TALK REVIEW : Prof Nalini Ratha on Trustworthy AI

The talk started with Biometrics and AI. Prof started discussion saying that Biometrics and AI are correlated. The need is how we can learn from Biometrics in our AI world. The topics covered in the discussion were Bias, Adversarial Attacks, Encrypted learning, Transparency. These were core pillars of the talk. He mentioned most of the deep learning architectures are opaque and non-intuitive. It has to be transparent so as the auditing of algorithms can be done with community. We agree that AI systems shows large growth but should be judged by rational behaviour. He explained the working of AI models that's how they are trained and related decision making. As we discussed in class already, he also mentioned about the racist behaviour of AI. The Amazon bias in recruitment, Google photos in African people, lawmakers as crime suspects and its penalty, working of facial recognition models on white people and prejudices. The word trust is built with time. Trust need to encompass wherever user is interacting the system. There are some issues with current AI systems. Trust building can be parametrised with Biometrics. There is a need of ubiquitous identity authentication. Also, this section of Artificial Intelligence deals with the security aspects. This is not about accuracy, it is about how one can fool the AI systems. This is analogous to the concept we studied in class where we try to fool ML algorithms by adding noise vector in the query image so as the algorithm refuses to detect the face in the query image. Prof. showed that how security cameras and systems were broken down when people changed their identities not the biometrics but face. Adversarial attacks are common to the community where a set of people want to break the system for their benefits. Bias can be added in training data itself such that it can be corrupted. Narrow AI is analogous to overfitting a huge and massive data such that it learns everything. Broad AI requires less training data such that it oversimplifies the decision making. And General AI is something which shows rational decision making. Dependable AI has different pillars mentioned as Beneficial AI, Responsible AI, Ethical AI, Trustworthy AI. These terms are modified by the time, but the underlying concept remains same. Big agencies and Governances issue the rules and laws which governs the AI Models. The t&c on AI system is defined by these bodies (of US, UK, Brazil, etc). Apart from this the final remark is that, the end user leveraging the services of AI system must know background

working of the AI system. Bias is always unwelcome unless intended. Sometimes bias is induced with error as well. The person injected bias in training data or the ML model will audit the working of the AI model correct, however this model may be not suitable for the large corpus. There are 180 different biases. Very important line mentioned in the presentation is “Although Machine Learning, by its very nature is a for of statistical discrimination, However, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage” Group fairness: (WAE) and (WYSIWYG). In final slides Prof talked about biometrics terminologies Sheep, Goats, Lambs and Wolves depending on the severity of match. (ref: https://arxiv.org/pdf/1209.6189.pdf).