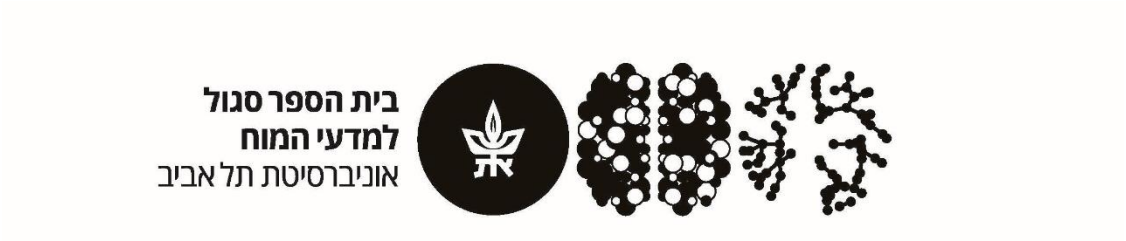


Sagol School of Neuroscience

The School of Psychological Sciences,

Gershon H. Gordon Faculty of Social Sciences

# Multidimensionality in human reinforcement learning: The dynamics of credit assignment to outcome-irrelevant task representations

By

**Inbal Alon**

**311148126**

The thesis was carried out under the supervision of

**Dr. Nitzan Shahar**

September 2021

# Acknowledgments

I thank my supervisor Dr. Nitzan Shahar for his invaluable advice and support during my MSc study. I would also like to thank all the wonderful lab colleagues I had the privilege to work with, and especially Dr. Maayan Pereg, Shay Kahan and Yaniv Davidi for their dedicated help with the design, development, and operation of the experiment. I am also thankful to the Sagol School of Neuroscience and its staff members for the opportunity, the support, and the academic guidance. Finally, I would like to express my gratitude to all my loved ones for their love, faith, and encouragement during the past few years. Thank you all.

# Abstract

**Background:** Goal-directed behavior can be acquired through trial and error, while holding minimal knowledge regarding the structure of the environment. Using reward feedback, humans create stimulus-action-outcome associations that help them predict future events and adapt their behavior accordingly, an ability that is assumed to rely on attentional control mechanisms. Recent studies suggest that credit might also be assigned to task representations that are irrelevant for the prediction of reward. The current study is aimed at further exploring this phenomenon.

**Method:** In the experiment, 203 participants performed a multiple-armed bandit task of four cards, each leading to reward on a drifting probability across trials. On each trial, participants were offered two randomly selected cards. They were asked to pick one out to gain, or to avoid loss of, monetary reward. Importantly, only the identity of the chosen card predicted reward and reward probability of each card was independent from the others. We assessed credit assignment to task components that do not predict reward using both sequential trial analysis and reinforcement learning modeling.

**Results:** We replicate previous findings by showing both outcome-irrelevant credit assignment to motor task representations and counterfactual credit assignment. We found that received credit was not only assigned to the chosen card, but also to the motor response used to select it. In addition, inverse credit was assigned to an unchosen card. We further show that not all unchosen cards are assigned with credit, but only unchosen cards that took part in the deliberation process.

**Conclusions:** We suggest that our findings reflect a control deficit in human reinforcement learning and hypothesize that, to reduce cognitive overloads, the learning system segregates active task representations based on premotor activity. That is, representations associated with motor execution are assigned with the received credit, while those associated with motor inhibition are assigned with the inverse credit.

# Table of Contents

# Introduction

Goal-directed behavior can be acquired through trial and error, while holding minimal knowledge regarding the structure of the environment. Using reward feedback, humans create stimulus-action-outcome associations that help them predict future events and adapt their behavior accordingly (1-5). Neurophysiological experiments show that sensory events that happen immediately after an action can affect the output of midbrain dopaminergic neurons that target cells in the striatum, tuning their sensitivity to descending cortical inputs (6,7). In turn, this dopamine-dependent cortico-striatal plasticity is thought to impact how the basal ganglia modulate cortical activity during action selection, by influencing the activity of thalamocortical connections (8-10).

Computational reinforcement learning (RL) studies offer a principled theory regarding this functionality. In line with many behavioral and physiological findings, they demonstrate how the dopamine signal conveyed by midbrain neurons is a physiological instantiation of a reward prediction error: the difference between the expected reward and the received reward (11-17). Learning is therefore conceptualized in model-free RL models by updating 'state representations', future reward values assigned to stimuli based on reward history, using a temporal difference algorithm. During decision-making, these values are used to favor previously rewarded actions (18).

One of the major challenges for this view, known as the structural 'credit assignment problem', is explaining how the network "knows" which representations should be reinforced when multiple features are concurrently present. Many studies suggest that credit is assigned to stimuli at the level of features (19-23). However, natural environments are complex and dynamic, and every stimulus can be considered by many features (i.e., color, shape, location, size) or even by combinations of features (e.g., a red flower, a square located to the right, etc.). On the other hand, working memory capacity is highly limited, allowing only a subset of available observations to be actively held in mind during learning and decision-making (24-29)

Addressing this problem, Niv et al. (2015) (30) studied participants' choice behavior in a multidimensional three-armed bandit task with probabilistic rewards. They found that a

model that combines value-based learning at the level of features with a decay mechanism that, regardless of reward, gradually decreases feature values of unselected options to 0, explained up to 70% of the choice variance. Their model successfully predicted participants' behavior even in games in which they performed randomly, suggesting that it captured meaningful aspects of the underlying learning and decision-making mechanisms. The extent to which participants were engaged in this learning was implicated in an attentional control network composed of the interparietal sulcus (IPS), precuneus and the dorsolateral prefrontal cortex (DLPFC).

In line with the neuronal findings of Niv et al. (2015) (30), studies show that during action selection the IPS, a cortical area associated with endogenous attention and visual feature search (31-34), encodes reward-related behavioral representations of presented stimuli (35-37). Connections between the IPS, the lateral orbitofrontal cortex (OFC) and the putamen signal the respective relevancy of behavioral task representations (38). In turn, neurons in the DLPFC, an area associated with cognitive control, were found to adaptively increase their activation in response to increasing working memory load and to respond selectively to attended perceptual dimensions of complex stimuli (39-50).

Although Niv et al. (2015) (30) offer a substantial evidence for an attentional control mechanism that reduces task dimensions during decision-making, they do not test the possibility that attentional control is applied during the learning stage. The contribution of reward prediction errors, putatively conveyed by midbrain dopaminergic projections, to the function of the DLPFC and IPS is yet to be explained. Earlier attentional RL models, on the other hand, suggested that teaching signals are weighted based on the learned predictive values of features, rather than through decaying the weights of unchosen features. In these studies, attentional weights were applied to the prediction error signal itself (51, 52).

In a more recent study, Shahar et al. (2019) (53) used a model that assigns values to outcome-irrelevant motor task representations and applies weights to them. Both model-agnostic and computational analyses showed that, in addition to an assignment to outcome-relevant features of the task, the model-free system assigns independent credit to motor outcome-irrelevant task representations. Moreover, the tendency to assign credit to outcome-irrelevant motor task representations was in negative correlation with the deployment of model-based strategies, implying that when working memory resources are

low, the ability to keep an accurate intrinsic- vs. extrinsic-state differentiation between task representations is attenuated, leading to higher involvement of outcome-irrelevant task representations on value-based choices. They suggested that the learning system has a requirement to segregate task representations (assumed to be actively held in working memory) into state-extrinsic/intrinsic information according to outcome relevancy.

Niv et al. (2015) (30) do not directly test how the model-free system deals with spatial-motor task representations of chosen stimuli. However, in line with Shahar et al. (2019) (53), their findings might imply that these representations, although irrelevant to the prediction of reward, are also highlighted during action selection.

Lastly, many studies show that the alternative option also plays a role in the decision-making process (72-75). When choosing between two alternatives, no (or very little) information is gained regarding the outcomes of the unchosen action. Nonetheless, it has been shown that people believe in an Illusory negative correlation between the outcomes of independent choice options, which leads them to irrational choices (54-56). This might imply that even when no new information is gained regarding the outcomes of an unchosen action, inverse credit is assigned to it.

In the current study, we aimed at further examine the dynamics of credit assignment to task representations that are actively held in working memory during decision-making, but do not predict reward. We analyzed participants' choice behavior in a multiple-armed bandit task of four fractal cards with probabilistic rewards, wherein only two cards were offered to the participant on each trial. This unique task setting allowed us to explore two types of credit assignment:

***Credit Assignment to motor task representations.*** Card choice itself, but not the motor effector response used to report it, predicted monetary outcome. This was emphasized in the instructions of the task. Following Shahar et al. (2019) (53), we hypothesized that credit could either be assigned to motor task representations independently (that is, to the keypress separately from the card), or as combination of card and keypress (for example, to Card A with right keypress). We assessed credit assignment to motor task representations using both sequential trial analysis and reinforcement learning modeling.

***Counterfactual Credit Assignment*.** Reward probability of the chosen card was independent of the rest of the cards. Therefore, received outcome did not reveal anything regarding the values of unchosen cards. In accordance with the findings of Marciano-Romm et al. (2016) (55) and Marciano (2019) (56), we hypothesized that inverse credit will be assigned to the offered card that was not selected. We assessed counterfactual credit assignment in a sequential trial analysis.

# Methods

**Participants.** The study was approved by the ethics committee of Tel Aviv University. Written informed consent was given for all participants. Our sample included 203 participants in total and 178 subjects after exclusion (76 females, 101 males, and one unspecified gender; mean age 26.08, range 18 to 51). The recruitment and participation were held online in English, via Prolific platform. Participation restrictions were age, normal or corrected vision, and no neurological or psychiatric diagnosis.

**Procedure.** Participants completed a multiple-armed bandit task of four cards and two reward conditions. Each card could lead to reward on an independent drifting probability across trials (Figure 1). Two cards were offered on each trial, and participants were asked to pick one out to gain or to avoid loss of a reward. The task was preceded by instructions, a short practice session and a test. At the end of the experiment, participants were paid a fixed amount (£2.5) plus a bonus (of £1 or £1.5) based on their performance.

**Reinforcement Learning Task.** Here, participants were asked to play a card game. The task started with written instructions accompanied by visual examples, a short practice session and a test to make sure all instructions were well understood. Then participants were announced that they were given "*100 game-coins to start with*". Each trial was preceded by 1 sec fixation point presented in the middle of the screen. Then two abstract fractal images were randomly assigned to one of two positions on the screen and were presented until response. If no response was received after 6 sec, the two stimuli disappeared and a message that asked the participant to respond faster appeared on the screen. Participants indicated a fractal choice by pressing the key that corresponds to the position of the fractal. After a choice has been made, the chosen card remained on the screen for 500 ms, while the other choice option disappeared. Then an outcome (1, 0, or -1 game-coin) appeared in the middle of the screen for 1 sec.

The task included four blocks of 50 trials, two of each reward condition. Every block started with a screen that introduced the subject with four cards and two possible outcomes. Positive blocks, in which the subject could either gain one game-coin or no-coin, were indicated by a green frame. Negative blocks, in which the outcome could either be a loss of

a game-coin or no-coin, were indicated by a red frame. Additional task information can be found in the Appendix.

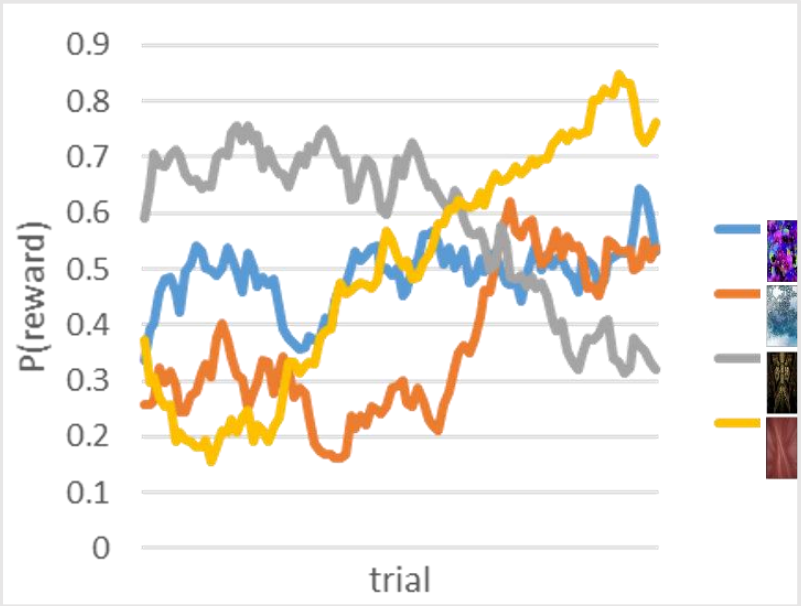


**Figure 1.** Reward probability of each card was drifted randomly across trials using a random walk algorithm.

**Computational Reinforcement Learning Models.** We fit three Q-Learning models to participants' choice behavior, wherein the value of each choice at each trial was predicted based on reward history. Q values of each choice were updated according to a prediction error signal using a temporal difference algorithm (12, 18). At the beginning of each game block, all Q values are initialized at 0. To select one of the two presented stimuli (S1, S2) on each trial n, their $Q_{net}$ values are entered into a Softmax probabilistic choice function as follows:

[1] P(choose $S_i$) = $\frac{e^{\beta Qnet(Si)}}{\sum_{j=1}^{2} e^{\beta Qnet(Sj)}}$

Where $\beta$ is the inverse temperature parameter that sets the level of noise in the decision process, with large $\beta$ corresponding to low decision noise and near-deterministic choice of the highest-value option, and small $\beta$ corresponding to high decision noise and nearly

random decisions. Each model differs from the others with respect to how motor elements of the task are integrated inside the $Q_{net}$ value of each choice.

***Model 1 (null model).*** Here, only card identity values were learned:

[1] $Q_{net} = Q_{card}$

After choosing a card and observing the received reward ($R^n \in \{0,1\}$ in positive conditions, $R^n \in \{-1,0\}$ in negative conditions), the value of the chosen card was updated according to a prediction error signal as follows:

[2] $Q_{card}(S_{chosen})^{n+1} = Q_{card}(S_{chosen})^{n} + \alpha(R^n - Q_{card}(S_{chosen})^{n})$

where α is the learning rate parameter, reflecting the rate at which new information is accumulated during the learning process. Therefore, this model has two free parameters, Θ = {α, $\beta$ }, which we fit to each participant separately from his data.

***Model 2 (Generalized motor credit assignment).*** Here, in addition to card identity values, we integrated separate values for each of the two available keypresses:

[3] $Qnet = Q_{card} + w*Q_{keypress}$

where w is a free parameter reflecting the weight the learner gives to outcome-irrelevant motor task elements. Thus, for Model 2, we also updated keypress values at the end of each trial as follows:

[4] $Q_{keypress}(S_{chosen})^{n+1} = Q_{keypress}(S_{chosen})^{n} + \alpha(R^n - Q_{keypress}(S_{chosen})^{n})$

This model has three free parameters, Θ = {α, $\beta, w$}.

***Model 3 (Combined motor credit assignment).*** Here, outcome-irrelevant motor values were assumed to reflect over-specific representations of the cards. Therefore, in addition to card identity values, we integrated separate values for each pairing of card and keypress as follows:

[5] $Qnet = Q_{card} + w_{comb}*Q_{card:keypress}$

where $w_{comb}$ indicates the weight the learner gives to over-specific card representations. That is, to the card when it is mapped to a specific keypress.

[6] $Q_{card:keypress}(S_{chosen})^{n+1} = Q_{card:keypress}(S_{chosen})^{n} + \alpha(R^{n} - Q_{card:keypress}(S_{chosen})^{n})$

This model has three free parameters, $\Theta = \{\alpha, \beta, w_{comb}\}$.

# Results

We studied a sample of 178 healthy subjects (ages 18 to 51 years) who had completed a multi-armed bandit task of 4 cards with drifting reward probabilities across trials. On each trial, two cards were randomly allocated to the right/left side of the screen. Participants were instructed to use a corresponding right/left key press to indicate a card selection to gain, or to avoid the loss of, monetary outcome. We were interested in participants' tendency to assign credit to task representations that are active during action selection but do not predict reward.

**Consecutive Trial Analysis.** We evaluated the tendency to assign credit to three task associations using model-agnostic measures. We examined whether reward history affected selection probability on the next trial (n + 1) in three consecutive trial analyses, as follows:

**Chosen card.** We analyzed choice probabilities in trials where the chosen card on trial *n* was reoffered on trial *n+1* and, importantly, the unchosen card was not reoffered. We calculated a mixed effect logistic regression (57) where previous outcome (rewarded vs. unrewarded), outcome condition (positive vs. negative reinforcement) and their paired interaction were entered as fixed effects, and subject as random effect, predicting the probability to select the chosen card from the preceded trial. We found a statistically significant main effect of reward ($p<0.001$) and outcome x condition interaction effect ($p<0.05$), showing a larger effect of reward on card stay probability in positive reinforcement blocks. Although chosen card alone predicted reward in the task, regardless of condition, reward increased card stay probability on the *n+1* trial by 24.19% on average in positive reinforcement blocks, compared to 20.63% in negative reinforcement blocks.

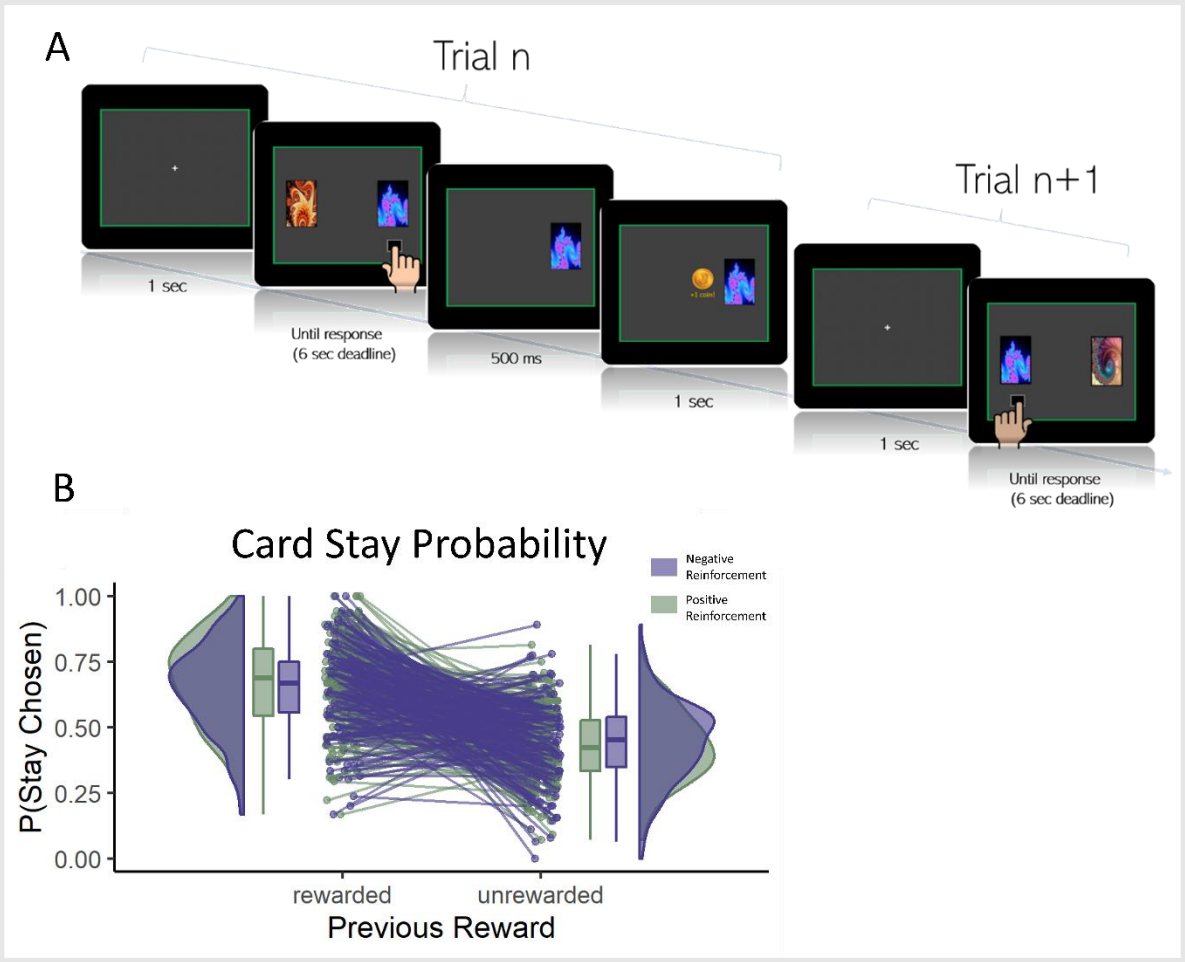


**Figure 2.** Effect of reward on the chosen card. **(A)** We analyzed the chances of choosing a previously chosen card as a function of previous outcome. **(B)** We found a substantial reward effect ($p<0.001$) showing greater chances of repeating a choice when it was rewarded compared to when it was unrewarded and an outcome x condition interaction effect ($p<0.05$) with larger reward effect on card stay probability in positive vs. negative reinforcement blocks.

**Keypress.** We analyzed choice probabilities in trials where both chosen and unchosen cards from trial *n* were not reoffered on trial *n+1*. We calculated a mixed effect logistic regression where previous outcome (rewarded vs. unrewarded), condition (positive vs. negative reinforcement) and their paired interaction were entered as fixed effects, and subject as random effect, predicting the probability to use the same keypress from the preceded trial. We found a statistically significant main effect of reward ($p<0.01$) and no outcome x condition interaction. Similarly in both block conditions, reward increased keypress

selection probability on the *n+1* trial by 5.77% on average, suggesting that credit for reward was assigned to some extent to the motor response used to indicate a card choice.

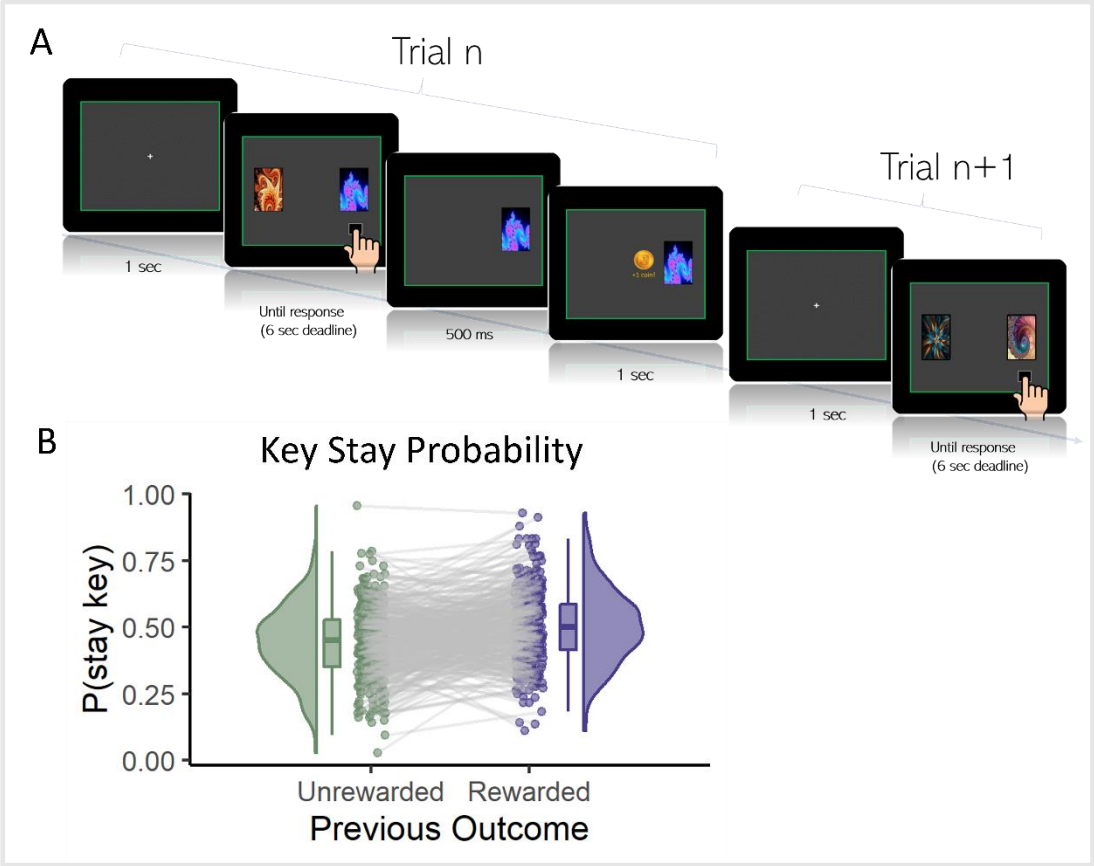


**Figure 3.** Effect of reward on the chosen keypress. **(A)** We analyzed the chances of choosing a previously chosen keypress as a function of previous outcome. **(B)** We found a substantial reward effect ($p<0.01$) showing greater chances of repeating a keypress following rewarded vs. unrewarded trials.

**Unchosen Card.** We analyzed choice probabilities in trials where the unchosen card on trial *n* was reoffered on trial *n+1* and, importantly, the chosen card was not reoffered. We calculated a mixed effect logistic regression where previous outcome (rewarded vs. unrewarded), condition (positive vs. negative reinforcement) and their paired interaction were entered as fixed effects, and subject as random effect, predicting the probability to select the unchosen card from the preceded trial. We found a statistically significant main reward effect (, $p<0.001$) and no outcome x condition interaction. Similarly in both block conditions, reward decreased unchosen card selection probability on the *n+1* trial by 5.23% on average.

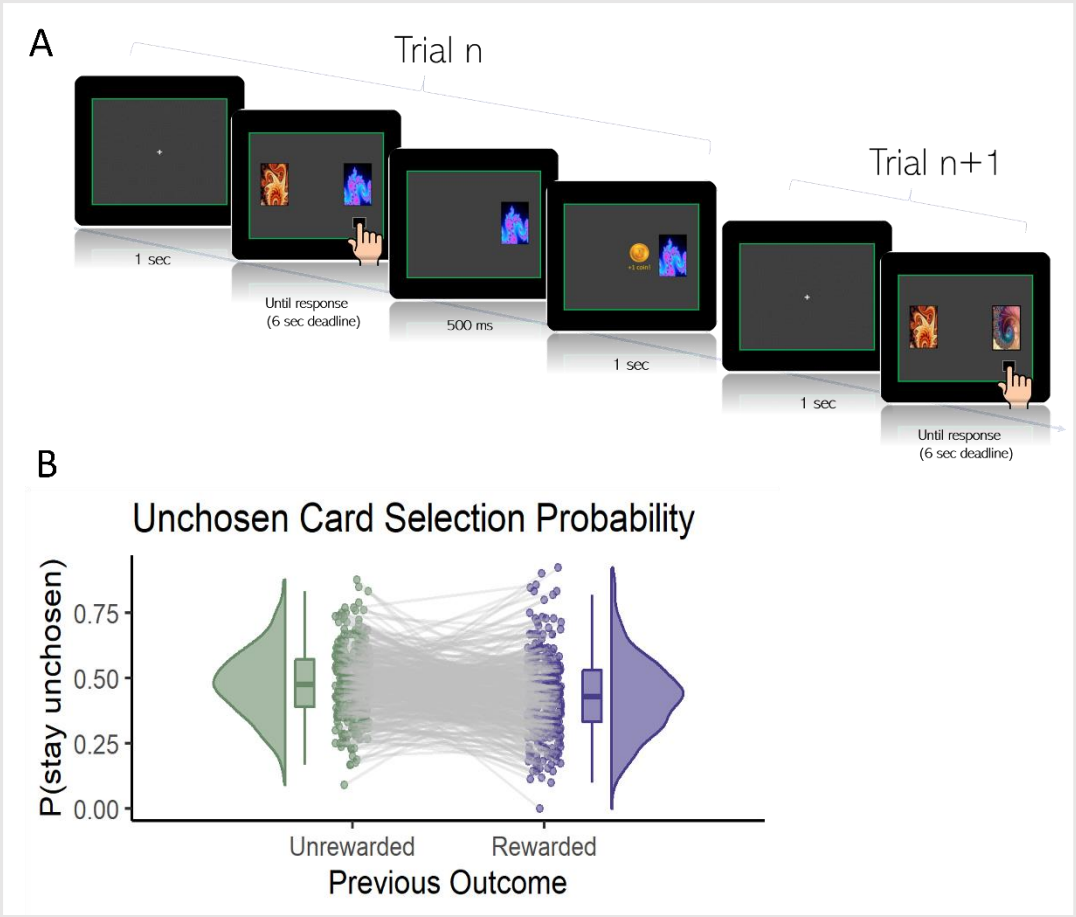


**Figure 4.** Effect of reward on the unchosen card. **(A)** We analyzed the chances of choosing a previously unchosen card as a function of previous outcome. **(B)** We found a substantial reward effect ($p<0.001$) showing lower chances of selecting an unchosen card following a rewarded vs. unrewarded trial.

**Computational Modeling.** We used computational modeling to further explore the learning process of outcome-irrelevant motor task representations. We fit 3 computational reinforcement learning models to participants' choice behavior:

1) **Model 1** ("null model") did not include outcome-irrelevant learning. Only card values were integrated in the decision-making process.

2) **Model 2** included the same card learning as Model 1, with additional separate state action values for motor response learning.

3) **Model 3** had the same fractal learning as Model 1 but assumed that the model-free system also holds separate state action values for each pairing of card and motor response.

Therefore, separate state action values for each combination of card and keypress were also integrated in the decision-making process.

We performed a model-fitting procedure using Laplace approximation method (65). We calculated the Bayesian Inference Criteria (BIC) score of each participant for each model. Then, we compared models' fit by averaging the BIC scores of each model. The BIC is a criterion for model selection among a finite set of models, which penalizes the number of model parameters to avoid overfitting. Lower scores (with 10 points difference or more) are considered strong evidence for better fit (66). All three models had similar BIC scores with best fit to the null model ($\Delta BIC_{int} < 4.5$, see table 1), suggesting that the addition of motor task representations did not improve models' ability to predict participants' choices.

| | Model1 | Model2 | Model3 |
|---|---|---|---|
| *Mean BIC* | 255.34 | 257.75 | 259.79 |

**Table 1.** We calculated the Bayesian Inference Criteria (BIC) score of each participant for each model. We compared models' fit by averaging the BIC scores of each model. All three models had similar BIC scores, suggesting that the addition of motor task representations did not improve models' ability to predict participants' choices.

# Discussion

Human cognition is marked by a sophisticated ability to attribute outcomes and events to previous choices. Understanding the mechanisms that govern this ability has been a major focus in the field of cognitive neuroscience. While many studies show that attentional control mechanisms contribute greatly to this ability by reducing task dimensions to only those which predict the outcome (30-52), other findings suggest that credit is also assigned to task associations that do not predict the outcome (53-56).

In the current experiment, we studied participants' choice behavior in a reinforcement learning task. The task included four cards, each leading to reward with an independent drifting probability. Importantly, only two of the cards appeared on the screen on each trial, placed randomly in one of two locations on the screen. Participants were asked to pick one out as fast as possible by pressing the key that corresponds to that location. Therefore, to succeed in the task, participants had to keep track of the changing reward probabilities of all four cards, while paying attention to the two available cards and the keypress each is mapped to on each trial. After a choice has been made, the chosen card alone remained on the screen, and the participants were introduced with the received outcome.

First, we replicate the findings of Shahar et al. (2019) (53) by showing credit assignment to outcome-irrelevant motor task representations. Even though participants were told that only the card they choose predicts the outcome, independently of the motor response used to report it, our model-agnostic results suggest that credit was also assigned to the keypress.

Second, we replicate the findings of Marciano-Romm et al. (2016) (55) and Marciano (2019) (56), by showing both factual and counterfactual credit assignment to the cards. Although only the outcome of the chosen card was revealed (and no new information was gained regarding the values of unchosen cards), our model-agnostic results demonstrate that participants learned through double updating: while credit for received outcome was assigned to the chosen card, inverse credit was assigned to an unchosen card.

Notably, the current study further shows that not all unchosen cards are assigned with inverse credit, but only the one that was attended in the preceded trial. If participants had

believed that there exists a negative correlation between the outcomes of the cards (55-56), we would expect that inverse credit will be assigned equally to all unchosen cards. However, it seems that only the values of cards that took part in the deliberation process were updated, which points to the involvement attentional control mechanisms in this phenomena. Therefore, we suggest that our findings reflect a control deficit in human reinforcement learning, rather than a mere belief in anti-correlations in the environment. We speculate that under cognitive overloads, the ability to hold in mind accurate action-outcome associations deteriorates, allowing task components that do not predict reward to be considered as part of the higher-level context that led to reward.

During the deliberation stage of two options and until an action is selected, it has been shown that cortical units that correspond to both options are activated. After action selection and until motor execution, neural activations of the unchosen action are inhibited, while activity in neurons that correspond to the chosen action is maintained and even amplified. This was found in dorsal premotor (PMd) neurons involved in motor planning and in other cortical areas (58-61) and supported by psychophysical experiments (62- 64).

Taken together with both of our findings, it might be that to reduce cognitive overloads, the learning system segregates active task representations based on premotor activity. That is, representations associated with motor execution are assigned with the received credit, while those associated with motor inhibition are assigned with the inverse credit. This explanation is in line with other finding that suggested that motor demands constrain the creation of abstract rule making (71).

There are limitations to our study as well. First, the motor response used to select a card is mapped to its location on the screen. Therefore, the spatial location and the motor response are perfectly confounded in our task. Future work should be made to be able to differentiate between these two dimensions. Second, our model-agnostic analyses only take into account the last received outcome. However, human reinforcement learning operates over larger number of preceding observations, not just the last action performed. This could be another interesting venue for future work.

**Conclusion**

In sum, as was found in previous studies, we show both credit assignment to outcome-irrelevant motor task representations and counterfactual credit assignment so that inverse credit is assigned to an unchosen action. We further show that only attended unchosen actions are assigned with inverse credit, suggesting that this effect reflects an attentional control deficit in human reinforcement learning. Finally, we hypothesize that, to reduce cognitive overloads, the learning system segregates active task representations based on premotor activity. That is, representations associated with motor execution are assigned with the received credit, while those associated with motor inhibition are assigned with the inverse credit.

# References


1. Shepard RN (1987) Toward a universal law of generalization for psychological science. Science 237:1317–1323, doi:10.1126/science.3629243, pmid:3629243.

2. Jones M, Cañas F (2010) Paper presented at CogSci 2010: The Annual Meeting of the Cognitive Science Society (August, Portland, OR), Integrating reinforcement learning with models of representation learning.

3. Collins A, Koechlin E. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. PLoS Biol 2012;10:e1001293. pmid:22479152

4. Donoso M, Collins a. GE, Koechlin E. Foundations of human reasoning in the prefrontal cortex. Science (80-) 2014;1481.

5. Badre D, Kayser AS, D'Esposito M, Esposito MD. Frontal cortex and the discovery of abstract action rules. Neuron 2010;66:315–26. pmid:20435006

6. Peak, J., Hart, G., & Balleine, B. W. (2019). From learning to action: the integration of dorsal striatal input and output pathways in instrumental conditioning. *European Journal of Neuroscience*, *49*(5), 658-671.

7. Surmeier, D. J., Shen, W., Day, M., Gertler, T., Chan, S., Tian, X., & Plotkin, J. L. (2010). The role of dopamine in modulating the structure and function of striatal circuits. *Progress in brain research*, *183*, 148-167.

8. Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience*, *9*(1), 357-381.

9. DeLong, M. R., & Wichmann, T. (2007). Circuits and circuit disorders of the basal ganglia. *Archives of neurology*, *64*(1), 20-24.

10. Parent, A., & Hazrati, L. N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain research reviews*, *20*(1), 91-127.

11. Barto AG (1995) in *Models of information processing in the basal ganglia*, Adaptive critic and the basal ganglia, eds Houk JC, Davis JL, Beiser DG (MIT, Cambridge, MA), pp 215–232.

12. Sutton RS, Barto AG (1998) *Reinforcement learning: an introduction* (MIT, Cambridge, MA).

13. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci **16**:1936–1947, pmid:8774460.

14. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science **275**:1593–1599, doi:10.1126/science.275.5306.1593, pmid:9054347.

15. Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129-141.

16. Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in cognitive sciences*, *12*(7), 265-272.

17. Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154.

18. Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, 64-99.

19. Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, *84*(3), 555-579.

20. Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860-12867.

21. Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human neuroscience*, *5*, 189.

22. Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P., & Dolan, R. (2012). Serotonin selectively modulates reward value in human decision-making. *Journal of Neuroscience*, *32*(17), 5833-5842.

23. Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, *81*(1), 195-206.

24. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

25. Fuster JM. The prefrontal cortex. New York: RAVEN Press, 1980.

26. Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

27. Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*(1), 122.

28. Callicott, J. H., Mattay, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A., ... & Weinberger, D. R. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral cortex*, *9*(1), 20-26.

29. Yun, R. J., Krystal, J. H., & Mathalon, D. H. (2010). Working memory overload: fronto-limbic interactions and effects on subsequent working memory function. *Brain imaging and behavior*, *4*(1), 96-108.

30. Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145-8157.

31. Culham, J. C., & Kanwisher, N. G. (2001). Neuroimaging of cognitive functions in human parietal cortex. *Current opinion in neurobiology*, *11*(2), 157-163.

32. Chica, A. B., Bartolomeo, P., & Valero-Cabré, A. (2011). Dorsal and ventral parietal contributions to spatial orienting in the human brain. *Journal of Neuroscience*, *31*(22), 8143-8149.

33. Liu, T., Hospadaruk, L., Zhu, D. C., & Gardner, J. L. (2011). Feature-specific attentional priority signals in human cortex. *Journal of Neuroscience*, *31*(12), 4484-4495.

34. Wei, P., Müller, H. J., Pollmann, S., & Zhou, X. (2011). Neural correlates of binding features within-or cross-dimensions in visual conjunction search: an fMRI study. *Neuroimage*, *57*(1), 235-241.

35. Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233-238.

36. Peck, C. J., Jangraw, D. C., Suzuki, M., Efem, R., & Gottlieb, J. (2009). Reward modulates attention independently of action value in posterior parietal cortex. *Journal of Neuroscience*, *29*(36), 11182-11191.

37. Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *science*, *304*(5678), 1782-1787.

38. Hunt, L. T., Dolan, R. J., & Behrens, T. E. (2014). Hierarchical competitions subserving multi-attribute choice. *Nature neuroscience*, *17*(11), 1613-1622.

39. Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(1), 49-62.

40. Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B. J., Servan-Schreiber, D., & Noll, D. C. (1994). Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI. *Human brain mapping*, *1*(4), 293-304.

41. Owen, A. M., Roberts, A. C., Polkey, C. E., Sahakian, B. J., & Robbins, T. W. (1991). Extra-dimensional versus intra-dimensional set shifting performance following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia*, *29*(10), 993-1006.

42. Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. *Nature*, *380*(6569), 69-72.

43. Dias, R., Robbins, T. W., & Roberts, A. C. (1996). Primate analogue of the Wisconsin Card Sorting Test: effects of excitotoxic lesions of the prefrontal cortex in the marmoset. *Behavioral neuroscience*, *110*(5), 872.

44. Dias, R., Robbins, T. W., & Roberts, A. C. (1997). Dissociable forms of inhibitory control within prefrontal cortex with an analog of the Wisconsin Card Sort Test: restriction to novel situations and independence from “on-line” processing. *Journal of Neuroscience*, *17*(23), 9285-9297.

45. Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, *393*(6685), 577-579.

46. Birrell, J. M., & Brown, V. J. (2000). Medial frontal cortex mediates perceptual attentional set shifting in the rat. *Journal of Neuroscience*, *20*(11), 4320-4324.
47. Everling, S., Tinsley, C. J., Gaffan, D., & Duncan, J. (2002). Filtering of neural signals by focused attention in the monkey prefrontal cortex. *Nature neuroscience*, *5*(7), 671-676.

48. Lebedev, M. A., Messinger, A., Kralik, J. D., Wise, S. P., & Schultz, W. (2004). Representation of attended versus remembered locations in prefrontal cortex. *PLoS biology*, *2*(11), e365.

49. Dalley, J. W., Cardinal, R. N., & Robbins, T. W. (2004). Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neuroscience & Biobehavioral Reviews*, *28*(7), 771-784.

50. Buchsbaum, B. R., Greer, S., Chang, W. L., & Berman, K. F. (2005). Meta-analysis of neuroimaging studies of the Wisconsin Card-Sorting task and component processes. *Human brain mapping*, *25*(1), 35-45.

51. Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review*, *82*(4), 276.

52. Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review*, *87*(6), 532.

53. Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., ... & NSPN Consortium. (2019). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences*, *116*(32), 15871-15876.

54. Gu, R., Lei, Z., Broster, L., Wu, T., Jiang, Y., & Luo, Y. J. (2011). Beyond valence and magnitude: a flexible evaluative coding system in the brain. *Neuropsychologia*, *49*(14), 3891-3897.

55. Marciano-Romm, D., Romm, A., Bourgeois-Gironde, S., & Deouell, L. Y. (2016). The Alternative Omen Effect: Illusory negative correlation between the outcomes of choice options. *Cognition*, *146*, 324-338.

56. Marciano, D., Krispin, E., Bourgeois-Gironde, S., & Deouell, L. Y. (2019). Limited resources or limited luck? Why people perceive an illusory negative correlation between the outcomes of choice options despite unequivocal evidence for independence. *Judgment and Decision Making*, *14*(5), 573.

57. D. Bates, M. Maechler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48 (2015).

58. Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801-814.

59. Cisek, P., & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, *45*(5), 801-814.

60. Klaes, C., Westendorff, S., Chakrabarti, S., & Gail, A. (2011). Choosing goals, not rules: deciding among rule-based action plans. *Neuron*, *70*(3), 536-548.

61. Pastor-Bernier, A., & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *Journal of Neuroscience*, *31*(19), 7083-7088.

62. Coallier, É., Michelet, T., & Kalaska, J. F. (2015). Dorsal premotor cortex: neural correlates of reach target decisions based on a color-location matching rule and conflicting sensory evidence. *Journal of neurophysiology*, *113*(10), 3543-3573.

63. McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, *19*(1), 22-24.

64. Gallivan, J. P., Logan, L., Wolpert, D. M., & Flanagan, J. R. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nature neuroscience*, *19*(2), 320.\

65. Gallivan, J. P., Stewart, B. M., Baugh, L. A., Wolpert, D. M., & Flanagan, J. R. (2017). Rapid automatic motor encoding of competing reach options. *Cell reports*, *18*(7), 1619-1626.

66. Q. J. M. Huys et al., Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS Comput. Biol. 7, e1002028 (2011).

67. G. Schwarz, Estimating the dimension of a model. Ann. Stat. 6, 461–464 (1978).

68. Collins AGE, Frank MJ. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychol Rev 2013;120:190–229. pmid:23356780

69. Badre D, Frank MJ. Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. Cereb Cortex 2011:1–10.

70. Frank, M. J., & O'Reilly, R. C. (2006). A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavioral neuroscience*, *120*(3), 497.

71. Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision*, *1*(4), 252.

72. Collins, A. G. E., & Frank, M. J. (2016). Motor demands constrain cognitive rule structures. *PLoS computational biology*, *12*(3), e1004785.

73. Boninger, D. S., Gleicher, F., & Strathman, A. (1994). Counterfactual thinking: From what might have been to what may be. *Journal of personality and social psychology*, *67*(2), 297.

74. Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135-157.

75. Roese, N. J. (1997). Counterfactual thinking. *Psychological bulletin*, *121*(1), 133.

76. Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and social psychology review*, *12*(2), 168-192.

# Appendix

**Additional task information.** The task included 4 blocks of 50 trials each, two blocks per condition (100 trials per condition, 200 trials in total). In the positive reward condition, which was framed as a "good card deck", the better scenario was the gain of a game-coin (+1) compared to a non-profit outcome (0). In the negative reward condition (framed as a "bad card deck") on the other hand, the better scenario was a non-profit outcome (0), compared to a loss of a game-coin (-1).

The two fractal card alternatives on each trial were presented on a grey background with a colored frame. To remind participants the condition they were at, "good card deck" blocks were indicated by a green frame, and "bad card deck" blocks were indicated by a red frame. The order of block conditions was alternated and counterbalanced between participants. To avoid imbalance between the negative and positive reward scenarios at the beginning of the task, participants were announced that they are given "*100 game-coins to start with*".

Before performing the task, participants received on-screen instructions informing them that on each trial, one of the fractals would result in rewards more often than the others (the exact probability was not mentioned): *"After selecting the card, you will see an outcome in the middle of the screen, as shown below. The outcome very much depends on the card you chose. Some cards in the deck are better than others. Your job is to find out which card is the best in the deck at any time and choose it.* Expected value for each fractal constantly changed using a random walk algorithm. That is, the probability of each fractal selection to result in the preferred outcome was slightly changed by the computer on each trial. Participants were told that: *"How good a card is can change along the game, somewhat similar to the value of market products that sometimes worth more and sometimes less. ".* Then, they went through a short test to make sure that all instructions were well understood.

In the task, on each trial, two fractal cards were randomly assigned to one of two positions on the screen: left or right. The left position was indicated by the 'S' response-key, and the right position was indicated by the 'K' response-key. Participants were instructed to use the key that corresponds to these locations to indicate their fractal choice. Fractal itself, but not the motor effector response used to report a fractal selection, predicted monetary outcome. Therefore, motor response on each trial was an outcome-irrelevant element of

the task. This was emphasized in the instructions of the task: *“Only the cards predict an outcome. The response key that was used to select a card does not influence the chances of winning. “*.

Each trial started with a 1 sec fixation point presented at the center of the screen, followed by the two fractal cards that were presented until response. After a choice was made (with a 6 sec deadline), the chosen stimulus remained on the screen for 500 ms, while the other stimulus disappeared. Then the outcome (0, -1, or 1 game-coin) appeared in the middle of the screen for 1 sec. If no response was received after 6 sec, the two stimuli disappeared from the screen, and the message: *‘No response. Please respond faster.'* appeared.

**Parameter recovery results.** We assessed the ability of each model to recover individual parameters by measuring the correlation between real and recovered parameters of 200 agents that completed the task with changing number of trials (70), as follows:

Model 1

| trials | alpha | beta |
|---|---|---|
| 50 | 0.765 | 0.471 |
| 300 | 0.984 | 0.935 |
| 1000 | 0.991 | 0.932 |

Model 2

| trials | alpha | beta | W |
|---|---|---|---|
| 50 | 0.871 | 0.591 | 0.600 |
| 300 | 0.922 | 0.775 | 0.798 |
| 1000 | 0.924 | 0.746 | 0.901 |

Model 3

| trials | alpha | beta | Wcomb |
|---|---|---|---|
| 50 | 0.916 | 0.617 | 0.545 |
| 300 | 0.963 | 0.584 | 0.752 |
| 1000 | 0.994 | 0.603 | 0.835 |

# תקציר

**רקע:** התנהגות מוכוונת-מטרה יכולה להרכש על-ידי למידה מניסוי וטעייה, עם ידע מינימלי אודות מבנה הסביבה. באמצעות תגמול, בני-אדם מייצרים אסוציאציות גירוי-פעולה-תוצאה שעוזרות להם לחזות אירועים עתידיים ולשנות את ההתנהגות שלהם בהתאם. מחקרים מציעים כי יכולת זו מסתמכת על מנגנוני שליטה קשבית. מחקרים נוספים מראים שערך ניתן גם לייצוגי מטלה שאינם רלוונטיים לחיזוי התגמול. מטרת המחקר הנוכחי היא לחקור את הנטייה לתת ערך לרכיבי מטלה שאינם מנבאים תגמול.

**שיטה:** בניסוי, 203 נבדקים מילאו מטלת multiple-armed bandit הכוללת ארבעה קלפים שכל אחד מהם מוביל לתגמול בהסתברות משתנה בין טריילים. בכל טרייל, הוצעו לנבדקים שני קלפים שנבחרו באופן אקראי על-ידי המחשב. הם התבקשו לבחור קלף אחד מבין השניים על מנת להרוויח, או להמנע מהפסד, של תגמול כספי. זהות הקלף הנבחר בלבד היוותה גורם מנבא לתגמול, ואילו הסתברות התגמול של כל קלף הייתה בלתי-תלויה בהסתברות התגמול של שאר הקלפים. בחנו את הנטייה לתת ערך לרכיבי מטלה שאינם חוזים תגמול באמצעות ניתוח טריילים עוקבים וגם באמצעות מודליים חישוביים של למידת חיזוקים.

**תוצאות:** שיחזרנו ממצאים קודמים בכך שהראנו שקיימת השמת ערך לייצוגי מטלה מוטוריים שאינם רלוונטיים לחיזוי תגמול וגם השמת ערך נגדית. קרדיט עבור התגמול ניתן לא רק לקלף הנבחר, אלא גם לפעולה המוטורית שבאמצעותה נבחר הקלף. בנוסף, קרדיט הפוך ניתן לקלף שלא נבחר. עוד אנו מראים, שקרדיט הפוך אינו ניתן לכל הקלפים, אלא רק לאלו שלקחו חלק בתהליך ההתלבטות.

**מסקנות:** אנו משערים שעל מנת להפחית עומס קוגניטיבי, מערכת הלמידה מבדילה ייצוגי מטלה אקטיביים על בסיס פעילות קדם-מוטורית. כלומר, ייצוגי מטלה המקושרים לביצוע מוטורי מקבלים ערך בהתאם לתגמול שניתן, ואילו ייצוגי מטלה המקושרים לאינהיביציה מטורית מקבלים את הערך ההפוך.

ספר סגול למדעי המוחבית

בית הספר לפסיכולוגיה,

הפקולטה למדעי החברה ע"ש גרשון גורדון

# רב-מימדיות בלמידה מבוססת חיזוקים: הדינמיקה של השמת משקל יתר לייצוגי מטלה שאינם רלוונטיים לחיזוי תגמול

מאת

**ענבל אלון**

**311148126**

החיבור בוצע בהנחייתו של

**ד"ר ניצן שחר**

ספטמבר 2021