DISSERTATION PROPOSAL

Behnam Mohammadi

"Human-AI-Interaction: Implications for Marketing and Policy-Making"

Wednesday, November 30, 2022
3:00pm
Tepper 4242

Chapter 1

Regulating eXplainable AI (XAI) May Harm Consumers

Joint work with Nikhil Malik, Tim Derdenger, and Kannan Srinivasan

Note: This paper is submitted to Marketing Science and is under review.

Recent AI algorithms (e.g., deep neural networks) are black box models whose decisions are difficult to interpret even by their developers. This intractability reduces the trust in model outputs and impedes successful implementation of responsible and ethical AI. eXplainable AI (XAI) is a class of methods that seek to address lack of AI interpretability by explaining to customers their AI decisions, e.g., decision to reject a loan application. Following the recent legislations such as GDPR in the EU and FCRA in the U.S., the common wisdom is that regulations mandating fully transparent XAI lead to greater social welfare. Our paper challenges this notion through a game theoretic model of a policy-maker who maximizes social welfare, firms in a duopoly competition that maximize profits, and heterogenous consumers. The results show that XAI regulation may be redundant. In fact, mandating fully transparent XAI may make firms and consumers worse off. This reveals a tradeoff between maximizing welfare and receiving explainable AI outputs. Our work extends the existing literature on method and substantive fronts, and introduces and studies the notion of XAI *fairness*, which is shown to be impossible to guarantee even under mandatory XAI. We also discuss the regulatory and managerial implications of my results for policy-makers and businesses, respectively.

Chapter 2

Estimating Consumer Preferences under Bounded Rationality in Static and Dynamic Choice Settings

Readers: Alan Montgomery and Tim Derdenger

This is a work in progress. The theoretical model and estimation methods are finished. More work needs to be done on the second part, i.e., application in recommender systems.

The second chapter deals with a model of human decision-making under limited resources such as time; for example, watching a movie/show from a list of recommended items on Netflix. I show that a tradeoff exists between the expected utility gains of "better" options and the information cost of finding them, a situation which is called *bounded rationality*. This leads to rational but suboptimal behavior, e.g., not watching the movie/show that could have satisfied the viewer the most. I translate this model into an optimization problem whose solution is a choice probability with Gibbs distribution. Previous standard structural logit models are a special case of this model. I provide a method to make inferences about individual consumer preferences in static and dynamic cases, and prove that an optimal consumer policy exists even when her decisions are suboptimal. The results of this line of research can be useful for improving the performance of personalized recommender systems that depend on choice models. Intuitively, acknowledging humans' imperfections and mistakes in decision-making should change the balance of *exploration* vs. *exploitation* in favor of more exploration.

Chapter 3

Simulating Heterogenous Consumer Behavior via Large Language Models

This is a work in progress.

Large Language Models (LLMs) such as OpenAI's GPT-3 have been recently receiving much attention from practitioners and academics. These models often use state-of-the-art *transformer* techniques to generate human-like texts, something that used to be impossible just a couple of years ago. Moreover, our interaction with such AI models has shifted from traditional fine-tuning to "prompting" via regular text input, making them approachable to a much wider audience. But since the training datasets of LLMs often consist of the entire internet, the models may reflect various types of human biases, e.g., associating high-paying jobs with males more than females. While many researchers focus on mitigating the inherent biases of LLMs, in this chapter I work on *utilizing* the bias to simulate human behavior for marketing purposes. My preliminary results show that LLMs can replicate not only some of the most famous behavioral economics studies such as Kahneman and Tversky's *loss aversion* theory, but also the results of behavioral studies that they had not seen in their dataset. This could be useful for applications in which involving a large and diverse group of humans is costly, unethical, or impossible. For example, suppose a job advertisement for a position that must appeal to a broad audience of different races and genders. One could simulate the responses of thousands of individuals from diverse backgrounds and iteratively improve the advertisement. Furthermore, one could obtain preliminary results in conjoint analyses or brand personality studies. The goal of this study is to show when (for which use cases) and how (under what conditioning) LLMs can be helpful for simulating human behavior.