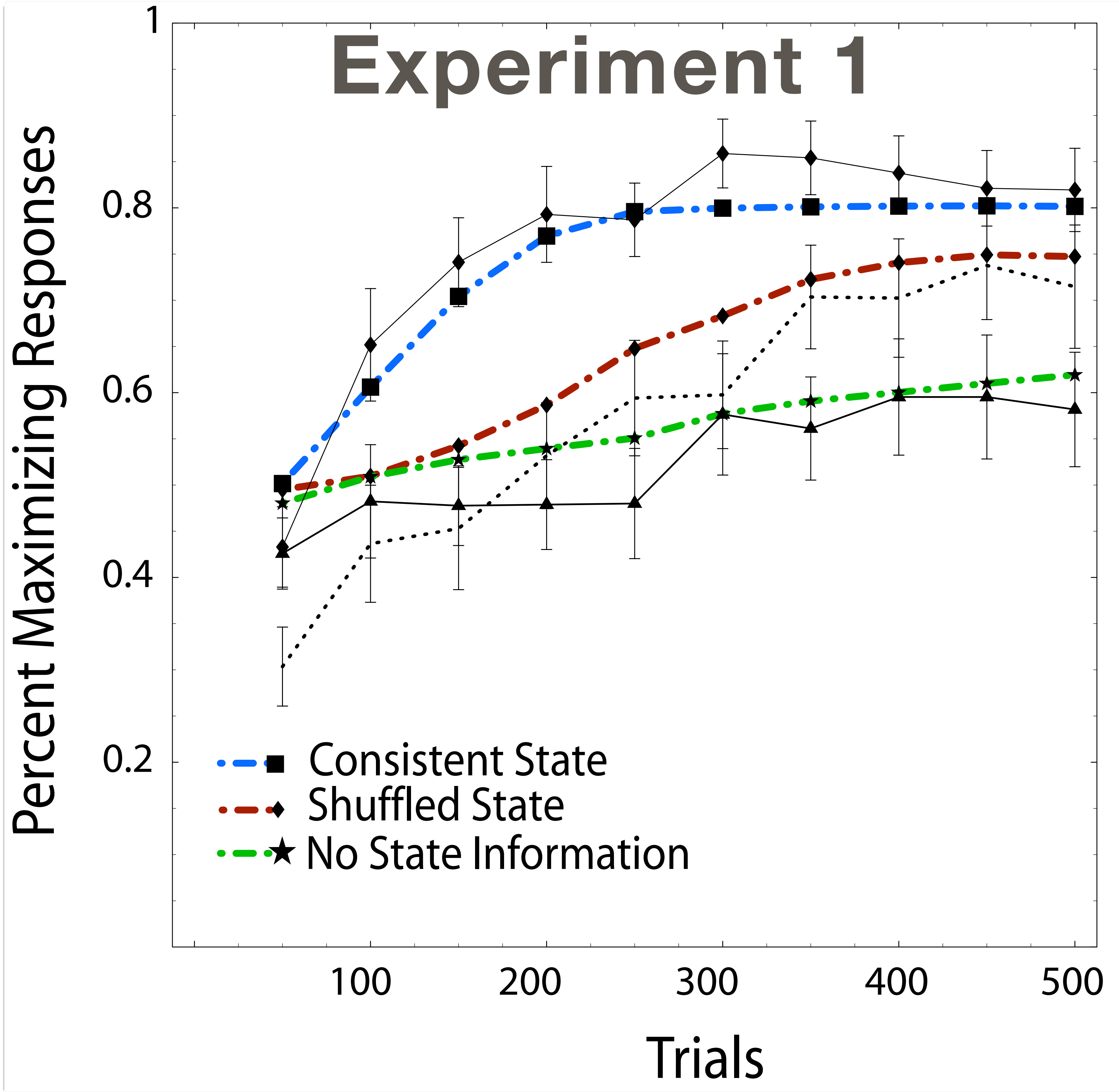
Modeling Analyses

Error Term During Learning

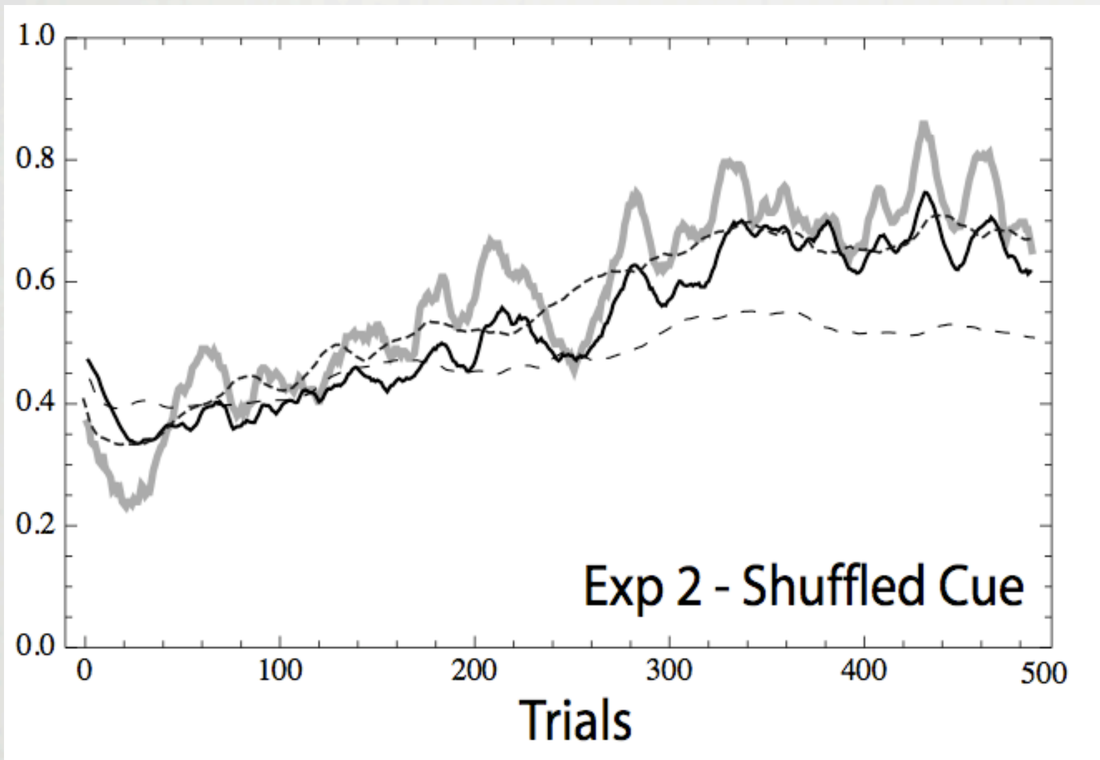
$$\delta = [r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

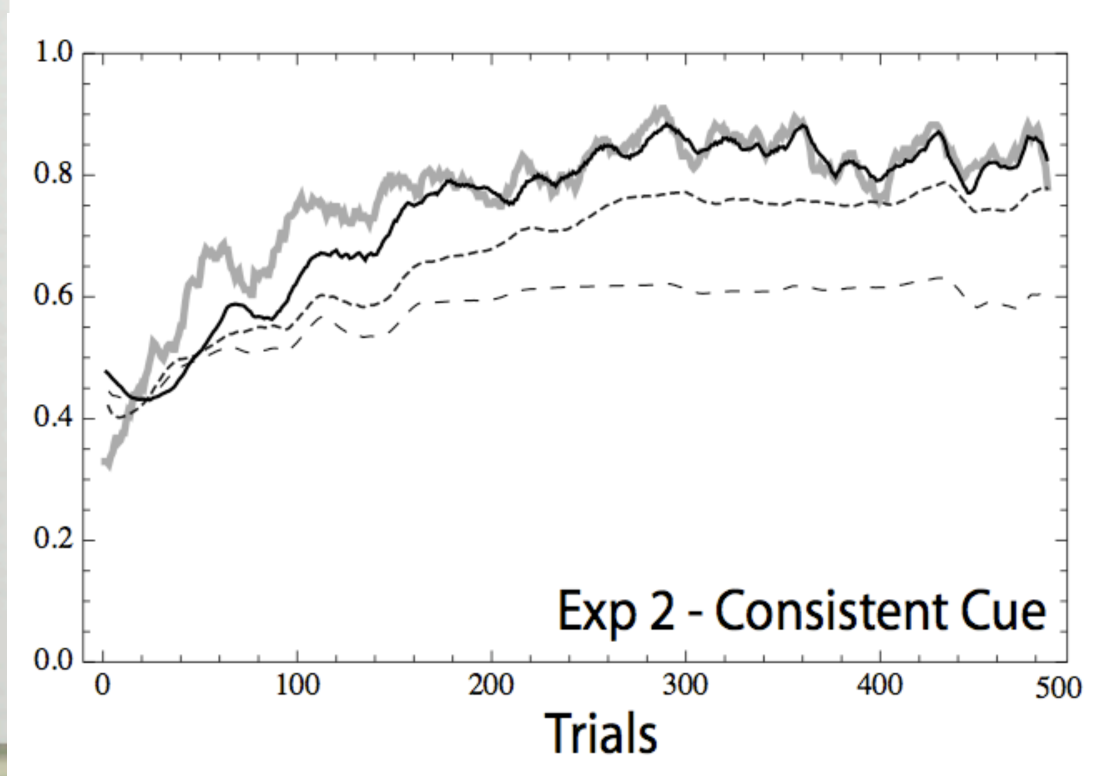
Choice

$$P(a_i) = \frac{e^{Q(s_t, a_i) \cdot \tau}}{\sum_{j=1}^{2} e^{Q(s_t, a_j) \cdot \tau}}$$

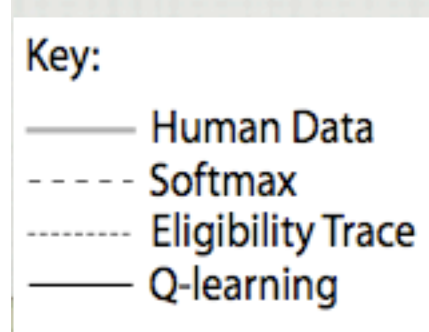

Experiment 1
Percent Maximizing Responses
1
0.8
0.6
0.4
0.2
100
200
300
400
500
Trials
Consistent State
Shuffled State
No State Information

Are cues simply memory for recent actions?



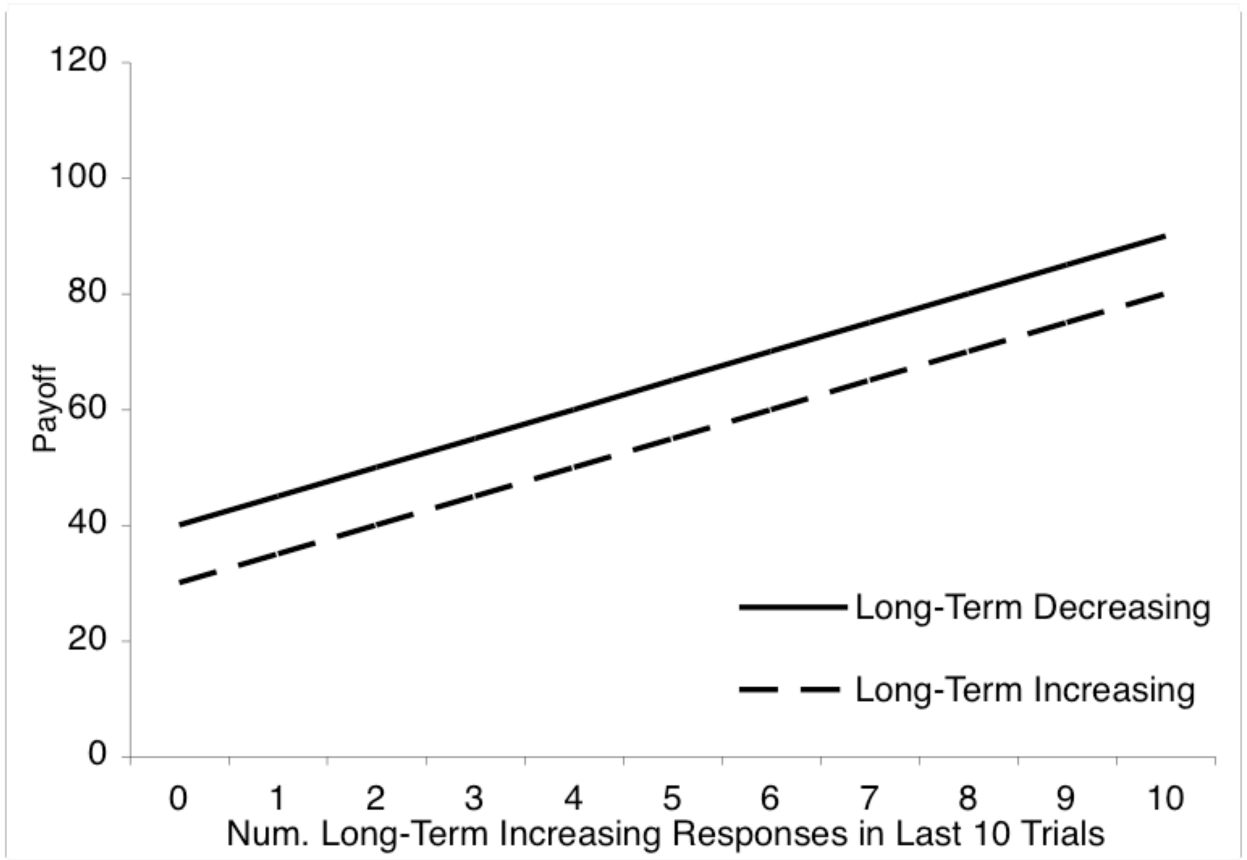


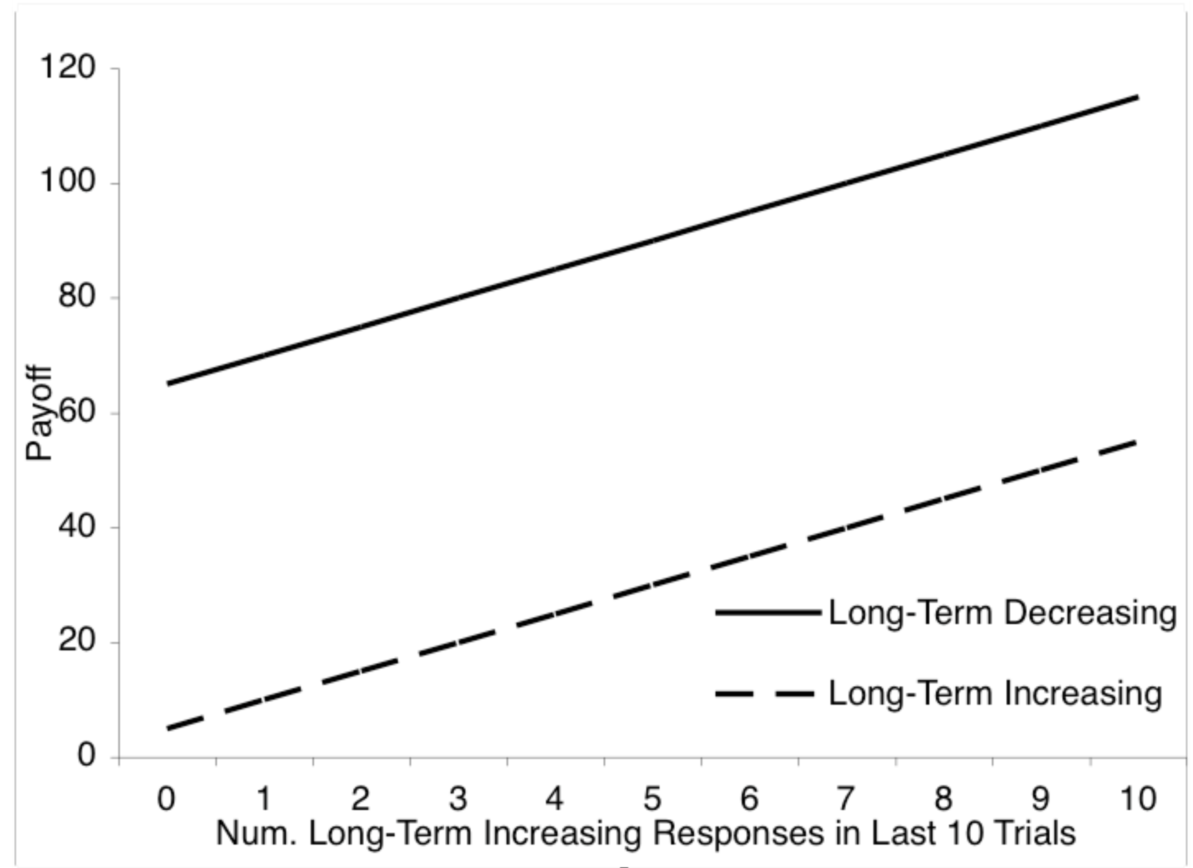
- Memory for recent actions/ observation can often help disambiguate current state, (c.f., McCallum, 1993)
- Tested a model *not* based on look ahead RL methods but using eligibility traces (Bogacz, et al., 2007)
- Overall, eligibility trace model under predicts performance in the task, owing to the generalization afforded by the function approximation scheme



Ok, but it can't all be good, right?

- State cues structure learning by helping participant disentangle the relative value of particular actions as a function of their state
- In addition, we find that models which allow generalization/ extrapolation of the experienced value from one state to another provides a good account of what people are doing.
- However, if this is true, then it may be possible to trip people up...





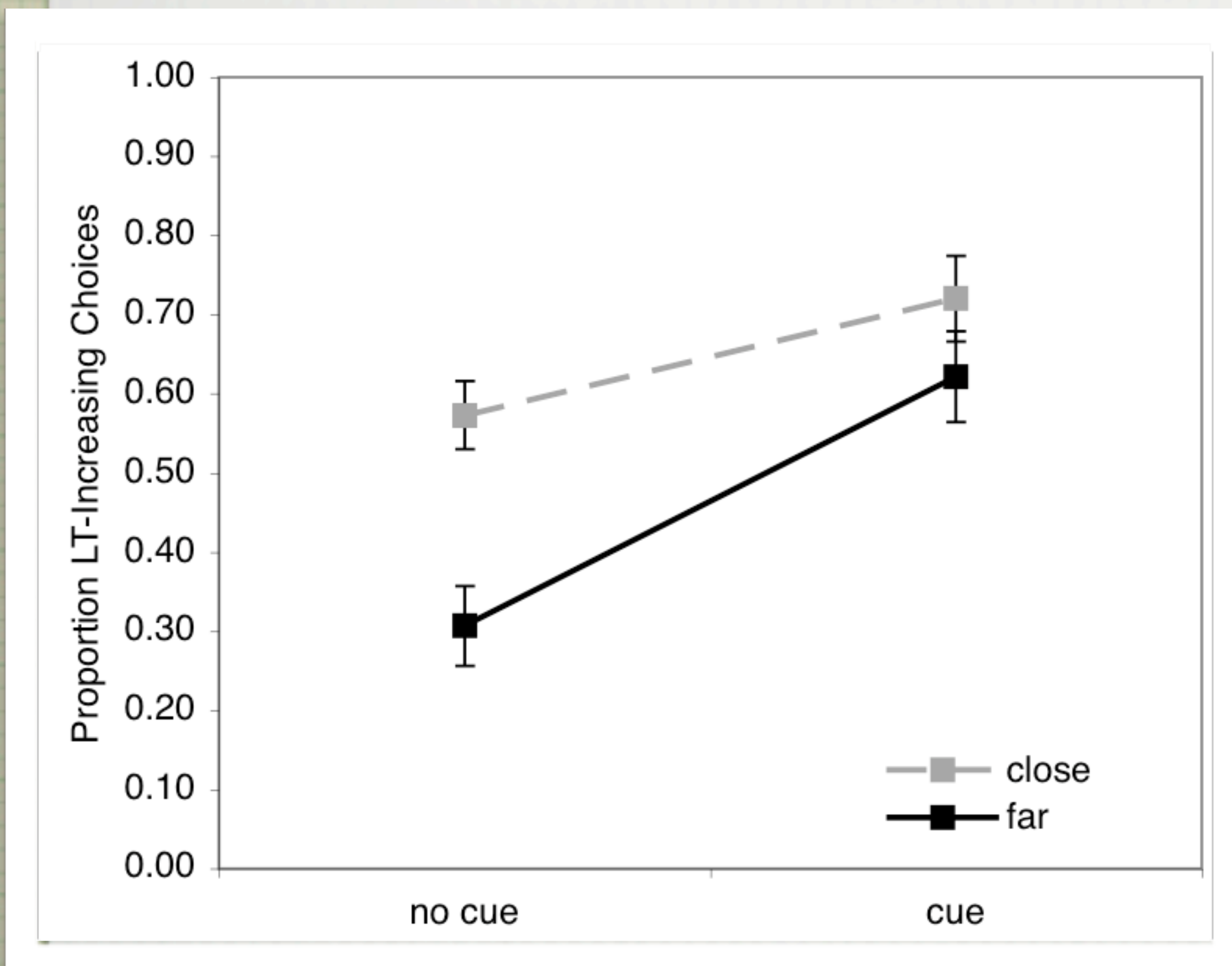
“Close” Curves

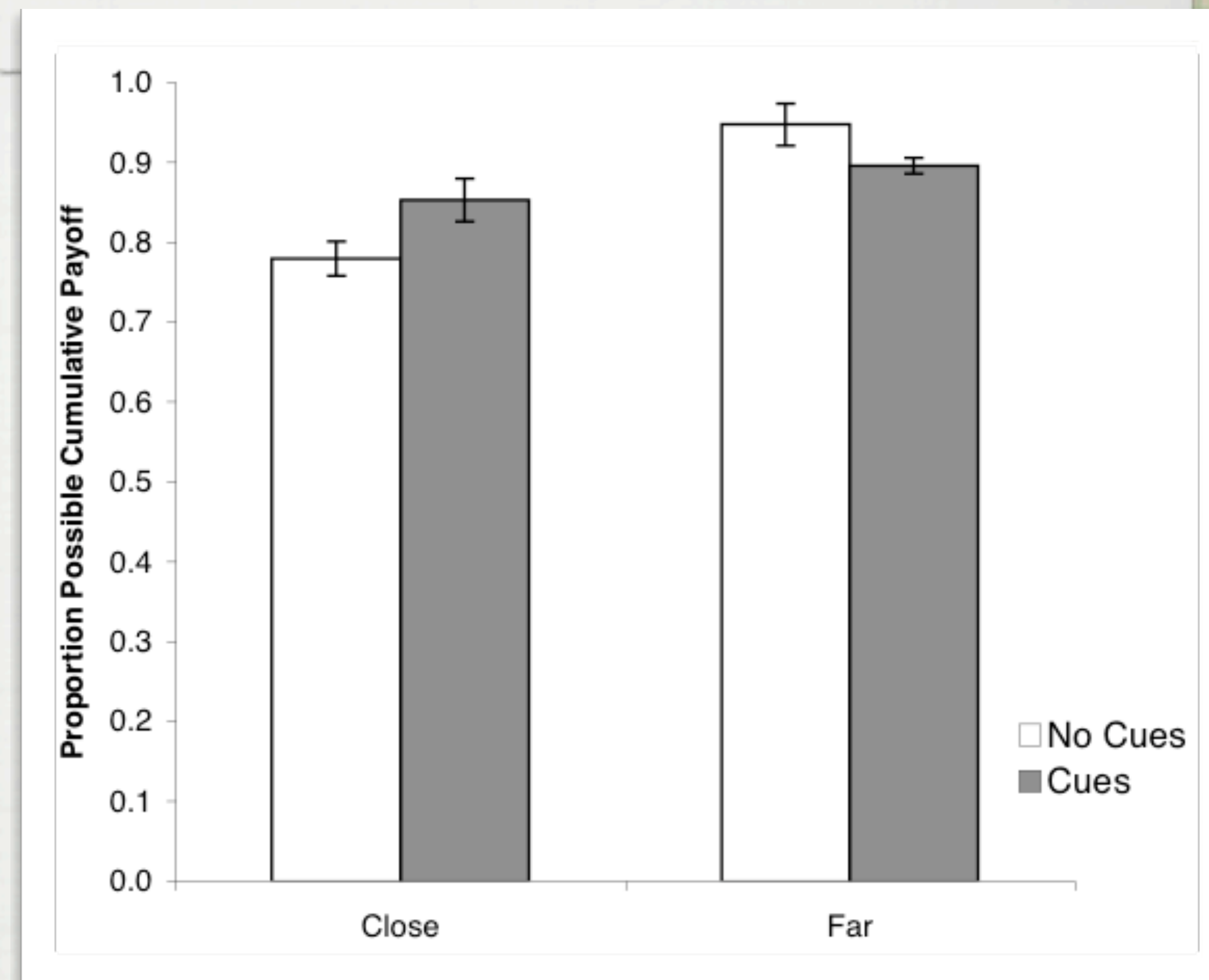
Far Apart

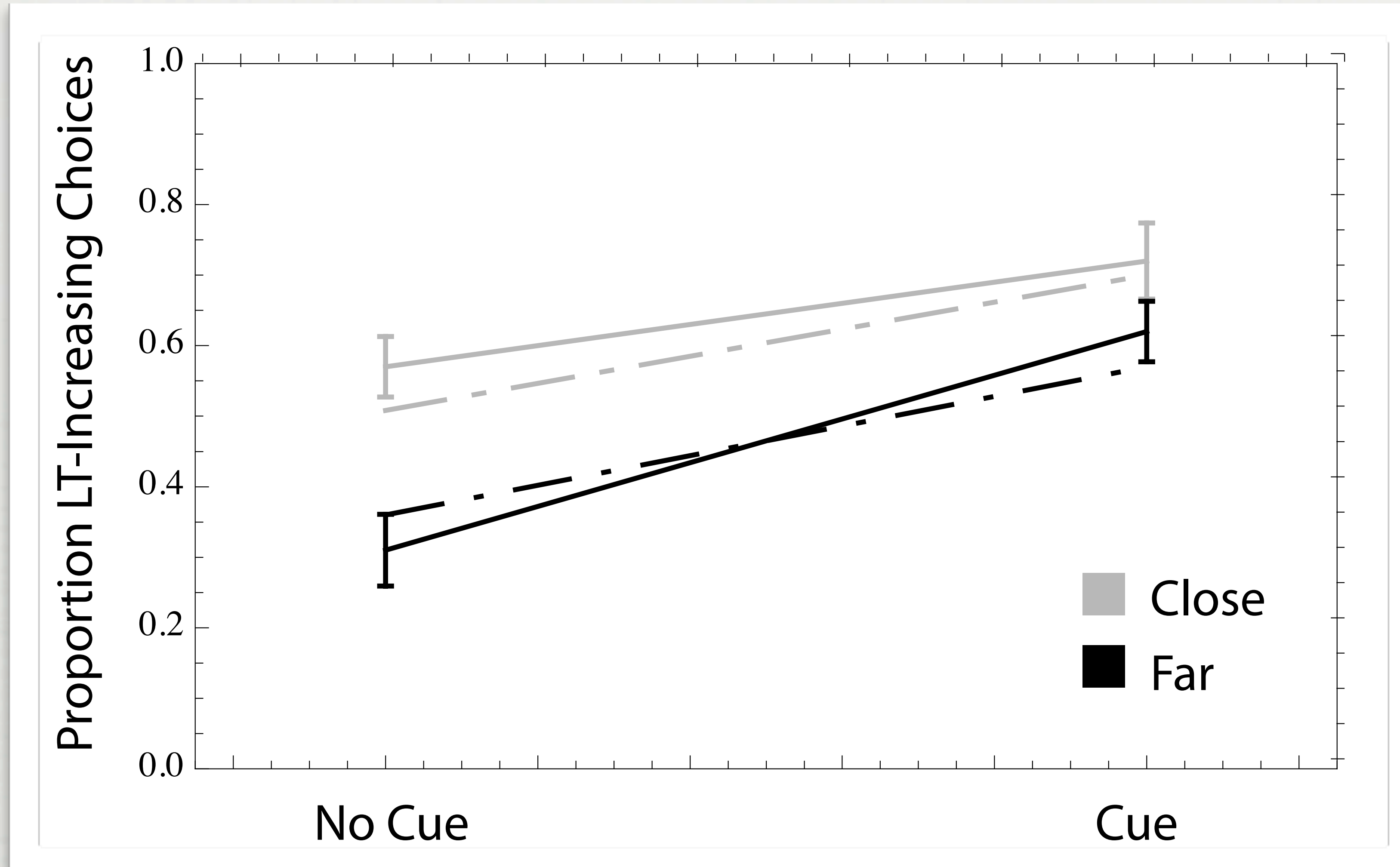
Experiment 2

Otto, Gureckis, Markman, Love, under review

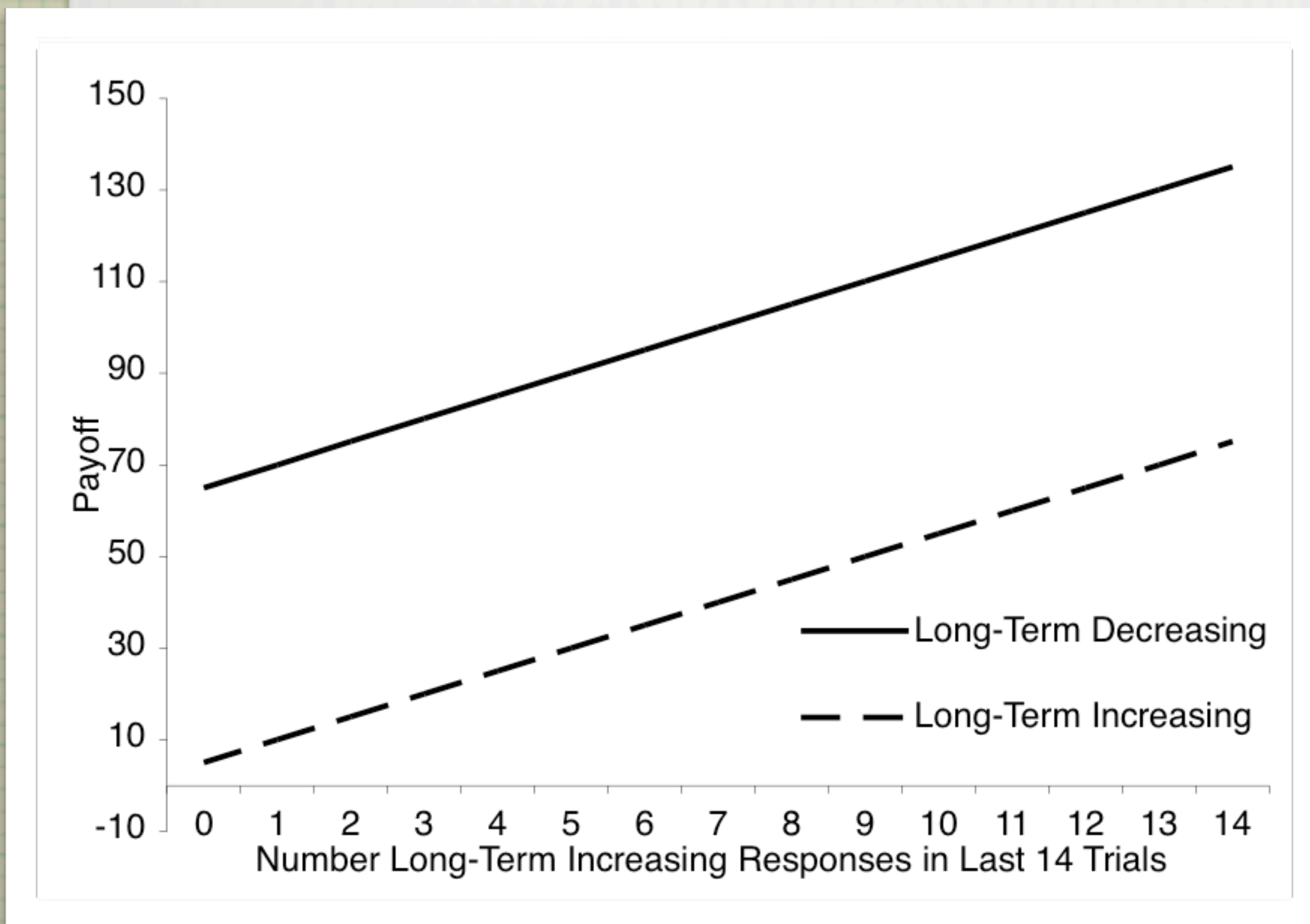
Close - No Cue	Close - Cues	Far - No Cue	Far - Cues
Close curves (optimal is the LT increasing option)	<--- Same but with state cues	Optimal is actually choosing the other option (no conflict)	<---- Same but with state cues

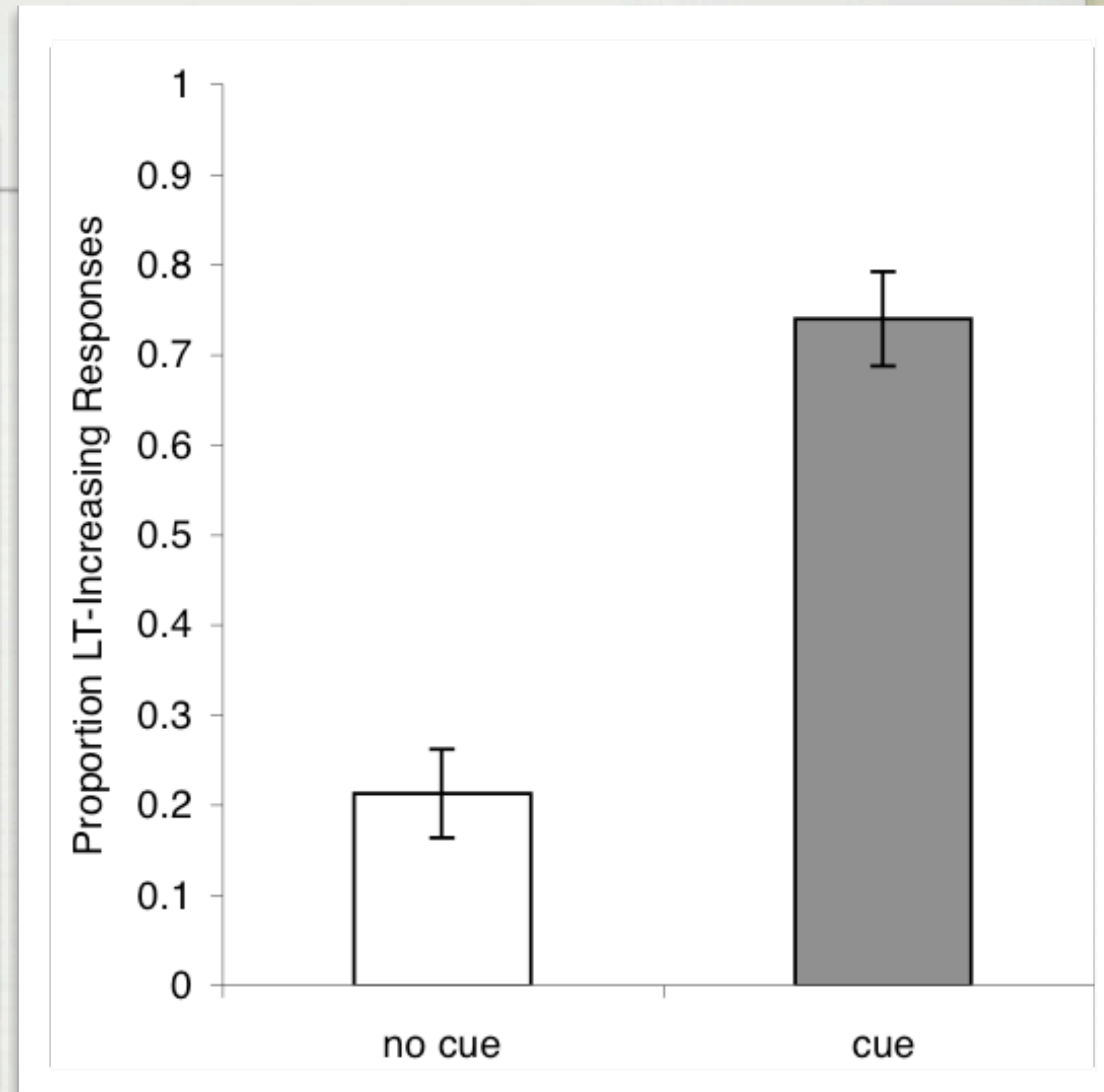

Proportion LT-Increasing Choices
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
no cue
cue
close
far


Proportion Possible Cumulative Payoff
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Close
Far
No Cues
Cues



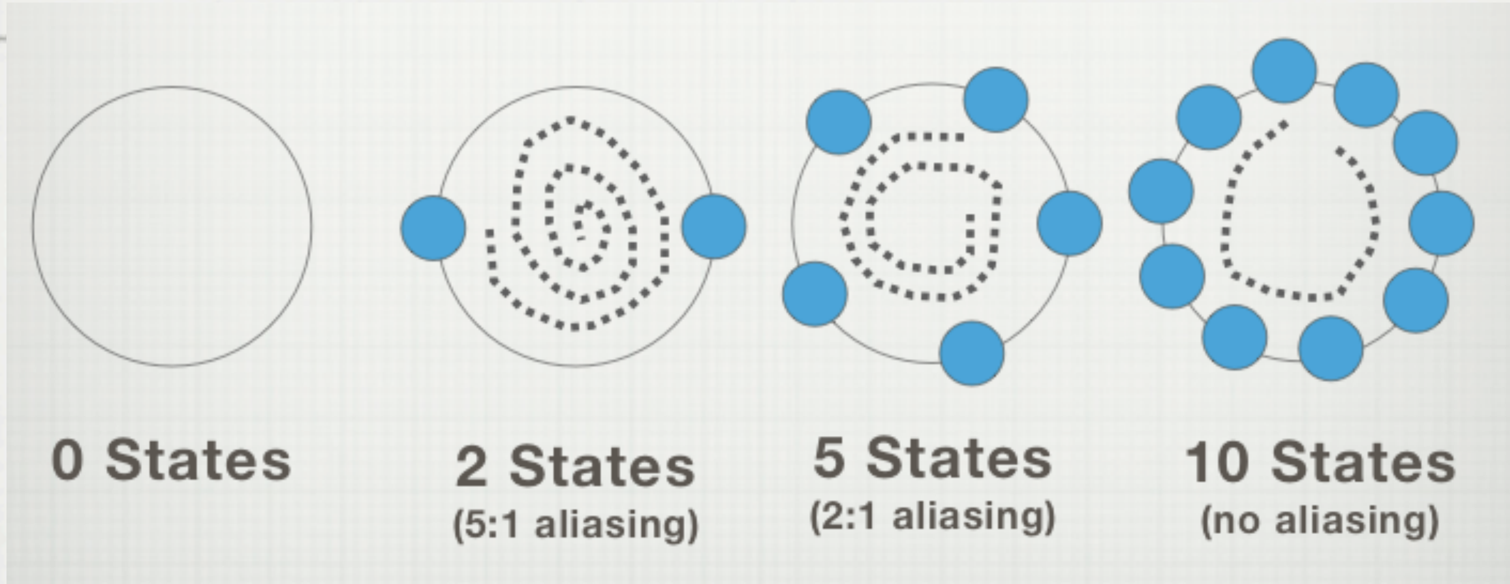
** Dashed lines are model fits with single set of parameters across all four conditions

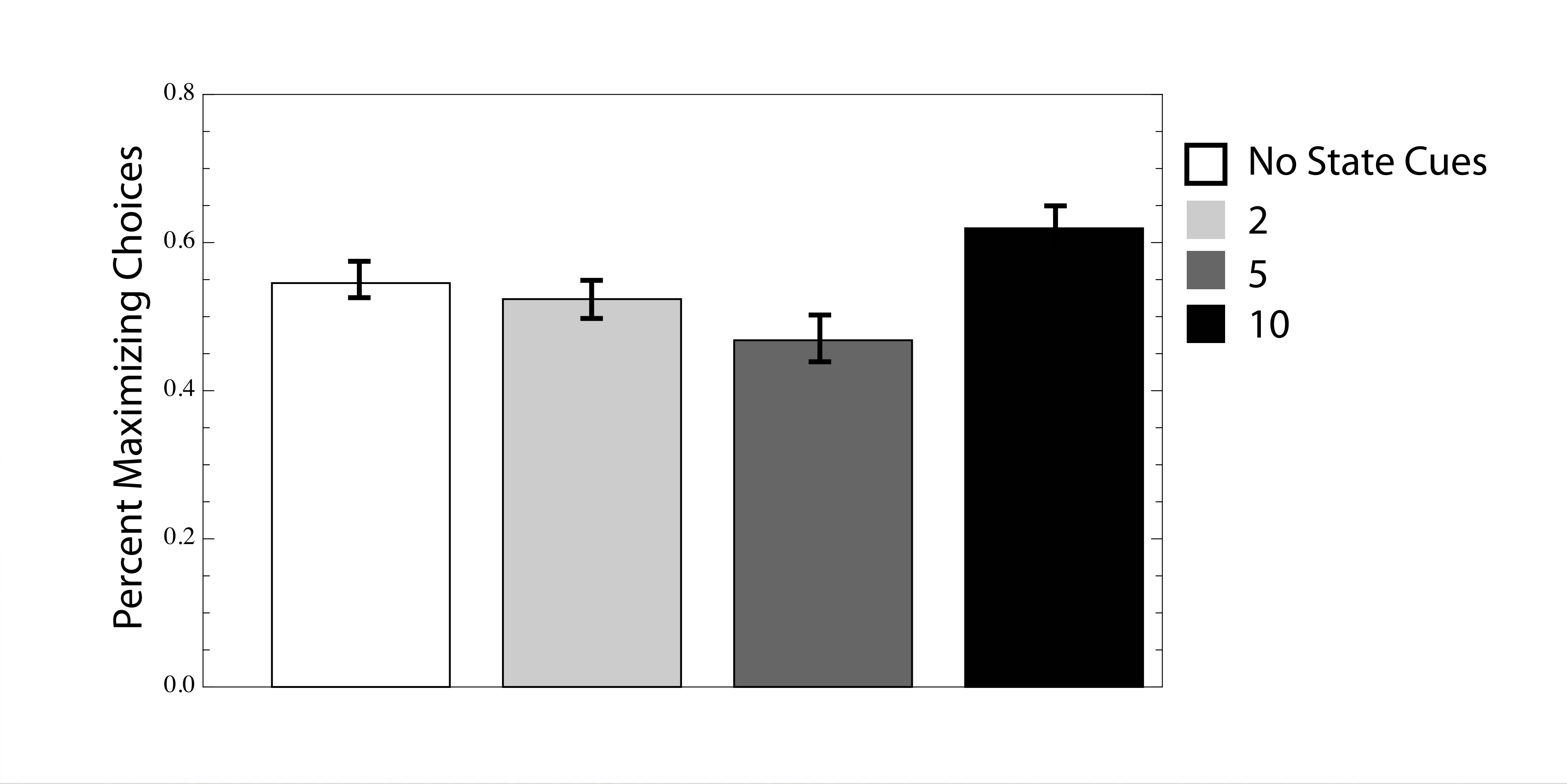




Add more states to the far apart task so the curves cross over again

Perceptual Aliasing and State "Blending"





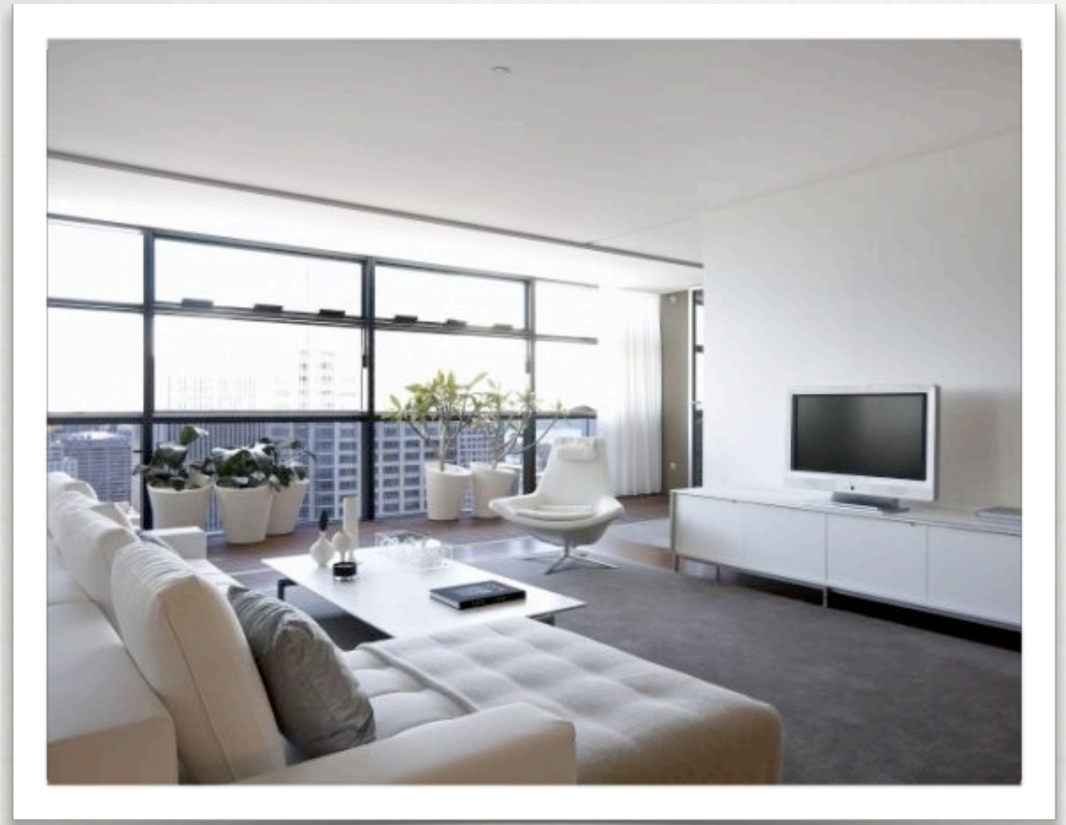
Future Directions and Challenges

- Instead of *giving* people the representation of the task, see how they learn it directly from experience

- State representation as categorization: how learning is integrated with respect to these representations (Redish, et al., 2007; Veksler, Gray, & Schoelles, 2007)

- Use more sophisticated machine learning methods (such as those from this morning) to infer *latent* state representations to figure out how prior knowledge and experience interact in dynamic tasks

Take Home Message

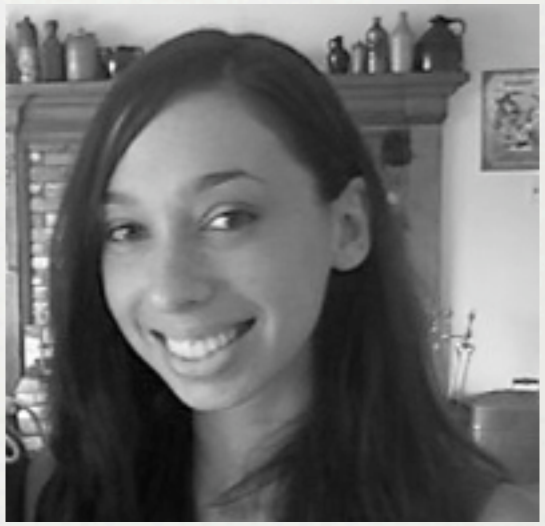


- Just as in artificial learning systems, the state structure the learner adopts, or is given, strong limits performance

- Some decision making problems with valuing short/long-term rewards may reflect bad representations of the task environment rather than poor update/ valuations

- **Keep the couch!** Another example of how machine learning concepts and frameworks can lead to further insights about human learning!

Thanks!



BRAD LOVE
@ TEXAS

ROSS OTTO
@ TEXAS

LISA ZAVAL
@ NYU