Learning and reasoning in a complex environment

David L Barack and C Daniel Salzman
Columbia University

Keywords: Reasoning, Cognition, Decision making, Exploration, Foraging

Summary

Whether negotiating social spaces, solving puzzles, or planning research, humans reason and learn in complex environments with many actions and states. From learning social hierarchies to planning a foraging route, nonhuman primates face similar problem spaces with many actions and states. Here, we report on monkeys playing a simplified version of the game Battleship, designed to investigate learning these complex environments. Monkeys visually uncovered one shape per trial over multiple choices. Each shape occurred at a specific location, but shapes could overlap. Choices could be made in any order and revealed either a piece of the colored target or a white background. Trials ended once the shape was fully uncovered.

We focus on three questions: were monkeys able to perform this task?; what were monkeys learning during this task (e.g., shapes or movement sequences)?; and what computations were monkeys using to learn? First, monkeys were adept learners, as assessed against the optimal choice that maximized expected reward at each choice in a trial (average number of trials to 75% optimal choices = 511 ± 117 trials, n = 7 shape sets). Second, monkeys used different patterns to reveal the shape across learning, indicating that the shapes, and not movement sequences, were learned. Third, while initial model fits to the behavior show that starting information bonus models fit better than basic e-greedy reinforcement learning or gradient search models (AIC; $\text{aic}_{\text{e-greedy}} = 6.18 \times 10^3$; $\text{aic}_{\text{start}} = 1.96 \times 10^3$; $\text{aic}_{\text{gradient}} = 8.48 \times 10^3$), simulated agents using these choice strategies were unable to learn the shapes (average number of choices to learn > 20,000 trials). We predict that further behavioral modelling will show that shape information, computed as the difference between shape entropy and the entropy of merely finding a shape piece, a variant of the Kullback-Leibler divergence, helped drive choice behavior.

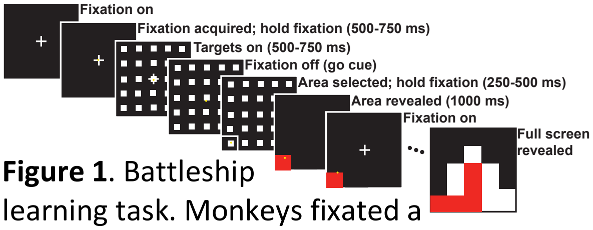

Figure 1. Battleship learning task. Monkeys fixated a central cross, maintaining fixation as targets appeared (small white squares). Once fixation extinguished (go cue), monkeys were free to choose any of the available targets. To register a choice, fixation was maintained on a target. The chosen grid square was then revealed, displaying either a part of a shape (red) or nothing (white). After a free look period, the fixation cross reappeared and the revealed grid squares stayed on the screen.

Additional Detail

Monkeys learned the locations of shapes on a grid over the course of many trials (Fig. 1). On each trial, monkeys made a series of oculomotor decisions to reveal a single shape. Monkeys could explore, only registering a choice upon fixating a target for a variable fixation time (250-500 ms). Trials ended once the full shape had been revealed. Small rewards were delivered for each piece of shape uncovered, and a large reward inversely proportional to the total number of choices needed to finish revealing a shape was delivered at the end of the trial. Each particular shape always appeared at the same location, with different locations for different shapes, and the shapes could partially overlap. Shape color varied randomly from trial to trial, and different shapes were randomly interleaved.

A sample shape set for a 5x5 grid is depicted in Fig. 2A. This design results in a value map that gradually changes over time as the monkey uncovers pieces of the shape; Fig. 2B depicts the state of the value map at the start of a trial (warmer colors = greater expected reward). At each time point, the value map defines the set of optimal choices, so performance can be assessed across choices and trials. Monkeys were proficient at learning this particular shape set (Fig. 2C), achieving approximately 80% optimal choices, defined as choosing one of the grid squares with the highest expected value on the value map given what's been revealed so far.

The learning curve for a particular 'H' shape is depicted in Fig. 2D. Using a changepoint detection test, which evaluates the log-odds ratio of a changepoint in a time-series (Gallistel, Fairhurst, and Balsam), on the cumulative number of choices to fully reveal the ‘H’ shape, the monkey learned the shape after only 34 trials (red circle, Fig. 2D). However, even after this trial, changes in the variance of the number of choices to reveal the shape remained (note changes in the deviation from the mean in the learning curve plotted in Fig. 2D).

Using the changepoint detection test on the variance (Inclan and Tiao), six learning intervals were uncovered between variance changepoints (Fig. 2E), indicating that even after learning the shape, the monkey continued to refine its choice behavior. Each trial was then coded by the pattern used to uncover the shape, counting from the first piece of the uncovered shape. Plotting the frequency of each unique pattern used to uncover the shape in each variance interval separately revealed that the monkey used a unique pattern for each successful trial in the two intervals before learning as well as the first interval after (Fig. 2F, top three panels). Only in later intervals did a preferred movement pattern begin to emerge, as revealed by peaks in the distribution of pattern frequency, but even in those intervals there was still residual variability in the patterns used (Fig. 2F, bottom three panels).

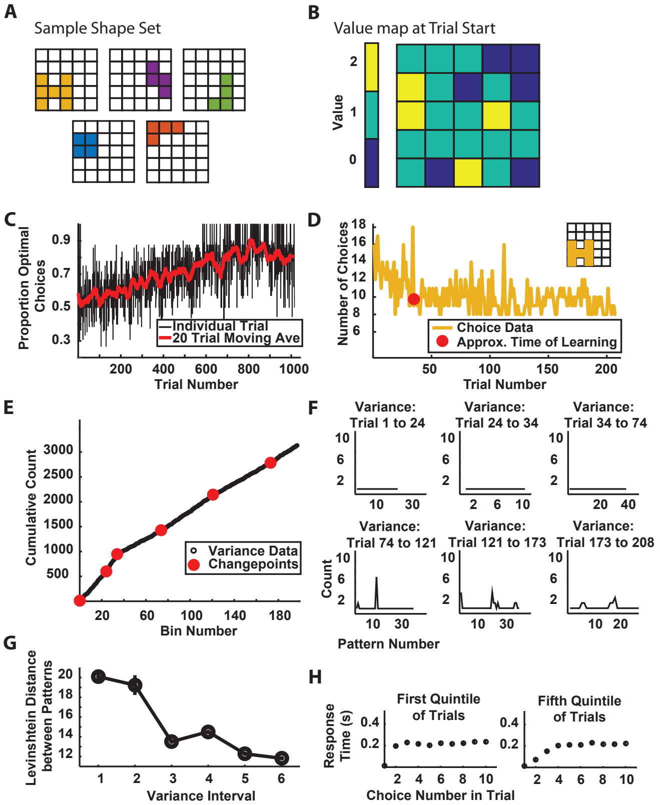

Figure 2. Data from sample shape set on a 5x5 grid with five shapes. See text for details. G, H: error bars occluded by data points.

The conclusion that shapes and not movement sequences were learned was further supported by analyzing the Levinshtein distance between patterns, a measure of pattern similarity, which showed much less similarity between patterns at the beginning of learning and much more by the end (Fig. 2G). Response time analysis suggests that monkeys may have been begun to settle on a particular movement sequence as learning proceeded. Choices were categorized as single-step, when the monkey made a single saccade to register a choice during a trial, or multi-step, when the monkey made a number of saccades before choosing. Examination of single-step response times in the first quintile of trials compared to the last quintile revealed increasing linearization in the time for each successive response over the course of learning (Fig. 2H), consistent with past findings that each next step in an eye movement sequence takes longer to execute (Zingale and Kowler).

Modeling of choices behavior on this task revealed that more traditional reinforcement learning algorithms are miserable at learning the shapes. Initial model fits to the behavior show that a starting information bonus model, which artificially adds a small value to all grid squares at the beginning of learning, fit better than a basic e-greedy reinforcement learning model (Sutton and Barto) or a gradient search model, which adds a small value bonus to grid squares adjacent to squares with pieces of shapes (AIC; $aic_{e\text{-}greedy} = 6.18 \times 10^3$; $aic_{start} = 1.96 \times 10^3$; $aic_{gradient} = 8.48 \times 10^3$). However, simulated agents using these choice strategies were unable to learn the shapes (models failed to converge after 20,000 trials), presumably because these algorithms represent the state space at the finest grain (3 states (revealed, piece of shape; revealed, no shape piece; unrevealed) for a 5x5 grid is 3^{25} = 847,288,609,443 states), requiring an enormous number of samples to learn to navigate the space. We predict that further behavioral modelling will show the influence of shape information throughout shape learning. In particular, we predict that either the difference between the entropy of getting a particular piece of a shape given a location and the entropy of getting a piece of a shape given a location and a shape, which includes shape information and is a variant of the Kullback-Leibler divergence (Johnson et al.), or that the entropy over different patterns of evidence which monkeys will track, a form of infotaxis (Vergassola, Villermaux, and Shraiman), will drive learning during the task.

References

Gallistel, C. R., S. Fairhurst, and P. Balsam. 2004. 'The learning curve: implications of a quantitative analysis', *Proceedings of the National Academy of Sciences of the United States of America*, 101: 13124-31.

Inclan, Carla, and George C Tiao. 1994. 'Use of cumulative sums of squares for retrospective detection of changes of variance', *Journal of the American Statistical Association*, 89: 913-23.

Johnson, Adam, Zachary Varberg, James Benhardus, Anthony Maahs, and Paul Schrater. 2012. 'The hippocampus and exploration: dynamically evolving behavior and neural representations', *Frontiers in human neuroscience*, 6.

Sutton, Richard S., and Andrew G. Barto. 1998. *Reinforcement learning : an introduction* (MIT Press: Cambridge, Mass.).

Vergassola, Massimo, Emmanuel Villermaux, and Boris I Shraiman. 2007. '‘Infotaxis’ as a strategy for searching without gradients', *Nature*, 445: 406.

Zingale, Carolina M., and Eileen Kowler. 1987. 'Planning sequences of saccades', *Vision Res*, 27: 1327-41.