Chapter 1

Introduction

Predicting human behavior is a challenging task when designing algorithms that support humans and interact with them. Human actions are often guided by heuristics and biases that sometimes lead to decisions that appear irrational [1]. *Heuristics* are rules of thumb that allow us to make decisions in complex environments without systematically considering all of the available information. Heuristics allow us to navigate ever-changing and unpredictable situations, but can occasionally lead to *biased* actions that are suboptimal in some way. In this dissertation, I investigate computational models of bias in human behavior grounded in psychological theories. These models use precise, testable methods to simulate how behavior emerges from underlying cognitive mechanisms and provide important insights into designing algorithms that can simulate and adapt to human bias. I examine heuristics and bias along several dimensions, specifically considering two single agent and one multi-agent scenario where algorithms have the potential to augment human performance and mitigate errors. I study bias in human behavior at different abstract levels of cognition and on tasks of varying complexity. My investigation starts from modeling how an agent attends to sounds in an auditory attention task. I then consider modeling an agent exhibiting confirmation bias in a probabilistic learning task, and,

finally, I analyze heuristics and their effectiveness in a multi-agent voting scenario. In my research, I use behavioral data and relevant psychological theories to inform computational methods based on artificial intelligence techniques such as constraint satisfaction problems, instance-based learning, and reinforcement learning. My experimental results show that these methods are effective in modeling specific aspects of human performance and can be embedded in more general cognitive architectures. The resulting analyses also provide valuable insights for the design of new algorithms to simulate and predict human behavior.

1.1 Background

Understanding and predicting human behavior is important for designing new algorithms and systems that support human performance. Behavioral experiments and psychological theories have guided the design of many AI technologies. For example, Plonsky et al. [2] showed how machine learning features drawn from psychological theories could improve the performance of a random forest algorithm in predicting human choices. Harman et al. [3] conducted an experiment to show how trust in recommender systems evolves differently when its output is personalized, compared to when it is not. Some recent work has started to examine voting in the context of human behavior. Fairstein et al. [4] modeled user's votes as a tradeoff between the attainability of the candidate (obtained through poll data) and their utility. Behavioral data were used to train the model for each participant and then predict their voting behavior. Mennle et al. [5] used an experiment to show that human behavior in a resource allocation task could be explained by greedy search algorithms that performed well on average.

Behavioral experiments are used to create new cognitive models and psychological theories that can support the development of technologies and recommendations to

augment human performance. Implementing a cognitive model as an algorithm requires representing a clearly defined theory about a cognitive mechanism (i.e., memory, learning, attention, etc.) in a transparent and precise way. Cognitive models represent psychological theories as an algorithm that mimics some aspect of measurable human performance, such as reaction time or accuracy. Comparing the cognitive model's output to that of human behavior can help to identify the strengths and weaknesses of the theory being modeled [6]. A variety of computational methods have been successfully applied to modeling cognition. For example, instance-based learning methods have been used to model a variety of learning mechanisms associated with dynamic decision making, including recognition-based retrieval and learning to adapt to the environment [7]. Reinforcement learning methods, including utility learning, have been effective for modeling how people choose between strategies in problem-solving tasks [8]. Algorithms have been used to represent the accumulation of information in a drift diffusion decision process as a Markov chain [9]. Constraint programming has been applied to models of skilled behavior [10] and learning [11].

It is important in the development of computational models to evaluate their performance in a cognitive architecture that models cognitive functions in the context of the mind as a whole. Cognitive architectures represent an area of research in general AI that seeks to model the human mind, including how it reasons and learns, and how it perceives the environment through senses, such as hearing or vision. Several cognitive architectures, including ACT-R [12] and SOAR[13] have been to test different cognitive theories. Such architectures use methods including instance-based learning [7], reinforcement learning [8] and Bayesian inference [12] to represent cognitive processes, such as learning and memory. By comparing their output to real human behavior, it is possible to identify the strengths and weaknesses of a model and draw conclusions about human behavior [12, 7, 14].

1.2 Contributions

Bias can present itself in many forms and many different environments. For example, biases in attention can affect how quickly or accurately a person will attend to an object. Other types of bias, such as confirmation bias, can affect how people encode or remember information. Biases can also result from aggregating reported preferences across groups, such as in voting. Different computational methods lend themselves to modeling different types of bias, depending on the underlying theoretical components being modeled. Therefore, it was important to consider a broad range of biases and the situations in which they may occur and to compare multiple computational methods for modeling these. In the following chapters, I consider several novel methods for analyzing and modeling heuristics and bias at different cognitive levels of abstraction and task complexity. The first level considers an agent's attentional bias in a spatial auditory attention task. The second level is a more complex task, modeling an agent's decision-making capabilities in a probabilistic learning task. Finally, the third level considers a multi-agent voting scenario to identify heuristics observed in approval voting, analyze their effectiveness, and show when they may lead to bias.

1.2.1 Bias in a Spatial Auditory Attention Task

Computational cognitive models are important for advancing scientific understanding of cognition, as well as providing a basis for designing systems that people interact with. While many aspects of cognition, such as visual attention, have been well explored, spatial auditory attention represents an area of cognition that has received limited consideration from the research community. Spatial auditory attention is unique from visual attention, allowing us to sense things at a distance and out of sight. Computational models provide one approach for analyzing the unique properties of spatial auditory attention. Such models form the basis for designing new

interfaces and spaces for situations where sound is important, e.g., immersive virtual environments. Chapter 3 outlines two approaches I developed for modeling attentional bias and response times in a spatial auditory attention task and shows how I incorporated these into a cognitive architecture. The first model represents the attentional mechanisms as a combination of top-down (i.e., goal-driven) attention and bottom-up (i.e., salient) attention. In contrast, the second considers how attention develops over time. I test both approaches on behavioral data in a spatial auditory attention task and show that they are comparable in terms of modeling. I then integrate the models into a cognitive architecture, making them accessible to other researchers investigating how spatial auditory attention operates in the context of the human mind as a whole.

I consider two models that focus on different aspects of attention and require very different computational approaches. The first model represents attentional bias as a combination of top-down and bottom-up attention and is well suited to being modeled as a constraint satisfaction problem (CSP) [15]. CSPs are an AI framework for representing real-world problems as a combination of variables and constraints that define which assignments to those variables are allowed [16]. This approach also allows rapid testing of a variety of hypotheses about how top-down and bottom-up attention combine to produce the attentional bias observed when subjects respond to spatial sounds coming from different locations. The second model focuses on the dynamic process of how information is accumulated over time, resulting in accurate or inaccurate responses. The drift diffusion model is one computational approach to modeling information accumulation processes in binary choice tasks [17]. A random walk algorithm is used to represent a stochastic amount of information gained towards one of two options. Once enough evidence has been accumulated to reach a defined threshold, then the choice is made. I implement the first application of the drift-diffusion model in the context of auditory attention. This approach allows modeling

the speed-accuracy tradeoff that is observed in subjects' responses to the spatial auditory attention task. Together, these methods address two important aspects of spatial auditory attention: including how attention can be modeled as a combination of attentional processes, and how information is accumulated over time. Both models were tested by comparing their output to the responses of participants in the spatial auditory attention task. My results show that both the constraint model and the drift-diffusion model effectively simulate the underlying cognitive mechanisms, resulting in emergent behaviors that replicate human responses. To make the models accessible to other researchers investigating spatial auditory attention, I integrate them into the ACT-R cognitive architecture as an extension to its Audio Module [18]. This work added fundamental capabilities for representing spatial sounds in ACT-R and modeling responses to them. This interdisciplinary effort was funded by NIH grant number R01-DC015736, and completed in collaboration with Dr. Edward J. Golob at UT San Antonio, and my Ph.D. advisor, Dr. Brent Venable.

1.2.2 Confirmation Bias in Probabilistic Learning

Many aspects of decision aids, including their accuracy, personalization features, and design of warnings, can affect whether or not a person will trust the recommendations provided or ignore them and make their own decision. This can have consequences for the success of the human-AI team. Therefore, it is important to be able to understand and model how the decision-making process may be affected by various design decisions. In Chapter 3, I model a behavioral task where people must make a decision when an automated system is sometimes providing incorrect feedback. I compare four cognitively inspired models of deliberate and intuitive decision-making based on two computational methods, including instance-based learning and utility learning.

Both computational methods simulate different types of decision-making behavior.

The instance-based learning technique [7] is used to model intuitive decision making. This method represents past experiences as instances made up of different attributes. Each instance is stored with an associated activation value that is based on the recency or frequency of the instance being retrieved. When a new decision needs to be made, the model attempts to retrieve a matching instance that received positive feedback and has a sufficiently high activation. If it succeeds, the same decision is made as before. Otherwise, the choice is made at random. I develop two models using an instance-based learning approach. The first instance-based model weights each model equally, while the second assigns more weight to an attribute that decision-makers determined was important to make correct decisions in the past [19, 20]. The final two models are developed using a hybrid approach that combines instance-based learning and utility learning. Utility learning [8] is a reinforcement learning technique that can be used to simulate a decision as a set of states that represent each step in the decision-making process. When the model receives positive feedback about the final decision, a reward is propagated back to each state that resulted in that decision. I use utility learning to model how a decision-maker learns the important cues by first following a deliberate, systematic consideration of the available information. The first hybrid model simulates heuristic [21] called *take the best*, making the decision based on the first cue recalled that differentiates between the two choices. The second hybrid model simulates a more informed approach, making a choice based on the first three cues that can be recalled and differentiate between each option.

By comparing four models based on these methods, I show that human behavior in a binary choice probabilistic learning task is best modeled using an instance-based approach that weights factors learned from past experience more highly than the feedback provided by the system. By comparing two groups of subjects with this model, I show that providing warnings of potentially incorrect feedback can increase this biased weighting in favor of past experience. This research was conducted at the

Naval Research Laboratory, in collaboration with Dr. Noelle Brown and Dina Acklin. It was funded by ONR Contract N0001416WX00044.

1.2.3 Heuristics and Biases in Voting

In addition to biases that occur when individuals make decisions, heuristics and biases also present themselves when groups make decisions. The field of computational social choice (COMSOC) considers issues involved in aggregating preferences across groups, such as the computational complexity of a single agent misreporting their true preferences to manipulate the outcome of a vote [22]. Less work has considered how heuristic decision making may affect the outcome of various voting rules. Many of the existing analyses assume that if the voting rule used to aggregate preferences is sufficiently complicated, people will default to using a truthful strategy, voting for the candidates they like the best [23]. Some voting rules, such as approval voting, may offer voters the option of several sincere voting strategies, allowing them to vote for one or more of their preferred candidates [24]. Analyses of approval voting often assume that voters will vote truthfully for all candidates for which they have positive utility, even when other sincere strategies exist.

In my research, I challenge this assumption and design a behavioral experiment to explore the sincere heuristics that people use in a variety of voting scenarios [25, 26]. For each voting scenario in the experiments, participants are assigned a preference ranking for each candidate, represented as the utility earned if that candidate were to win. They are presented with information about the aggregated votes so far and asked to vote for as many of the candidates as they wished. By comparing the maximum expected utility to the utility earned by each voting heuristic observed in participant's behavior, I show how effective each heuristic is at maximizing the utility of the voter. Through analysis of the results obtained from an experiment run on the Mechanical Turk platform, I show that people often do vote sincerely for one or more

of their preferred candidates, but rarely for all candidates for which they have positive utility. The majority of participants also do not vote for the candidates that would maximize their utility. This work presents a first step in understanding the behavioral component in approval voting, providing many new insights for the design of more realistic simulation tools that predict outcomes in approval voting and other voting rules. This research was done in collaboration with Dr. Jason Harman of Louisiana State University, Dr. Nicholas Mattei of Tulane University, and my Ph.D. advisor, Dr. Brent Venable.

1.3 Thesis Structure

The following chapters examine novel approaches for using computational methods to model bias in tasks of varying levels of abstraction and complexity. I will first introduce preliminaries in Chapter 2. In Chapter 3, I show how two modeling paradigms, including constraint satisfaction problems and the drift-diffusion model, can be used to model different aspects of attentional bias in a spatial auditory attention task. In Chapter 4, I present four cognitively-inspired models of behavior in a probabilistic learning task and show how an instance-based approach that weighs past experiences more highly was the best for simulating human behavior in this task. Finally, in Chapter 5, I present a novel behavioral experiment that examines sincere voting behavior in approval voting and shows that people often do not vote truthfully for all candidates with positive utility, nor do they always vote optimally. Together, these cognitive models and behavioral experiments can inform the design of new algorithms and technologies that adapt to human biases and behavior.