Modeling Bias Reduction Strategies in a Biased Agent

Jaelle Scheuerman
Naval Research Laboratory
Stennis Space Center, MS, USA
Tulane University
New Orleans, LA, USA
jscheuer@tulane.edu

Dina Acklin
Naval Research Laboratory
Stennis Space Center, MS, USA
Louisiana State University
Baton Rouge, LA, USA
dackli1@lsu.edu

Abstract

Costly mistakes can occur when decision makers rely on intuition or learned biases to make decisions. To better understand the cognitive processes that lead to bias and develop strategies to combat it, we developed an intelligent agent using the cognitive architecture, ACT-R 7.0. The agent simulates a human participating in a decision making task designed to assess the effectiveness of bias reduction strategies. The agent's performance is compared to that of human participants completing a similar task. Similar results support the underlying cognitive theories and reveal limitations of reducing bias in human decision making. This should provide insights for designing intelligent agents that can reason about bias while supporting decision makers.

1 Introduction

Decision makers are often relied upon to make quick and accurate decisions while processing non-trivial amounts of data. Due to limitations of memory [Gigure and Love, 2013] and a tendency to recall the most recent and readily available instances of an event [Kahneman and Tversky, 1972], people often must rely on learned biases and heuristic processes. Heuristic decision making uses few cognitive resources and often seems so successful that it can feel intuitive and accurate. This can result in confirmation bias, which is a tendency to value evidence that confirms our currently held beliefs and discount anything that challenges those beliefs. Because confirmation bias can lead to costly errors, it is important to understand the cognitive processes that allow bias to persist and develop strategies that can be used to support decision makers in systematically considering all available information, especially when it contradicts their existing bias.

2 Current Progress

Our present work aims to model the cognitive processes involved in decision bias. We designed an intelligent agent that uses bias reduction strategies when making decisions in the face of biased and sometimes incorrect information. The agent was developed using the cognitive architecture, ACT-R version 7.0, which provides many computational mechanisms to model human behavior using current theories of cognition. It is well suited to modeling decision making tasks [Lebiere and West, 1999; Gonzalez *et al.*, 2007] and provides modules that correspond to specialized cognitive systems. These include a declarative memory module that contains facts (called chunks) that the agent knows, a procedural memory module that provides a pattern matching rule-based production system allowing the agent to adapt to current circumstances, and a goal module that represents top-down control guiding the agent's behavior [Anderson *et al.*, 2004].

Modeling decision making in an ACT-R agent affords many benefits. Since the agent's behavior can be observed and measured, its performance can be compared to that of its human counterparts. This can reveal weaknesses in the underlying theories that make up the agent's design and lead to increased understanding of the underlying cognitive processes [Gazzinga *et al.*, 2002].

Our agent was designed to simulate a human participating in a matchmaker task. The task was structured as a repeated binary choice where the agent decided whether or not a suggested match was compatible with a bachelor's preferences. After the agent completed a set of trials, we measured the average performance of the agent to the mean performance of human participants.

2.1 Matchmaker Task

The matchmaker task was based on an experiment conducted by co-author, Dina Acklin, on 200 participants. It was designed to assess the effectiveness of bias reduction strategies during a probabilistic learning task. In the experiment, participants were intentionally biased to believe certain incorrect information is important to making a decision. They also received feedback that was not completely accurate, reflecting common conditions in real world learning. The matchmaker task addresses three issues that may affect systematic processing of information including continued hypothesis generation, attention to contradicting information, and knowledge of erroneous feedback.

In the experiment, participants were asked to match a bachelor to potential partners based on compatibility factors such as hair color, hobby and entertainment preference. Subjects were first biased to believe an incorrect factor was important to the bachelor. Participants then completed a baseline phase where they were asked to make good matches for bachelors

over 30 trials without feedback to measure how often they choose matches that include the biased factor. The subsequent learning phase consists of 60 trials where the participants received feedback about whether or not their choice was correct. The feedback was probabilistic such that 25% of the time it was inaccurate. Finally, participants completed a test phase of 30 trials without feedback to provide a measure of bias after receiving feedback. Because the critical factor was always presented, a participant using a systematic approach should show increased accuracy on trials that pair the incorrect biased factor (entertainment preference) with a non-critical factor (such as hobby). Experimental results gathered from 200 participants showed significant increases in accuracy from the baseline phase (M = 0.51) to the learning phase (M = 0.56) to the test phase (M = 0.60).

2.2 Designing the Matchmaker Agent

The agent was designed to simulate a human using a systematic approach. The agent generates hypotheses and recognizes when feedback contradicts its bias. As it proceeds through the matchmaker task, the agent generates hypotheses about each compatibility factor's importance and stores these as chunks in declarative memory. Bias towards a particular compatibility factor is simulated by ACT-R's base activation value which is calculated using the Bayesian learning formula:

$$A_i = ln \sum_{j=1}^{n} t_j^{-d} \quad (1)$$

In equation 1, *n* represents the number of past references to a chunk *i*, *t* represents the time since the *j*th reference and *d* represents the decay rate.

When a new match is presented to the agent, production rules in procedural memory fire and the agent retrieves a hypothesis chunk from declarative memory, given the probability computed using the softmax function:

$$P_i = \frac{e^{A_i}/t}{\sum_j e^{A_j}/t} \quad (2)$$

There is a high probability that the hypothesis with the highest activation value is retrieved. However, since memory retrieval is probabilistic, equation 2 includes a noise parameter *t* set to a default recommended value of 0.25 to include some variability in the retrieval process [Bothell, 2016].

The agent compares the retrieved hypothesis to the current match. If the match is congruent with the current hypothesis, then the activation level of that hypothesis will increase and the agent will choose to pair the current match with the bachelor. If it receives positive feedback, the bias level of the current hypothesis increases further. However, if negative feedback is received or if a match is incongruent with this hypothesis, then the agent will generate or retrieve alternative hypotheses that align with the conflicting information and the activation value of the alternative hypotheses will increase. This leads to an increase in the agent's accuracy over time and simulates attention to feedback that contradicts the bias. During the test phase, our agent showed 60% accuracy, which is equal to that of the average human participant.

3 Future Work

Further work is needed to understand the various systematic approaches participants may use to lead to better performance levels in the test phase. In the matchmaker experiment, human participants could be divided into low-, medium- and high-performing groups, based on their accuracy in the final test phase. High performers (n=31) achieved at least 80% accuracy, medium performers (n=117) achieved between 50% and 79% accuracy, and the low performers (n=54) achieved less than 50% accuracy. To test the effectiveness of different systematic approaches, the agent will be modified to simulate various strategies and compare performance with human participants. Approaches that lead to greater accuracy should provide insights into effective strategies for bias reduction in human decision making.

Future work will also incorporate the model into an intelligent agent that interacts with humans to support decision making. Systems that can understand and reason about bias will be better equipped to assist users who must systematically parse through vast amounts of data to make decisions.

Acknowledgments

The design of the agent was performed at the Naval Research Laboratory under funding number N0001417WX00111 from the Office of Naval Research to Noelle Brown of the Naval Research Laboratory.

References

[Anderson *et al.*, 2004] John R. Anderson, Dan Bothell, Mike D. Byrne, Scott A. Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

[Bothell, 2016] Dan Bothell. ACT-R 7 Reference Manual, 2016.

[Gazzinga *et al.*, 2002] Michael S. Gazzinga, Richard B. Ivry, and George R. Mangun. *Cognitive Neuroscience: The biology of the mind*. Norton & Co., New York, 2nd edition edition, 2002.

[Gigure and Love, 2013] Gyslain Gigure and Bradley C. Love. Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences*, 110(19):7613–7618, May 2013.

[Gonzalez *et al.*, 2007] Cleotilde Gonzalez, Christian Lebiere, and M Martin. Instance-based decision making model of repeated binary choice. In *Proceedings of the 8th International Conference on Cognitive Modeling*, Ann Arbor, Michigan, USA, 2007.

[Kahneman and Tversky, 1972] Daniel Kahneman and Amos Tversky. Subjective Probability: A Judgment of Representativeness. In *The Concept of Probability in Psychological Experiments*, pages 25–48. Springer Netherlands, 1972.

[Lebiere and West, 1999] Christian Lebiere and Robert L. West. A dynamic ACT-R model of simple games. In *Proceedings of the Twenty-first Conference of the Cognitive Science Society*, pages 296–301, Mahwah, NJ, 1999.