University of Pittsburgh | School of Computing and Information

Proposal Defense

Doctor of Philosophy in Information Science

"**Countering system-induced bias and stereotypes via human-centric interactive AI systems**" by **Yongsu Ahn**

Date: December 12, 2024
Time: 12:30 – 3:30 p.m.
Place: Room 502, Information Sciences Building, 135 N. Bellefield Ave Pittsburgh, PA 15260

Committee:

- (Chair) Dr. Yu-Ru Lin, Associate Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. Peter Leonid Brusilovsky, Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. Rosta Farzan, Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. Adam Perer, Assistant Professor, School of Computer Science, Carnegie Mellon University
-

Abstract:

Artificial intelligence has unprecedentedly penetrated in various domains ranging from high-stakes decisions to everyday information use. A noteworthy aspect of AI's advancement lies in its capability of better capturing the characteristics of groups and individuals; however, there is an increasing concern of system-induced bias and stereotypes. Recent studies evidenced the negative consequences of AI that systematically discriminate against certain groups or overgeneralize individuals' characteristics based on their demographics. It is challenging for AI developers, decision makers, data subjects and users to be aware of such potential impacts on the AI-driven process and decisions. When such systematic effects perpetuate in decision tools and information services, some groups and individuals may systematically be given unfavorable decisions or information. How can these various stakeholders make informed practice and use of AI-driven systems against those negative impacts? Despite a great advance in automated methods, stakeholders are not provided with guidance on how to understand and mitigate biases and stereotypes or mediums to intervene in the process. In this context, AI systems need to be designed with better transparency to promote better understanding and ways to allow experts and users to incorporate their insights and preferences in the loop.

In my PhD dissertation, I aim to explore potential design issues and visualizations for interactive AI systems to help inform users of system-induced bias and stereotypes. In the study, I tackle the problem of system-induced bias and stereotypes particularly in recommender systems impacting thousands of people daily by tailoring content across various domains, such as news, entertainment, and e-commerce, to keep individual users informed and engaged. First, I investigate ways to better understand system-induced bias and stereotyping problems in recommender systems. This includes developing a unified framework to comprehensively measure how bias, stereotyping, and miscalibration intertwine and lead to disparate impact over groups and individuals. Next, I propose to make a participatory turn in designing an interactive recommender interface to promote user-driven valuation and autonomy over algorithmic effects. Through a participatory design approach, I find that users perceive algorithmic effects in varied ways, viewing them not solely as harmful or valuable depending on objectives such as personalization or diversity. Guided by these findings and users' design preferences, I develop an interactive interface that enables users to view, interpret, and adjust

algorithmic effects in their recommendations. The evaluation demonstrates that the interface not only enhances users' understanding of potential algorithmic effects but also empower them to mitigate and align them with their preferred values, such as personalization or diversity. Together, these studies promote the capability of recommender systems with a greater degree of user autonomy and help prevent unintended reinforcement in subsequent feedback loops.