Positive Reward Bias on Human Reinforcement Learning under Increased Dopaminergic Neurotransmission

Arnaud Zalta (arnaud.zalta@ens.psl.eu)
Laboratoire de Neurosciences Cognitives et Computationnelles,
Institut National de la Santé et de la Recherche Médicale,
Département d'Études Cognitives, École Normale Supérieure,
Université PSL, Paris, France.

Vasilisa Skvortsova (v.skvortsova@gmail.com)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
University College London, London, UK.

Samuel Hewitt (s.hewitt.17@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
University College London, London, UK.

Michael Moutoussis (m.moutoussis@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
University College London, London, UK.

Matthew M. Nour (matthew.nour@psych.ox.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
University College London, London, UK.
Wellcome Trust Centre for Neuroimaging, University College London, London, United Kingdom.
Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.

Raymond J. Dolan (r.dolan@ucl.ac.uk)
Max Planck UCL Centre for Computational Psychiatry and Ageing Research,
University College London, London, UK.
Wellcome Centre for Human Neuroimaging, University College London, London, UK.

Charles Findling (charles.findling@internationalbrainlab.org)
University of Geneva, Switzerland

Tobias U. Hauser (tobias.hauser@uni-tuebingen.de) *
Department for Psychiatry and Psychotherapy, German Center for Mental Health (DZPG),
University of Tübingen, Germany.

Valentin Wyart (valentin.wyart@ens.psl.eu) *
Laboratoire de Neurosciences Cognitives et Computationnelles,
Institut National de la Santé et de la Recherche Médicale,
Département d'Études Cognitives, École Normale Supérieure,
Université PSL, Paris, France.

* : equal senior authorship

Abstract

Learning action values is key to maximize their effective payoff in uncertain reward environments. But how does dopamine affect this reinforcement learning (RL) process in humans? To test the hypothesis that increases in sustained dopamine concentration levels trigger a positive reward bias on human RL, we administered dopamine precursor L-DOPA to healthy adult volunteers performing a restless two-armed bandit task during a double-blind randomized placebo-controlled study. We found that L-DOPA decreases switching between volatile choice options. Using computational modelling, we show that L-DOPA decreases the learning rate and precision of RL but does not affect the policy used to choose between options. These learning effects of L-DOPA are best explained by a positive reward bias on recurrent neural networks (RNNs) trained to perform the same task.

Keywords: **human reinforcement learning; dopamine; computation noise; recurrent neural network**

Introduction

Dopamine has been described as crucial for reward-guided learning. The phasic mesolimbic dopamine release received abundant evidence to implement the reward prediction error of temporal difference-based reinforcement learning (TD-RL) algorithm (Schultz, 2015). By contrast, dopamine brain concentration levels have been correlated with motivation and parameters of choice policies, including exploration (Chakroun et al., 2023; Howard et al., 2017; Niv, 2007). However, recent work has shown that random noise in TD-RL explains a large fraction of the human decision variability otherwise attributed to exploration (Findling et al., 2019). To investigate possible effects of increased dopaminergic neurotransmission on human TD-RL, we administered L-DOPA to healthy adult volunteers performing a restless two-armed bandit task in a double-blind, randomized, between-subject, placebo-controlled study.

Methods and Results

Population and protocol. In total, 58 healthy participants were included in the study (n = 28 for placebo group, n = 30 for L-DOPA group; between-subject; all males, 27,75 ± 5,9 years; double-blind design). The participants reported no history of neurological or psychiatric disease, and no family history of psychotic disorders. They reported no addiction to psychoactive drugs, nor history of psychotropic medication. Before taking part in the study, all participants provided informed written consent and passed a medical check. The procedures were approved by the local ethics committee.

After ingestion of ascorbic acid (placebo; group in grey) or L-DOPA (Modopar: 150mg L-DOPA + benserazide; group in red), completion of medical checks and other tasks unrelated to this study, participants performed a restless two-armed bandit task (Fig. 1a) (96 trials/block; 2 blocks). In each trial, participants were asked to choose one of two shapes to receive its currently associated reward (1-99 points). Participants were asked to maximize their monetary payoff. They were asked to favor precision over speed, and no time limit was imposed on the latency of their responses. (Fig. 1b).

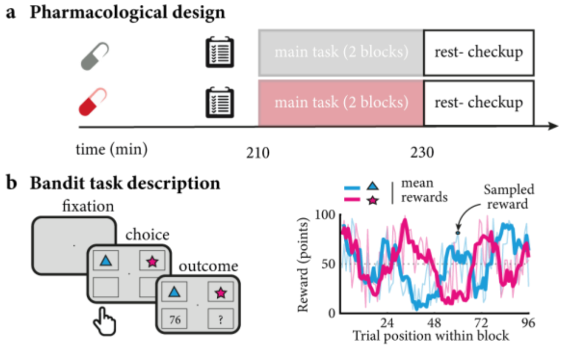

Figure 1. *Protocol and behavioral task*

Behavioral results. As expected, the probability to choose the same arm as in the previous trial grew as a function of the obtained reward (Fig. 2a; mixed-effects ANOVA, $F(7,399) = 215.2$, $p < 0.001$). Interestingly, placebo and L-DOPA groups differed with respect to this psychometric curve ($F(1,57) = 16.0$, $p < 0.001$), an effect which depended on the magnitude of the obtained reward (interaction: $F(7,399) = 3.02$, $p < 0.01$). Participants under L-DOPA repeated more their last choice than under placebo following lower-than-average rewards (rank-sum tests: $p < 0.05$, $z(57) > 2.26$, all other bins data : $z(57) < 1.81$, $p > 0.05$, $BF < 1.93$). To investigate whether this tendency to switch less (repeat more) following smaller rewards under L-DOPA is aligned with individual differences in this behavioral metric, we applied a Principal Component Analysis (PCA) on this metric for the placebo group (step #1), and reconstructed the scores of the first component (PC1, 64% expl. var.) for the L-DOPA group (step #2). Finally, we pooled the two groups and applied a median split to PC1 scores (step #3). Like L-DOPA, PC1 was associated with individual differences in the probability to repeat the last choice following smaller rewards. Moreover, PC1 scores differed significantly between the L-DOPA and placebo groups (rank-sum test between placebo and L-DOPA: $p<0.01$, $z= -2.73$) (Fig. 2a, inset)

Reinforcement Learning model. To capture the suboptimal variability of human decisions, we fitted a noisy TD-RL model (Findling et al., 2019) composed of four free parameters: (1) a learning rate α that controls the update of option values following each obtained reward; (2) a decay rate δ that controls the exponential forgetting of unchosen option values; (3) a learning noise ζ that controls the inverse precision of the TD-RL process; (4) a choice temperature τ that generates exploration through a ‘softmax’ choice policy. (Fig. 2b). To ensure that the inclusion of learning noise (controlled by its Weber fraction ζ -M1&M2) were necessary to fit participants’ choices but not asymmetry in TD-RL (positive and

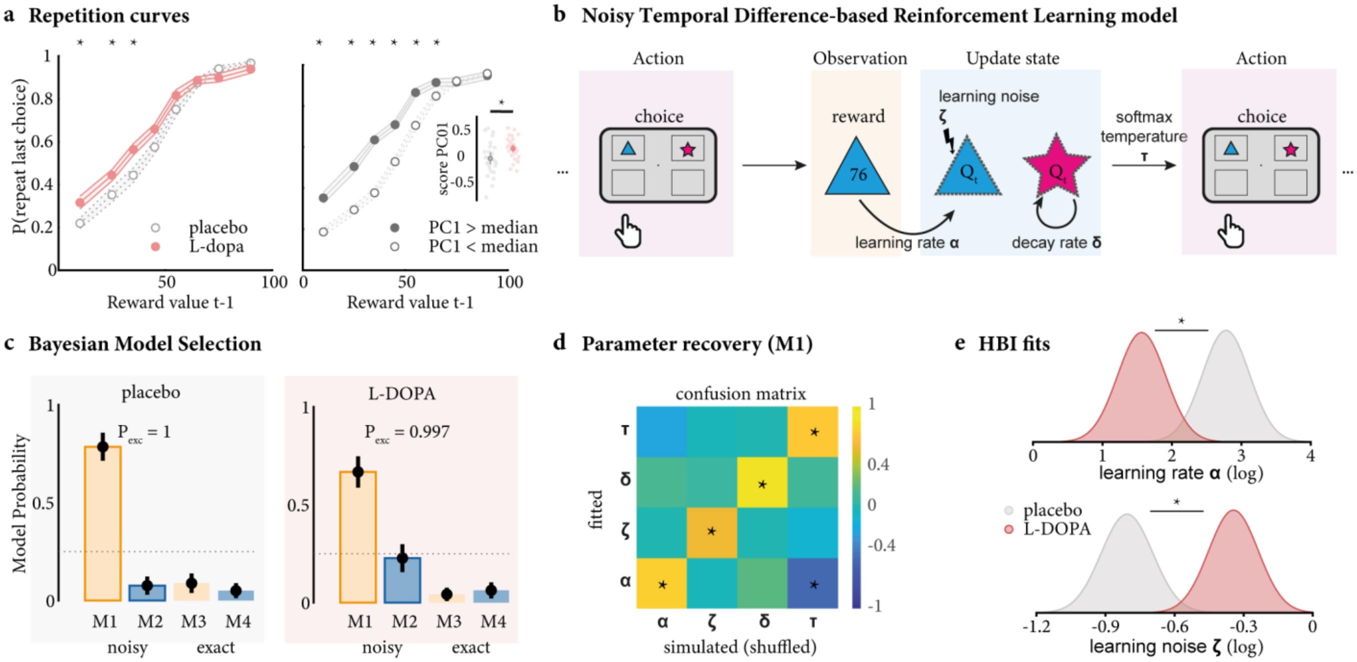

Figure 2. *Behavioral results & noisy temporal difference-based Reinforcement Learning model*

negative learning rate α added in M2&M4 (Lefebvre et al., 2017)), we performed random-effects Bayesian model selection (Rigoux et al., 2014). The first model M1 described above outperformed the other three models for both placebo and L-DOPA groups (Fig. 2c.; exceedance of P> 0.997). Critically, we performed standard parameters recovery to validate our fitting procedure of the winning model M1 (stars show significant correlations; p<0,01) (Fig. 2d).

Then, using a Hierarchical Bayesian Inference (HBI) procedure for parameter fitting at the group level (Piray et al., 2019), we observed that α is significantly lower (two-sample *t* test, p = 0.02, z(56) = -2.49) and ζ significantly higher (p <0.01, z(56) = 2.98) in the L-DOPA group compared to the placebo group. We did not find any difference across groups for δ and τ.

Recurrent Neural Networks (RNNs). Finally, we trained and tested 10 RNNs corrupted by computation noise in the recurrent layer (Findling & Wyart, 2020) on the same task as humans (Fig. 3a). We then fitted global – not structural (weights) – parameters of the trained RNNs to human behavior, including two key parameters: (1) an input *bias* (β_{in}) that controls the reward received by the RNN as input; (2) an input *gain* (γ_{in}) that controls the magnitude of the reward received by the recurrent layer (Fig. 3a). Using the same HBI procedure (Piray et al., 2019), we found that L-DOPA is associated with a positive increase in input bias β_{in} (p = 0.011, z(56) = 2.63) but no change in input gain (γ_{in} : p = 0.37, z(56) = -0.91, Fig. 3b). To determine whether a positive input bias explains L-DOPA effects on TD-RL, we simulated RNNs with varying input bias β_{in}. These simulations were then fitted using the noisy TD-RL model M1 to investigate the relation between TD-RL and RNN parameters. We found that a positive input bias β_{in} on RNN computations reproduces the effect of L-DOPA on the rate and precision of TD-RL (Fig. 3c). In other terms, applying a positive reward bias at the input of the recurrent layer implementing RL decreases its learning rate and precision. Importantly, affecting the input gain did not produce the same effects (data not shown).

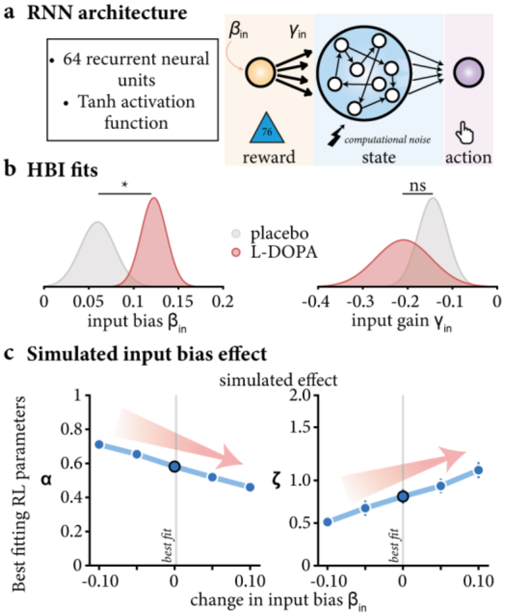

Figure 3. *Recurrent Neural Networks (RNNs)*

Conclusion

Increases in sustained dopamine concentration levels decrease human switching between volatile choice options, especially following smaller rewards, by decreasing the rate and precision of human reinforcement learning. These learning effects of L-DOPA are best explained by a positive reward bias in RNNs trained in the same conditions.

Acknowledgments

AZ is supported by the Fondation pour la Recherche Médicale (FRM) (SPF202209015740). VW has received funding from the European Research Council (grant agreement ERC-StG-759341). TUH is supported by a Sir Henry Dale Fellowship (211155/Z/18/Z; 211155/Z/18/B; 224051/Z/21) from Wellcome & Royal Society, a grant from the Jacobs Foundation (2017-1261-04), the Medical Research Foundation, a 2018 NARSAD Young Investigator grant (27023) from the Brain & Behavior Research Foundation, and a Philip Leverhulme Prize from the Leverhulme Trust (PLP-2021-040). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 946055). TUH is also supported by the Carl-Zeiss-Stiftung. TUH consults for limbic ltd and holds shares in the company, which is unrelated to the current project.

References

Chakroun, K., Wiehler, A., Wagner, B., Mathar, D., Ganzer, F., van Eimeren, T., Sommer, T., & Peters, J. (2023). Dopamine regulates decision thresholds in human reinforcement learning in males. *Nature Communications 2023 14:1*, *14*(1), 1–14. https://doi.org/10.1038/s41467-023-41130-y

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9

Findling, C., & Wyart, V. (2020). Computation noise promotes cognitive resilience to adverse conditions during decision-making. *BioRxiv*, 2020.06.10.145300. https://doi.org/10.1101/2020.06.10.145300

Howard, C. D., Li, H., Geddes, C. E., & Jin, X. (2017). Dynamic Nigrostriatal Dopamine Biases Action Selection Article Dynamic Nigrostriatal Dopamine Biases Action Selection. *Neuron*, *93*. https://doi.org/10.1016/j.neuron.2017.02.029

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9. https://doi.org/10.1038/S41562-017-0067

Niv, Y. (2007). Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and motivation? *Annals of the New York Academy of Sciences*, *1104*, 357–376. https://doi.org/10.1196/annals.1390.018

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). *Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies*. https://doi.org/10.1371/journal.pcbi.1007043

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews*, *95*(3), 853–951. https://doi.org/10.1152/physrev.00023.2014