RESEARCH ARTICLE

Stress enhances model-free reinforcement learning only after negative outcome

Heyeon Park[1¤], **Daeyeol Lee**[2,3], **Jeanyung Chey**[1]*

1 Department of Psychology, Seoul National University, Seoul, Korea, **2** Department of Neuroscience, Department of Psychiatry, Yale School of Medicine, New Haven, United States of America, **3** Department of Psychology, Yale University, New Haven, United States of America

¤ Current address: Department of Public Health Medical Services, Seoul National University Bundang Hospital, Seongnam, Korea
* jychey@snu.ac.kr

Abstract

Previous studies found that stress shifts behavioral control by promoting habits while decreasing goal-directed behaviors during reward-based decision-making. It is, however, unclear how stress disrupts the relative contribution of the two systems controlling reward-seeking behavior, i.e. model-free (or habit) and model-based (or goal-directed). Here, we investigated whether stress biases the contribution of model-free and model-based reinforcement learning processes differently depending on the valence of outcome, and whether stress alters the learning rate, i.e., how quickly information from the new environment is incorporated into choices. Participants were randomly assigned to either a stress or a control condition, and performed a two-stage Markov decision-making task in which the reward probabilities underwent periodic reversals without notice. We found that stress increased the contribution of model-free reinforcement learning only after negative outcome. Furthermore, stress decreased the learning rate. The results suggest that stress diminishes one's ability to make adaptive choices in multiple aspects of reinforcement learning. This finding has implications for understanding how stress facilitates maladaptive habits, such as addictive behavior, and other dysfunctional behaviors associated with stress in clinical and educational contexts.

OPEN ACCESS

Citation: Park H, Lee D, Chey J (2017) Stress enhances model-free reinforcement learning only after negative outcome. PLoS ONE 12(7): e0180588. https://doi.org/10.1371/journal.pone.0180588

Editor: Thomas Boraud, Centre national de la recherche scientifique, FRANCE

Received: March 3, 2017

Accepted: June 16, 2017

Published: July 19, 2017

Copyright: © 2017 Park et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (No. 2011-0005029). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Reward-seeking behaviors can be described by two different computational principles that might be supported by distinct neuroanatomical substrates [1–4]. On the one hand, a goal-directed controller selects behaviors expected to produce the best outcomes according to the knowledge of the decision-maker's environment and motivational state. The process by which the knowledge is updated and outcomes expected from alternative actions are derived from this knowledge is referred to as model-based reinforcement learning (RL). On the other hand, a habit controller relies on the expected values of outcome adjusted incrementally by trial and error, and results in automatic and less computationally demanding action selection. Accordingly, these goal-directed and habit systems might favor different actions, when the motivational status of the actor or the properties of environment change rapidly. However, precisely how the balance between these two controllers is adjusted across different behavioral settings remains poorly understood.

Stress might influence the arbitration between a goal-directed and a habit controller during decision making. Previous studies showed that stress causes humans to repeat behavior previously learned despite environmental changes [5–9] and tends to impair episodic memory while enhancing sensory processing [10–12], raising the possibility that stress might promote a switch from a high-order cognitive control to a simpler stimulus-response mapping. However, these previous studies have not examined precisely whether and how different aspects of learning and action selection are influenced by stress in a dynamic environment.

One of the critical issues regarding the relationship between stress and decision-making is how stress has an impact on the trade-off between habit and goal-directed behaviors. More specifically, whether stress leads to more habitual behaviors by either selectively weakening the process of goal-directed behaviors, by merely strengthening the process of habit or both. It is also possible that the effect of stress on behavior might vary depending on whether the result of previous behavior was positive or negative. Indeed, previous studies have suggested that stress might differently influence decisions depending on the valence of the outcome [13–15]. Considering that the neural circuits of reward processing frequently reflects valence-dependent activity of outcome [16, 17], it is possible that stress alters the neural processing of reinforcement and punishment differentially. In other words, stress may boost the neural signals related to decision-making differentially depending on the valence of outcomes. Finally, it remains unclear whether persistent behaviors resulting from stress simply reflects a decrease in the ability to incorporate the information about environmental changes, as quantified by the rate of learning, rather than changes in the nature of RL itself.

In the present study, we investigated the effects of stress on multiple aspects of RL such as model-free and model-based tendency according to the valence of the outcome and the learning rate. Participants were assigned to either a stress or a control condition before performing a multiple-stage decision-making task designed to distinguish model-based behavior from model-free RL behavior. In this task, reward probabilities associated with different choices were periodically reversed. By applying computational models to choice data, we quantified the extent to which choices were influenced by model-free vs. model-based RL, and dissociated the RL processing according to whether decision was followed by positive or negative outcome. Also, how the learning rate was affected by stress was examined.

Materials and methods

This study was approved by the Seoul National University Institutional Review Board (SNUIRB), and all participants provided written informed consent.

Participants

Fifty six healthy undergraduate students participated in this study (29 women, 27 men; age, 20.36 ± 1.91; body mass index 21.00 ± 2.52). Individuals who met any of the following criteria were excluded from participation: history of head injury, treatment with psychotropic medications, steroids, or any other medication that affects the central nervous system or the endocrine systems, current medical illness, self-report of mental disorder or substance abuse, existence of current stressful episode or major life event. Also, smokers and women taking oral contraceptives were excluded from the study due to possible effects of nicotine and oral contraceptive on the neuroendocrine stress response [18, 19]. Although gender could affect the hypothalamus-pituitary-adrenal cortex responsiveness to psychosocial stress differently, it has been demonstrated that there were no differences in salivary cortisol response between men and women in the luteal phase [18]. Therefore, women in the late luteal phase (after Day 21 and before the start of the next cycle) of the menstrual cycle were included in this study. Participants were asked to refrain from caffeine and physical exercise during the 6 hours prior to participation, and then were randomly assigned to the stress and the control conditions. Age (t_{50} = 1.23, p = .226), body mass index (t_{50} = -.20, p = .846), and perceived stress during the past month (t_{50} = .71, p = .483), assessed with Perceived Stress Scale [20], were not significantly different between participants in the two conditions. Four participants (two from each condition), who continued to choose the same action in more than 95% of the trials during the task, were excluded from the analysis, since this reflected lack of learning.

Stress protocol

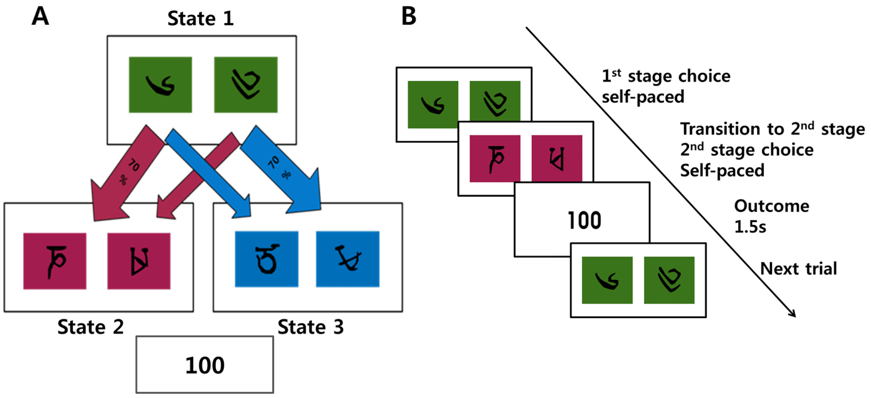
The socially evaluated cold pressor test (SECPT) [5, 21] was administered to the participants in the stress condition (15 women and 13 men). They immersed one hand (left-handed, right; right-handed, left) up to and including the wrist for 3 minutes (2 participants did it for 2 minutes which was their limits) into ice water (0 ~ 2 °C). During hand immersion, they were recorded on video by an unfamiliar person. Participants in the control condition (14 women and 14 men) submerged one hand up to and including the wrist for 1 minute in warm water (36 ~ 38 °C), and they were not recorded on video. To assess whether the treatments were successful, participants were required to report subjective stress on the visual analogue scale (VAS), with the lower and upper bound of the scale marked with numbers 0 and 100, representing a range from "no stress" to "the most stressful." All experiments took place between 1:00 P.M. and 5:40 P.M. to control for diurnal rhythm of the stress hormone (cortisol). Ten minutes after the cessation of the SECPT or the control procedure, the participants performed a two-stage reversal learning task described below.

Behavioral task

We used a two-stage reversal learning task which combined a reversal learning paradigm with the two-stage Markov decision task developed by Daw and his colleagues [22] (see Fig 1 for details). The two-stage Markov decision task has been used to distinguish the contribution of model-free and model-based RL to action selection. We also adopted the reversal learning paradigm, so that the participants were faced with a changing environment, and their choices in response to discrete environmental change could be investigated. The task consisted of six blocks of 40 trials, totaling 240 trials without any breaks. There was no explicit cue for block transition.

Each trial required two successive choices. In the first stage ("state 1"), participants chose between two options, represented by figures similar to Tibetan characters in green-colored boxes. The first-stage choice led probabilistically to one of the two second-stage states ("state 2" and "state 3"), represented by different colors (pink and blue). Each of the first-stage options was associated strongly (with a chance of 70%) with one of the two states in the second-stage, and this contingency was fixed throughout the experiment (Fig 1A). In the second-stage, subjects made another binary choice, and this second choice was linked to either 100 or 0 points depending on reward probability that was predetermined (Fig 1C). The assignment of two colors (pink or blue) to state 2 and 3 was counterbalanced across subjects, and the locations of two options in each state were randomized from trial to trial.

The reward probabilities for the two options in the second stage changed from block to block, employing the reversal learning paradigm as shown in Fig 1C. In the first block, both states 2 and 3 had one option leading to 60% chance of reward while the other leading to 20% chance. Therefore, in block 1, the two options in the first stage were equally favorable. In block



C

	STATE 2		STATE 3	
	option 1	option 2	option 1	option 2
block1	0.6	0.2	0.6	0.2
block2	0.8	0.2	0.2	0.2
block3	0.2	0.2	0.8	0.2
block4	0.8	0.2	0.2	0.2
block5	0.2	0.2	0.8	0.2
block6	0.8	0.2	0.2	0.2

Fig 1. Task design. (A) Task structure. Choice in the first stage leads probabilistically to different states in the second stage. Each stimulus in the second-stage resulted in either 0 or 100 points with different probabilities. (B) Timeline of events in a single trial. (C) Reward-probabilities of four options in stage 2.

https://doi.org/10.1371/journal.pone.0180588.g001

2, however, the two options in state 2 were rewarded with 80% and 20%, respectively, while both options in state 3 were rewarded with 20%. Therefore, it was more advantageous to choose the option more strongly associated with state 2 in the first-stage. In the following blocks, the advantageous choice in the first stage ("state 1") alternated as the reward probabilities of the options were switched between two states of the stage 2 after each block transition.

Prior to the experiment, the participants were informed that the reward probabilities for different choices in second stage would change, and that the probabilities of the transitions from the first state to different states in the second stage were fixed throughout the experiment. A practice session was given to familiarize the participants with the structure of the task. The practice session comprised of thirty trials, with five trials in each block.

Behavioral analyses

A series of two-tailed t-tests were used to examine whether there were differences in task performance between the two conditions. As different measures of performance, we analyzed the average response time to make a choice in the first stage, total points (cumulated reward), and overall probability of selecting the advantageous option (the option more strongly associated with "state 2" in block 2, 4, and 6, and the option more strongly associated with "state 3" in block 3 and 5) in the first stage. Next, a mixed-design ANOVA with outcome type (rewarded or unrewarded), and transition type (common or rare) as within-subjects factors, and treatment (stress or control condition) as between-subjects factors was used to examine whether staying probabilities (the probability of choosing the same option as in the preceding trial) in the first-stage varied significantly with stress, reward on previous trial, and transition type in previous trial. The data were analyzed using the IBM SPSS statistics 21 software.

Computational modeling

We used a RL model to characterize the trial-by-trial choice dynamics. Various different RL algorithms have been proposed to predict the reward from each option. In this study, we adopted the modified version of the Q-learning model since it performed better than the standard RL model to account for choice behaviors [23]. In the Q-learning model, action values are updated via a simple Rescorla-Wagner (RW) rule [24], and therefore, for a simple binary choice, the value function, $V_t(x)$, for option x can be updated after each trial t according to the following:

$$v_{t+1}(x) = V_t(x) + \alpha(R_t - V_t(x)) \tag{1}$$

where R_t means the outcome of the action at the trial t. This is equivalent to the following.

$$V_{t+1}(x) = (1 - \alpha)V_t(x) + \alpha R_t \tag{2}$$

In the present study, this RL model was modified to quantify model-free and model-based choice behaviors in the first stage of the task [23, 25]. In the model, the action values are updated according to the following:

$$V_{t+1}(x) = \begin{cases} (1 - \alpha)V_t(x) + \alpha\kappa_+, & \text{if it is rewarded,} \\ (1 - \alpha)V_t(x) + \alpha\kappa_-, & \text{if it is unrewarded} \end{cases} \tag{3}$$

where α was the learning rate for the selected option. The parameter κ_+ represented the strength of reinforcement by the reward outcome, and κ_- represented the strength of punishment by the no-reward outcome. In the present study, this model was expanded to update the value function for the choice in the first stage differently depending on the type of state transition in the same trial. Namely, the perturbation term κ was duplicated to reflect the components expected from model-free (κ^{mf}) and mode-based (κ^{mb}) RL model. For example, if reward occurred after a common transition, the value function of the option participants chose in the first stage ("state 1") was updated by $\kappa_+^{mf} + \kappa_+^{mb}$, since in this case, both model-free and model-based algorithms would attribute the positive outcome to the chosen action. By contrast, if reward occurred after a rare transition, the value function for the option selected in the first stage ("state 1") was updated by κ_+^{mf}, while the value function for the option unselected was updated by κ_+^{mb}, since in this case, model-free algorithms would attribute this positive outcome to the option chosen and the model-based learning would attribute the positive outcome to the option unchosen. Similarly, if reward did not occur after common transition, the value function of the chosen option was updated by $\kappa_-^{mf} + \kappa_-^{mb}$. If there was no reward after a rare transition, the value function for the chosen option was updated by κ_-^{mf}, while the value function for the other option was updated by κ_-^{mb}.

We found that for some subjects, the value of α and κ parameters estimated using the above equations were not stable, since the value of κ could increase in order to compensate a vanishingly small value of the learning rate. Therefore, model parameters were estimated using the following equation, which is mathematically equivalent to (3).

$$V_{t+1}(x) = \begin{cases} \gamma V_t(x) + \Delta_+, & \text{if it is rewarded,} \\ \gamma V_t(x) + \Delta_-, & \text{if it is unrewarded} \end{cases} \tag{4}$$

where, γ = 1 –α and represents a decay (or discount) factor, a weighting parameter given to the previous value estimate, and $\Delta = \alpha\kappa$, represents the change in the value function determined by the participant's choice and its outcome [25]. In other words, Δ_+^{mf}, Δ_+^{mb}, $\Delta\_^{mf}$, and $\Delta\_^{mb}$ replaced $\alpha\kappa_+^{mf}$, $\alpha\kappa_+^{mb}$, $\alpha\kappa\_^{mf}$, and $\alpha\kappa\_^{mb}$, respectively. In this RL model, the tendency to switch away from the unrewarded action corresponds to $\Delta\_ < 0$ while the tendency to stay with the same action regardless of no-reward corresponds to $\Delta\_ > 0$. More specifically, if $\Delta\_^{mf}$ and $\Delta\_^{mb}$ are negative, their magnitudes quantify how strongly model-free and model-based RL predict the tendency to switch to a different option after no reward.

The probability of choosing each option was given by the probability from softmax function related to the difference between the value functions. In other words, denoting the first stage actions by a_1 and a_2,

$$P_t(a_1) = 1/(1 + \exp(-(V_t(a_1) - V_t(a_2)))) \tag{5}$$

It should be noted that this model does not require any inverse temperature to determine the randomness in the participant's choices, since this can be changed by the magnitude of other model parameters (Δ's). This model is similar to the model used in Daw and his colleagues (2011), except that the value functions for unchosen actions decay gradually.

Parameters of the models were estimated separately for each participant. To maximize the log-likelihood of the data for each subject, we used the Nelder-Mead simplex algorithm [26]. We constrained discount factor to lie between zero and one, and allowed four change parameters to float arbitrarily. Model fitting was iterated 500 times with randomly chosen initial values in order to minimize the risk of finding a local but not global optimal solution.

A series of two-tailed t-tests were used to examine whether there were differences in parameter estimates of the RL models between the two conditions. Also, to test whether stress independently influences the learning rate and the weight of model-free & model-based RL, we performed regression on the model-free or model-based parameter estimate with a decay parameter and a treatment (stress vs. control) for each individual. The data were analyzed using the IBM SPSS statistics 21 software.

Results

Effects of stress on decision-making performance

We analyzed the choice behaviors of 52 participants (26 in each condition) during the two-stage reversal learning task following either stress-inducing or control treatment. Task performance of each participant is in S1 Table. As expected, participants in the stress condition rated the hand immersion as significantly more stressful (two-tailed t-test: $t_{50} = 8.61$, $p < 0.001$) than participants in the control condition. Average reaction times for choices in the first stage did not differ significantly for stress and control conditions (two-tailed t-test, $t_{50} = 1.6$, $p =$ .116). By contrast, total earnings were significantly lower in the stress condition than in the control condition ($t_{50} = 2.52$, $p = .015$). Also, the probabilities of selecting the advantageous option in the first stage were lower in the stress condition than those in the control condition ($t_{50} = 2.90$, $p = .006$).

In order to examine how stress might alter the model-free and model-based RL overall, we analyzed the stay-vs.-shift behavior in the first stage. Model-free RL assumes that participants select action solely based on previous choice outcome (reward or no-reward), whereas model-based RL assumes that they choose the optimal actions using their knowledge of the task structure. Therefore, participants relying on model-based RL would tend to stay with the same action even after no reward if this was preceded by a rare transition. By contrast, participants behaving strictly according to model-free RL would choose the same option in the first stage as in the previous trial when the same choice was rewarded in the previous trial, regardless of

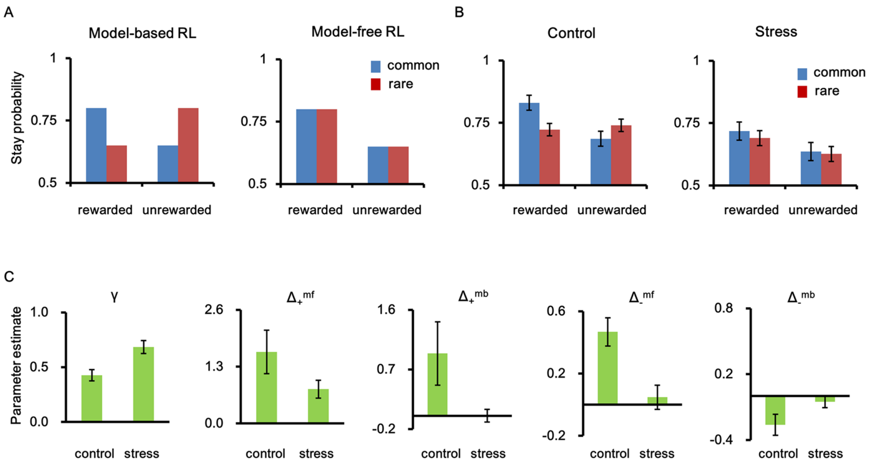


Fig 2. The effects of stress on decision-making. (A) The hypothetic results of stay-shift analysis expected for model-based (left) and model-free (right) reinforcement learning. (B) The behavioral results of stay-shift analysis. Participants' task performance in control condition showed characteristics of both model-free and model-based influences, while stressed participants showed stronger characteristic of model-free reinforcement learning. The stress × reward × transition interaction, $p = .004$. (C) The results of parameter estimation of a reinforcement learning model. Stress heightened the discount factor, γ (which means stress declined the learning rate) ($p = .002$) and boosted only model-free tendency to switch to a different option after no reward ($\Delta\_^{mf}$) ($p < .001$). Error bars represent SEM. Δ_+^{mb} $\Delta\_^{mb}$, and Δ_+^{mf} are parameters of the RL model which indicate the model-based tendency after reward, the model-based tendency after no-reward, and the model-free tendency after reward, respectively.

https://doi.org/10.1371/journal.pone.0180588.g002

whether the outcome was preceded by a common or rare transition (Fig 2A). The results from the mixed-design ANOVA revealed a significant main effect of outcome ($F_{(1,50)} = 18.44$, $p < 0.001$), reflecting the pattern predicted for model-free RL. Moreover, a significant reward × transition type interaction ($F_{(1,50)} = 14.06$, $p < 0.001$) showed that there was also a significant effect of model-based RL. More importantly, a significant stress × reward × transition type interaction ($F_{(1,50)} = 8.86$, $p = 0.004$) demonstrated a modulatory role of stress in the coordination of model-free and model-based performance in the task (Fig 2B). Neither stress × reward ($F_{(1,50)} = 0.06$, $p = 0.81$) nor stress × transition type interactions ($F_{(1,50)} = 0.09$, $p = 0.76$) were significant.

Effects of stress on reinforcement learning model parameters

In order to test multiple factors involved in decision making, and how they are modulated by stress, we applied the RL model that differentiates the model-free and model-based components of action selection. We found that the maximum likelihood estimates of the model parameter for the effect of negative outcome in the model-free learning ($\Delta\_^{mf}$) was significantly positive in the control condition ($t_{25} = 5.17$, $p < 0.001$) (Table 1). Presumably, the positive value of this parameter indicates that the subjects tended to stay with the same option in the first stage even when the previous outcome was negative. More importantly, the value of this parameter was significantly reduced in the stress condition ($t_{50} = 3.52$, $p = 0.001$) (Table 1, Fig 2C), suggesting that the tendency to avoid the option with the negative outcome was strengthened by stress. There was no significant difference in the model-free effect of positive outcome (Δ_+^{mf}) between the two conditions. Also, stress did not alter the parameters associated with model-based RL (Δ_+^{mb} and $\Delta\_^{mb}$). Instead, we found that the discount factor, γ, was significantly higher in the stress compared to the control condition ($t_{50} =$ _3.31, $p = 0.002$). The discount factor determines how rapidly the previous value function is forgotten, and is related to

Table 1. Best-fitting parameter estimates, shown as median plus quartiles across conditions.

	γ^{*}	Δ_{+}^{mf}	Δ_{+}^{mb}	$\Delta\_^{mf**}$	$\Delta\_^{mb}$
CONTROL					
25th	0.23	0.70	0.10	0.12	-0.41
Median	0.47	1.09	0.33	0.48	-0.14
75th	0.60	1.67	0.80	0.78	0.04
T	8.41**	3.30*	2.00	5.17**	-2.74
STRESS					
25th	0.56	0.04	-0.23	-0.10	-0.13
Median	0.73	0.50	-0.04	0.02	-0.01
75th	0.95	1.06	0.18	0.14	0.11
T	11.50**	3.84*	.02	.62	-.89

Notes: Δ_{+}^{mb} and $\Delta\_^{mb}$ are parameters which represent the model-based tendency after reward and no-reward, respectively. Δ_{+}^{mf} and Δ_{+}^{mf} are parameters which indicate the model-free tendency after reward and no-reward, respectively.

* $p < 0.01$

** $p < 0.001$. T is the *t* value from the paired t-test which was performed to investigate whether each parameter was significantly different from zero.

https://doi.org/10.1371/journal.pone.0180588.t001

the learning rate α ($\gamma = 1\_\ \alpha$). Thus, it appears that stress boosted only model-free tendency to switch to a different option after no reward and decreased the learning rate, i.e., the ability to incorporate new information into decision-making.

We performed additional analyses in order to clarify further how the changes in model-free RL were related to stress treatments. First, to test whether stress independently influences the learning rate and the weight of model-free RL after negative outcome, we performed regression on the model-free parameter estimate ($\Delta\_^{mf}$) with a decay parameter and a treatment (stressed or not) for each subject. The results confirmed that the effect of stress on model-free RL after negative outcome ($\Delta\_^{mf}$) was significant (B = -0.433, SE = 0.133, $\beta = -0.459$, $p = 0.002$) even after controlling for the effect of stress on the learning rate (B = 0.049, SE = 0.218, $\beta = 0.032$, $p = 0.822$). Second, we conducted the ANCOVA with the probability of the advantageous action as a covariate. In the present study, a reversal learning component was incorporated into the two-stage decision-making task [22]. During a two-stage decision task with reversal, subjects with model-free tendency, who simply choose the advantageous option in each block, might appear to have a model-based tendency [27]. The ANCOVA results showed that the effect of stress on the strength of model-free RL after receiving negative outcome was significant even after controlling for the probability of the advantageous action ($F_{(1,49)} = 8.01$, $p = 0.007$). Taken together, these results suggest that stress increased the contribution of model-free RL only after negative outcomes.

Discussion

In this study, we found that stress impaired the reward-seeking behavior and demonstrated that the inferior performance under stress might be due to at least two different mechanisms. First, stress increased the influence of the model-free reinforcement learning, particularly the likelihood of switching to an alternative choice when the previous choice led to an undesirable outcome. Second, stress decreased the learning rate, namely, the degree to which new information is incorporated into trial-by-trial decision making. These findings suggest that maladaptive choice behavior under stress might be attributable to both a slower learning rate and the strengthening of model-free RL after a negative outcome.

It has not been investigated clearly whether stress leads to more habitual behaviors by selectively weakening the process of goal-directed behaviors, by merely strengthening the process of habit, or both. In order to investigate the effect of acute stress on the distinct contributions of habit and goal-directed processing, recent researches have tried to use computational modeling for reinforcement learning (RL) to separate the habit and goal-directed processing into two RL algorithms, model-free and model-based, respectively. In previous computational studies, however, the effects of acute stress on the two RL were inconsistent [28, 29]. Otto and his colleagues showed that stress-related physiological (cortisol) response was negatively correlated with model-based but not model-free contributions. However, their study did not demonstrate the effect of stress on decision making itself. Also, Radenbach and his colleagues reported the effect of stress on the ratio of model-based RL to model-free RL, but without clearly separating out the effects of stress on model-based RL from those of model-free RL[29]. Thus, how acute stress facilitates habit or model-free choice behavior remained incompletely understood.

In this study, we investigated the effects of stress on model-free and model-based RL using a 2-step decision task incorporating the reversal learning paradigm and showed that stress increased the model-free RL without altering the strength of model-based RL. These results suggested that stress-enhancement of habit behavior may not be merely compensatory by-product of impaired model-based RL behavior. Also, habitual processing might be strengthened by stress because stress disrupts inhibition of the model-free processing which could be a default model of RL[30]. Furthermore, we differentiated the model-free tendency to make a shift following no-reward (lose-switch) and to stay following a reward (win-stay), and showed that stress increased the model-free RL after no-reward selectively without affecting the model-free RL after reward. These results suggest that stress may disproportionately boost the neural processing of decision-making involved in model-free learning from negative outcomes. Our findings are consistent with previous studies showing there are separate neural processing for reinforcement and punishment [16, 17, 31].

Although the 2-step decision task has been designed to distinguish model-free and model-based RL, a recent study revealed that the original task does not lead to significant difference in performance (points or income) predicted by model-based vs. model-free RL approach, through a computational simulation [27]. Therefore, we incorporated a reversal learning paradigm into the original task, which produced more consistent difference in the performance for the two RL strategies. However, in a reversal learning task, decision-makers can learn that there are two distinct latent states of the task and rely on such inference about the current latent state to make their choices [32, 33]. The decision-maker who infers and uses latent state of the task could outperform a standard model-free RL and looks like a model-based decision-maker, even without using the knowledge of the transition structure linking the first actions to the states of the second stage [34]. In this study, stress enhanced only model-free RL without impairing model-based RL. We conducted the ANCOVA with the probability of selecting the advantageous action as a covariate, which would reflect the tendency to make choices based on the inference about a latent state. The results from this analysis showed that the effect of stress on the strength of model-free RL after receiving negative outcome was significant even after controlling for the probability of the advantageous action. Therefore, stress might increase the contribution of model-free RL regardless of its effect on the ability to make choices based on the inferred state of the environment.

Also, we found that stress decreased the learning rate during a reward-based choice task. In the RL model, the learning rate reflects how quickly the valuation of selected action is updated by the difference between the prediction and the actual outcome, referred to as the prediction error [2]. Therefore, it represents how rapidly new information from the environment is incorporated in subsequent actions [35–37]. For adaptive decision making, it is critical to utilize new information efficiently and to avoid maladaptive perseverative behaviors when faced with environmental change. Decision-makers with low learning rate would fail to switch their behaviors flexibly in response to unexpected changes in the real world. It is possible that a decrease in learning rate under stress may be an important factor contributing to stress-induced alteration in RL. However, we could not examine the effect of stress on the distinct learning rate of model-free and model-based RL, because we estimated a single learning rate from observed choices and rewards for each subject. Further investigations are necessary to clarify whether stress changes learning rate during both model-free and model-based RL.

Conclusions

This study characterized the effect of stress on adaptive decision making, by providing participants with a changing environment where their choice behaviors were modeled in a computational framework of reinforcement learning. We found that stress facilitated the habitual, model-free RL process to shift away from unrewarded action, and that it also interrupted the subjects from incorporating new information into their subsequent choices. These findings provide insight as to the mechanism by which stress diminishes the ability to behave flexibly in reward-based decision making, and have significant implications for understanding and treating stress-related maladaptive conditions characterized by enhanced habit behavior such as addiction and impulse control disorders.

Supporting information

S1 Table. Task performance for each subject.
(XLSX)

Author Contributions

Conceptualization: HP JC DL.

Formal analysis: HP DL.

Funding acquisition: JC.

Investigation: HP JC.

Methodology: DL HP.

Project administration: JC.

Resources: JC.

Software: HP DL.

Supervision: JC.

Validation: DL JC HP.

Visualization: HP.

Writing – original draft: HP.

Writing – review & editing: JC DL HP.

References

1. Doll BB, Simon DA, Daw ND. The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology. 2012.

2. Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press; 1998.
3. Lee D. Decision making: from neuroscience to psychiatry. Neuron. 2013; 78(2):233–48. https://doi.org/10.1016/j.neuron.2013.04.008 PMID: 23622061
4. Penner MR, Mizumori SJ. Neural systems analysis of decision making during goal-directed navigation. Progress in neurobiology. 2012; 96(1):96–135. https://doi.org/10.1016/j.pneurobio.2011.08.010 PMID: 21964237
5. Schwabe L, Wolf OT. Stress prompts habit behavior in humans. The Journal of Neuroscience. 2009; 29 (22):7191–8. https://doi.org/10.1523/JNEUROSCI.0979-09.2009 PMID: 19494141
6. Dias-Ferreira E, Sousa JC, Melo I, Morgado P, Mesquita AR, Cerqueira JJ, et al. Chronic stress causes frontostriatal reorganization and affects decision-making. Science. 2009; 325(5940):621–5. https://doi.org/10.1126/science.1171203 PMID: 19644122
7. Schwabe L, Dalm S, Schächinger H, Oitzl MS. Chronic stress modulates the use of spatial and stimulus-response learning strategies in mice and man. Neurobiology of learning and memory. 2008; 90 (3):495–503. https://doi.org/10.1016/j.nlm.2008.07.015 PMID: 18707011
8. Schwabe L, Oitzl MS, Philippsen C, Richter S, Bohringer A, Wippich W, et al. Stress modulates the use of spatial versus stimulus-response learning strategies in humans. Learning & Memory. 2007; 14(1–2):109–16.
9. Schwabe L, Schächinger H, de Kloet ER, Oitzl MS. Corticosteroids operate as a switch between memory systems. Journal of cognitive neuroscience. 2010; 22(7):1362–72. https://doi.org/10.1162/jocn.2009.21278 PMID: 19445601
10. Kim J, Diamond DM. The stressed hippocampus, synaptic plasticity and lost memories. Nature Reviews Neuroscience. 2002; 3(6):453–62. https://doi.org/10.1038/nrn849 PMID: 12042880
11. Henckens MJ, Hermans EJ, Pu Z, Joëls M, Fernández G. Stressed memories: how acute stress affects memory formation in humans. The Journal of Neuroscience. 2009; 29(32):10111–9. https://doi.org/10.1523/JNEUROSCI.1184-09.2009 PMID: 19675245
12. Ferragud A, Haro A, Sylvain A, Velazquez-Sanchez C, Hernandez-Rabaza V, Canales J. Enhanced habit-based learning and decreased neurogenesis in the adult hippocampus in a murine model of chronic social stress. Behavioural brain research. 2010; 210(1):134–9. https://doi.org/10.1016/j.bbr.2010.02.013 PMID: 20153381
13. Porcelli AJ, Delgado MR. Acute stress modulates risk taking in financial decision making. Psychological Science. 2009; 20(3):278–83. https://doi.org/10.1111/j.1467-9280.2009.02288.x PMID: 19207694
14. Lighthall NR, Gorlick MA, Schoeke A, Frank MJ, Mather M. Stress modulates reinforcement learning in younger and older adults. Psychology and aging. 2013; 28(1):35. https://doi.org/10.1037/a0029823 PMID: 22946523
15. Cavanagh JF, Frank MJ, Allen JJ. Social stress reactivity alters reward and punishment learning. Social cognitive and affective neuroscience. 2011; 6(3):311–20. https://doi.org/10.1093/scan/nsq041 PMID: 20453038
16. Jensen J, Smith AJ, Willeit M, Crawley AP, Mikulis DJ, Vitcu I, et al. Separate brain regions code for salience vs. valence during reward prediction in humans. Human brain mapping. 2007; 28(4):294–302. https://doi.org/10.1002/hbm.20274 PMID: 16779798
17. Bromberg-Martin ES, Matsumoto M, Hikosaka O. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron. 2010; 68(5):815–34. https://doi.org/10.1016/j.neuron.2010.11.022 PMID: 21144997
18. Kirschbaum C, Kudielka BM, Gaab J, Schommer NC, Hellhammer DH. Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. Psychosomatic medicine. 1999; 61(2):154–62. Epub 1999/04/16. PMID: 10204967.
19. Mendelson JH, Sholar MB, Goletiani N, Siegel AJ, Mello NK. Effects of low- and high-nicotine cigarette smoking on mood states and the HPA axis in men. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology. 2005; 30(9):1751–63. Epub 2005/05/05. https://doi.org/10.1038/sj.npp.1300753 PMID: 15870834; PubMed Central PMCID: PMCPMC1383570.
20. Cohen S, Williamson GM. Perceived stress in a probability sample of the United states In: Spacapan S, Oskamp S, editors. the social psychology of health: claremont symposium on applied social psychology Newbury Park, CA: Sage; 1988.
21. Schwabe L, Haddad L, Schachinger H. HPA axis activation by a socially evaluated cold-pressor test. Psychoneuroendocrinology. 2008; 33(6):890–5. Epub 2008/04/12. https://doi.org/10.1016/j.psyneuen.2008.03.001 PMID: 18403130.
22. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69(6):1204–15. https://doi.org/10.1016/j.neuron.2011.02.027 PMID: 21435563

23. Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci. 2009; 29(31):9861–74. Epub 2009/08/07. https://doi.org/10.1523/JNEUROSCI.6157-08.2009 PMID: 19657038.

24. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory. 1972; 2:64–99.

25. Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci. 2004; 7(4):404–10. Epub 2004/03/09. https://doi.org/10.1038/nn1209 PMID: 15004564.

26. Lagarias JC, Reeds JA, Wright MH, Wright PE. Convergence properties of the Nelder—Mead simplex method in low dimensions. SIAM Journal on optimization. 1998; 9(1):112–47.

27. Akam T, Costa R, Dayan P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. PLoS Comput Biol. 2015; 11(12):e1004648. https://doi.org/10.1371/journal.pcbi.1004648 PMID: 26657806

28. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proceedings of the National Academy of Sciences. 2013; 110(52):20941–6.

29. Radenbach C, Reiter AM, Engert V, Sjoerds Z, Villringer A, Heinze H-J, et al. The interaction of acute and chronic stress impairs model-based behavioral control. Psychoneuroendocrinology. 2015; 53:268–80. https://doi.org/10.1016/j.psyneuen.2014.12.017 PMID: 25662093

30. Lee Sang W, Shimojo S, O'Doherty John P. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron. 2014; 81(3):687–99. https://doi.org/https://doi.org/10.1016/j.neuron.2013.11.028 PMID: 24507199

31. Worbe Y, Palminteri S, Savulich G, Daw N, Fernandez-Egea E, Robbins T, et al. Valence-dependent influence of serotonin depletion on model-based choice strategy. Molecular psychiatry. 2016; 21(5):624–9. https://doi.org/10.1038/mp.2015.46 PMID: 25869808

32. Costa VD, Tran VL, Turchi J, Averbeck BB. Reversal learning and dopamine: a Bayesian perspective. Journal of Neuroscience. 2015; 35(6):2407–16. https://doi.org/10.1523/JNEUROSCI.1989-14.2015 PMID: 25673835

33. Hampton AN, Bossaerts P, O'doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. The Journal of Neuroscience. 2006; 26(32):8360–7. https://doi.org/10.1523/JNEUROSCI.1010-06.2006 PMID: 16899731

34. Akam T, Costa R, Dayan P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-step Task. bioRxiv. 2015:021428.

35. Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nature neuroscience. 2007; 10(9):1214–21. Epub 2007/08/07. https://doi.org/10.1038/nn1954 PMID: 17676057.

36. Bernacchia A, Seo H, Lee D, Wang XJ. A reservoir of time constants for memory traces in cortical neurons. Nature neuroscience. 2011; 14(3):366–72. Epub 2011/02/15. https://doi.org/10.1038/nn.2752 PMID: 21317906; PubMed Central PMCID: PMCPMC3079398.

37. Simon DA, Daw ND. Environmental statistics and the trade-off between model-based and TD learning in humans. Advances in Neural Information Processing Systems. 2011; 24:127–35.