Research Statement

Bishwamittra Ghosh

PhD Candidate
School of Computing
National University of Singapore (NUS)
https://bishwamittra.github.io

The last decade has witnessed significant progress in machine learning with a host of applications of algorithmic decision-making in different safety-critical domains, such as college admission, recidivism prediction, employee hiring, and clinical processes. In these domains, the decisions by machine learning classifiers have far-reaching consequences, potentially influencing human lives in various ways. Consequently, fairness and interpretability of the deployed classifiers have paramount importance to the end-users, who expect to have a fair decision from an algorithm and an interpretable decision that they can understand. My research vision is to improve the safety-critical properties of machine learning, such as fairness and interpretability, and enhance the applicability of machine learning for social good.

Towards **fair and interpretable machine learning**, my research so far has focused on the verification and identification of the unfairness of classifiers and designing interpretable classifiers with higher scalability. Fairness verification gives a formal certificate of whether a classifier has achieved the desired level of fairness on specified data distribution. In identifying the source of unfairness, we compute fairness influence functions of input features as their contribution towards the bias/unfairness of the classifier. In interpretable machine learning, we focus on designing an interpretable rule-based classifier alternative to black-box classifiers. In particular, we prioritize a scalable learning framework, which achieves competitive accuracy and interpretability while enabling learning on million samples. In the technical aspects of my research, I have closely integrated formal methods with machine learning and designed frameworks that can withstand with improved scalability and performance guarantees.