Computational Methods for Simulating Biased Agents

Doctoral Consortium

Jaelle Scheuerman
Tulane University
New Orleans, LA
jscheuer@tulane.edu

ABSTRACT

I present an overview of my research which investigates how models of human behavior can inform the design of new algorithms and interfaces. Specifically, I show how precise, testable computational methods and behavioral experiments can be used to simulate heuristics and bias in human attention and decision making.

KEYWORDS

cognitive architectures, decision making, voting, heuristics, bias

ACM Reference Format:
Jaelle Scheuerman. 2020. Computational Methods for Simulating Biased Agents. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 2 pages.

1 MOTIVATION

Human behavior is guided by a variety of cognitive biases. Often, these allow us to navigate the complex and uncertain environments faced in real life, but sometimes bias can lead to systematic and costly errors. Behavior-informed algorithms have the potential to identify situations where systematic errors are likely to occur and provide recommendations to mitigate these errors. My work uses insights from computational cognitive models and behavioral experiments to design algorithms and interfaces that account for and adapt to human behavior.

2 BACKGROUND

Understanding and predicting human behavior is important for designing new algorithms and systems that support human performance. For example, behavioral experiments have guided the design of recommender systems [5] and new machine learning algorithms [8], as well as models for analyzing voting behavior [2] and heuristics used in resource allocation [7].

Behavioral experiments are also used to create new cognitive models and psychological theories that support the development of technologies that augment human performance. Cognitive architectures, such as ACT-R (Adaptive Control of Thought-Rationale) [1] use precise methods such as instance based learning [4], reinforcement learning [6], and constraint programming [12] to represent cognitive processes in the context of the human mind. These architectures seek to explain how intelligent behavior emerges from cognitive mechanisms. They represent psychological theories as algorithms that mimic some aspect of measurable human performance, such as reaction time or accuracy. Cognitive architectures have been used to simulate behavior in human-machine interaction tasks, such as interactive virtual tutors [9], and to teach robots about human behavior [13].

3 MY WORK

My dissertation examines novel approaches for using computational methods to model bias in human behavior and presents new cognitive models and behavioral experiments that inform the design of algorithms and technologies that can adapt to human biases and behavior. Several applications are examined over the course of three projects. The first project uses constraint satisfaction problems to model the underlying cognitive processes that result in attentional bias in a spatial auditory attention task. The results are integrated into ACT-R so that researchers can more easily model human performance tasks where sound is important [12]. The second project investigates how computational cognitive models can give insights into the design of recommender systems. Using experimental data from a probabilistic learning task, I compare four cognitively inspired methods for modeling confirmation bias in situations where an algorithm may give incorrect feedback [10]. In the final project, I design an experiment to examine heuristics used in different approval voting environments. These results augment the theoretical models of approval voting that have been developed in the area of computational social choice [11]. These projects are described further in the following sections.

3.1 Modeling Bias in Spatial Auditory Attention

Cognitive architectures provide a framework for developing models of users interacting with everything from mobile phone interface [14] to interactive tutors [9]. However, much of the research has focused on modeling aspects of cognition associated with traditional computer interfaces. This includes identifying items on a screen using models of visual attention and perception, or modeling the motor skills required for mouse and key presses. Less has been done to integrate other cognitive functions, such as spatial auditory attention, which affects how quickly and accurately we attend to the sounds around us. There are many situations where it would be useful to simulate auditory attention. For example, hospital emergency rooms use auditory alarms to convey important information and it would be helpful to understand when these alarms will be heard or go ignored. Behavioral experiments have shown that response times to spatial sounds are dependent on the spatial location of the sound [3]. This attentional bias can be modeled as a combination of top-down, or goal driven processes and bottom-up, or

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

salient, processes. In this project, I use the AI framework of constraint satisfaction problems to model the combinatorial structure of attentional bias in spatial auditory attention, and incorporate the resulting model into ACT-R [12].

3.2 Confirmation Bias in a Probabilistic Learning Task

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 how confident it is in the recommendation. The user may choose to use the recommender system's suggestion or choose some other option. For this project, I compare four different cognitive models of a task where users are required to choose between two alternatives. The participants and the model receive feedback from a computer program that was sometimes incorrect. By comparing the models, I show that using an instance based learning approach that weights past experiences highly is effective in simulating users in this task. Using this model I show that when users are warned about the potential for incorrect feedback, they weight past experiences much higher than when they are not warned [10]. This has implications in the design of systems that make potentially inaccurate recommendations when faced with uncertain data. Further work must be done to ensure that warnings about recommendations and feedback can be used effectively, such as first addressing a user's preconceived biases. For example, the cognitive model described 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, leading to better engagement with the system.

3.3 Heuristics and Biases in Uncertain Approval Voting Environments

Many real-world situations involve multiple agents participating in collective decision-making tasks. This usually involves aggregating preferences through a vote to choose the alternative that best reflects the preferences of the group. Agents may vote with their true preference, use heuristics (such as voting for the current leader in a poll), or vote strategically to attain a better outcome. In real world voting scenarios, people often do not have complete information about other voter preferences and it can be computationally complex to identify a strategy that will maximize their expected utility. In such scenarios, it is often assumed that voters will vote sincerely rather than expending the effort to strategize. This project examines voting behavior in approval voting elections. In an approval election, voters can try to maximize their utility or use a heuristic. Several sincere heuristics are possible, including voting completely truthfully (for all candidates for which the voter has some positive utility) or voting for their top x candidates with the highest utility. I present a behavioral experiment to examine the use and effectiveness of sincere heuristics in multi-winner approval voting scenarios with missing votes. The results show that people generally vote sincerely, but use different underlying heuristics that depended on features of the voting scenario including the number of winners and whether or not there is a strong preference for or against a particular candidate [11]. This work provides key insights on human behavior in voting environments and can inform the development of more realistic simulation tools and accurate predictions of election outcomes where approval voting is used.

4 FUTURE WORK

My work so far has focused on modeling heuristics and biases that can affect how users interact with AI systems. Going forward, I plan to use the resulting cognitive models and insights to design new algorithms, and validate their effectiveness for engaging users and mitigating bias. I am also exploring other areas, such as interactive machine learning, where behavioral insights could improve the transparency and explainability of working with these systems.

REFERENCES

[1] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. 2004. An integrated theory of the mind. *Psychological review* 111, 4 (Oct. 2004), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

[2] Roy Fairstein, Adam Lauz, Reshef Meir, and Kobi Gal. 2019. Modeling People's Voting Behavior with Poll Information. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems.*

[3] Edward J. Golob, K. Brent Venable, Jaelle Scheuerman, and Maxwell T. Anderson. 2017. Computational modeling of auditory spatial attention. In *Cog Sci 2017*. London, UK.

[4] Cleotilde Gonzalez, Javier F. Lerch, and Christian Lebiere. 2003. Instance-based learning in dynamic decision making. *Cognitive Science* 27, 4 (2003), 591–635. https://doi.org/10.1207/s15516709cog2704_2

[5] Jason L. Harman, John O'Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. 2014. Dynamics of Human Trust in Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 305–308. https://doi.org/10.1145/2645710.2645761 event-place: Foster City, Silicon Valley, California, USA.

[6] M.C. Lovett. 1998. Choice. In *The atomic components of thought.* Routledge, 255–296.

[7] Timo Mennle, Michael Weiss, Basil Philipp, and Sven Seuken. 2015. The Power of Local Manipulation Strategies in Assignment Mechanisms. In *Twenty-Fourth International Joint Conference on Artificial Intelligence.* https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/11433

[8] Ori Plonsky, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. 2017. Psychological Forest: Predicting Human Behavior. In *Thirty-First AAAI Conference on Artificial Intelligence.* https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14925

[9] Steven Ritter, John R. Anderson, Kennet R. Koedinger, and Albert Corbett. 2007. Cognitive tutor: Applied research in mathematics education. 14, 2 (2007), 249–255. http://act-r.psy.cmu.edu/?post_type=publications&p=14352

[10] Jaelle Scheuerman and Dina Acklin. 2017. Modeling bias reduction strategies in a biased agent. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press, Melbourne, Australia, 5205–5206.

[11] Jaelle Scheuerman, Jason Harman, Nicholas Mattei, and K. Brent Venable. 2017. Heuristic Strategies in Uncertain Approval Voting Environments *(AAMAS '20)*. Auckland, New Zealand.

[12] Jaelle Scheuerman, K. Brent Venable, Maxwell T. Anderson, and Edward J. Golob. 2018. Modeling spatial auditory attention in ACT-R: a constraint-based approach. In *Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures*, Vol. 145. 797–804. https://doi.org/10.1016/j.procs.2018.11.028

[13] J. Gregory Trafton, Laura M. Hiatt, Anthony M. Harrison, Franklin P. Tamborello, II, Sangeet S. Khemlani, and Alan C. Schultz. 2013. ACT-R/E: An Embodied Cognitive Architecture for Human-robot Interaction. *Journal of Human-Robot Interaction* 2, 1 (Feb. 2013), 30–55. https://doi.org/10.5898/JHRI.2.1.Trafton

[14] Maria Wirzberger and Nele Russwinkel. 2015. Modeling Interruption and Resumption in a Smartphone Task: An ACT-R Approach. *Journal of Interactive Media* 14, 2 (2015), 147–154. https://doi.org/10.1515/icom-2015-0033