Noisy inference of value signals in frontal cortex drives exploration during reward-guided learning

Charles Findling
(charles.findling@gmail.com)
Laboratoire de Neurosciences Cognitives
ENS, PSL Research University

Vasilisa Skvortsova
(vasilisaskv@gmail.com)
Laboratoire de Neurosciences Cognitives
ENS, PSL Research University

Rémi Dromnelle
(remi.dromnelle@gmail.com)
Institut des Systèmes Intelligents et de Robotique
CNRS, UMR 7222

Nicolas Chopin
(nicolas.chopin@ensae.fr)
Center for Research in Economics and Statistics
CNRS, UMR 9194

Stefano Palminteri
(stefano.palminteri@ens.fr)
Laboratoire de Neurosciences Cognitives
ENS, PSL Research University

Valentin Wyart
(valentin.wyart@ens.fr)
Laboratoire de Neurosciences Cognitives
ENS, PSL Research University

Abstract:

When tracking rewarded stimulus-response associations in volatile environments, humans make a surprisingly large number of seemingly suboptimal decisions, which do not maximize expected outcome. These 'exploratory' decisions have been assigned either to information seeking or to stochasticity in response selection. We reasoned that a fraction of exploratory decisions could be due to random noise in the inference process driving learning, noise which is otherwise assumed to be negligible. Accounting simultaneously for these different sources of exploration in reinforcement learning revealed that more than half of exploratory decisions are due to inference noise alone. This computational dissection of exploration is supported by neuroimaging data, which shows a dissociation in the relationship between choice behavior and two brain regions associated with exploration: fluctuations in anterior cingulate activity co-vary with inference noise during learning, whereas frontopolar activity drives exploration during choice. Together, these findings indicate that exploration in reward-guided learning is driven to a large part by random errors in inference, unbeknownst to the decision-maker.

Keywords: reinforcement leaning; inference noise; exploration-exploitation trade-off; fMRI

In uncertain environments, decision-makers learn rewarding actions by trial-and-error to maximize their expected payoff. Well-acknowledged reinforcement learning (RL) models propose to track value signals associated with each possible action (Sutton & Barto, 1998). Importantly, the large trial-to-trial choice variability in reward-guided learning is classically captured in RL models by a 'softmax' action selection rule modeling noise in the decision process. This choice variability captured by the softmax leads to occasional choices which do not maximize expected payoff. These occasional choices are often attributed to an 'exploration' process driven by the need to reduce uncertainty regarding recently unchosen options. Standard RL models typically assume all choice variability is captured by this softmax rule and that the mental computations involved in the update of the value signals – i.e., the inference process – are noise-free.

However, it has recently been shown that mental inference suffers from a substantial amount of noise, responsible for a dominant fraction of human choice variability (Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016). An intriguing possibility is that the inference process at the heart of reward-guided learning might be subject to the same kind of noise - i.e., random deviations from RL computations. Crucially, these random deviations reflecting computational imprecisions would capture seemingly exploratory trials unbeknownst to the decision-maker and formerly ascribed to noise in the decision process.

To determine whether, and to what extent, inference noise accounts for exploration during reward-guided learning, we derived a theoretical formulation of reinforcement learning which accounts for random noise in its core computations. We then quantified the extent to which exploration is triggered *incidentally* by a noisy inference process rather than *intentionally* by modulations of the choice process (Gershman, 2018; Wilson, Geana, White, Ludvig, & Cohen, 2014). Lastly, we identified the neural correlates of inference noise in the human brain using functional magnetic resonance imaging (fMRI).

Results

Task We designed a restless, two-armed bandit game divided in 8 short blocks in which N=29 human subjects were asked to maximize their monetary payoff. On each trial, subjects chose one of the two options and received its associated outcome. The payoffs that could be obtained from either option (from 1 to 99 points) were sampled from distributions whose means drifted independently across trials. Additionally, among the 8 blocks, 4 were 'factual' blocks where subjects only observed the outcome of the chosen option and the 4 others were 'counterfactual' blocks where subjects also observed the outcome of the unchosen option. A strong prediction concerning 'counterfactual' blocks stems from the fact there is no uncertainty on the unchosen option.

Computational Model To characterize the origin of exploratory decisions in this task, we derived a reinforcement learning model in which the update rule is corrupted by random inference noise (Figure 1a). As in existing theories, decision noise is modeled with a softmax decision rule.

Importantly, although inference and decision noises both capture exploration as defined by standard (noise-free) reinforcement learning, the two components of exploration make different predictions regarding the temporal structure of decisions across successive trials. Indeed, inference noise corrupts the latent value signals which are gradually updated across trials, and which are used to drive successive decisions. Therefore, for the same fraction of exploratory decisions simulated either using inference or decision noise, inference noise engenders larger dependencies across successive choices, which are not predicted by decision noise (Figure 1b).

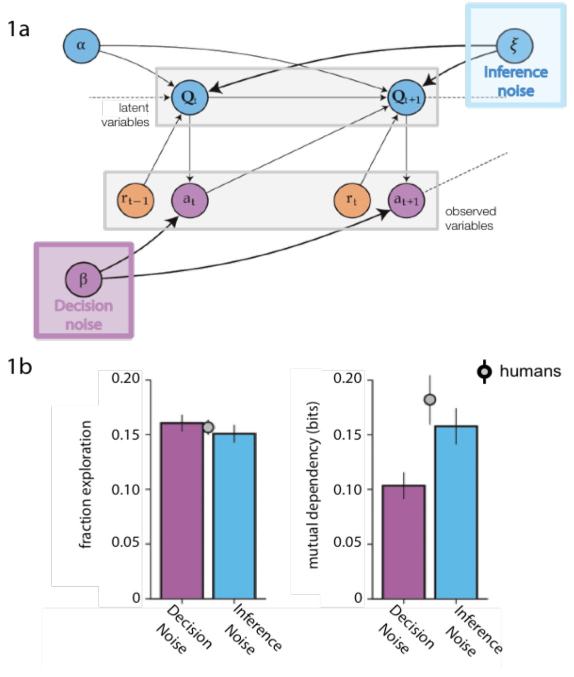

Figure 1: Model predictions and validation (1a) Reinforcement learning model with inference and decision noise. (1b) Inference noise predicts a larger mutual dependency across trials than decision noise.

Dominant contribution of inference noise to exploration We performed Bayesian model selection to characterize the contributions of both candidate sources of behavioral variability to exploratory decisions. Using particle filters to obtain estimates of the model evidence, we found that a reinforcement learning model corrupted by inference noise explained human behavior significantly better than a standard (noise-free) reinforcement learning model in the 'factual' (exceedance p > 0.95) and 'counterfactual' blocks (exceedance p > 0.999). Furthermore, we found that, in 'factual' blocks, assuming a softmax decision rule lead to better explain human performances (exceedance p > 0.999). Considering now the 'counterfactual' blocks, there is, by definition, no incentive to make exploratory decisions given that there is no uncertainty about the unchosen option. Therefore, theoretically, subjects should rely on a greedy, argmax decision rule rather than an exploratory, softmax decision rule. In accordance with this prediction, we found that a purely greedy, argmax decision rule captured the choice process better than a stochastic softmax decision rule (exceedance p > 0.999). To further quantify the respective contributions of inference and decision noises in both blocks, we estimated the trial-to-trial trajectories of the latent value signals corrupted by noise conditioned on all decisions made by each subject in every block (Lindsten, 2013). We then assessed the fraction of exploratory decisions that could be explained solely by inference noise. We found that inference noise explained about 61% of exploratory decisions in 'factual' blocks and 86% in 'counterfactual' ones. This is in agreement with the previous finding of the absence of the softmax decision rule in 'counterfactual' blocks. Interestingly, plotting the raw spread of choice variability due to inference noise and decision noise revealed that only decision noise was significantly reduced in the 'counterfactual' condition relative to the 'partial' condition (t(28) = 4.6, p < 0.001). The inference noise was not different between the two conditions (t(28) = 1.2, p = 0.24), consistent with our hypothesis that noise-driven exploration does not aim explicitly at reducing uncertainty about unchosen

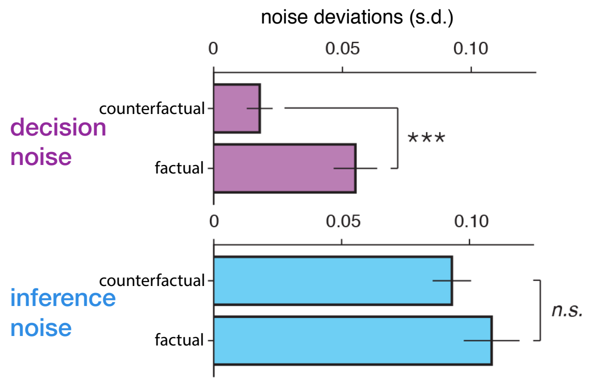

Figure 3: Estimates of decision and inference noise in factual and counterfactual blocks (top: decision noise; bottom: inference noise).

options, but rather reflects a computational constraint on the underlying learning process (Figure 2).

Dissociating inference noise from heuristics in learning One important possibility is that part of this inference noise is caused not by random deviations from the RL rule, but by systematic deviations around this hypothesized rule. In other words, subjects might be using a different learning scheme than the hypothesized reinforcement learning, which would then be fitted as inference noise. To test this hypothesis, we tested N=30 additional subjects in the 'counterfactual' condition where the choice variability has been established above to be solely caused by inference noise (this results was replicated on this second dataset). In this second experiment, subjects played the same blocks of trials twice such that we could measure the consistency of their decisions across the two blocks and decompose the inference noise into a bias and a variance component. Indeed, systematic deviations tend to increase the consistency of decisions across repeated blocks, whereas random deviations tend to decrease decision consistency.

The procedure to obtain the bias/variance trade-off quantifying the respective amounts of deterministic and stochastic deviations is the following: we fitted the reinforcement learning model with inference noise to each subject, and then simulated the model where inference noise was split in two additive terms, a first systematic bias term which was duplicated in the two repetitions of the same block, and a second random, variance term, which was sampled independently in the two repetitions. We varied the relative bias-variance trade-off from zero (pure variance) to one (pure bias) for the simulations of each subject, and found that the trade-off that best accounted for the observed consistency of human decisions across repeated blocs was of 31.8% - indicating that more than two thirds of inference noise are not assignable to any systematic deviations from the RL rule across the two repetitions of the same block. This supports our hypothesis that most of the estimated inference noise truly reflects the limited precision in reinforcement learning updates, and not a systematic deviation between the assumed learning scheme used by the subjects and the one effectively used by the subjects (which may vary across subjects).

Neural correlates of inference noise in the frontal cortex To analyze the neural markers of inference noise, we recorded BOLD fMRI data while subjects performed the first 'factual'/'counterfactual' task. We focused our fMRI analyses on a subset of frontal regions of interest (ROIs) previously highlighted

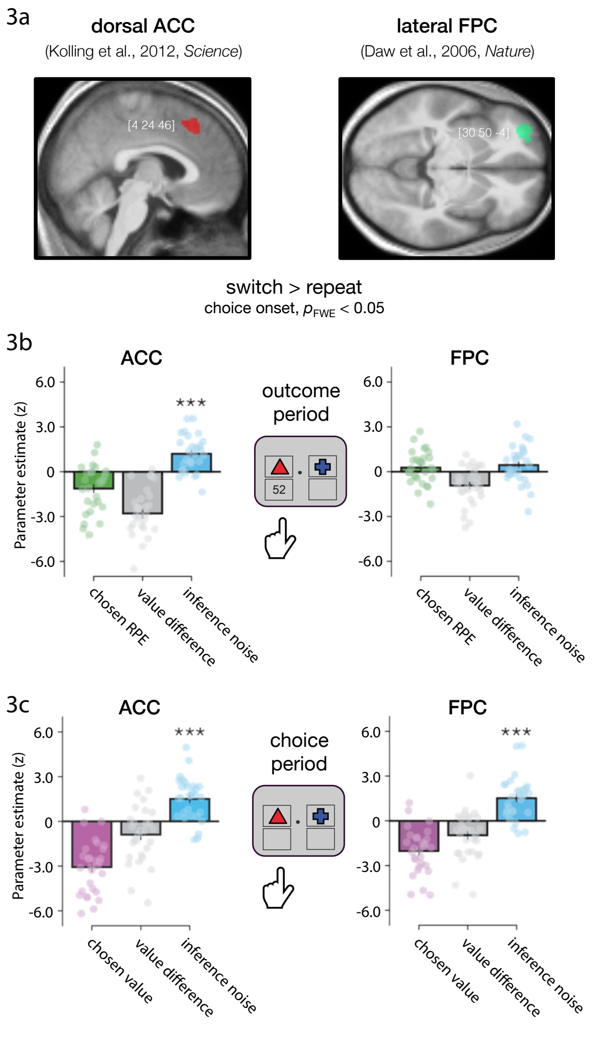

Figure 2: Neural correlates of inference noise in human frontal cortex. (3a) Regions of interest defined using switch > repeat contrast ($p_{\mathrm{FWE}} < 0.05$). (3b-c) Parameter estimates for parametric regressions of BOLD activity in the two ROIs involving the inference noise at outcome (3b) and choice (3c).

in the exploration-exploration trade-off (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) and monitoring uncertainty during reward-guided learning (Behrens, Woolrich, Walton, & Rushworth, 2007; Kolling, Behrens, Mars, & Rushworth, 2012). The ROIs (Figure 3a) were defined with an independent analysis using a switch/stay model-free contrast.

To assess whether activity in these two brain regions co-varied with inference noise at each learning step, we regressed outcome-locked BOLD activity in these ROIs against three parametric regressors derived: the chosen prediction error (chosen RPE), the absolute distance between option values (value difference), and the magnitude of inference noise (inference noise) – Figure 3b. This revealed only the ACC correlated significantly with the inference noise - ACC: t(28) = 5.409, p < 0.001; rFPC: t(28) = 1.938, p = 0.063. We further investigated whether the co-variation of ACC activity with inference noise observed during the learning step - i.e., locked to outcome, is also present during the subsequent choice period when it drives exploratory decisions. To do so, we regressed BOLD activity in the pre-defined ROIs locked to the choice onset against three parametric regressors: the relative value of the chosen option (chosen value), the absolute distance between option values (value difference), and the magnitude of inference noise (inference noise) – Figure 3c. Interestingly, we found that inference noise propagated in time to the choice period in the ACC (t(28) = 5.589, p < 0.001), but also spread to FPC (t(28) = 5.745, p < 0.001). This observation suggests that inference noise is not solely driven by neural variability in response to outcome presentation, but also by neural variability in the maintenance of value signals in the choice period following the learning step.

Dissociating the contributions of frontal cortex to exploration Our fMRI results so far indicate that inference noise initially reflected in the ACC at each learning step, subsequently co-varied with FPC activity during choice. To assess whether and how these two prefrontal regions influence subjects' decision to stay or switch away from the previously chosen option, we ran a logistic regression analysis. The regressors were the theoretical relative value, the trial-by-trial residual estimates of the fMRI activity in the target regions (ACC, FPC) after regressing out the relative value, as well as their interactions with the relative value. We reasoned that a region modulating the *neural gain of learning* should affect the slope of the sigmoid function (interaction term), whereas a region involved in trading off exploration against exploitation should shift the sigmoid function (main effect term). This logistic regression showed that ACC activity negatively affected the neural gain of learning in both 'factual' and 'counter factual' conditions (t(28) = -3.45, p = 0.0017; t(28) = -3.71, p < 0.001). By contrast, FPC activity biased decisions toward more exploration but only in the 'factual' condition when no information about the unchosen option was given ('factual' condition: t(28) = -2.40, p = 0.02; 'counterfactual' condition: t(28) = -0.05, p = 0.96).

Discussion

In this study, we show that more than half of the overall exploration is triggered by random errors in the learning process rather than by an active drive to reduce uncertainty during choice. Model-based neuroimaging suggests distinct roles of the dorsal ACC and lateral FPC in exploration: in contrast to the FPC, our findings suggest that the dorsal ACC does not control the exploration-exploitation trade-off. Instead, by modulating the neural gain of learning, the dorsal ACC triggers a computationally cheap and previously underestimated form of exploration.

References

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Drugowitsch, J., Wyart, V., Devauchelle, A. D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *92*(6), 1398–1411.

Gershman, S. J. (2018). Uncertainty and exploration. *bioRxiv*, 265504.

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, *336*(6077), 95–98.

Lindsten, F. (2013). Backward Simulation Methods for Monte Carlo Statistical Inference. *Foundations and Trends in Machine Learning*, *6*(1), 1–143.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.

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.