Hierarchical Instance-Based Learning for Decision Making from Delayed Feedback

Tailia Malloy
Social and Decision Sciences
Carnegie Mellon University
Pittsburgh, PA, USA

David Hagmann
Department of Management
The Hong Kong University of
Science and Technology, Hong Kong

Cleotilde Gonzalez
Social and Decision Sciences
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

In real-world decision making, outcomes are often delayed, meaning individuals must make multiple decisions before receiving any feedback. Moreover, feedback can be presented in different ways: it may summarize the overall results of multiple decisions (aggregated feedback) or report the outcome of individual decisions after some delay (clustered feedback). Despite its importance, the timing and presentation of delayed feedback has received little attention in cognitive modeling of decision-making, which typically focuses on immediate feedback. To address this, we conducted an experiment to compare the effect of delayed vs. immediate feedback and aggregated vs. clustered feedback. We also propose a Hierarchical Instance-Based Learning (HIBL) model that captures how people make decisions in delayed feedback settings. HIBL uses a super-model that chooses between sub-models to perform the decision-making task until an outcome is observed. Simulations show that HIBL best predicts human behavior and specific patterns, demonstrating the flexibility of IBL models.
Keywords: Instance-Based Learning, Hierarchical Learning, Decision Making, Delayed Feedback

Introduction

Many consequential decisions in our lives requires us to make multiple decisions before we receive direct feedback on our choices. For example, firefighters place multiple controlled burns to prevent wild fires, but only later observe whether a fire starts across a large area (i.e., aggregated feedback) or in specific smaller regions (i.e., clustered feedback). This type of decision making adds uncertainty regarding the relationships between our actions and the outcomes we observe. Generally, uncertainty has been a well studied phenomenon in decision-making literature. Some of our decisions are rarely informed by objective descriptions of probabilities and values, especially those that are highly dependent on our own preferences and on personal experiences. An introductory college course with an engaging professor, for example, can inspire students to pursue a degree in the major (Chambliss & Takacs, 2014) and a pleasant experience in a restaurant or an airline is likely to make us regular customers. When we have the opportunity to repeatedly expose ourselves to an experience, we can learn the distribution of the outcomes (Hertwig et al., 2004).

The idea that recent outcomes have an immediate effect on how we feel about an option has a long history (Thorndike, 1898). Decision rules that incorporate the last outcome (such as win-stay, lose-shift) are common when choosing from unknown options, as well as when engaging in strategic interactions with others (Robbins, 1952). When individuals are motivated to reproduce successes, they can inherently be biased against risky options that could appear unfavorable in small samples (Denrell & March, 2001). Even in tasks that are repeated frequently, people can draw premature conclusions about the "goodness" of an option (Sims et al., 2013), potentially due to an increased cognitive load associated with overthinking, which can negatively impact behavior (Gray et al., 2006). This type of rumination has also been shown to negatively impact performance independently of attention (Hitchcock et al., 2022), indicating that it may be difficult to counteract the effect in decision making.

A commonly observed phenomenon in decision making from description and experience is the so-called Description-Experience Gap (DEG), which describes the reversal of risk preferences towards risk seeking when decision makers learn outcome probabilities through experience, compared to when they are directly observed (e.g., (Fox & Hadar, 2006); (Hertwig, 2015);(Martin et al., 2014)). We may be tempted to believe that after enough experience or with enough information, we may consider different options as if we had a description of outcomes available to us. However, observations of repeated decision making have demonstrated that this DEG continues to exist even as more experience is gathered, representing a Repeated DEG (RDEG) (Lejarraga & Gonzalez, 2011). While repeating choices and continuing to observe outcomes does not reduce this DEG alone, some properties of repeated decision tasks have shown reductions in RDEG, such as framing outcomes as losses compared to gains (Gonzalez & Mehlhorn, 2016). This observation raises the question of what other changes in decision-making tasks might also reduce this RDEG.

We propose that the RDEG is driven, in part, by a lack of exploration following unfavorable outcomes. Specifically, we believe that this is the result of selecting between different temporary strategies that are used in a hierarchical structure to guide decision making between outcome observations. Previous research in predicting human behavior has applied Hierarchical Reinforcement Learning (HRL) and found close connections to multiple features of human decision making (Botvinick, 2012; Botvinick & Weinstein, 2014; Eckstein & Collins, 2020). However, hierarchical structure has yet to be applied to cognitive models of decision making that rely on a memory, rather than the policy based method that defines agent behavior in RL models.

Cognitive models of decision making in utility-based tasks, such as binary choice tasks, allow us to compare different theories for the mechanisms underlying human decision making (Gonzalez & Dutt, 2011). One type of cognitive model is

Instance-Based Learning (IBL), which is inspired by a theory of dynamic decision making. IBL models have been applied to the type of binary choice tasks used in this work, though typically with immediate feedback (Gonzalez & Dutt, 2011; Lejarraga et al., 2012). IBL models have also been used to represent the cognitive mechanisms of updating past experiences based on outcomes observed later, called credit assignment (Nguyen et al., 2023). While these previous methods have attempted to account for how humans assign credit in long episodic tasks with a reward signal at the end, they do not directly account for individual differences in behavior between observations of outcomes. This is because they function by assuming that the outcome of an option is equal to their predicted value of that option.

The Hierarchical IBL (HIBL) model presented in this work resolves the issue of differences in behavior between observations of outcomes by using a super-model that selects between sub-agents to perform tasks in between observations of outcomes. Thus, the super-model does not observe a delay in the outcome observed, and the individual differences in behavior are explained based on the differences between these sub-models. In this paper, we explore the behavior of human participants when feedback is delayed, including both aggregated feedback, which displays the sum of all choices made, as well as a novel clustered feedback condition, which displays the outcomes of each individual choice made after a delay. We compare the ability of the HIBL model with a standard flat IBL model in their ability to explain individual differences in choice, by comparing how well both models are able to predict the behavior of individual participants.

In the following sections, we first describe in detail the choice tasks that we use before providing background information about the IBL model and how our proposed HIBL model functions. Then, we describe the experiment performed to compare human decision making with different timescales of feedback. After this, we compare the behavior of the HIBL and IBL models in simulation with the behavior of humans. Finally, we perform model tracing to predict the behavior of individual participants and demonstrate that the HIBL model is better able to account for individual differences in behavior and more accurately predict individuals' choices. We conclude with a discussion of the importance of these results from the perspective of both human behavior and cognitive modeling of decision making.

Binary Choice Task with Aggregated and Clustered Feedback

Binary choice tasks involve decision makers selecting between two options, often labeled as A&B or 1&2. Descriptive binary choice tasks include the probabilities of choice outcomes, and have long demonstrated a risk-averse effect by which participants prefer lower utility guaranteed outcomes over higher utility outcomes that have some probability of low payoff (Kahneman & Tversky, 2013). Experiential binary choice tasks involve selecting between two options without outcome probabilities, learning utility through repeated choice (Barron & Erev, 2003), which has demonstrated a reversal of preference toward riskier options compared to descriptive choice tasks (Erev & Barron, 2005). In repeated choice decision making this is the RDEG mentioned previously. One explanation for this RDEG is that experiences of low utility outcomes prevent decision makers from fully exploring options to learn their true expected utilities due to the counteracting goals of exploration (selecting different choices) and exploitation (selecting the best estimate option (Hertwig & Erev, 2009)).

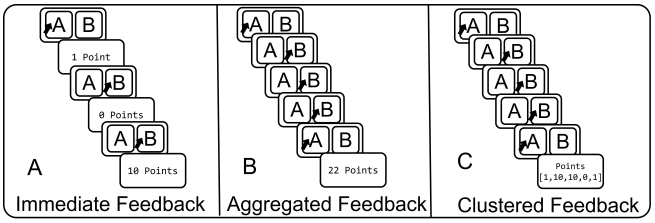

Figure 1: **Left:** Immediate Feedback **Middle:** Delayed Aggregate Feedback **Right:** Delayed Clustered Feedback

In this work, we investigate how exploration and exploitation impact RDEG using two different types of delayed feedback. This compares traditional "immediate feedback" (Figure 1.A) with "aggregated feedback" (Figure 1.B) of the sum of the choice outcomes, and with "clustered feedback" (Figure 1.C) of the outcomes of a series of choices only after all choices in the cluster have been made. We predicted that with immediate feedback, fewer participants would choose the risky option in the absence of a description than if a description is available because they would get discouraged from observing bad outcomes early on and stop exploring a risky option further. We propose that decision-makers use a hierarchical strategy to select between different levels of exploration and exploitation, only adjusting these after an outcome is observed. This strategy would reduce the gap between the proportion of risky choices in descriptive versus experiential decision making when feedback is delayed.

Instance-Based Learning

Instance-Based Learning Theory (IBLT) is a general algorithm that aims to represent a comprehensive human cognitive process of experiential choice in dynamic tasks (Gonzalez, 2024). IBLT is grounded on mathematical expressions of human memory retrieval which have been developed and validated in a large number of empirical studies. Models developed based on IBLT (i.e., IBL models) have successfully accounted for human behavior in a variety of contexts, ranging from abstract repeated-choice tasks to more naturalistic decision tasks of search and choice (Nguyen & Gonzalez, 2022), continuous control tasks (Gonzalez et al., 2003), and phishing email identification (Malloy & Gonzalez, 2024). Given IBLT's previous success in capturing human behavior in immediate feedback settings (see summaries in (Gonzalez

et al., 2013; Gonzalez, 2017), we are interested in exploring the ability of these models to explain and predict behavior in delayed feedback settings.

IBLT proposes that an agent makes a prediction of the potential outcome of a decision based on similar instances stored in memory. Decisions are stored as instances in memory. To predict the utility of each potential action, the past utilities of similar instances are weighted by their instance's probability of retrieval from memory to generate estimates of the expected utility. The agent then selects the option that has the highest estimated utility.

Activation

The contribution of a retrieved instance's utility to the final estimated utility of an option depends on the memory activation of the instance. The activation of an instance A_i reflects how readily it comes to mind relative to the current choices considered. The activation, $A_i(t)$ of instance i at time t is a sum of three components, $A_i(t) = B_i(t) + M_i + \varepsilon$, the base-level activation, the partial matching correction, and the activation noise. This equation comes from the ACT-R cognitive architecture (Anderson & Lebiere, 2014). The full equation for activation in IBL models is given by:

$$A_i(t) = \ln\left(\sum_{t' \in \mathcal{T}_i(t)} (t-t')^{-d}\right) + \mu \sum_{j \in \mathcal{F}} \omega_j (S_{ij} - 1) + \sigma\xi \quad (1)$$

In this equation, d is the decay parameter that scales the activation of an instance based on how far in the past it occurred. μ is the mismatch penalty that weights the sum of all feature similarities S_{ij}, which are themselves weighted by the values ω_j which determine the relevance of specific features. A logistic distribution centered on zero ξ adds random noise from Gaussian $\mathcal{N}(-1, 1)$ and is scaled by the σ parameter.

Probability of Retrieval

Once the activations of all relevant instances have been calculated, they are used to calculate the probability of retrieval of the instance. A parameter, the temperature, or τ, is used in constructing this probability. For a given option being considered, k, let $\mathcal{M}_k$ be the set of all matching instances. Then the probability of retrieval of instance $i \in \mathcal{M}_k$ at time t is

$$P_i(t) = \frac{e^{A_i(t)/\tau}}{\sum_{i' \in \mathcal{M}_k} e^{A_{i'}(t)/\tau}} \quad (2)$$

Blending

The expected utility of an option is calculated as an average of the utilities of the instances weighted by their probability of retrieval. The blended value at time t, $V_k(t)$ of the various utilities in the instances for this option. When the outcome is immediately observed, u_i corresponds to the utility observed. this gives blended value as:

$$V_k(t) = \sum_{i \in \mathcal{M}_k} P_i(t) u_i \quad (3)$$

When the feedback is delayed, u_i is initially set to be the predicted value V_{k_i} of the option that was selected. This creates a delayed response in memory that can be later resolved according to the utility that is observed and the credit assignment method.

Delayed Feedback

As described in IBLT, feedback is often delayed (Gonzalez et al., 2003). The theory proposes that instances are initially stored with the "expectation" created by blending (i.e., expected utility). Then, when the outcome of decisions is experienced, the theory proposes that a credit assignment mechanism is used to update expectations with experienced values. Previous research has explored the use of temporal difference error, borrowed from reinforcement learning research (Sutton et al., 1999), as a credit assignment method in IBL models (Nguyen et al., 2023). Other approaches have applied linear credit assignment before adjusting credit based on the attribute weights w_j of the features in previous instances (Malloy et al., 2025).

This linear application of credit is similar to research into how humans adjust past memories, particularly when unexpected events occur (Bein et al., 2023). However, the exact formulation of the credit assignment is ill-defined in IBLT. In the HIBL model described in detail in the next section, the issue of delayed feedback is partially addressed by using a super-agent that does not experience a delay, by selecting between different sub-agents that do experience this delay.

Hierarchical Instance Based Learning Model

Hierarchical models of decision making typically function by instantiating multiple models and having a super-agent that selects between different sub-agents to perform the decision making task (Rasmussen et al., 2017; Eppe et al., 2022). In Reinforcement Learning (RL), this has previously been applied to decision making tasks that have an explicit hierarchical structure (Pateria et al., 2021), such as where the agent takes an initial action in the environment which determines which actions are available to them in the subsequent time step. However, these models have also been applied to cases without explicit hierarchical structures (Botvinick, 2012). Possible explanations for why these models would be useful in cases that do not have an explicit hierarchy of decisions include the improved learning efficiency of these models (Nachum et al., 2019), or human decision makers who apply unnecessarily more complex reasoning (Botvinick, 2012).

Importantly, RL research has shown that HRL is useful with long time horizons and significantly delayed rewards (Krishnan et al., 2016). In this work, we assume that the delay in reward observations that requires multiple actions to be selected before an outcome is observed encourages hierarchical reasoning by allowing participants to choose exploration-exploitation strategies and adjusting these after outcomes are observed. The Hierarchical IBL (HIBL) model we present in this work is shown in Figure 2 where the model has the structure that allows it to be applied to the specific binary

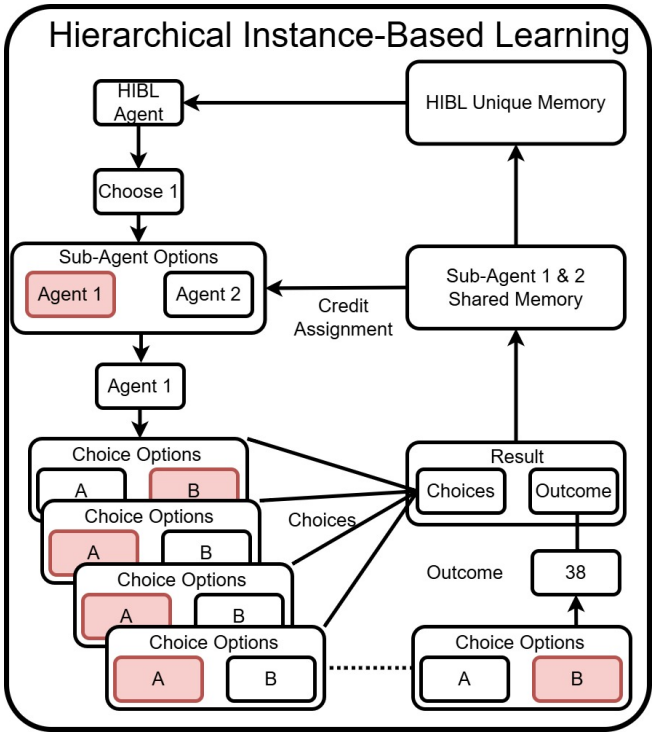

Figure 2: Schematic of the Hierarchical IBL model and an example of processioning the aggregated feedback condition.

choice task we use. This model consists of a single super-agent and two sub-agents, though the number of sub-agents was selected to keep the choice size the same for all models under comparison instead of corresponding to the choices.

HIBL Model Super-Agent

The super-agent in the HIBL model has a choice selection that represents the selection of one of the two agents to perform the ten choices of the experiment before the next outcome observation, in the two delayed feedback settings. In the immediate feedback setting, the super-agent selects from the sub-agents on each time-step. Since the super-agent only makes selections once every ten steps in the two delayed conditions, it does not directly observe a delay in the outcome of its actions, and is not required to assign credit onto the outcomes of choices that it selects. Once the super-agent selects its choice, Agent 1 in the example shown in Figure 2, that agent takes actions in the environment until the next outcome.

HIBL Model Sub-Agents

When the sub-agents are making choices in the environment with delayed outcomes, they use the same standard method of using predicted expected values as the outcome observation before adjusting these instances later. This means they are required to assign credit to past actions; in both sub-agent models in this work we use the same linear assignment that is used in the standard IBL model. Additionally, the memories of both sub-agents are conformed to be equivalent after the credit assignment takes place, so that they can update their behavior based on observations of the environment.

If the sub-agents in this hierarchical structure had exactly the same parameters and memory, then their behavior would be identical, meaning that the structure would have no specific effect. The actions of sub-agents can be made variable in a variety of ways. In this work, we explore two different methods of altering the behavior of these sub-agents so that the top level agent is making a meaningful choice between agents. Firstly, the sub-agents have different instances pre-populated into their memory at initialization, meaning they will have different initial predictions for the rewards associated with the two options. Specifically, this is done by pre-populating Agent 1 with optimistic outcomes of the risky choice, and pessimistic outcomes of the risky choice for Agent 2. In addition, both agents are initialized with different parameters for noise, temperature, and decay. Specifically, the optimistic Agent 1 had parameters selected so that on average it would produce more consistent behavior, while the pessimistic Agent 2 had them selected to produce more sporadic behavior. This means that even though the memories of both agents become more similar, their long-term behavior will remain variable.

Experiment

Participants We recruited 401 participants from Amazon Mechanical Turk (55.86% male, mean age 32.37) and asked them to make 110 choices between two options, labeled "Option A" and "Option B." One option represented a safe payoff and the other a binary lottery. The position of the safe and risky options on the right and left side of the screen was counterbalanced. The safe option provided a payoff of 4 units with certainty; the lottery returned a payoff of 10 units with 50% probability and 0 otherwise (providing an expected value of 5 units). Participants earned $1 for 400 units. The complete instructions and interface are presented on OSF[1].

Experimental Design and Procedure. Participants were randomly assigned to conditions in a 2x3 design. In the first dimension, we varied whether participants were given an explicit description of the available options. In the description condition, the potential outcome and probabilities were explicitly stated in the instructions and were present on the buttons throughout the experiment. In the experience condition, the buttons were labeled only as "Option A" and "Option B," without information about the options themselves. Participants learned about outcomes and their likelihood only from the outcome feedback of the choices they made.

In the second dimension, we varied whether participants observed the outcome of their choice immediately after choosing one of the options ("Immediate Feedback") or whether they only received feedback after every 10 periods in the form of a list of individual choice outcomes ("Clustered Feedback") or as a sum of all choice outcomes ("Ag-

[1]Link to Anonymous OSF Repository

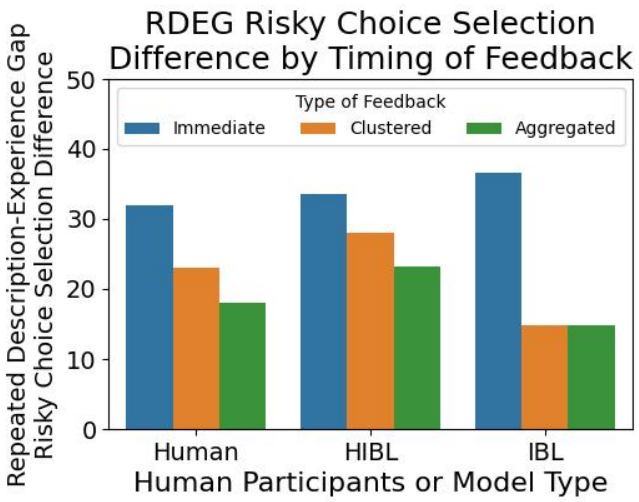

Figure 3: RDEG metric across model types and conditions.

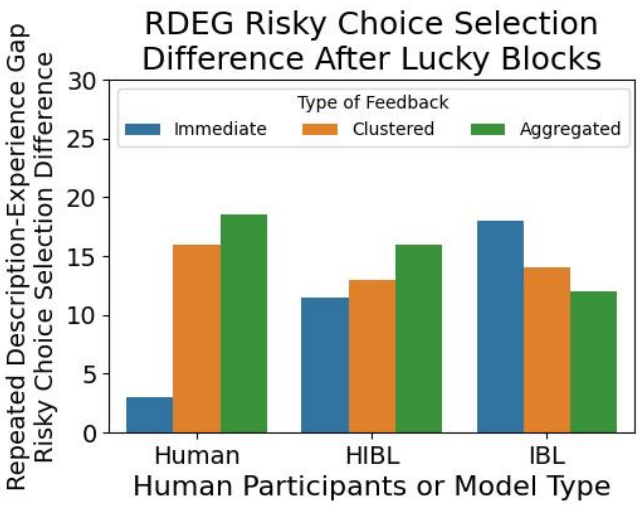

Figure 4: RDEG metric across model types and conditions. This analysis is limited to decisions made in a 10 trial period after a block with 6 or more high valued outcomes.

gregated Feedback"). We chose the number of 10 periods to ensure that participants were likely to observe high and low outcomes in the risky option and to see that the safe option always returned the same result.

Results

We are interested in comparing the differences in the selection of risky choices between the description and experience conditions of our experiment. For this, we use a Repeated Description-Experience Gap (RDEG) metric by calculating the difference in risky choice selection between description and experience conditions. This value was calculated by taking the average percent of risky choices in the description condition $p_d(r)$ and subtracting it by the average percent of risky choices in the experience condition $p_e(r)$. This gives the description-experience risky choice selection difference as $p_d(r) - p_e(r)$, which is plotted on the y-axis of Figures 3, 4, and 5 in percentage points (pps). For simplicity, this value is referred to as the RDEG.

While the RDEG is positive in humans for all conditions, it is highest when feedback is immediate (32 pps), lower with clustered feedback (24 pps), and lowest with aggregated feedback (18 pps). The specific ordering of these gaps also corresponds to the amount of information that participants observe about the outcome of their choices meaning that individual bad outcomes are hidden in the aggregate feedback. As a result, the less a participant observes about the outcome of their choices, the more they prefer the risky option.

Another important behavior of human participants is how they respond to lucky high-valued outcomes, and whether this is different between description and experience settings. This is shown in Figure 4 which displays the same RDEG metric, but limits the comparison to blocks of 10 time-steps immediately after a participant or model observed 6 or more high-valued lucky outcomes in a 10 time-step period. From this we can see that lucky outcomes make little difference in decision making between description and experience for immediate feedback, but they result in a larger RDEG for clustered and aggregated feedback. This is an interesting result, as it shows the opposite trend of the overall human behavior. However, in the same way that hiding bad outcomes in the aggregate produced riskier behavior, hiding lucky outcomes in the aggregate tempers much riskier behavior.

Simulated Behavior

We first compare the performance of the simulated IBL and HIBL models with the behavior of human participants and split this analysis between the three types of feedback (Immediate, Clustered, and Aggregated). We are most interested in the gap between the number of risky choices made in the description and experience conditions. These values are shown on Figure 3 as the percentage point difference describing the RDEG in each type of feedback. For the HIBL model, there is a roughly similar gap as in human behavior, and the same relationship of a decreasing gap is observed as information lesses. Meanwhile, for the IBL model the RDEG is higher for the immediate feedback and lower for both delayed feedback conditions. While the simulated IBL and HIBL model corresponds reasonably well with human behavior generally, when we limit the analysis to behavior after lucky outcome periods, as shown in Figure 4, neither the HIBL or IBL replicate the effects shown in human participants. In the next section we perform model tracing to see if either of these models can replicate this trend.

Model Tracing

Model tracing is done by iteratively replacing the memory of an IBL or HIBL model with the choices and outcome observations of an individual participant and then using that model to predict the next action that the human will take (R. Thom-

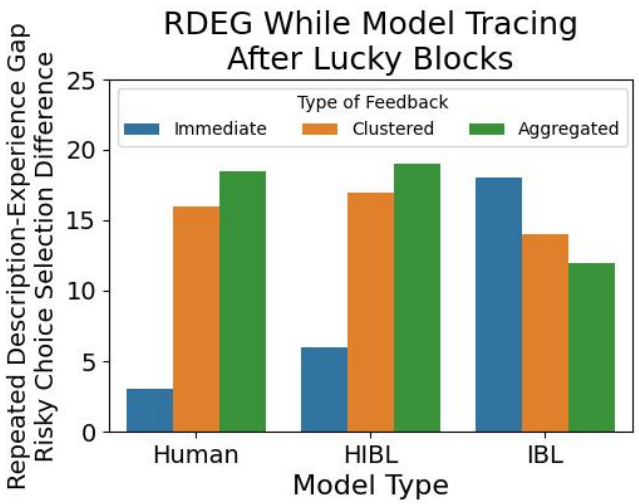

Figure 5: RDEG metric across model types and conditions. This analysis is limited to decisions made in a 10 trial period after a block with 6 or more high valued outcomes.

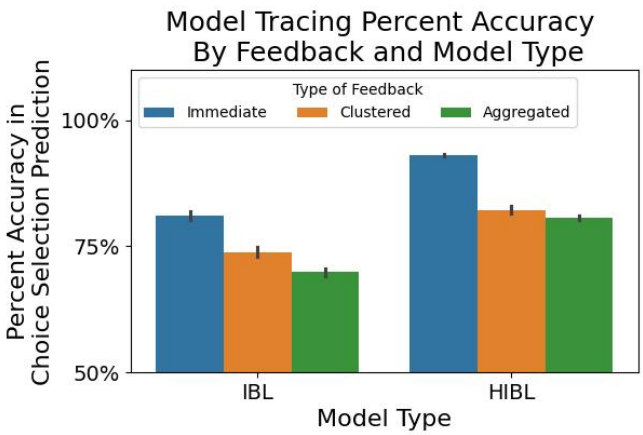

Figure 6: Percent accuracy in choice selection prediction of both types of model, for each condition. Error bars represent 95% confidence intervals at the time-step level.

son et al., 2021; Cranford et al., 2024; R. H. Thomson et al., 2024). Model Tracing can optionally involve fitting the parameters of IBL models to individual behavior by varying these parameters and using the values that produce the highest proportion of correct predictions. This approach is used in our comparison of IBL and HIBL models by testing all combinations of parameter values in steps of 0.1 in a range of 0-1. An important note of this is that the sub-agents in the HIBL model do not have their parameters adjusted during the fitting stage so that the same parameters are being fit between the IBL and HIBL models for a fairer comparison of model tracing accuracy.

The first important result of model tracing is visible in Figure 5, which shows that both human participants and HIBL models show a higher RDEG in the two delayed feedback conditions. This behavior was not observed in the IBL or HIBL models when human behavior was merely simulated. However, by tracing the behavior of real human participants, we observe that the HIBL model behavior in terms of the RDEG following lucky periods was closer to that observed in humans. The standard IBL model shows a reverse of the trend we observe in humans, with a higher RDEG after lucky blocks for immediate compared to delayed feedback.

The average accuracy of the model in predicting each of the decisions of each individual, is shown in Figure 6. The results indicate a higher percent accuracy in choice selection prediction in the HIBL model for all types of feedback, with the difference between the model's accuracy in the immediate condition being the largest difference. This makes sense as the behavior of the HIBL model replicated how humans responded to periods of lucky outcomes, while the standard IBL model showed a reversal of the observed trend.

An important point of this model fitting is that the IBL and HIBL models fit their parameters using the same method of testing all combinations of parameters in steps of 0.1 in the range of 0-1, and they both used the same parameters that define IBL model behavior in this type of task, decay, temperature, noise, and the number of pre-populated instances in model memory prior to the start of the task. However, the HIBL model did require a definition for the behavior of the two sub-agents that would differentiate their choices, which was done according to the specifics of this decision making task. We felt that this was a reasonable assumption that kept the number of fit parameters between models the same. In the following section we discuss potential future work that can investigate how these sub-agents can be trained to further support the predictions of the HIBL model.

Discussion

In this work, we introduce the Hierarchical IBL model for predicting human behavior from various different timings and presentations of choice outcomes. We showed through simulation of HIBL models that this approach produces more human-like behavior as measured by both the proportion of risky choices and the relationship between high-value lucky-reward observations and the amount of risky choices made subsequently. Additionally, we showed through model tracing that the HIBL model is better able than IBL to predict the behavior of individual participants in settings with delayed feedback. These results demonstrate that the HIBL approach is valuable for understanding how humans may make decisions when feedback is significantly delayed. Specifically, this analysis suggests that human decision makers may construct multiple lower-level strategies that they choose from.

Another result we observed was the difference in how participants responded to lucky outcomes between different types of feedback. In conditions with experience, getting lucky in the immediate feedback condition resulted in a reduction of the description-experience gap, whereas this was not as clear in the clustered and delayed feedback settings.

This observation represents the inverse of the previously described effect whereby individual bad outcomes were hidden in the aggregate feedback resulting in a relatively higher risk selection, here individual good outcomes were also hidden in the aggregate resulting in relatively lower risk selection.

While the model fitting we performed on individuals took reasonable measures to make the comparison between IBL and HIBL as fair as possible, as noted previously it is possible that the predefined behavior of the sub-agents contributed to the high accuracy of the HIBL model. Additionally, it is conceivable that a credit assignment method exists that could reproduce similar behavior while using a standard IBL model.

We plan to investigate these possibilities further in future work that will incorporate choice features into these tasks. Contextual bandit tasks have been used to understand how human attention impacts utility learning based on choice attributes (Niv et al., 2015). This has been extended into transfer of learning domains (Malloy et al., 2023), as well as delayed feedback (Malloy et al., 2025), but as of yet these two concepts have not been explored simultaneously.

Acknowledgments

This research was sponsored by the Army Research Office and accomplished under Australia-US MURI Grant Number W911NF-20-S-000, and the AI Research Institutes Program funded by the National Science Foundation under the AI Institute for Societal Decision Making (AI-SDM), Award No. 2229881.

References

Anderson, J. R., & Lebiere, C. J. (2014). *The atomic components of thought*. Psychology Press.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of behavioral decision making*, *16*(3), 215–233.

Bein, O., Gasser, C., Amer, T., Maril, A., & Davachi, L. (2023). Predictions transform memories: How expected versus unexpected events are integrated or separated in memory. *Neuroscience & Biobehavioral Reviews*, 105368.

Botvinick, M. (2012). Hierarchical reinforcement learning and decision making. *Current opinion in neurobiology*, *22*(6), 956–962.

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480.

Chambliss, D. F., & Takacs, C. G. (2014). *How college works*. Harvard University Press.

Cranford, E. A., Lebiere, C., Gonzalez, C., Aggarwal, P., Somers, S., Mitsopoulos, K., & Tambe, M. (2024). Personalized model-driven interventions for decisions from experience. *Topics in Cognitive Science*.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization science*, *12*(5), 523–538.

Eckstein, M. K., & Collins, A. G. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences*, *117*(47), 29381–29389.

Eppe, M., Gumbsch, C., Kerzel, M., Nguyen, P. D., Butz, M. V., & Wermter, S. (2022). Intelligent problem-solving as integrated hierarchical reinforcement learning. *Nature Machine Intelligence*, *4*(1), 11–20.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, *112*(4), 912.

Fox, C. R., & Hadar, L. (2006). "decisions from experience"= sampling error+ prospect theory: Reconsidering hertwig, barron, weber & erev (2004). *Judgment and Decision making*, *1*(2), 159–161.

Gonzalez, C. (2017). 13 decision-making: A cognitive science perspective. *The Oxford handbook of cognitive science*, 249.

Gonzalez, C. (2024). Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments. *Perspectives on Psychological Science*, *19*(5), 860–873.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, *118*(4), 523.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Gonzalez, C., & Mehlhorn, K. (2016). Framing from experience: Cognitive processes and predictions of risky choice. *Cognitive science*, *40*(5), 1163–1191.

Gonzalez, C., Meyer, J., Klein, G., Yates, J. F., & Roth, A. E. (2013). Trends in decision making research: How can they change cognitive engineering and decision making in human factors? In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 57, pp. 163–166).

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychological review*, *113*(3), 461.

Hertwig, R. (2015). Decisions from experience. *The Wiley Blackwell handbook of judgment and decision making*, *2*, 239–267.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534–539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517–523.

Hitchcock, P., Forman, E., Rothstein, N., Zhang, F., Kounios, J., Niv, Y., & Sims, C. (2022). Rumination derails reinforcement learning with possible implications for ineffective behavior. *Clinical Psychological Science*, *10*(4), 714–733.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127). World Scientific.

Krishnan, S., Garg, A., Liaw, R., Miller, L., Pokorny, F. T., & Goldberg, K. (2016). Hirl: Hierarchical inverse reinforcement learning for long-horizon tasks with delayed rewards. *arXiv preprint arXiv:1604.06508*.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational behavior and human decision processes*, *116*(2), 286–295.

Malloy, T., Du, Y., Fang, F., & Gonzalez, C. (2023). Accounting for transfer of learning using human behavior models. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 11, pp. 115–126).

Malloy, T., & Gonzalez, C. (2024). Applying generative artificial intelligence to cognitive models of decision making. *Frontiers in Psychology*, *15*, 1387948.

Malloy, T., Seow, R., & Gonzalez, C. (2025). Modeling attention during dimensional shifts with counterfactual and delayed feedback. In *arxiv.*

Martin, J. M., Gonzalez, C., Juvina, I., & Lebiere, C. (2014). A description–experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making*, *27*(4), 349–362.

Nachum, O., Tang, H., Lu, X., Gu, S., Lee, H., & Levine, S. (2019). Why does hierarchy (sometimes) work so well in reinforcement learning? *arXiv preprint arXiv:1909.10618*.

Nguyen, T. N., & Gonzalez, C. (2022). Theory of mind from observation in cognitive models and humans. *Topics in cognitive science*, *14*(4), 665–686.

Nguyen, T. N., McDonald, C., & Gonzalez, C. (2023). Credit assignment: Challenges and opportunities in developing human-like ai agents. *arXiv preprint arXiv:2307.08171*.

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.

Pateria, S., Subagdja, B., Tan, A.-h., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensive survey. *ACM Computing Surveys (CSUR)*, *54*(5), 1–35.

Rasmussen, D., Voelker, A., & Eliasmith, C. (2017). A neural model of hierarchical reinforcement learning. *PloS one*, *12*(7), e0180234.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*(5), 527–535.

Sims, C. R., Neth, H., Jacobs, R. A., & Gray, W. D. (2013). Melioration as rational choice: sequential decision making in uncertain environments. *Psychological review*, *120*(1), 139.

Sutton, R. S., Barto, A. G., et al. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, *11*(1), 126–134.

Thomson, R., Cranford, E., & Lebiere, C. (2021). Achieving active cybersecurity through agent-based cognitive models for detection and defense. In *Proceedings of the 1st international conference on autonomous intelligent cyber-defence agents (aica 2021).*

Thomson, R. H., Cranford, E. A., Tucker, G., & Lebiere, C. (2024). Comparison of cognitively-inspired salience and feature importance techniques in intrusion detection datasets. In *Assurance and security for ai-enabled systems* (Vol. 13054, pp. 186–196).

Thorndike, E. L. (1898). *Animal intelligence.* JSTOR.