

Human and Optimal Exploration and Exploitation in Bandit Problems

Shunan Zhang, Michael D. Lee, Miles Munro, University of California, Irvine



UCIRVINE

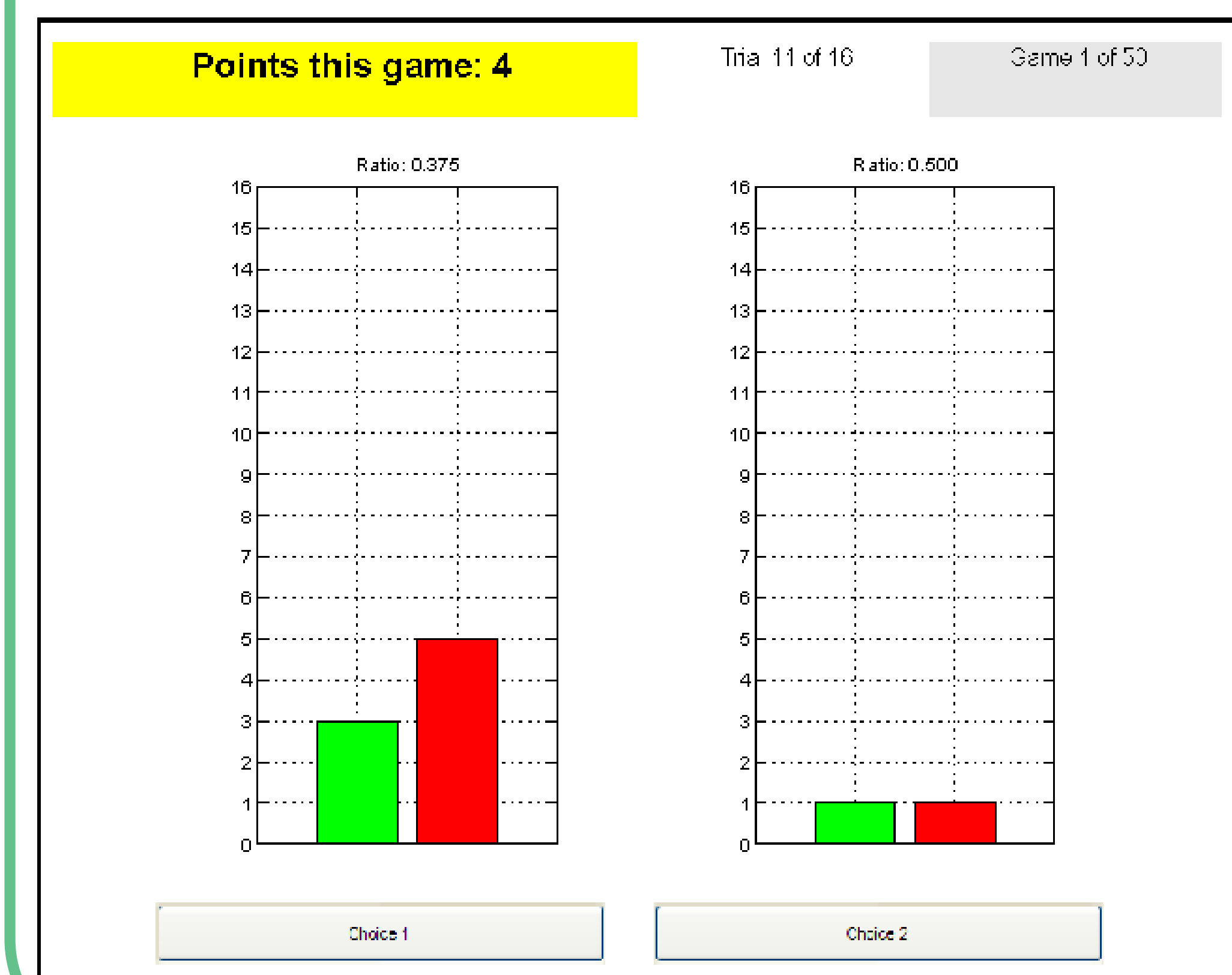
Bandit Problems

A decision-maker must choose between a set of alternatives – each of which has a fixed but unknown rate of reward – to maximize their total number of rewards over a short sequence of trials.

Requires balancing the need to search for highly-rewarding alternatives (exploration) with the need to capitalize on those alternatives already known to be reasonably good (exploitation).

Interested to know if people switch between exploration and exploitation in short-horizon bandit problems, where the changing rewards rates are no longer a confounding concern, as a contrast to when people have long-horizon bandit problems.

User Interface



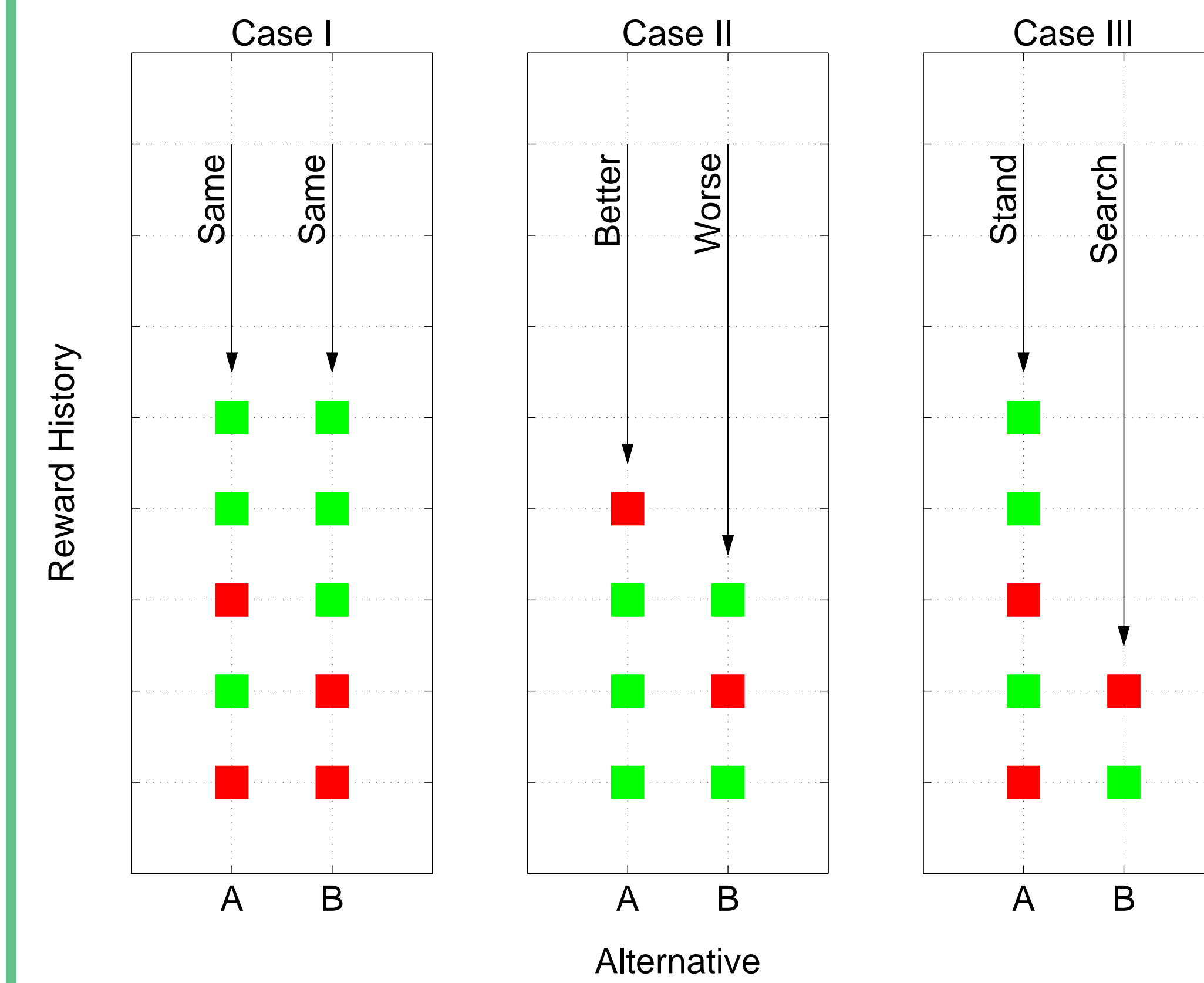
ϵ -first Model

ϵ -first model (Sutton and Barto, 1988) assumes two distinct stages in bandit problems.

- Exploration: alternatives are chosen at random
- Exploitation: the alternative with the best observed ratio of successes to failures is chosen

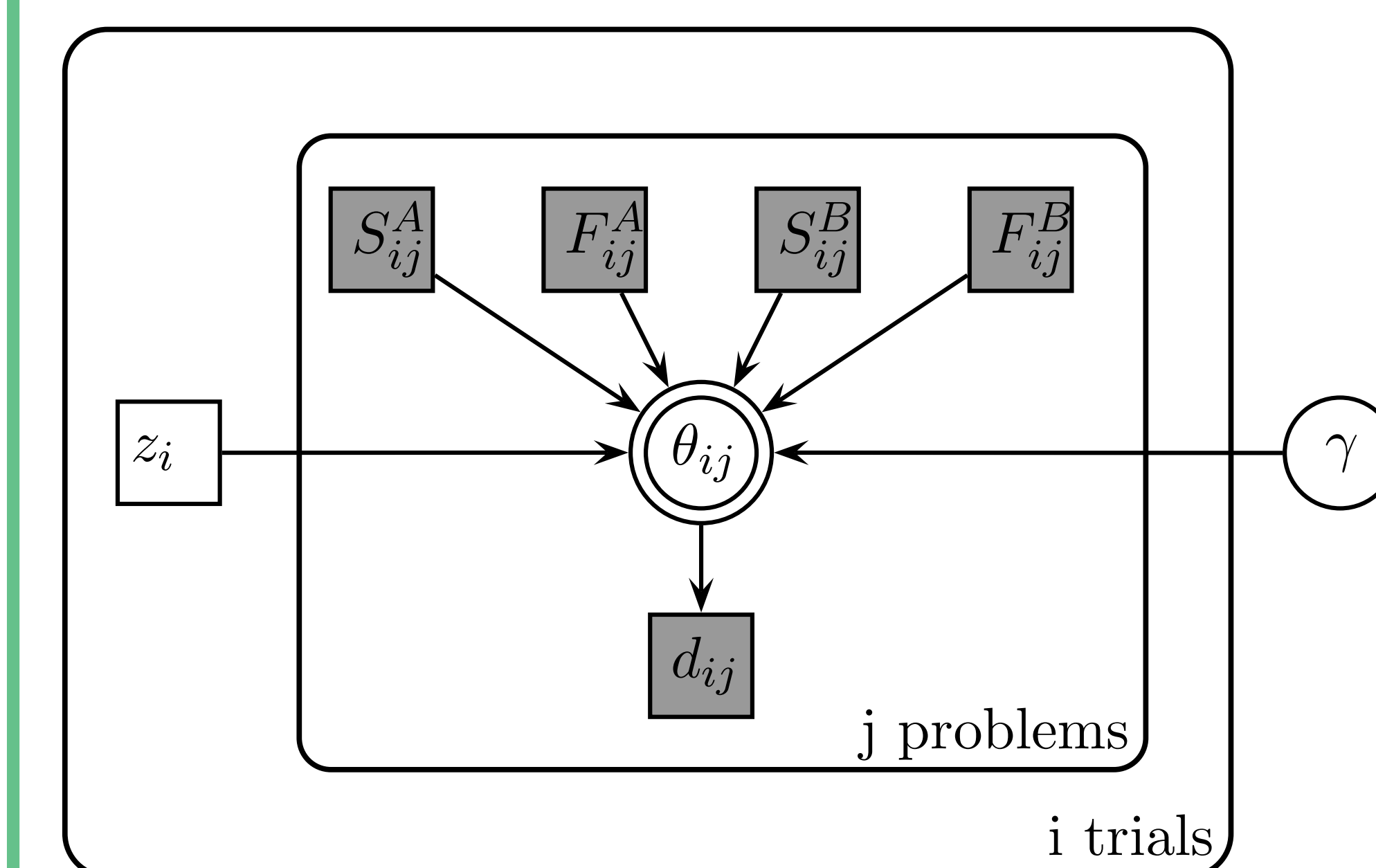
Demarcation between stages is determined by a parameter which corresponds to the trial at which exploration stops and exploitation starts.

Our Model



- The *Same*: same number of observed successes and failures for both alternatives.
- The *Better-Worse*: one has more successes and fewer failures than the other alternative.
- The *Search-Stand*: one has more successes but also more failures than the other.

Graphical Model Implementation



$$\theta_{ij} = \begin{cases} 1/2 & \text{if A is same} \\ \gamma & \text{if A is better} \\ 1 - \gamma & \text{if A is worse} \\ \gamma & \text{if A is search and } z_i = 0 \\ 1 - \gamma & \text{if A is search and } z_i = 1 \\ \gamma & \text{if A is stand and } z_i = 1 \\ 1 - \gamma & \text{if A is stand and } z_i = 0. \end{cases}$$

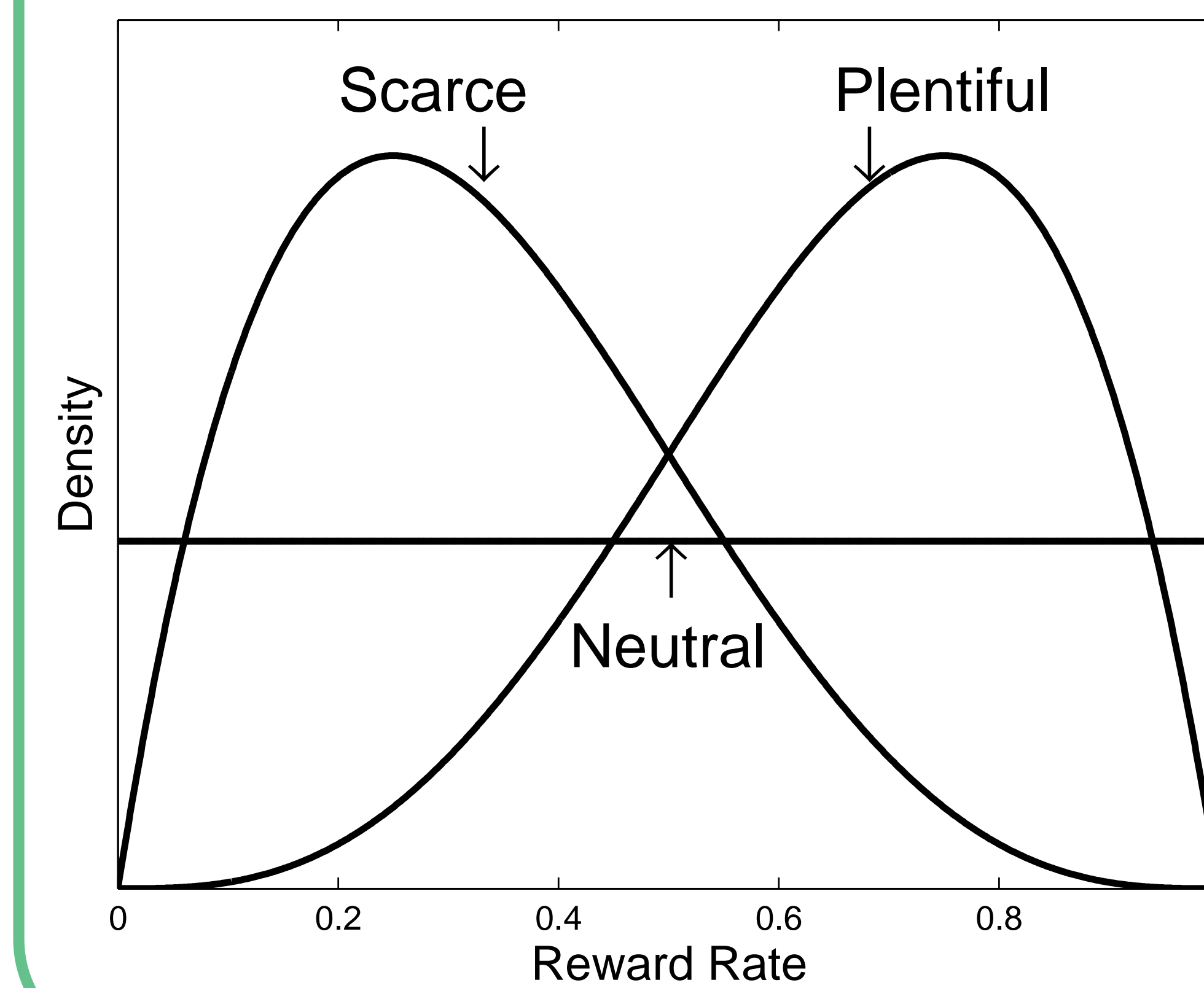
The graphical model has two parameters: switch point τ and accuracy of execution γ .

Human Experiment

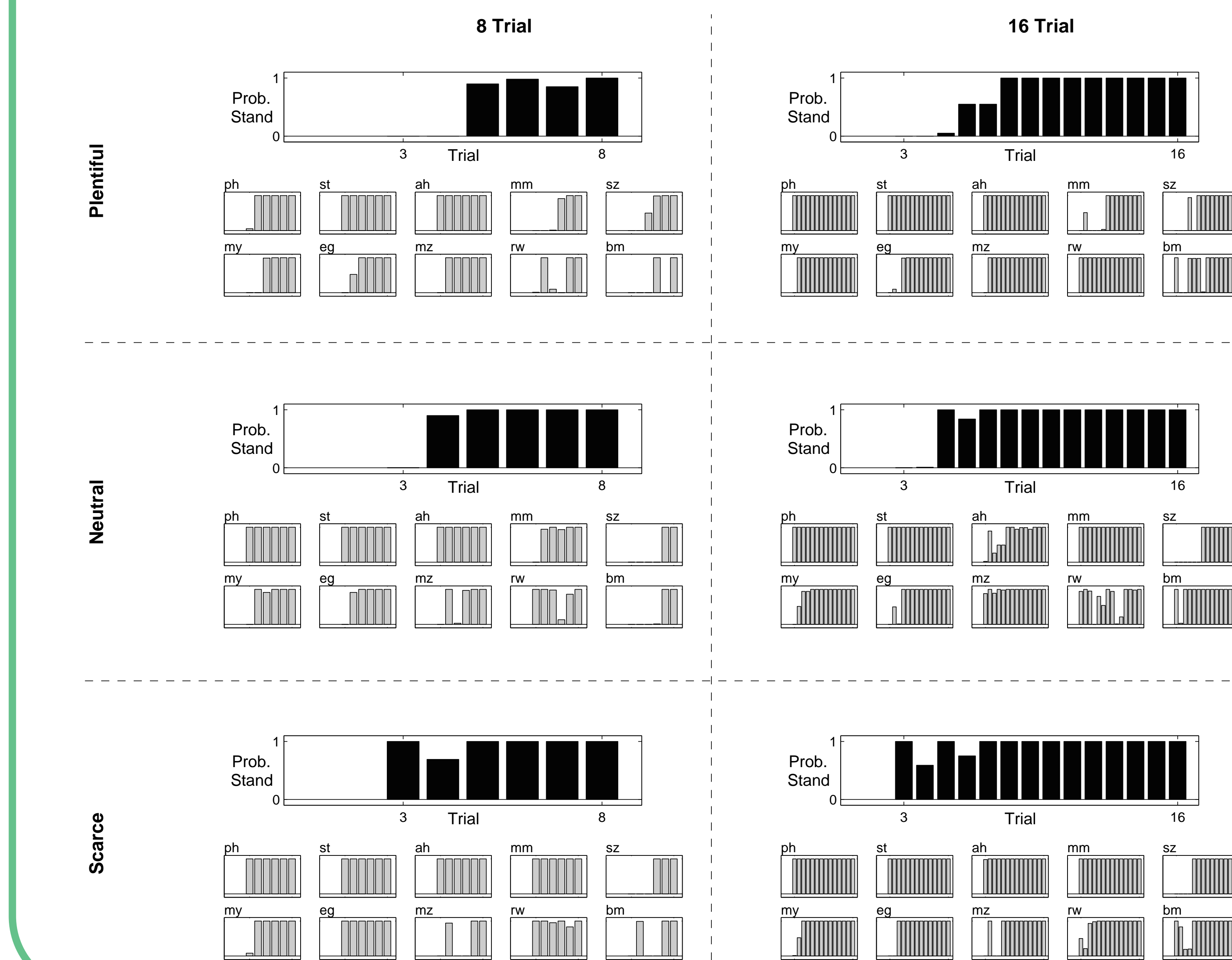
Subjects: 10 naive subjects: 6 males, 4 females.

Stimuli: two-armed, varied in terms of two trial sizes (8 trials and 16 trials) and three different environmental distributions

Procedure: within-participant data were collected on 50 problems for all six conditions. Order of conditions was randomized for each subject.



Analysis



Optimal Player

Given the environmental distribution and the trial size, we can implement the optimal decision-making process using approach well understood in the reinforcement learning literature.

Optimal behavior was calculated for all problems completed by subjects.

Descriptive Adequacy

DM	Plentiful		Neutral		Scarce	
	8	16	8	16	8	16
Optimal	.95	.93	.95	.94	.92	.90
PH	.96	.94	.92	.92	.84	.90
ST	.99	.87	.94	.84	.93	.80
AH	.89	.89	.76	.75	.71	.73
MM	.92	.88	.92	.93	.90	.94
SZ	.92	.94	.95	.92	.88	.91
MY	.94	.95	.92	.93	.89	.88
EG	.94	.91	.90	.90	.85	.89
MZ	.97	.91	.92	.88	.93	.86
RW	.89	.90	.86	.80	.84	.80
BM	.93	.88	.92	.87	.89	.90

Thanks

This work was supported by the Air Force Office of Scientific Research (FA9550-07-1-0082) to Michael Lee and Mark Steyvers.

We thank Jun Zhang for suggesting the type of model studied in this paper.

References

- [1] Steyvers, M., Lee, M. D., & Wagenmakers, E-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*.
- [2] Sutton, R. S., & Barto, A. G. (1988). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.