Introduction

- Classical conditioning studies have shown that fear is most easily acquired and difficult to extinguish to certain biologically relevant classes of stimuli (Pavlovian triggers; 1).
- These include dangerous animals (e.g., snakes), threatening con-specifics (2), and racial out-group members (3).
- We asked how different Pavlovian triggers would impact goal-directed aversive reinforcement learning (RL) across three experiments.

Prediction: Enhanced RL when the Pavlovian trigger stimuli is most predictive of shocks, and impaired RL when the neutral stimuli is most predictive of shocks, reflecting competition between instrumental and Pavlovian systems.

Methods

Proportional two-choice shock avoidance task: Mixed 2 (Group: PavlovianToNeutral/NeutralToPavlovian) * 2 (Phase: Learning/Reversal) design. After 35 (of 70) trials, the stimulus – shock contingency switched: P(Shock | A, Learning) = .25 => P(Shock | A, Reversal) = .75

Experiment 1: Pavlovian Neutral (European subjects)

Experiment 2: Pavlovian Neutral (Modern Racism Scale, learning phase)

Experiment 3: Pavlovian Neutral (Modern Racism Scale, reversal phase)

Results

Reinforcement Learning Model

Experiment 1 (n = 32)

Q-learning model (Exp. 1):
\[ Q_{\text{snake}}(t + 1) = Q_{\text{snake}}(t) + \alpha * \delta(t) \]
\[ \delta(t) = R(t) - Q_{\text{snake}}(t) \]

Softmax decision function:
\[ P_{\text{snake}}(t) = \frac{\exp((Q_{\text{snake}}(t) - \text{Pavlovian}_{\text{snake}})/\beta)}{\exp((Q_{\text{snake}}(t) - \text{Pavlovian}_{\text{snake}})/\beta) + \exp((Q_{\text{shock}}(t))/\beta)} \]

Conclusions

- Confirming the predictions, Pavlovian trigger stimuli had a strong and lasting impact on instrumental behavior. This impact could have important implications for social behavior.
- The results are consistent with a view of aversive Pavlovian triggers as eliciting behavioral inhibition that is transferred to instrumental behavior (Pavlovian to instrumental transfer).
- The RL model provides a parsimonious account of how both enhancement and impairments can result from the congruency or incongruity between an action (i.e., avoiding shocks) and the intrinsic value of Pavlovian triggers.

References