COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Harmanpreet Kaur

Where are the Humans in Human-AI Interaction: The Missing Human-Centered Perspective on Interpretability Tools for Machine Learning

Tuesday, May 30, 2023
3:00pm – 5:00pm
Ehrlicher Room (North Quad 3100)
Hybrid – Zoom

ABSTRACT: Effective human-machine collaboration is bound by how well the human and the machine counterparts understand each other. With continued advancements in AI and ML, the machine can not only process information at higher speeds, but can also infer user preferences and filter content accordingly. These systems are now routinely deployed in real-world settings, including sensitive domains like criminal justice and healthcare. Given their rising ubiquity, understanding how AI and ML work is a prerequisite for responsibly designing, deploying, and using these systems. With interpretability and explainability approaches, AI- and ML-based systems can now offer explanations for their outputs to aid human understanding. Though these approaches rely on guidelines for how humans explain things to each other, they ultimately solve for improving an artifact (e.g., an explanation or explanation system).

My thesis makes the argument that helping people understand AI and ML is as much a human problem as a technical one. Detailing this missing piece, I present work that shows that current interpretability and explainability tools do not meet their intended goals for a key stakeholder, ML practitioners, who end up either over- or under-utilizing these tools. Investigating the reasons behind this behavior, I apply the cognitive model of bounded rationality to the human-machine setting. Under this model of decision-making, people select plausible options based on their prior heuristics rather than internalizing all relevant information. I find significant evidence showing that interpretability tools exacerbate the application of bounded rationality. As a solution, I present a new framework for re-imagining interpretability and explainability based on sensemaking theory from organizational studies. This Sensible AI framework prescribes design guidelines grounded in nuances of human cognition---facets of the individual and their environmental, social, and organizational context.

CHAIR: Prof. Eric Gilbert and Prof. Cliff Lampe