Chapter 4

Confirmation Bias in Probabilistic Learning Tasks

In the previous chapter, we considered bias in auditory attention. In this chapter, we continue exploring bias, but in a different setting. When people are confronted with large amounts of data or uncertainty, they rarely have the time or cognitive resources to conduct a systematic analysis of all available information. Instead, the person will usually make a decision using a heuristic, which may lead to bias and potentially costly errors. A recommender system can try to mitigate potential bias by examining all the relevant information and providing feedback or recommendations to the user to guide their decision. In some cases, the recommender system may inform the user of how confident it is in the recommendation. The user may choose to use the recommender system's suggestion or choose some other option. To understand the role of bias in this type of interaction, we designed and compared four different cognitive models of a task where participants are required to choose between two alternatives. They receive feedback from a computer program that is sometimes incorrect. To evaluate the models, we computed their accuracy in reproducing behavioral data. We show that using an instance-based learning approach that weights past experiences highly

is effective in simulating users in this task. Using this model, we show that when users are warned about the potential for incorrect feedback, they weight past experiences much higher than when they are not warned, leading them to ignore feedback more often than they should [19]. This has implications in the design of systems that make potentially inaccurate recommendations when faced with uncertain data. The work presented in this chapter was originally published in the *Proceedings of the 16th International Conference on Cognitive Modeling* [20] and the *Proceedings of the 2017 International Joint Conference on Artificial Intelligence* [71], and presented at the *Society for Judgement and Decision Making Annual Conference 2018* [72]. [1]

4.1 Introduction

In this chapter, we model the cognitive mechanisms involved in a decision making task where an agent makes decisions in an uncertain environment where they receive potentially incorrect feedback. In an ideal world, decision-makers would have the time and cognitive resources to review all relevant factors necessary to make an informed decision. However, this is often not realistic. In many real-world situations, relevant information may be missing or incorrect, and time pressure does not allow for a full systematic analysis of all available data. Situations like these often lead to the use of heuristics, or rules of thumb, in important decision-making tasks. Although heuristics can lead to accurate decisions in many realistic scenarios [21], they can also lead to systematic errors and biases [73].

Decision support systems have the potential to help users reduce or remove cognitive biases in decision making, either by offering recommendations or showing only the information relevant to a decision [74]. Making sure such systems are effective requires an understanding of the underlying cognitive mechanisms and behavior of a user [3, 75] through behavioral experiments, user testing a prototype, simulation, or

[1] Code relating to this project can be found at: https://github.com/jaelle/matchmaker

some combination of these.

As discussed in 2.1, cognitive architectures offer one approach to simulating users. These provide a combination of cognitive theories and computational tools that can be used to model a user completing a particular task [76]. They simulate human behavior using cognitive mechanisms and theories, which provides insights into the underlying strategies that users employ and can help to identify the strengths and weaknesses of a particular design choice.

In this chapter, we examine how displaying warnings about possibly incorrect feedback can affect a decision maker's ability to learn from new information and make effective decisions. In situations where the system may provide some inaccurate information, it may seem advisable to warn the user so that they can more carefully scrutinize the available information [77]. Similar to what was done in Chapter 3, in the context of spatial auditory attention, we simulate users completing a probabilistic learning task using ACT-R. Our results show that warnings may not always improve decision making and can instead cause a user to give more weight to their past experiences and ignore new information, even when it is not rational to do so.

4.2 Background

Two theories have been proposed to explain how people use their past experiences to make decisions, including instance-based learning (IBL) [7] and utility learning theory [8]. In this chapter, we demonstrate how ACT-R can be used to compare instance-based learning and utility learning models of subjects' decision-making process in a probabilistic learning task. We show which models are useful in modeling behavior when people are warned about potentially incorrect feedback and what those models can tell us about the underlying cognitive process involved in processing such warnings.

4.2.1 Instance-Based Learning

Instance-based learning methods propose that humans adapt their decision-making strategies in response to past experiences. Initially, decision-makers may use a heuristic approach and make a choice based on a simple rule. As they gain more experience and encounter similar situations, they adjust their strategies to choose actions that previously resulted in favorable outcomes [32].

When modelers implement instance-based learning in ACT-R, they store decision making experiences as instances in declarative memory (introduced in Section 2.1). Instances are made up of three components, including the situation, the decision, and the outcome. The situation includes the attributes relevant to the decision, chosen by the modeler. The decision would be how the model responded in that situation. Finally, the outcome can be represented as 1 if it were positive, or 0 in the case of a negative outcome.

For example, if the agent is asked to respond to a query of $2 + 2$, then the situation would be the attributes making up the query (i.e., operator=sum, num1=2, num2=2). They may respond correctly or incorrectly, based on the knowledge that has been added by the modeler or learned from the environment. If the agent responds with 4 and receives positive feedback, a new instance would be stored, that contains the situation (operator=sum, num1=2, num=2), the decision (decision=4), and the outcome (outcome=1). This information is stored as a tuple in memory as (operator=sum, num1=2, num2=2, response=4, outcome=1). If the agent incorrectly responded 5 and received negative feedback, a tuple (operator=sum, num1=2, num2=2, response=4, outcome=0) would be stored in memory.

If the agent is asked about $2 + 2$ in the future, it will make a retrieval request from ACT-R, which would return instances with sufficiently high enough activation, calculated using base-level learning as described in Equation 2.2. Other components of activation, such as partial matching and spreading activation, may also be used.

When decisions are made by retrieving instances from memory, they are considered to be intuitive. Modelers often simulate more deliberate decision making processes using a hybrid approach that combines utility learning and instance-based learning [31, 78].

4.2.2 Hybrid Models with Utility Learning

Utility learning models simulate a deliberate decision-making process where the agent can try different approaches to solving a problem. Over time, the agent will learn to use actions that produce the best outcome. A modeler can create several production rules that represent different ways of solving a problem or making a decision. The agent will then complete a task by choosing a rule that best fits the current situation. As long as a rule exists that matches the current situation, the agent will continue to retrieve and act upon it. Once the agent receives feedback, the utility value for each rule that led to that feedback will be updated using the time discounted equation described in Equation 2.7. As the agent encounters new trials, it uses feedback to refine its decision-making process and adapt to the environment.

By combining procedural memory with instance-based learning, modelers can simulate more deliberate decision making based on explicit rules. With utility learning, the modeler can provide the agent with some basic decision-making rules. Using these rules as a basis, the agent can simulate how a decision-maker tries different approaches to learn the one that works best in the given situation.

4.3 Related Work

In the following subsections, we discuss related work that has applied cognitive models to learning and decision making.

4.3.1 Cognitive Models for Predicting Human Behavior

Several examples exist where decision support system interfaces and their predictions were improved by incorporating theories and models of cognition. In a recent human choice prediction competition, Plonsky et al. [2] showed that machine learning algorithms incorporating cues based on observed cognitive processes were more effective than systems that did not. Cognitive models have also been used to generate synthetic data used to train machine learning algorithms that predict human behavior [79] or generate accurate predictions on their own. Lebiere, Christian et al. [80] demonstrate how cognitive models can use spatial information gathered from real-world data to generate recommendations for users during a natural disaster or other emergencies. In Cranford et al. [81], an IBL model was developed in ACT-R to simulate how humans might learn to react to deceptive strategies in a cyber attack. By observing the cognitive model's behavior, researchers optimized an algorithm that decides when to present deceptive warnings to reduce the likelihood of an attack.

The models presented in this chapter are, to the best of our knowledge, the first to model confirmation bias in a setting where a system warns its users about potentially incorrect feedback.

4.3.2 Cognitive Models of Learning

Researchers have used instance-based learning (IBL) to model a variety of tasks involving learning and decision making. Cleotilde Gonzalez et al. [82] reviews many of the dynamic decision making tasks that were successfully modeled with IBL, such as stock portfolio management, and hiring decisions. Instance-based learning has also been used to model decisions involving probability adjustment and resource allocation as instinctive choices made without deliberation or explicit awareness in [83].

Utility learning is another common approach to modeling learning in humans. Researchers have used utility learning to effectively model a variety of experiments

conducted to study behavior in recurrent choice and skill learning [84]. It has also been used to model heuristics. In [85], the authors created an ACT-R model that used utility learning to show how feedback causes heuristics to become the dominant strategy.

Researchers in artificial intelligence have incorporated both utility learning and instance-based learning into autonomous systems and robots for a variety of purposes. In [86], a classification system was designed to recognize facial expressions. It used a hybrid cognitive model that incorporated both instance-based learning and utility learning to identify which features were necessary for classifying an emotion. In [31], another hybrid cognitive model used utility learning and instance-based learning to simulate how young children learn to follow the gaze of others. A robot was programmed with the resulting model to show awareness about its environment and the people in it.

The work presented in this chapter is novel since it compares both instance-based and hybrid models to determine the efficacy of using cognitive models of memory and learning to simulate humans making decisions in the face of unreliable information.

4.4 Methods

We test four approaches to modeling confirmation bias in the probabilistic learning task. The first approach utilizes spreading activation and instance-based learning to apply a higher weight to a cue that participants were biased to believe was essential for making correct decisions. The second approach is based on the heuristic strategy, *take the best*. In *take the best*, decision-makers order a set of cues and then try each one in order until finding one that can differentiate between two choices [1]. Using procedural memory and utility learning, we created a model that learns which factors are important for distinguishing between two choices and then makes decisions based

on between 1-3 of those factors while ignoring the remaining information.

In this section, we first describe the behavioral task being modeled and then present the four cognitive models that we have designed using either IBL or a hybrid approach that incorporates utility learning. These models are used to simulate decision making in a probabilistic learning task. We compare each model to the results of the behavioral experiment described in the next section and find that the IBL models are more accurate at simulating decision making in our task than the hybrid models. Using the IBL approach, we show that spreading activation (see Section 2.1.2) can be used to model situations where a person applies a disproportionate amount of attention to some feature to make decisions.

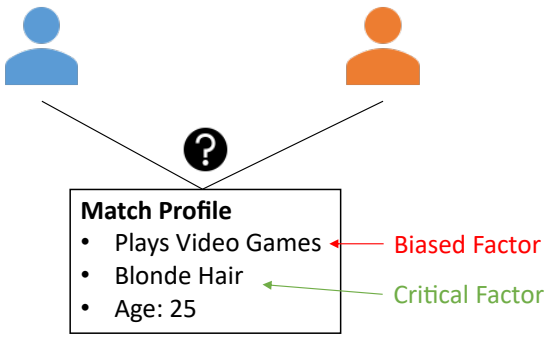

Figure 4.1: In the behavioral task, participants play matchmaker and choose a matching profile that best suits a particular bachelor. They are initially biased to believe that the bachelor only cares about what the potential match does for entertainment. However, the critical factor important to making correct matches is actually hair color.

4.4.1 Behavioral Experiment

We designed a behavioral task to examine the effect of warnings on the effectiveness of decision-makers. In the task, subjects (N=99), were asked to play the role of a matchmaker and choose one of two bachelors (Frank or James) for a given match, based on the match's attributes (hair color, hobby, entertainment preference, age, and drinking habits). Initially, subjects were given example profiles that they were told each bachelor would like. The examples led participants to an incorrect biased

belief that entertainment preference was critical to choosing the correct bachelor. For the remainder of the experiment, hair color was critical to making accurate matches.

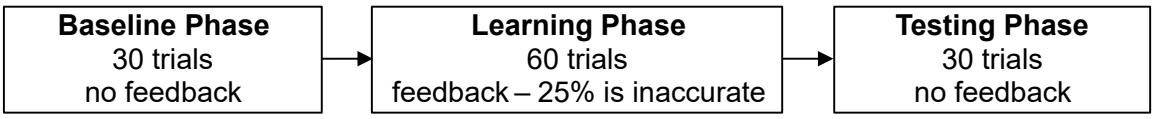

Figure 4.2: Subjects completed three phases in the experiment.

Figure 4.2 illustrates how the experiment proceeded. After viewing the examples, subjects completed a Baseline phase of 30 trials, without feedback, to test whether or not they were making matches based on the entertainment attribute. After that, subjects began the Learning phase, lasting 60 trials. Feedback in the Learning phase was incorrect 25% of the time. 51 of the subjects received a warning about the potentially incorrect feedback, and the remaining 48 did not. Following every trial in the Learning phase, the Warned group was asked to rate their level of trust in the feedback. After completing the Learning phase, subjects completed a Testing phase of 30 trials to test whether or not they were able to overcome the initial bias to make accurate decisions.

Throughout the experiment, subjects respond to three types of trials: Congruent, Incongruent, and Irrelevant.

Congruent trials pair the learned bias with the correct value for the critical attribute. An example of a congruent trial is $entertainment = videogames$ being paired with $haircolor = blonde$. Participants have been biased to believe that James likes matches who like video games, and in reality, he likes matches with blonde hair, so the biased and critical factor is congruent.

Incongruent trials pair the learned bias with an incorrect value for the critical attribute. An example of an Incongruent trial is $entertainment = videogames$ being paired with $haircolor = brown$. James does not like matches with brown hair, but the participant may believe this is a good match for him, using the established bias

that James likes matches who like video games.

Finally, irrelevant trials pair the critical attribute with an entertainment preference that is not related to the participants' learned bias. An example of an Incongruent trial is $entertainment = sports$ (sports is not a learned value for the biased attribute) paired with $haircolor = blonde$. If participants have correctly learned that hair color is critical for making correct matches, they can correctly make this match. However, they cannot make the match based on the biased entertainment factor established at the start of the experiment.

From the responses, it is possible to analyze how well subjects mitigated the bias and increased their match accuracy. Participants can make correct responses for each trial type by matching on the critical factor (hair color). If participants make matches based on the established biased factor (entertainment), they will be correct on Congruent trials and incorrect on Incongruent trials. Finally, their bias was not useful for making correct decisions in Irrelevant trials.

4.4.2 Model Design

Using ACT-R, we design cognitive agents to test four different theories about the strategies that may be used to describe how human subjects respond to warnings about incomplete information and can be tested on the collected behavioral data.

Instance-Based Learning Agents

We developed two agents (IBL Model 1 and IBL Model 2) that used instance-based learning models to simulate intuitive decision making in the experimental task.

Both IBL models are made up of four production rules, which allow the agent to complete the same behavioral task as the human subjects. Two rules simulate a heuristic that an incorrect feature (entertainment preference) is critical for choosing the correct bachelor. One rule will choose to assign a match to James if the match

```
(p match-entertainment-james
   =goal>
    isa match
    hair =hair
    entertainment videogames
   ==>
   +goal>
    name james
    entertainment videogames
    hair =hair)
```

Figure 4.3: Example of an ACT-R production rule. The rule checks to see if the current match's entertainment attribute value is videogames. If so, ACT-R creates a new instance in the goal buffer, pairing James with match and its associated attributes for entertainment and hair.

prefers video games (an example of the model code is provided in Figure 4.3). The other rule will choose to assign the match to Frank if the match prefers sports. If the match prefers neither sports nor video games, a third rule will assign the match to either Frank or James with 50% probability. These rules emulate how the human subjects were biased to believe that entertainment was critical to making a successful match. In addition to these heuristic rules, a fourth rule is present that assigns a match to a bachelor chosen based on previous experiences, using the instance-based learning approach described in Section 4.2.1. This rule looks for instances stored in memory and chooses the one with the highest activation value. The two IBL models differ in how each calculates activation.

IBL Model 1. This model simulates how people make decisions based on past experiences. The activation value is determined as described in Equation 2.1, using the partial matching (Equation 2.3) to account for similarity, spreading activation (Equation 2.4) to model the effect of context, and base-level learning (Equation 2.2) to account for how often and recently an instance has been retrieved. In this model, spreading activation models attention being spread equally among each cue in a decision making scenario.

IBL Model 2. This model was designed to test the theory that people give a higher attentional weight to the attribute that they were first biased to believe led to

correct matches. By default, spreading activation in ACT-R distributes attentional weights evenly among all attributes in an instance. In order to model more attention applied to the entertainment preference attribute e, the Equation 2.4 was modified to include a biased weight parameter, represented by b in the adjusted equation below:

$$S_i = bS_{e,i} + \sum_{j \notin e} W_j S_{ji}, \text{ where } W_j = 1/n \text{ and } S_{j,i} = S_{e,i} = S - \ln(fan_j). \quad (4.1)$$

Hybrid Agents

In order to simulate more deliberate decision making, we developed two models that use instance-based learning in conjunction with utility learning to simulate using a deliberate process to examine one or more features systematically before making a decision. Over time, these models learn which features are critical for making correct decisions and begin to ignore the remaining information.

Both models include rules that encode features and assign a match to a bachelor, given some encoded features. Encoding rules are triggered if features in the current trial are not yet encoded. Matching rules are triggered if at least one feature is encoded. The model retrieves the rule with the highest utility. Throughout the task, it is possible to select both encoding and matching rules. Following the approach outlined in Equation 2.7, the model retrieves the rule with the highest utility, and if multiple rules have the same utility, then one is retrieved probabilistically. As soon as a matching rule's utility surpasses that of a feature encoding rule's utility, the model tries to retrieve a past instance that shares those features, using an instance-based learning approach. If the retrieval request fails, then the model will continue encoding features and using matching rules to retrieve similar matches from declarative memory. If the model is unsuccessful in retrieving a memory by the time all features are encoded, then it defaults to matching the bachelor randomly.

Once the model chooses which bachelor is a match for the current trial, it receives

feedback and updates the utilities of all the production rules that led to that outcome. Over many trials, the model learns which features are important for receiving positive feedback, allowing it to make the decision based on a few high utility features and ignoring the rest.

The two models differ only in their implementation of the *take the best* heuristic.

Hybrid Model 1. The first model simulates a traditional approach to *take the best*, looking at each feature in turn and choosing the match once a feature successfully acts as a cue to retrieve a memory that previously resulted in positive feedback.

Hybrid Model 2. Sometimes, a matching memory may be available in declarative memory, but matching on a single feature is not sufficient to achieve an activation value that exceeds the retrieval threshold. In this situation, Hybrid Model 1 will continue encoding and matching until it has tried all combinations of up to three features. This approach allows the agent to simulate a more systematic approach to examining more information than is possible with a traditional *take the best* heuristic.

4.5 Experimental Design and Results

We will now discuss the experimental design used to compare the four agents and the results.

4.5.1 Instance-based Learning Agents

Both ACT-R agents completed 120 trials in the same order presented to human subjects. First, the agent completed 30 trials of a Baseline phase to establish the bias to make matches based on entertainment preference. After the Baseline phase was complete, the agent proceeded to the Learning phase, which consists of 60 trials. During this phase, the agent received inaccurate feedback about whether or not the match was successful 25% of the time. During the learning phase, the agent used

Parameter	Values Tested
ans	0.25, 0.5, 0.75, 1, 1.25, 1.5
mp	1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3
rt	0, 0.5, 1, 1.5, 2
bw (only in IBL Model 2)	0.2, 1, 5, 10

Table 4.1: Parameter values tested for IBL models.

an instance-based learning rule to base its decisions on previous experiences, rather than the initial biased rules. Finally, upon completing the Learning phase, the agent advanced to the Testing phase and did not receive any more feedback. This phase was used to evaluate the final decision-making strategy.

Model Parameters

ACT-R provides several parameters that can be adjusted to fine-tune the performance of the model. For both IBL models, five of these parameters were chosen using a parameter search over several possible values. First, noise (ans), which is used to add noise to the activation values of the instances, making the retrieval process stochastic. Second, a mismatch penalty constant (mp) was adjusted to represent the degree of similarity that must exist between a chunk in the buffer and those retrieved from declarative memory. If mp is substantially large, the chunks must be very similar to be retrieved. If it is a smaller value, then the chunks can be less similar. The retrieval threshold parameter (rt) represents how high the activation value must be for a chunk to be retrieved [87].

The parameter search for the IBL Model 2 required the sixth parameter to be added (bw). As described in Section 4.4.2, this parameter adjusts the amount of attentional weight applied to the entertainment preference attribute.

For every combination of parameter values described in Table 4.1, the associated model was run 100 times to simulate 100 agents completing the behavioral task. We calculate R^2 and $RMSE$ for each run by comparing its average performance with the

average performance of the Warned and Unwarned human subjects. The results are discussed in Section 4.5.

A grid search was used to identify the parameter combinations that led to the best fit of both the Warned and Unwarned group. Given the limited number of combinations, the search tried all combinations of parameter values described in Table 4.1 for the IBL models and 4.5 for the hybrid models. We discuss below the results corresponding to the parameters that led to the highest R^2 values for each model for the Warned and Unwarned groups.

Condition	Model	Congruent	Incongruent	Irrelevant
Warned	**Human**	**0.92**	**0.18**	**0.60**
Warned	IBL Model 2	0.91	0.18	0.60
Warned	IBL Model 1	0.78	0.27	0.55
Warned	Hybrid Model 2	0.96	0.10	0.49
Warned	Hybrid Model 1	0.5	0.48	0.51
Unwarned	**Human**	**0.86**	**0.32**	**0.62**
Unwarned	IBL Model 2	0.82	0.32	0.61
Unwarned	IBL Model 1	0.78	0.34	0.61
Unwarned	Hybrid Model 2	0.96	0.10	0.49
Unwarned	Hybrid Model 1	0.5	0.48	0.51

Table 4.2: Comparison of the models' and human's accuracy for both Warned and Unwarned groups over each trial type. Models are sorted by best to worst fit for each condition. Human performance is in bold. IBL Models replicated human behavior more closely than Hybrid Models, with IBL Model 2 performing the best.

IBL Model 1. This model implemented instance-based learning with ACT-R's implementation of spreading activation, as described in Section 2.1.2. The resulting R^2 values for the best fitting parameters were relatively good. The parameters leading to the best R^2 and $RMSE$ values are reported in Table 4.3. We also examined the resulting accuracy of the model and human behavior to see how well the model was able to simulate human performance in the Congruent, Incongruent, and Irrelevant trials. Table 4.2 shows that IBL Model 2 reproduced a proportion of accurate responses very similar to human data in both the Warned and Unwarned conditions and across all trial types.

Model	ans	mp	rt	R^2	RMSE
Warned	0.5	1.25	0	0.87	0.12
Unwarned	1.5	2.75	1.5	0.91	0.08

Table 4.3: IBL Model 1 - Parameter values that resulted in the best fit (lowest $RMSE$) for the Warned and Unwarned group.

IBL Model 2. This model implemented instance-based learning with a modified implementation of spreading activation, as described in Section 2.1.2. The resulting R^2 and $RMSE$ measures for the best fitting parameters are reported in Table 4.4 and a comparison of each model's accuracy to that of humans is found in Table 4.2.

Model	ans	mp	rt	bw	R^2	RMSE
Warned	1.25	1.75	1	10	0.95	0.08
Unwarned	0.75	2.25	2	1	0.93	0.08

Table 4.4: IBL Model 2 - Parameter values that resulted in the best fit (lowest $RMSE$) for the Warned and Unwarned group.

4.5.2 Hybrid Agents

For both hybrid models, the values of the five parameters were set using a grid search over several possible values.

Like the IBL models, noise (ans) and retrieval threshold (rt) are free parameters in the hybrid models. Three additional free parameters were also fit for the hybrid models. Spreading activation weight (ga) represents the total amount of attentional weight that gets divided amongst each attribute (S in Equation 2.4). The utility noise parameter (egs) represents how much noise gets added to the production utilities. Finally, we fit the reward parameter, which represents a value that gets propagated to the production rules that led to a correct response, using the function described in 2.7.

For every combination of parameter values described in Table 4.5, the associated model was run 100 times to simulate 100 agents completing the behavioral task. We

Parameter	Values tested
egs	0.5, 1, 1.5, 2, 2.5, 3
reward	1, 2, 3, 4, 5, 6, 7, 8
ans	0.25, 0.5, 0.75, 1, 1.25, 1.5
rt	0, 0.5, 1, 1.5, 2

Table 4.5: Parameter values tested for hybrid models.

calculate R^2 and $RMSE$ for each run by comparing its average performance with the average performance of the Warned and Unwarned human subjects. The results are described in Section 4.5.

The results showed that the hybrid learning models using instance-based learning combined with utility learning were not as effective at modeling human performance.

Hybrid Model 1. Utility learning was used to model the *take the best* heuristic, choosing the first recalled cue that differentiates between two bachelors. The parameters that achieved the highest R^2 and $RMSE$ values for each condition are reported in 4.6. Table 4.2 compares the accuracy of the model to that of Warned and Unwarned participants.

Model	egs	reward	rt	ans	ga	R^2	RMSE
Warned	6	8	1.5	0	0.25	0.83	0.27
Unwarned	6	8	1.5	0	0.25	0.80	0.23

Table 4.6: Hybrid Model 1 - Parameter values that resulted in the best fit (lowest $RMSE$) for the Warned and Unwarned group.

Hybrid Model 2. This model implemented a modified version of *take the best*, where the agent chose the matching bachelor based on the first combination of up to three cues that differentiated between them. This approach was a better fit than the first hybrid model but was still not as good as pure instance-based learning approaches. The parameters that achieved the the highest R^2 and $RMSE$ values for each group are reported in Table 4.7, while Table 4.2 compares the accuracy for each trial type.

Model	egs	reward	rt	ans	ga	R^2	RMSE
Warned	1	6	1	0.3	0	0.85	0.18
Unwarned	1	6	1	0.3	0	0.78	0.21

Table 4.7: Hybrid Model 2 - Parameter values that resulted in the best fit (lowest $RMSE$) for the Warned and Unwarned group.

4.6 Discussion

In the previous sections, we described and examined four approaches to modeling a probabilistic learning task. These included using instance-based learning methods to model intuitive decision making, and hybrid models that used instance-based learning and utility learning to use a more deliberate approach to learn which features were necessary for making correct decisions.

IBL Model 1 used instance-based learning to model how participants make decisions using an intuitive approach that evenly distributed attentional weight across all cues in the current trial. This approach did better than both hybrid models, indicating that the intuitive decision making modeled by instance-based learning is a better explanation for human behavior in this task, compared to the more deliberate decision making modeled by the hybrid models. However, this model did not perform as well as IBL Model 2, which indicates that using spreading activation, which evenly distributes attentional weights across all attributes, is not sufficient to model biased human behavior.

For IBL Model 2, we simulated participants using an intuitive approach but weighting the entertainment cue more than other cues. In Incongruent trials, participants who matched based on entertainment found that doing so did not result in positive feedback. Since the Warned group knew that some feedback was incorrect, they may have discounted this feedback and continued to weight the biased factor highly when making decisions. The experimental data supported this notion since the Warned group achieved lower accuracy in the Testing phase (59%) than the Un-

warned group (64%) on the Incongruent trials. IBL Model 2 represented this by applying a higher weight (*bw*) to the spreading activation supplied by the entertainment attribute.

When comparing the values for the *bw* parameter, the value that best fit the Warned group was much higher than the one that led to the best fit in the Unwarned group (10 vs. 1). The Warned group's large *bw* value indicates that the Warned group was weighting entertainment higher than the Unwarned group and less effective than the Warned group at using the provided feedback to overcome that bias.

The hybrid models were designed to simulate someone deliberately examining one cue after another using a strategy based on the *take the best* heuristic. Using 1-3 cues, the hybrid agents attempted to retrieve a memory of receiving correct feedback. If successful, the agent answered as it did before. If they could not recall a matching instance, the agents would randomly choose which bachelor to assign to the match. Over time, utility learning updated the utilities for production rules that considered cues that are effective for making correct decisions.

Both hybrid models were not very successful at modeling human decision making. Modeling a *take the best* approach with only one cue was not a good explanation of human behavior, and it led to responses that were correct only about 50% of the time for all trial types. Hybrid Model 2 performed a little better, simulating high accuracy on the Congruent trials, low accuracy on the Incongruent trials, and chance accuracy on the Irrelevant trials. However, this model was more biased than its human counterparts. Like human participants, the models were initially biased to believe that entertainment was the most critical factor in making correct matches. It also received positive feedback in Congruent trials, reinforcing the utility of the production rule to recall instances from their entertainment attribute. In Incongruent trials, the utility function reduced the utility of using the entertainment rule, but this was not sufficient to overcome the bias induced at the start of the experiment, given

the parameter combinations we tested. Comparatively, human participants showed more variability in their choices than the hybrid agents, which led to less biased responses overall.

4.7 Summary and Future Directions

We analyzed instance-based learning [7] and utility learning methods [8] for simulating users completing a probabilistic learning task. We compared the output of four models to that of the human data to identify the underlying strategies that are most likely to be used when a user receives a warning about possibly incorrect feedback, compared to when they are not. Our results show that the instance-based learning models simulated human participants much more accurately than the hybrid models. In particular, an IBL model that incorporated a biased attentional weight for the entertainment factor achieved the best fit when compared to human subjects. These results indicate that participants were using an intuitive process for making decisions, rather than the more deliberate approaches and that they were weighting entertainment more highly than other cues in the decision process. Both groups exhibited this bias, but it was especially prominent for the Warned group, as indicated by the higher *bw* values in the model simulating this group.

We considered how people would respond when a recommendation system gives feedback that might be wrong. In a behavioral experiment, we compared a group of participants who were warned about possibly incorrect information with those who were not. Experimental results showed that the Warned group was significantly less accurate at making correct matches than those who were not.

These results are the first step in analyzing how warnings affect how users' ability to process feedback and are significant for considering the design of feedback and warnings in automated recommendation systems.

A next step could be to consider how bias and trust affect the sensitivity to feedback in overcoming bias. During the experiment, warned participants rated (on a scale of 1-5) how much they trusted the feedback they received after each trial in the Learning phase. Overall, they rated Incongruent trials with an average rating of 2.5 and Congruent trials with an average rating of 4. This lack of trust seems to have led the warned participants to ignore feedback more often when it did not align with their preconceptions, leading to more weight applied to the entertainment factor when making decisions. It would be interesting to see if we observe this effect in unbiased participants, where they cannot rely on a preconceived bias to make decisions and must instead use environmental cues (such as feedback) to learn what is important for making correct choices.

In the future, we will also investigate if warnings about recommendations and feedback can be used effectively, such as first addressing a user's preconceived biases. The cognitive model proposed here could be incorporated into an adaptive recommendation system that can tailor feedback and warnings to an individual user to correctly calibrate their trust in the system and lead to better engagement with recommendation systems.