2018 Research Interest/Project Ideas

Elias Bareinboim

http://causalai.net

==

Fairness and Bias Analysis in Automated Decision-Making

AI plays an increasingly prominent role in modern society since decisions that were once made by humans are now being delegated to automated systems. These systems are currently in charge of deciding bank loans, criminals' incarceration, and the hiring of new employees, and it's not hard to envision a future where they will underpin most of the society's decision infra-structure. Despite the high stakes entailed by this task, there is still almost no understanding of some basic properties of such systems, including due to issues of fairness, accountability, and transparency. We currently do not fully grasp, for instance, how to design systems that abide by the decision-constraints agreed by society, including obvious ones such as avoiding to reinforce previous biases and to perpetuate current discriminatory practices, both of which that could be present in the training data. In this project, we plan to develop the mathematical foundations for assisting the data scientist in analyzing the existence, and possibly the "magnitude," of unfairness in an already deployed decision-system. Further, this framework will also guide a system's designer in the process of selecting a fairness criterion in its to-be-deployed system while ascertaining an established level of fairness and accuracy.