Analogous neural representations underlying risky decision making in deep reinforcement learning agents and humans

T Alexander Price (alexander.price@neuro.utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

Andrew Liu (a.liu@utah.edu)
Math Department, 155 South 1400 East, JWB 233
Salt Lake City, UT 84112 USA

Rhiannon L Cowan (rhiannon.cowan@utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

Tyler S Davis (tyler.davis@hsc.utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

Niloufar Shahdoust (niloufar.shahdoust@utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

Shervin Rahimpour (Shervin.rahimpour@hsc.utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

Ben Shofty (ben.shofty@hsc.utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

John D Rolston (jrolston@bwh.harvard.edu)
Department of Neurosurgery, 60 Fenwood Road,
Boston, MA 02115 USA

Elliot H Smith (e.h.smith@utah.edu)
Neurosurgery Department, 175 N Medical Dr.,
Salt Lake City, UT 84132 USA

Alla Borisyuk (borisyuk@math.utah.edu)
Math Department, 155 South 1400 East, JWB 233
Salt Lake City, UT 84112 USA

Abstract

We employed deep reinforcement learning to discover behavioral and neural strategies underlying a spectrum of performance on a risky decision-making task. Working backwards, we identified analogous behavior from a large cohort of neurosurgical patients from whom we recorded single neuron activity in decision making circuits. Examining low dimensional factors in neuron population activity uncovered temporal and trial factors differentiating task performance groups, with improved task performance being associated with more nonlinear neural representations of reward prediction.

Keywords: deep reinforcement learning, risky decision making, human single neurons

The Balloon Analog Risk Task (BART) is an ecologically valid decision-making task that models risk taking and impulsivity behavior (Lejuez et al., 2002). Despite BART's widespread use as a psychophysical paradigm, behavioral strategies and optimal performance are opaque (Schonberg et al., 2011). Here, we address this knowledge gap by training deep reinforcement learning agents on BART to gain insight into the diversity of neural and behavioral representations that underlie BART behavior. We sought to use this information to better understand how human BART participants may be understanding the task and how these internal representations may be impacting their performance. Clustering neural representations in agents revealed a spectrum of task strategies and neural representations related to risk that were subsequently uncovered in human neural representations.

Methods

BART participants were instructed to maximize the points earned throughout their session. Points were rewarded based on the size of the balloon when inflation ceases and receive no points if the balloon pops. In the version of BART that the humans completed, balloons are colored either gray, red, orange or yellow. The color of balloon, and the presence/absence of an indicator of a passive trial, cued patients to the potential for reward on each trial. There were five reward categories of balloons: gray (unrewarded passive trials), yellow, orange, red

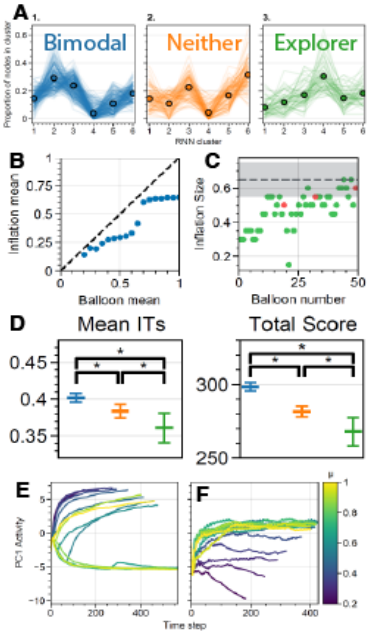

Figure 1: Actor-critic network behavior and representations. A) proportions of specific node types each type of agent had. Circles are means, each line is an individual agent. B) example of bimodal behavior. C) example of explorer behavior. D) left: Mean inflation time (IT), right: total score. E) PC1 activity for a bimodal agent (left) and not (right) across evaluation episodes.

(increasingly risky active trials – red balloons popped at the smallest size, yellow the largest) and yellow, orange or red balloons with a passive trial indicator which are represented as pink (rewarded passive trials).

Actor-Critic agents (a feed-forward layer connected to a RNN layer, and then split into actor and critic feed-forward layers, where each layer has N = 64 nodes) were trained on a variation of BART using a standard proximal policy optimization (PPO; Schulman et al., 2017). In this version of BART, for 50 consecutive balloons a mean balloon size, μ, was drawn uniformly between 0.2 and 1 and each balloon's maximum size drawn from N(μ,.05). Observations vectors were formed by the size of the balloon, the previous action and the previous reward while actions were drawn from policy outputs stochastically. Activity of the recurrent layer nodes was clustered using k-means (k=6 was selected as optimal after analysis comparing k's between [3, 21]). Next, the agents were clustered using k-means (k=3) based on the proportions of their recurrent layer nodes that fit into each of the six categories. Agents were classified as having bimodal behavior if any gap of mean IT between two consecutive μ conditions exceeded 0.15. To meet the qualifications of an explorer, an exponentially weighted moving average of balloon ITs must increase by 0.2 over the course of an episode. The time steps from the evaluative periods were concatenated into one matrix and Principal Component Analysis (PCA) was done. PC1 was then split back into vectors for the different μ conditions.

Single neuron activity was recorded from patients undergoing neuromonitoring for treatment of drug-resistant epilepsy using Behnke-Fried microwires (Misra et al., 2014). Single units were isolated by bandpass filtering between 0.25 and 7.5 kHz and sorting waveforms that crossed -3.5 times the root mean squared of the filtered signal using Offline Sorter (Plexon, Inc.; Dallas, TX). Humans were considered to use a bimodal strategy if the difference between the mean red and orange ITs was either three times greater or smaller than the difference between the mean orange and yellow ITs. To categorize participants as using an explorer strategy, a moving average of inflation durations was taken for each patient and each balloon color. If the final average was inflated for a longer duration than the first average and all other average values were neither inflated 20% longer than the final average nor 20% shorter than the first average, these patients were considered explorers. Cue-aligned firing rate data was loaded into a three-dimensional tensor (neurons x time x trials). Tensor Component Analysis (TCA) was used to find low dimensional, demixed factors from human neuron pseudoensembles from each behavioral strategy (Williams et al., 2018). Numbers of tensor components were pared down using the prescribed method of finding inflection points in TCA model error.

Results

Agent Results: Based on distributions of node activity clusters, agents were classified into three types, 1) bimodal (55%), 2) neither (30%) and 3) explorers (15%; Fig. 1A). Figures 1B and 1C show exemplary bimodal and explorer strategies, respectively. Digital agents using a bimodal strategy scored significantly higher compared to the other two strategies. They also had significantly longer inflation times than the other two groups (Fig. 1D). When PC1 is plotted for different μ, a clearly bimodal, divergent trajectory appeared (Fig 1E), whereas explorers encoded expected balloon size more linearly. This analysis showed a clear link between the RNN's representation of the task and the agent's performance.

Human Results: Forty-two human participants completed 45 sessions of BART with a mean±s.d. of 236.1± 25.9 trials of BART with a mean±s.d accuracy of 81.9%±6.9%. We recorded from 576 well-isolated units which were grouped into four anatomical areas: The Orbitofrontal Cortex (117 units), The Medial Frontal Cortex (54 units), the Anterior Cingulate Cortex (131 units), and the Mesial Temporal Lobe (274 units).

Human behavior clustered similarly to agent behavior (Bimodal: 57.8%, Neither 33.3% and Explorer: 8.9%), with *bimodal* humans achieving significantly higher scores and inflation times compared to the two other behavioral groups (Fig. 2A). Moreover, low dimensional factors of pseudoensemble activity in *bimodal* patients (Fig 2C) showed curvilinear representations across trial-averaged firing rate data in contrast to the more linear representations in explorer patients. The parabolic shape of the trial factors (Fig 2C, right) is consistent with a canonical representation of risk. Additionally, when we included all units from all patients, the factor upon which explorer units loaded negatively, also exhibited an increase in the time domain. These bimodal representations also appear to group the trials by risk category better than the explorers.

Here, we gained insight into the neural underpinnings of risky decisions using actor-critic networks, finding similar behavior and neural representations in human decision-making circuits. We found that in both humans and agents, more nonlinear encoding of reward probability resulted in improved task performance.

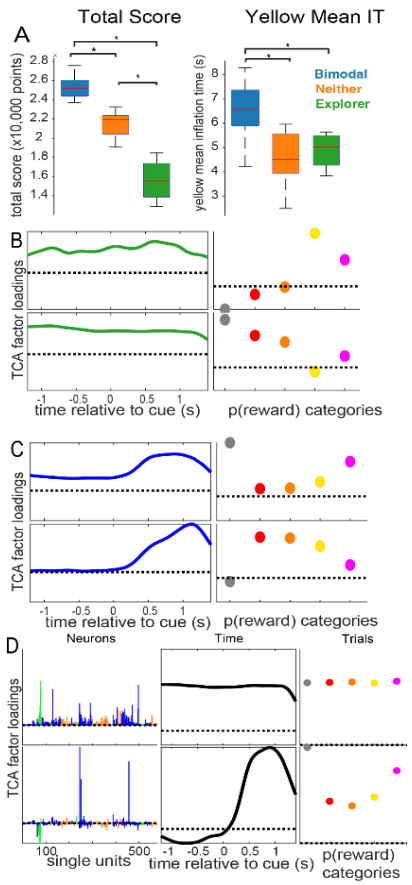

Figure 2: Human behavior and representation. A) Total score and IT boxplots for each bhavior category. B) Two of four TCA factors for cue-aligned firing rates for all units recorded from human *explorers*. C) Same as B but from *bimodal* humans. D) Two factor TCA for all units in cue-aligned trial averaged tensor. Neurons are pseudocolored by behavioral category. Note that green explorer units load negatively onto the second factor, the time factor mirrors the bimodal factor and the trial factor exhibits quadratic reward probability representation, i.e., risk.

We gratefully acknowledge support from a grant from the National Institute of Mental Health: R01MH128187.

References

Hassabis, D., Kumaran, D., Summerfield C., & Botvinick M. “Neuroscience-Inspired Artificial Intelligence.” *Neuron* 95, no. 2 (July 2017): https://doi.org/10.1016/j.neuron.2017.06.011

Lejuez, C. W., Read, J.P., Kahler, C.W., Richards, J. B., Ramsey, S.E., Stuart, G.L., Strong, D. R., & Brown, R. A., (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analog Risk Task (BART). *Journal of Experimental Psychology: Applied,* 8(2), https://doi.org/10.1037//1076-898X.8.2.75.

Misra, A., Burke, J., Ramayya, A., Jacobs, J., Sperling, M., Moxon, K., Kahana, M., Evans, J., & Shara, A. (2014) Methods for implantation of micro-wire bundles and optimization of single/multiunit recordings from human mesial temporal lobe. *Journal of Neural Engineering,* 11(2), 026013. https://doi.org/10.1088/1741-2560/11/2/026013.

Platt, M. L., & Huettel, S. A. (2008). Risky business: The neuroeconomics of decision making under Uncertainty. *Nature Neuroscience*, 11(4), https://doi.org/10.1038/nn2062

Schonberg, T., Fox, C. R., & Poldrack, R. A., “Mind the Gap: Bridging Economic and Naturalistic Risk-Taking with Cognitive Neuroscience.” *Trends in Cognitive Sciences* 15, no. 1 (January 2011): https://doi.org/10.1016/j.tics.2010.10.002.

Schulman J., Wolski F., Dhariwal P., Radford A., & Klimov O,. Proximal policy optimization algorithms. *arXiv* preprint arXiv:1707.06347, 2017.

Williams, A. H., Kim, T. H., Wang F., Vyas S.,. Ryu, S. I., Shenoy K. V., Schnitzer, M., Kolda, T. G., & Ganguli, S. “Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis.” *Neuron* 98, no. 6 (June 2018): https://doi.org/10.1016/j.neuron.2018.05.015