Social context affects policy and counterfactual learning in a two-armed bandit task

Amric Trudel* (amric.trudel@ens.psl.eu)
Laboratoire de Neurosciences Cognitives et Computationnelles,
École normale supérieure, Université PSL, INSERM, 75005 Paris, France

Centre de recherche en Épidémiologie et Santé des Populations
Université Paris-Saclay, UVSQ, INSERM, 94807 Villejuif, France

Bence C. Farkas* (bence.farkas@universite-paris-saclay.fr)
Direction Générale Adjointe Enfance Famille Santé,
Conseil Départemental des Yvelines, 78012 Versailles, France

Centre de recherche en Épidémiologie et Santé des Populations
Université Paris-Saclay, UVSQ, INSERM, 94807 Villejuif, France

Laboratoire de Neurosciences Cognitives et Computationnelles,
École normale supérieure, Université PSL, INSERM, 75005 Paris, France

Pierre O. Jacquet† (pjacquet@yvelines.fr)
Direction Générale Adjointe Enfance Famille Santé,
Conseil Départemental des Yvelines, 78012 Versailles, France

Centre de recherche en Épidémiologie et Santé des Populations
Université Paris-Saclay, UVSQ, INSERM, 94807 Villejuif, France

Laboratoire de Neurosciences Cognitives et Computationnelles,
École normale supérieure, Université PSL, INSERM, 75005 Paris, France

Valentin Wyart† (valentin.wyart@ens.psl.eu)
Laboratoire de Neurosciences Cognitives et Computationnelles,
École normale supérieure, Université PSL, INSERM, 75005 Paris, France

Direction Générale Adjointe Enfance Famille Santé,
Conseil Départemental des Yvelines, 78012 Versailles, France

*shared first authorship
†shared senior authorship

Abstract

Social contexts shape decision making under uncertainty. While reinforcement learning models explain how individuals learn from rewards, it remains unclear how this process is affected when rewards come from other agents. Here we introduce a novel social decision-making task based on the two-armed bandit paradigm, which allows the isolation of social learning mechanisms while keeping reward structure exactly matched between social and non-social contexts. Model-free analyses revealed increased behavioral switching and reduced blind repetition in the social condition. A reinforcement learning model incorporating a counterfactual learning parameter revealed that the social context primarily altered policy parameters and counterfactual updating, suggesting participants imagined a competitive dynamic between their partners. Our findings indicate that while learning mechanisms about chosen options remain relatively stable across conditions, social framing seems to shift how people explore and infer from unchosen outcomes. This advances our understanding of the cognitive architecture underpinning human social learning.

Keywords: behavior, social cognition, computational modeling, reinforcement learning

Introduction

Reinforcement learning models accurately describe the dynamics of reward learning and decision-making in both artificial and biological agents, including humans (Behrens et al., 2007; Dayan & Niv, 2008; Sutton & Barto, 2018) . In this work, we study how a social context, where subjects perceive the rewards as coming from a human agent, affects these learning and decision-making processes (Behrens et al., 2008). State-of-the-art computational models use parameters that account for the speed of update of the estimates, the exploration strategy (Schulz & Gershman, 2019), and more recently learning precision (Findling et al., 2019; Lee et al., 2023). We propose a model with a new parameter that accounts for counterfactual thinking about the unchosen option in a way that would be mostly plausible in a social context. As the current lack of appropriate non-social control conditions is an issue in the field of social cognition (Lockwood et al., 2020), we propose an experimental paradigm based on the restless two-armed bandit task with perfectly matched reward structure between social and non-social conditions.

Methods and Results

Task

We designed a repeated social cooperative game with three players, based on the two-armed bandit paradigm. At every trial, a **Chooser** decides to share an endowment of 100 points with one of his/her two **Partners**, who then decides how many points to return to the Chooser. The task comprises three conditions, each with two blocks of 64 trials. The **Non-Social Chooser** (NSC) condition is a regular restless two-armed bandit task (Findling et al., 2019). The **Social Chooser** (SC) condition has the exact same reward structure as its non-social counterpart, except that participants are led to believe that they are playing with other humans rather than with slot machines. The **Partner** (P) condition allows the participants to also play the role of the Partner and augments the credibility of the game. This task design allows a direct, within-subject comparison of identical social vs non-social decision tasks. Participants were randomly assigned to a block order that was either [SC, P, SC, P, NSC, NSC] or [P, SC, P, SC, NSC, NSC].

Model-free results

A logistic regression was fit to each subject (N=98) to characterize the probability of repeating any action given the reward it gave (Figure 1b). We used a three-parameter sigmoid function with (1) a **threshold** (2) a **slope**, and (3) a **base rate**, which captures the blind repetition rate. We found that the social condition was associated with a higher threshold (NSC: 0.387, SC: 0.444, $p<0.001$, $d=0.54$),

a lower slope (NSC: 14.6, SC: 10.0, p<0.001, d=0.50), and a lower base rate (NSC: 0.254, SC: 0.183, p<0.001, d=0.51). This indicates that subjects have a higher tendency to switch actions in the social condition, but we need finer modelling to tell if this is a learning or a policy effect.

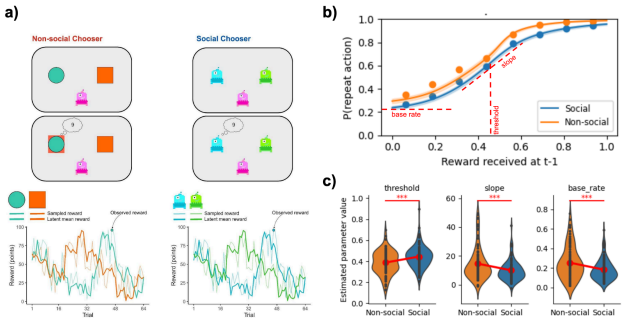

Figure 1: The Social Bandit task design and model-free results. *(a) Illustration of the Non-Social (NSC) and Social (SC) Chooser conditions, using the same reward trajectories. (b) Action repetition curve: probability of repeating the last action as a function of the last reward received (reward = points / 100). (c) Comparison of the action repetition curve parameters between the NSC and SC conditions.*

Computational modeling

We derived a Kalman filter model (Figure 2a), with five free parameters: a **learning rate (α)** that mediates the update of the values of each action; a **learning noise (ζ)** that regulates the Weber noise added to action-value updates; the **choice temperature (τ)** that tunes the level of exploration in the softmax policy; and the **blind repetition rate (ε)**. We added a **counterfactual learning rate (δ)** to account for how subjects update their estimate of the unchosen option at time *t*, where they realign it towards the reward received with the chosen option. This mechanism can be expected especially in the social condition, where a competition dynamic can be imagined between the two partners. Parameter recovery was adequate (Figure 2b).

Model parameters were fit for each subject, in each condition. Paired t-tests revealed that the social condition resulted in a higher reinforcement learning rate (α) (NSC: 0.690, SC: 0.739, p=0.03, d=0.25), a higher counterfactual learning rate (δ) (NSC: 0.059, SC: 0.090, p<0.001, d=0.50), a higher choice temperature (τ) (NSC: 0.081, SC: 0.118, p<0.001, d=0.69) and a lower blind repetition rate (ε) (NSC: 0.137, SC: 0.085, p<0.001, d=0.49). This suggests that conditions have a small effect on conventional reinforcement learning parameters (α, ζ), but a strong effect on counterfactual learning (δ) and choice policy parameters (τ, ε).

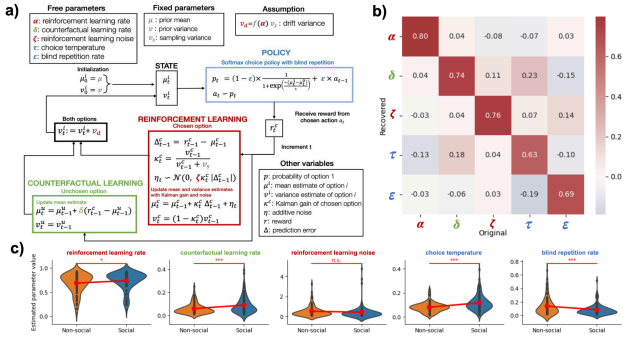

Figure 2: Computational modeling. *a) Description of the reinforcement learning model using Kalman filters: reinforcement learning parameters are in red, counterfactual learning in green, and policy in blue. b) Parameter recovery: correlation between the original parameters of the simulating models (columns) and the recovered parameters from the fitted models (rows). c) Comparison of the estimated parameter values across conditions.*

Conclusion

We provide evidence that a social setting gives rise to higher random exploration and lower blind repetition. We observed much smaller effects for value updates themselves, but our results suggest the existence of a counterfactual learning effect where participants in a social condition were more prone to see the partners as competitors who try to match the other's reward when they are not chosen.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche, grant ANR-22-CE28-0012-01 eLIFUN (JCJC) and a institutional grant awarded to the Département d'Études Cognitives of ENS-PSL (ANR-17-EURE-0017, EUR FrontCog).

References

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W.,& Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

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. https://doi.org/10.1038/nn1954

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

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

Lee, J. K., Rouault, M., & Wyart, V. (2023). Adaptive tuning of human learning and choice variability to unexpected uncertainty. *SCIENCE ADVANCES*.

Lockwood, P. L., Apps, M. A. J., & Chang, S. W. C. (2020). Is There a 'Social' Brain? Implementations and Algorithms. *Trends in Cognitive Sciences*, *24*(10), 802–813. https://doi.org/10.1016/j.tics.2020.06.011

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. A Bradford Book.