**Reinforcement Learning Modeling of the Iowa Gambling Task**

4조 202011003 강승수, 202211080 박시윤, 202311132 이상아

1. **Project overview**

This project aims to explain general human decision-making behavior by applying reinforcement learning models to data from the Iowa Gambling Task (IGT). Previous studies have typically used reinforcement learning in the context of psychological theories(Steingroever et al., 2018) or to examine the negative effects of specific substances(Kjome et al., 2010). However, these approaches are often limited to individual conditions or specific scenarios, making it difficult to capture universal patterns in human choice behavior.

In this project, we estimated representative learning parameters using Maximum Likelihood Estimation (MLE) based on the IGT choice data of 602 participants(Naidoo et al., 2019).

These parameters were then applied to Q-learning and Valence-Specific Q-learning (VS-Q learning) models to analyze how humans learn from and respond to positive and negative outcomes.

1. **Activities performed by week/team member**
2. Week 1

We implemented the core structure of the Iowa Gambling Task environment, the Q-learning model, and the parameter estimation framework using Maximum Likelihood Estimation (MLE). 강승수 implemented the Iowa Gambling Task environment and the Q-learning agent. 박시윤 developed the Maximum Likelihood Estimation (MLE) procedure. 이상아 was responsible for preprocessing the dataset of 602 participants.

1. Week 2

We have implemented the Valence-Specific Q-learning model and modified the maximum likelihood estimation code to enable more accurate parameter estimation, and also see the distribution of estimated parameters. 강승수 implemented the Valence-Specific Q-learning agent, revised the parameter estimation code using Maximum Likelihood Estimation (MLE), and developed code to examine the distribution of the model parameters. 박시윤 contributed to revising the MLE-based parameter estimation code and implemented the simulation framework for the Iowa Gambling Task (IGT). 이상아 supported the revision of the MLE-based parameter estimation code and conducted analysis on the properties of the dataset.

1. Week 3

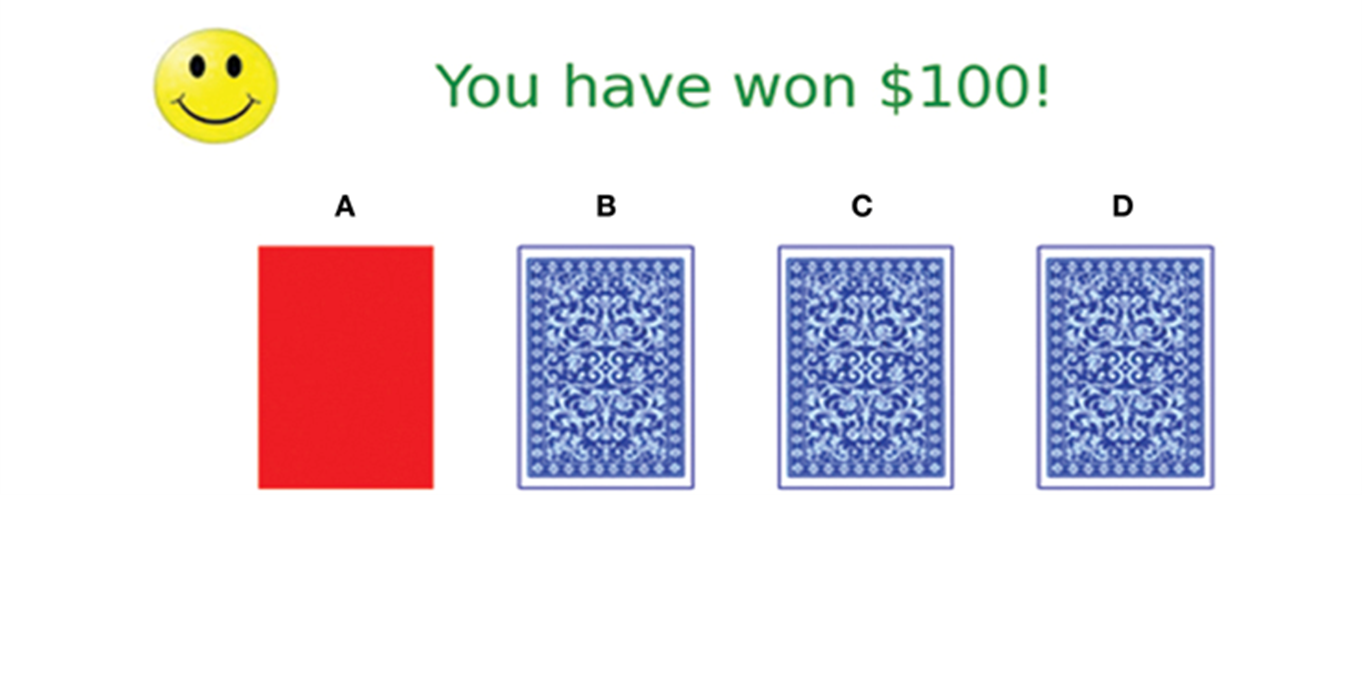
We first identified that the Iowa Gambling Task dataset included both successful and unsuccessful learners. Clustering was conducted based on α and β for the Q-learning model, and for the Valence-Specific Q-learning model. Representative parameter values were then extracted from each cluster by calculating the median of the respective parameters. This process allowed us to distinguish clusters that demonstrated effective learning behavior within the IGT environment. Furthermore, using the code we developed for direct participation in the Iowa Gambling Task, each team member individually completed the task. We estimated the learning parameters for each member and conducted personalized simulations within the same environment. This enabled us to observe how the reinforcement learning agents captured and modeled the decision-making behavior of each individual. 강승수 was responsible for clustering, writing the code to identify representative parameter values for participants who successfully learned the task. 박시윤 developed the classification algorithm to distinguish successful learners from the full dataset and visualized the clustering results. 이상아 aimed to analyze and interpret the relationship between the learning parameters and performance outcomes based on data from relevant research studies.

1. **Methodology**
2. IGT environment and model construction
3. IGT

The IGT is a psychological task designed to simulate real-life decision-making under uncertainty. Participants repeatedly select one card from four available decks, where each deck is associated with a distinct pattern of rewards and punishments. In the task, participants are presented with four decks of cards (A, B, C, and D). Each card in the decks represents a potential gain or loss in virtual money. The goal is to maximize the total amount of money by choosing cards from these decks.

However, some decks are ‘bad’ (e.g., decks A and B), offering high immediate rewards but large long-term losses. Other decks are ‘good’ (e.g., decks C and D), providing smaller immediate gains but lower long-term penalties. Participants are not told which decks are advantageous or disadvantageous; they have to learn this through experience and feedback from their choices.

The IGT has been widely used in psychology and neuroscience to assess decision-making, particularly in contexts involving risk, reward, and punishment.

**Figure 1. Iowa gambling task**

In the simulation of the IGT, each deck has its own reward structure defined as follows:

Gains:

Deck A: +100

Deck B: +10

Deck C: +5

Deck D: +5

Losses (randomly sampled upon each card selection):

Deck A: A value randomly chosen from {0, 150, 200, 250, 300, 350}, each with an equal probability of 1/6.  
Deck B: A value randomly chosen from {0, 1250}, with probabilities 0.9 and 0.1 respectively.

Deck C: A value randomly chosen from {0, 25, 75, 50}, with probabilities 0.5, 0.1, 0.1, and 0.3 respectively.

Deck D: A value randomly chosen from {0, 250}, with probabilities 0.9 and 0.1 respectively.

Using this deck structure, we built an IGT environment that allows users to perform the task and automatically record data such as choices, gains, and losses into a file. Our team members performed the IGT in this environment, and we stored the results.

1. Models used
2. Q-learning

We used Q-learning to model human decision-making behavior. Q-learning captures how individuals learn to choose actions based on feedback from previous experiences, with the goal of maximizing cumulative rewards.

Specifically, our model includes two key parameters:

α: the learning rate, which determines how much new experiences influence the update of Q-values.

β: the inverse temperature parameter, which controls the trade-off between exploration and exploitation in decision-making. A higher beta results in more deterministic (greedy) choices, while a lower beta leads to more exploratory behavior.

The Q-values, representing the expected value of each action, are updated using the following Q-learning update rule:

(: current state, : action, : reward, : next state, : a possible future action, : discount factor)

In our modeling, the softmax function was used to map Q-values into choice probabilities:

1. Valence-specific Q-learning

In addition to the standard Q-learning model, we also implemented Valence-specific Q-learning (VSQ).

The VSQ model separates learning rates for positive and negative outcomes:

: the learning rate for positive prediction errors (e.g., gains)

the learning rate for negative prediction errors (e.g., losses)

This distinction allows the model to reflect potential asymmetries in how participants adjust their expectations based on rewards versus punishments.

In the VSQ model, separate Q-values are maintained for positive and negative feedback. The combined value of an action is calculated as a weighted sum of these Q-values:

For a positive outcome (gain):

For a negative outcome (loss):

The model combines and into a single Q-value used for making decisions. The combination can be formulated as . w is a weight parameter that determines how much emphasis is placed on positive learning versus negative learning. When w = 1, the agent relies only on positive outcomes. When w = 0, the agent relies only on negative outcomes. When w = 0.5, both are weighed equally. This approach captures how people might combine positive and negative feedback when evaluating actions.

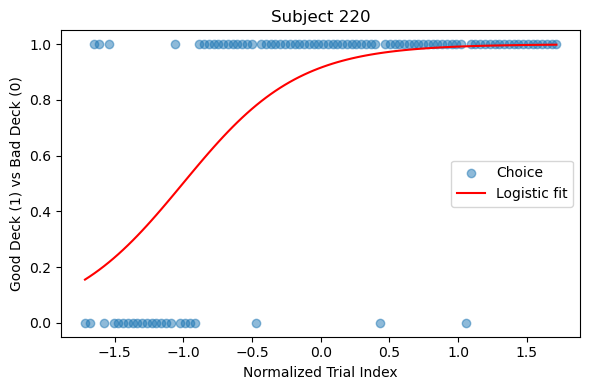
Action selection is still governed by the softmax function.

1. Classification of participants who successfully learned the IGT

We aimed to identify participants who successfully learned the Iowa Gambling Task (IGT) and fit reinforcement learning models to their data. Out of 602 participants, we classified those who demonstrated successful IGT learning.

Since decks A and B result in long-term losses and decks C and D result in long-term gains, we coded choices from decks A and B as 0, and choices from decks C and D as 1. A logistic function was then fitted to the data, with the x-axis representing trial number and the y-axis representing the coded deck choice (0 or 1).

As shown in Figure 2, participants whose fitted logistic function had a positive slope were considered to have successfully learned the IGT. Based on this criterion, 198 participants were identified as successful learners out of 602 participants. All subsequent analyses focused on the data from these 198 successful learners.

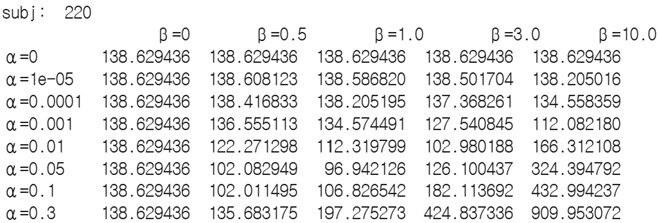


**Figure 2. Result of logistic function fitting for IGT learning (Subject 220)**

1. Parameter estimation

For each participant, we conducted parameter estimation by identifying the parameter values that minimized the negative log likelihood using MLE. To ensure the selection of appropriate initial parameter values for MLE, we first performed a grid search using a range of α values [0, 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.3] and β values [0, 0.5, 1.0, 3.0, 10.0].

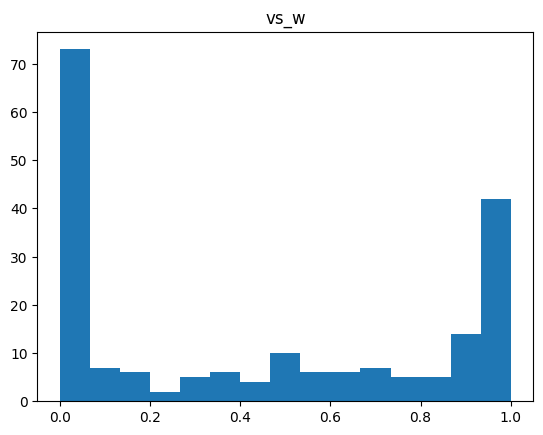
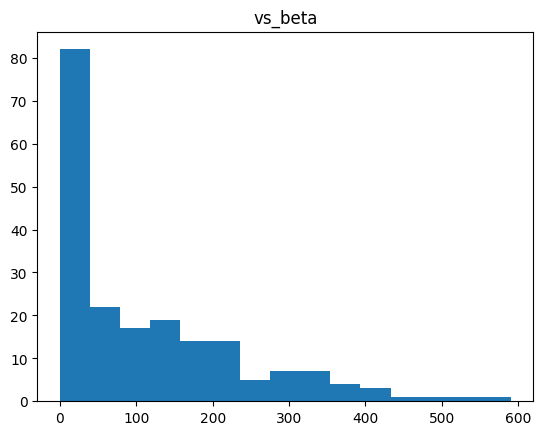
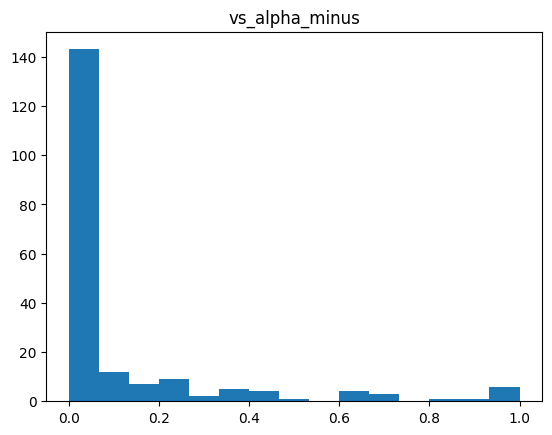
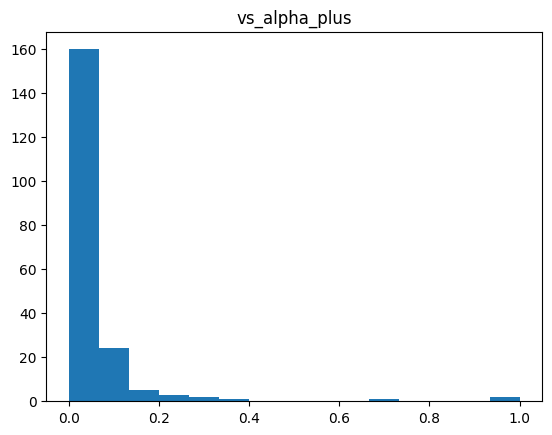
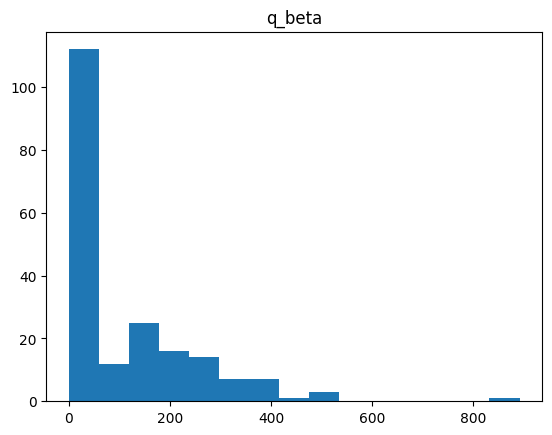
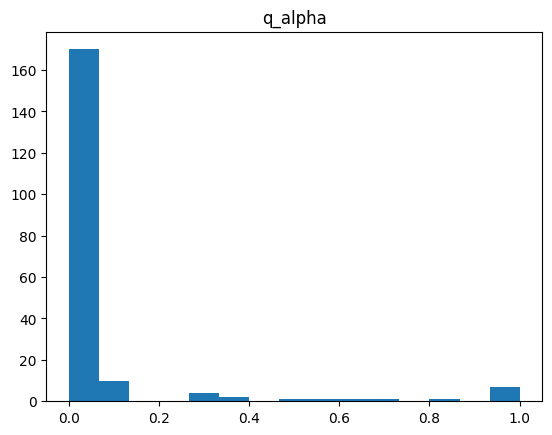
The grid search results, as illustrated in Figure 3, revealed that the negative log likelihood was lower for smaller values of α and β, and the corresponding parameter estimates were confirmed to be reasonable.

****

**Figure 3. Results of the grid search for subject 220**

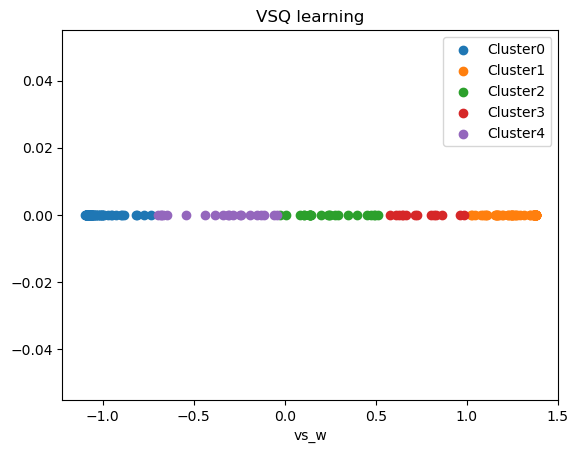
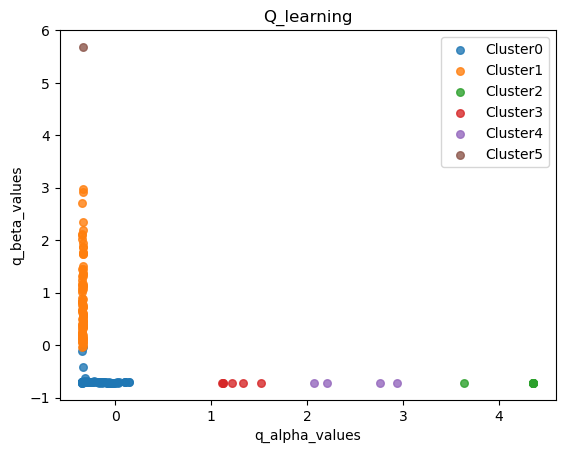
Based on the outcomes of the grid search, we initialized the parameters α and β to 0.0001 and 1.0, respectively. Considering the high magnitude of gains and losses in the Iowa Gambling Task (reaching up to 1250), we applied a scaling factor by dividing both gains and losses by 10. Using these settings, we performed MLE to fit the model. We conducted parameter estimation on the results from our own IGT sessions using the same method, by identifying the parameter values that minimized the negative log likelihood.

1. Subject clustering

The distribution of parameters estimated through MLE is shown in Figure 4. Since learning strategies may differ across individuals, we performed clustering based on the distribution of the parameters (α and β for Q-learning; w for valence-specific Q-learning). In the case of valence-specific Q-learning, while parameters such as α+, α-, β, and w are present, the distributions of α+, α-, and β are clustered near zero. We determined that using too many parameters for clustering could hinder the interpretability of the clusters. Therefore, clustering for the Valence-specific Q-learning model was conducted using only the w.

**Figure 4. Distribution of estimated parameters for all subjects (from the top left: Q-learning α and β, Valence-Specific Q-learning α+, α-, β, and w).**

We used the mean shift algorithm for clustering, and the results of the clustering are presented in Figure 5. We then conducted reinforcement learning model simulations for each cluster.



**Figure 5. Clustering results of parameters (left: Q-learning, right: valence-specific Q-learning)**

1. Simulation

We identified the median value of each parameter within each cluster and used these values to run simulations using the Q-learning and valence-specific Q-learning models that we implemented. Through the simulations, we were able to observe how scores and Q-values evolved over time steps. Following the same procedure, we conducted simulations using the parameters estimated from our own IGT data.

1. **Execution result and analysis**
2. Simulation of model using the estimated parameters
3. Q-learning

|  |  |
| --- | --- |
|  |  |
| Figure 6 : Q-learning cumulative rewards - Cluster 0(α : 0.00894, β : 1.00019) | Figure 7 : Q-learning Action Q value- Cluster 0(α : 0.00894, β : 1.00019) |
|  |  |
| Figure 8 : Q-learning cumulative rewards - Cluster 1(α : 0.00008, β : 201.29405) | Figure 9 : Q-learning Action Q value- Cluster 1(α : 0.00008, β : 201.29405) |

Clustering shown in figure 5 revealed two clearly distinguishable groups. Figure 6 and 8 show the cumulative rewards over steps within a trial for each cluster and Figure 7 and 9 depict how Q-values for each action evolve with step for each cluster. The two clusters exhibit starkly different learning dynamics.

In cluster 0, β is very small and α is approximately 0.008. The cumulative reward rises almost linearly and steadily, and the Q-values quickly identify action C as the optimal strategy, converging on it while action A and B early are early discarded. In other trial simulations, some agents also converged quickly to action D. These results indicate a group that learns rapidly and efficiently the best choice.

By contrast, Cluster 1 has a β of about 201 and a very small α. The reward goes up quickly at first because action B happens to give a big reward, but it soon drops sharply after a heavy penalty. Because the β is so large, the agents take longer to recognize action C as the optimal strategy and they are slower to realize that actions A and B are poor options. Consequently this group reaches the optimal behavior after considerable trial and error.

1. VSQ-learning

|  |  |
| --- | --- |
| 텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진  AI가 생성한 콘텐츠는 부정확할 수 있습니다. |  |
| Figure 10. VSQ-learning cumulative rewards - Cluster 0(α+ : 0.03280, α- : 0.00016, β : 115.71403, w : 0.01053) | Figure 11. VSQ-learning Action Q value - Cluster 0(α+ : 0.03280, α- : 0.00016, β : 115.71403, w : 0.01053) |
| 텍스트, 라인, 그래프, 도표이(가) 표시된 사진  AI가 생성한 콘텐츠는 부정확할 수 있습니다. | 텍스트, 도표, 라인, 그래프이(가) 표시된 사진  AI가 생성한 콘텐츠는 부정확할 수 있습니다. |
| Figure 12. VSQ-learning cumulative rewards - Cluster 1(α+ : 0.00631, α- : 0.18387, β : 6.0202, w : 0.96606) | Figure 13. VSQ-learning Action Q value - Cluster 1(α+ : 0.00631, α- : 0.18387, β : 6.0202, w : 0.96606) |

In VSQ-learning, agents were grouped into five clusters based on the parameter w. Higher w indicates greater sensitivity to positive rewards, while a lower w reflects greater sensitivity to penalties. Among these, the two clusters at the extremes were selected for comparison, as they account for the largest proportions and show the most noticeable differences.

Figure 10 and 12 display the changes in cumulative rewards over steps within a trial, while Figure 11 and 13 show how the Q-values for each action evolve over the steps. As clearly seen in these four figures, the cluster exhibits distinctly different learning dynamics.

In cluster 0, the w value is low, suggesting that the agent places significantly more weight on losses than on gains. In figure 10, the cumulative rewards increase steadily with each step, and in Figure 11, only the Q-value for action D, the favorable action, increases. This suggests that the agent, being highly sensitive to loss, quickly identifies and focuses on the strategy that minimizes penalties.

In contrast, Cluster 1 has a w value of approximately 0.96, indicating that the agent places far greater emphasis on gains than on losses. As a result, the agent tends to ignore negative feedback and learns in a reward-driven manner. In Figure 12, the cumulative rewards repeatedly rise and fall. Figure 13 shows that Q-values are not concentrated on any specific action, implying that the agent fails to form a clear strategy. This suggests that the agent, by focusing too heavily on rewards may repeatedly make poor decisions and fail to correct them effectively.

In summary, the comparison between the two clusters demonstrate that asymmetric value weighting significantly impacts learning performance and strategic stability. The loss-sensitive agent in Cluster 0 quickly establishes a clear strategy and learns stably. whereas the reward-sensitive agent in Cluster 1 tends to fail in forming consistent strategies.

1. Simulation using the estimated parameters for 강승수, 박시윤, 이상아
2. Q-learning

|  |  |
| --- | --- |
|  |  |
| Figure 14. Q-learning cumulative rewards - 강승수(α : 0.00847, β : 259.79646) | Figure 15. Q-learning Action Q value - 강승수(α : 0.00847, β : 259.79646) |
|  |  |
| Figure 16. Q-learning cumulative rewards - 박시윤(α : 0.086711, β : 0.80648) | Figure 17. Q-learning Action Q value - 박시윤(α : 0.086711, β : 0.80648) |
|  |  |
| Figure 18. Q-learning cumulative rewards - 이상아(α : 0.00023, β : 663.43) | Figure 19. Q-learning Action Q value - 이상아(α : 0.00023, β : 663.43) |

Figure 14, 16 and 18 show the cumulative rewards over steps within a trial, and Figure 15, 17 and 19 depict how Q-values for each action evolve with steps.

강승수 and 이상아 seems to belong to Cluster 1. Typically Cluster 1 is characterized by low α and high β. However, their α values are relatively higher than the median of cluster 1, and their β values are also high. The higher α likely enabled them to learn favorable choices more quickly than others in the same cluster. As seen in Figure 14 and Figure 18, after an initial drop, the cumulative rewards recovers rapidly and converges to a high level. And due to the high β, it took some time to converge on the action C, as shown in Figure 15, because of increased exploration resulting from the high β value.

In contrast, 박시윤 seems to belong to Cluster 0. Typically Cluster 0 is characterized by high α and low β. However, her α are relatively higher than the median of cluster 0 and her β are relatively lower than the median of cluster 0. As shown in Figure 16, after an initial drop, the cumulative rewards recovers rapidly and converges to a high level. And due to the low β and high α, the Q\_value of action C is rapidly increased. Because low β led to less exploration and quicker convergence and the relatively high α supported efficient learning of optimal action.

1. VSQ-learning

|  |  |
| --- | --- |
|  |  |
| Figure 20. VSQ-learning cumulative rewards - 강승수(α+ : 0.00630, α- : 0.00049, β : 738.15515, w : 0.95763) | Figure 21. VSQ-learning Action Q value - 강승수(α+ : 0.00630, α- : 0.00049, β : 738.15515, w : 0.95763) |
|  |  |
| Figure 22. VSQ-learning cumulative rewards - 박시윤(α+ : 0.00102, α- : 0.00015, β : 342.571, w : 0.04781) | Figure 23. VSQ-learning Action Q value - 박시윤(α+ : 0.00102, α- : 0.00015, β : 342.571, w : 0.04781) |
|  |  |
| Figure 24. VSQ-learning cumulative rewards - 이상아(α+ : 0.00023, α- : 0.0, β : 602.61926, w : 1.0) | Figure 25. VSQ-learning Action Q value - 이상아(α+ : 0.00023, α- : 0.0, β : 602.61926, w : 1.0) |

Figures 20, 22 and 24 show the cumulative rewards over steps within a trial and Figures 21, 23 and 25 depict how Q-values for each action evolve with steps.

강승수 and 이상아 seems to belong to Cluster 1. Typically Cluster 1 is characterized by high w. As a result, 강승수 and 이상아 tend to pay little attention to negative feedback and instead focus solely on reward. As shown in Figure 20 and 24, the cumulative rewards increase over steps, but the trend is somewhat unstable. In Figure 25, the Q-values are spread across multiple actions rather than being concentrated on a specific one. This suggests that the agent’s focus on rewards may have led to a lack of clear strategic direction.

In contrast, 박시윤 seems to belong Cluster 0 which is typically characterized by a low w value. As shown in Figure 22, the cumulative rewards increase steadily with each step, and in Figure 23, the Q-values for advantageous actions C and D rise quickly. This indicates that the loss-sensitive agent is able to quickly identify and adopt strategies that minimize penalties.

1. Limitations of Q-learning agents and Advantage of VSQ-learning agents

Q-learning agents tend to select optimal actions too quickly which makes their behavior appear unnatural when compared to the human learning process. In particular, the rapid convergence of action q-values in one direction seem overly mechanical and unrealistic.

In contrast, VSQ-learning agents better capture human-like characteristics such as focusing more on positive rewards or being sensitive to penalties. As a result their Q-values tend to rise more evenly across multiple actions, rather than being heavily biased toward a single one. This pattern may better reflect the decision-making processes observed in actual participants.

1. **Conclusion**

This project aimed to explain general human decision-making behavior by analyzing the Iowa Gambling Task (IGT) using reinforcement learning models. Based on data from 602 participants, learning parameters were estimated through Maximum Likelihood Estimation (MLE), and both Q-learning and Valence-Specific Q-learning (VSQ-learning) models were applied to simulate general patterns of reward and punishment learning in humans.

In the Q-learning analysis, participants were classified into two distinct groups depending on their learning rate (α) and exploration-exploitation control parameter (β). Cluster 0, characterized by high α and low β values , exhibited rapid convergence to optimal strategies with minimal exploration. In contrast, Cluster 1, which had low α and high β values, initially engaged in suboptimal actions and required more time and trial-and-error to learn the best choices.

The VSQ-learning model, which separates responses to rewards and punishments, revealed further asymmetries. Participants in Cluster 0, who were more sensitive to losses (low w values), quickly formed stable strategies by avoiding high-penalty actions. On the other hand, participants in Cluster 1, who placed greater weight on gains (high w values), tended to ignore negative feedback and showed unstable learning patterns, often repeating poor choices. These findings highlight that the value-weighting parameter w plays a crucial role in strategic learning performance.

Additionally, when the IGT data from the three team members were modeled, both the Q-learning and Valence-Specific Q-learning models effectively reflect the observed choice behaviors. This suggests that these reinforcement learning models are capable of capturing ‘individual’ decision-making dynamics.

In summary, this study demonstrates that reinforcement learning models can effectively explain human choice behavior and that individual differences can be quantitatively analyzed through parameter-based clustering and simulations. Future research may incorporate broader emotional responses, temporal discounting, and contextual information to further refine models of human decision-making.

1. **참고문헌**
2. Naidoo, L. (2019). Knowledge, attitudes and perceptions of Genomic Medicine among medical students, private medical practitioners and the general public in eThekwini, KwaZulu-Natal. https://doi.org/10.51415/10321/4428
3. Kjome, K L, Lane, S D, Schmitz, J M, Green, C, Ma, L, Prasla, I, . . . Moeller, F G 2010 Relationship between impulsivity and decision making in cocaine dependence. Psychiatry Research, 178: 299–304. DOI: http://dx.doi.org/10.1016/j.psychres.2009.11.024.
4. Steingroever, H, Šmíra, M, Lee, M D, and Pachur, T inpreparation “Do intuitive and deliberate decision makers perform differently on the Iowa gambling task?”
5. Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (2013). Insensitivity to future consequences following damage to human prefrontal cortex. In *Personality and Personality Disorders* (pp. 287-295). Routledge.
6. Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. *Psychological assessment*, *14*(3), 253.
7. Haines, N., Vassileva, J., & Ahn, W. Y. (2018). The outcome‐representation learning model: A novel reinforcement learning model of the iowa gambling task. *Cognitive science*, *42*(8), 2534-2561.