Comparing the Effects of Stress on Directed and Random Exploration

Kyle J. LaFollette
Case Western Reserve University
Cleveland, OH
kjlafoll@case.edu

Heath A. Demaree
Case Western Reserve University
Cleveland, OH
had4@case.edu

Abstract

Recent insights from artificial intelligence and computer science have proven invaluable for studying the explore-exploit dilemma, particularly in disentangling information-directed from random exploration strategies. Despite these advances, the algorithms used to solve this dilemma are largely devoid of affective parameters. In this study, we aimed to investigate the relationship between the cognitive computational substrates of the explore-exploit dilemma and a particularly powerful indicator of affective state: physiological stress reactivity. We first used a Bayesian generalized logistic regression model to evaluate propensities to choose more informative bandits over higher valued bandits. This revealed more directed exploration under stress, and no change in random exploration. We then considered three hierarchical Bayesian reinforcement learning models to determine how uncertainty associated with Kalman filter states can influence explorative behavior under stress versus no stress: a Thompson sampler, an upper confidence bound algorithm, and a hybrid of the two. We found that an upper confidence bound algorithm best explains exploration under stress, whereas a hybrid model may better capture exploration without stress. We discuss the implications of these findings for future modeling of the explore-exploit dilemma; models must be flexible to affective states such as stress if modeling of the explore-exploit dilemma is to progress beyond normative considerations.

Keywords: exploration-exploitation, reinforcement learning, decision making, stress

1 Introduction

Exploration versus exploitation decisions are fundamental to everyday human decision-making. From a computational perspective, these decisions arise from uncertainty in the value of multiple options. For example, when a person feels uncertain about whether s/he should continue working a demanding, perhaps unfulfilling job, s/he may choose to continue *exploiting* their current position, or *explore* alternative employers. However, not all exploration is value-driven: Some exploration is stochastic. A person may choose to explore new targeted job ads to gain information on other specific employers (i.e., *value-directed* exploration), or the choice may be reflexive of their overall uncertainty (i.e., *random* exploration).

Much of the explore-exploit dilemma has been studied under normative considerations, but what can we expect people to do under stress? Numerous animal models suggest that anxiety is negatively associated with exploration [6, 10]. In humans, Lenow and colleagues [5] discovered that both acute and chronic physiological stress – markers of environmental safety or quality – were associated with an increase in exploitation. Other studies have suggested that anxiety in humans is similarly inversely related to explorative behavior [2, 8]. Although it may appear from this literature that stress drives exploitation at the cost of exploration, few studies have considered disparate effects on directed versus random exploration. We hypothesized that directed exploration would be attenuated under stress, whereas random exploration would be unaffected. To test this hypothesis, we measured participants' preferences on a multi-armed bandit task immediately after an acute physiological and psychological stress manipulation.

2 Methods

Participants. We tested 29 participants in this study. Participants were recruited from a midwestern University and received partial course credit for their participation. All participants provided written informed consent in accordance with the university's institutional review board. In addition to course credit, participants were incentivized with up to $5 USD in performance bonuses.

Procedure. The study was conducted over two experimental sessions. The sessions were approximately 1.5-hours in duration and separated by 1-week. All participants completed a variation of the Maastricht Acute Stress Task (MAST; [7]) during each session. The MAST is a physiological and psychosocial stress manipulation which combines a cold pressor task (i.e. cold produced pain), serial subtraction task (i.e. social evaluation during arithmetic), and uncertainty (i.e. pseudorandom durations). During one session, participants alternated between immersing a foot into a bucket of 2-degrees Celsius ice water and counting backwards from 2043 in units of 17. During another control session, the same procedures were followed but with room temperature (17-degrees Celsius) water and serial subtractions of 2. Session order was within-subjects and counterbalanced.

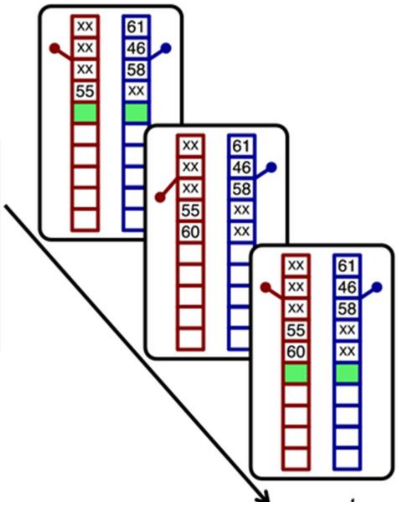

Figure 1. Horizon task design.

Participants completed 80 games of the Horizon Task (Figure 1 [9]) immediately after the MAST. On each game, participants chose between two one-armed bandits (i.e., slot machines), each of which provided points from different Gaussian distributions. The means of these two distributions ranged anywhere between 1 and 100 points, but their standard deviations were always equal to 8 points. Through sampling, participants were expected to learn the means of these two distributions and maximize their total reward. At the start of each game, participants were instructed to make four forced options (i.e., they were bound to select a specific bandit without any volition of their own). These forced choices set up the game to have unequal information between the two bandits (e.g., one bandit was forcibly chosen three times while the other was chosen once). After the forced choices, participants were free to choose either bandit either one time (i.e., short horizon 1) or six times (i.e., long horizon 6; participants may switch the bandits chosen throughout the 6 trials). These two horizon conditions manipulated the relative value of exploration and exploitation.

Behavioral Analysis. We fit a Bayesian generalized logistic regression model (i.e., Richards' curve) to participant choice data immediately following the four forced trials. This determined the probability of choosing the lesser known bandit on the first free-choice trial as a function of the difference in average bandit values δ with Gaussian noise: $choice \sim N\left(a + \frac{a-b}{1+e^{-\beta\delta}}, \sigma^2\right)$. a, b, and β were free to vary with stress condition (stress versus no stress). Following the informal observations of Wilson and colleagues [9], the growth rate β reflects the degree to which choice is influenced by random exploration, and the point of inflection $a - b$ reflects the influence of directed exploration.

RL Modeling. We modeled choice behavior using three hierarchical Bayesian reinforcement learning models; one using a Thompson sampler, another an upper confidence bound algorithm, and the last a hybrid of the two. Bandit values were updated using Kalman filtering equations, $Q_{t+1}(a) = Q_t(a) + \kappa_t\left(r_t - Q_t(a)\right)$ where $Q_t(a)$ is the expected value of the chosen bandit and κ_t is the Kalman gain. Likewise, variance was updated as a proxy for uncertainty, $\sigma_{t+1}^2(a) = \sigma_t^2(a) - \kappa_t\sigma_t^2(a)$. The Kalman gain is dynamic with learned uncertainty and measurement variance, $\kappa_t = \frac{\sigma_t^2(a)}{\sigma_t^2(a) + \rho^2(a)}$, where $\rho^2(a)$ is the measurement variance. Initial $Q_0(a)$ was fixed at 0 whereas initial $\sigma_0^2(a)$ was free to vary across subject and stress conditions. Choice probabilities $p(a)$ for the first model were calculated using a Thompson sampling policy to gauge random exploration, $p(a_t) = \Phi((Q_t(a) - Q_t(b))/\surd(\sigma_t^2(a) + \sigma_t^2(b)))$, where stochasticity is proportional to total uncertainty $\sqrt{\sigma_t^2(a) + \sigma_t^2(b)}$ across bandits a and b. The second model used an upper confidence bound algorithm instead for directed exploration, $p(a_t) = \Phi(((Q_t(a) - Q_t(b)) + \gamma(\sigma_t^2(a) - \sigma_t^2(b)))/\lambda)$, where γ reflects an information bonus for bandit a's uncertainty relative to b, and λ controls stochasticity. Both γ and λ were free to vary across subject and stress conditions. Last, the hybrid model combined these policies to assess the simultaneous influence of random and directed exploration, $p(a_t) = \Phi\left(\frac{(1-\omega)\left(Q_t(a) - Q_t(b)\right)}{\sqrt{\sigma_t^2(a) + \sigma_t^2(b)}} + \omega\left(\sigma_t^2(a) - \sigma_t^2(b)\right)\right)$, where ω is the weight of directed exploration on choice relative to random exploration [4]. ω was also free to vary across subject and stress conditions.

Model fitting. All models were fit using a Hamiltonian Monte Carlo No-U-Turn sampler, a MCMC method provided in the Stan probabilistic programming language. We collected 10,000 samples for each parameter across 4 chains run in parallel. The first 5,000 samples of each chain were discarded as warm-up. We compared the relative fits of models by approximating their point-wise out-of-sample predictive accuracy using leave-one-out cross validation (LOO-CV). All Stan models and analysis scripts are available at: https://github.com/kjlafoll/stress-ee-rldm2022.

3 Results

3.1 Behavioral

First, we checked for main effects of stress and horizon on the propensity to choose the explorative or lessinformed bandit. These behavioral analyses did not account for learning and as such were restricted to the first free-choice trial of each game. A frequentist logistic regression model revealed more explorative choices with longer horizons, $\beta = 0.458$, $z = 7.162$, $p < 0.001$, as well as after stress manipulation, $\beta = 0.402$, $z = 6.291$, $p < 0.001$ (Fig 2A). Second, we considered the propensity to explore as a function of the mean difference between both bandits prior to the first free-choice. A Bayesian generalized logistic regression model revealed disparate effects of stress on model parameters α and β, the lower asymptote and slope of the sigmoid, respectively (Figure 2B-C). Participants had a much greater α after the stress session, meaning that they had a greater propensity to choose a much lower-valued explorative bandit when under stress than when not. This suggests more directed exploration under stress. Conversely, participants had near equivalent β after stress relative to control, indicating similar accelerations in preference with changes in value. This suggests no change in random exploration between stress conditions.

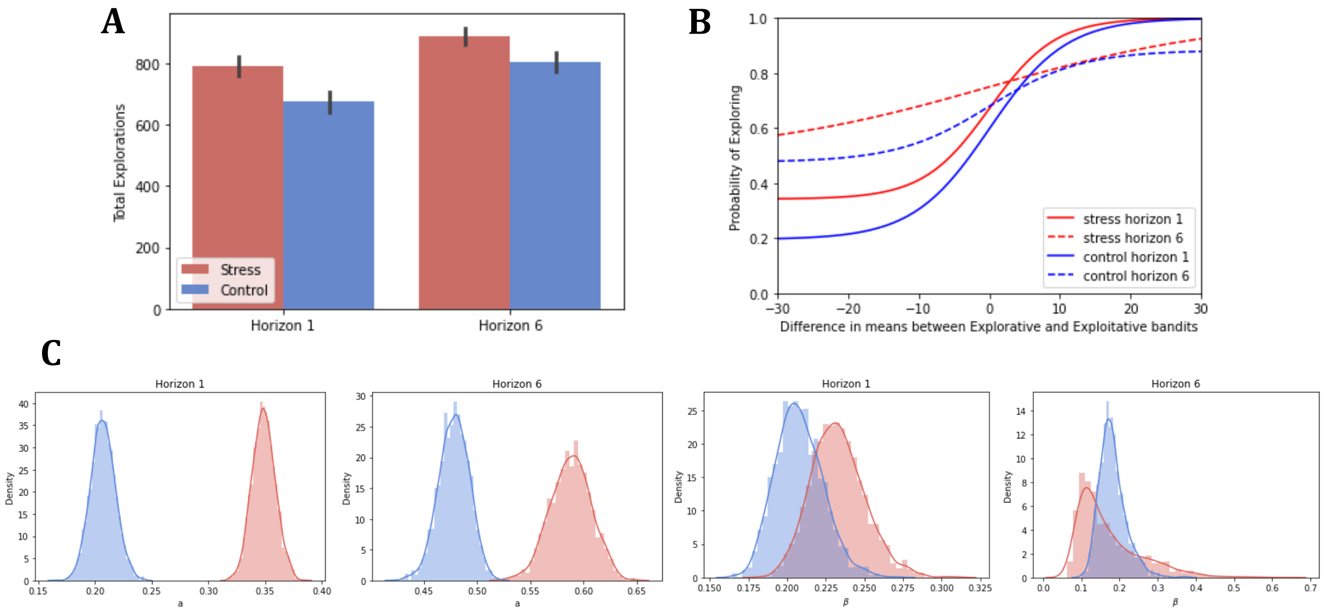

Figure 2. First free-choice behavior. (A) Total number of explorative choices made. (B) Logistic fits for probability of choosing explorative bandit as a function of mean value difference. (C) Posteriors over parameter means from Bayesian generalized logistic model for each stress condition.

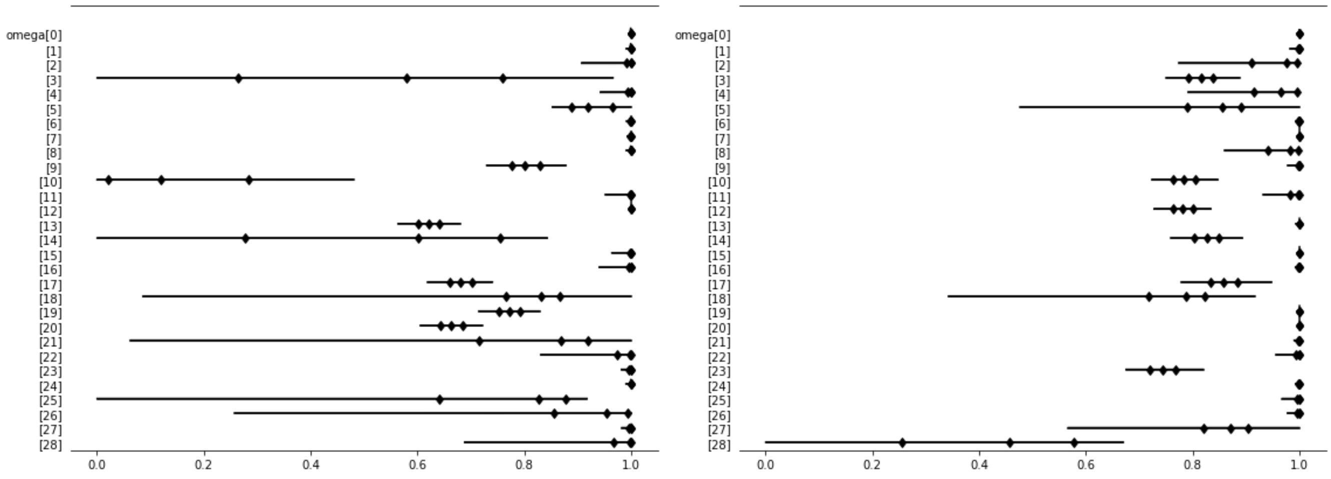

Figure 3. Results of hybrid model. Subject-level posteriors over ω weight for directed exploration for control (left) and stress (right) sessions.

3.2 Reinforcement Learning

The behavioral analyses were limited in that while they could elucidate horizon effects, they were restricted to the first free-choice and blind to learning processes. To extend our analysis to processes of reinforcement learning, we fit three hierarchical Bayesian model to all free-choice trials in long horizon 6 games (short games were excluded due to learning being a nonfactor). The winning model, a hybrid of a Thompson sampling policy and upper confidence bound algorithm, was selected for inference. Subject-level posteriors over the parameter weighting directed exploration relative to random exploration revealed substantially more directed exploration during the stress session than during control (Figure 3).

4 Discussion

In this study, we presented a two-factored approach to modeling the explore-exploit dilemma under acute physiological and psychological stress: a descriptive, generalized logistic regression model, and an explanatory hybrid reinforcement learning model. The models revealed that, counter to our hypotheses, directed exploration strengthened under stress. This finding conflicts with many behavioral analyses of exploration under stress. Nonetheless, it is a step forward toward disentangling directed and random exploration at a latent processing level, especially considering that some conflicts in the literature suggest that predicting explorative behavior is not so clear-cut. For example, Feldman-Hall and colleagues [3] reported that electrodermal activity – a gold-standard measure for stress reactivity – was associated with increased risk-taking behavior when probabilities were ambiguous or uncertain. A recent study by Bennett and colleagues [1] also concluded that anxiety correlates with informatic utility. Taken together, this suggests that stress may have a modulatory effect on explorative behavior dependent on context; it is important not to conflate threats of opportunity costs with elements of uncertainty when predicting exploration under stress. More research is needed on explorative behavior outside normative considerations such as stress in the face of uncertainty, and computational modeling can be a useful tool for formalizing new theoretical directions.

References

[1] Bennett, D., Sutcliffe, K., Tan, N. P.-J., Smillie, L. D., & Bode, S. (2021). Anxious and obsessive-compulsive traits are independently associated with valuation of noninstrumental information. Journal of Experimental Psychology: General, 150(4), 739–755.

[2] Biedermann, S. V., Biedermann, D. G., Wenzlaff, F., Kurjak, T., Nouri, S., Auer, M. K., Wiedemann, K., Briken, P., Haaker, J., Lonsdorf, T. B., & Fuss, J. (2017). An elevated plus-maze in mixed reality for studying human anxiety-related behavior. BMC Biology, 15(1), 125.

[3] Feldman-Hall, O., Glimcher, P., Baker, A. L., & Phelps, E. A. (2016). Emotion and decision-making under uncertainty: Physiological arousal predicts increased gambling during ambiguity but not risk. Journal of Experimental Psychology: General, 145(10), 1255-1262.

[4] Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. Cognition, 173, 34-42.

[5] Lenow, J. K., Constantino, S. M., Daw, N. D., & Phelps, E. A. (2017). Chronic and acute stress promote overexploitation in serial decision making. Journal of Neuroscience, 37(23), 5681-5689.

[6] Prut, L., & Belzung, C. (2003). The open field as a paradigm to measure the effects of drugs on anxiety-like behaviors: A review. European Journal of Pharmacology, 463(1–3), 3–33.

[7] Smeets, T., Cornelisse, S., Quaedflieg, C. W. E. M., Meyer, T., Jelicic, M., & Merckelbach, H. (2012). Introducing the Maastricht Acute Stress Test (MAST): a quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. Psychoneuroendocrinology, 37(12), 1998-2008.

[8] Walz, N., Mühlberger, A., & Pauli, P. (2016). A human open field test reveals thigmotaxis related to agoraphobic fear. Biological Psychiatry, 80(5), 390–397.

[9] Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. Journal of Experimental Psychology: General, 143(6), 2074-2081.

[10] Zweifel, L. S., Fadok, J. P., Argilli, E., Garelick, M. G., Jones, G. L., Dickerson, T. M. K., Allen, J. M., Mizumori, S. J. Y., Bonci, A., & Palmiter, R. D. (2011). Activation of dopamine neurons is critical for aversive conditioning and prevention of generalized anxiety. Nature Neuroscience, 14(5), 620–626.