"Online Fairness Auditing through Iterative Refinement"

Pranav Maneriker, PhD Candidate
The Ohio State University

Professor Srinivasan Parthasarathy
The Ohio State University

Pranav Maneriker is a PhD Candidate at The Ohio State University. His research interests are in natural language processing and graph representation learning with applications to trustworthy computing. His work has been featured in various leading conferences including ACM SIGKDD, EMNLP, and the Web Conference (including a best paper award). Prior to Ohio State, Pranav received his BTech in Computer Science from IIT Kanpur and spent a couple of years at Adobe Research. He is a recipient of the 2023 Graduate Research Award at CSE@Ohio State University.

Srinivasan Parthasarathy is a Professor in Computer Science and Engineering at The Ohio State University. His research interests are broadly in the space of data mining, database systems, network science and high performance computing. A recent focus of his is in the space of responsible data science - where he co-leads a community of practice on this theme. His work has been featured in leading conferences and journals in the field including 15 best paper awards or best-of-conference selections at venues such as VLDB, ACM WWW, ACM WSDM, ACM SIGKDD, IEEE ICDM, SIAM SDM, ACM BCB and ISMB. Prior to Ohio State Prof. Parthasarathy was at Intel Research and at the University of Rochester (where he received his PhD). Prof. Parthasarathy is a fellow of the IEEE, the Asia Pacific Artificial Intelligence Association and the Risk Institute.

Abstract: A sizable proportion of deployed machine learning models make their decisions in a black-box manner. Such decision-making procedures are susceptible to intrinsic biases, which has led to a call for accountability in deployed decision systems. In this work, we present a first-of-its-kind practical system (AVOIR) for the auditing of (deployed) black-box AI models against various fairness criteria. AVOIR permits the user to flexibly design and specify the fairness criterion (including non-binary and intersectional notions of fairness) they want to audit within a custom domain specific language (DSL). It then relies on the mathematical notion of confidence sets to facilitate the runtime monitoring of probabilistic assertions over said fairness metrics on the corresponding decision functions associated with modern AI models. AVOIR enables the automated inference of probabilistic guarantees and the visual exploration of fairness violations aligned with recent regulatory requirements. In this talk we will describe the AVOIR system and then we will illustrate through case studies how AVOIR can help detect and localize fairness violations and help model designers and ML engineers ameliorate such issues.

July 27, 2023 at 12:00 PM ET

Join Zoom Meeting

https://osu.zoom.us/j/94365864594?pwd=cjhkMURQT0ZoWGZFZjYvMnVaek5mQT09