Research Statement

I am Abdus Salam Azad, a first-year Ph.D. student at the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. My research interest spans broadly in the field of Artificial Intelligence (AI). Use of AI or Machine Learning (ML) has become ubiquitous in various domains of our day-to-day life and society, including personal assistants, health-care, e-commerce, finance, justice, education, social networks, transportation, and what not. With this ubiquity, the question of trust, bias, transparency, and safety of these models have become more important than ever. Several recent incidents indicate that the straightforward use of ML models can raise serious concerns.

In 2014 the Houston Independent School District(ISD) was sued when it attempted to fire 85% of its teachers based on the performance scores predicted by a software. The software did not provide any explanations behind the scores it assigned, nor its authors revealed the inner logics of the software. The court ruled against Houston ISD as the firing violated the civil rights of the teachers [1]. In May 2017, an independent report claimed that a proprietary software, with an unknown working procedure, used by a court is unjustly biased against black prisoners [2]. Amazon tried building an AI for hiring people, which was found biased against women [3]. Now, with the increasing use of ML in each and every sector, such issues are bound to increase. Then, what could be done to make the models more transparent and trustworthy?

One answer lies in the field of Explainable AI or, Interpretable ML. We can use ML models that are inherently interpretable (e.g., shallow decision tree or, sparse linear models) or, we can explain each prediction a model makes. Both of them provide a way for people to understand the behavior or reasoning behind an ML system and thus increase credibility. However, most of the recent success in AI has been mainly due to Deep Learning, i.e, hard-to-understand large neural networks, and the typical interpretable models are no match to them. Hence, explaining the individual predictions of such models is the only practical option we have right now. The governments and international bodies are also acknowledging the importance of explanations by gradually incorporating the 'Right to Explain' within their laws [4].

However, is that enough? No! We can inspect explanations only for a limited number of predictions and it can be risky to put trust in the entire model based on that limited information. Hence, we need more concrete proofs or, quantified measures of trust. We need to design methods that can mathematically or at least empirically, 'prove' certain properties of an ML model e.g, not being biased in terms of gender, race, or ethnicity. Hence, the long-term vision of my current research interest is **"to design methods that can formally verify AI"**.

To design methods that can formally verify AI or AI systems (i.e., systems that internally use AI or ML models), the most natural starting point is to view the problem from the perspective of Formal Methods: the field of computer science dealing with mathematical methods for specification, development and verification of software and hardware systems. The central theme of formal methods can be informally stated as "How can we formally 'prove' that a system follows certain properties?". The current state-of-the-art in the field of formal methods can not readily solve the problem of AI verification. Nonetheless, the vast literature of formal methods can identify the challenges we need to overcome and offer prospective solution principles [5]. Hence, one interesting research direction I would like to explore is to leverage the ideas, methods, and tools from the domain of formal methods to verify AI systems. Another domain of particular interest is explainable AI. Research on interpretable models or, explaining predictions of ML models will certainly lead us closer to formally verifying AI systems.

Finally, I would like to mention that I come from Bangladesh and I've worked as a lecturer at the Department of Computer Science and Engineering in Bangladesh University of Engineering and Technology (BUET) since 2014. Currently, I am on a study leave and I will rejoin there as an Assistant Professor after completing my Ph.D. As a country, Bangladesh is adopting automation more and more in Government, industry, and NGOs. As BUET is the premier engineering university of Bangladesh, it has been part of my responsibility to actively participate (e.g., design, develop, monitor, and audit) in some of such projects. I believe, with time, the use of AI in Bangladesh will only increase. And the question of trusting AI systems will also become much more important, especially when used by law enforcing agencies, judiciary systems, and other appropriate bodies. A research towards my intended direction will allow me to serve my country for the transparent use of AI and this prestigious fellowship can allow me the freedom to work in that direction, especially with the problems those are unique to the developing or underdeveloped nations.

References

[1] "Houston schools must face teacher evaluation lawsuit." `https://www.courthousenews.com/houston-schools-must-face-teacher-evaluation-lawsuit/`. Accessed: 2018-10-26.

[2] "Rise of the racist robots – how ai is learning all our worst impulses." `https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses`. Accessed: 2018-10-26.

[3] "Amazon built an ai tool to hire people but had to shut it down because it was discriminating against women." `https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10`. Accessed: 2018-10-26.

[4] "Right to expalanation." `https://en.wikipedia.org/wiki/Right_to_explanation`. Accessed: 2018-10-26.

[5] S. A. Seshia, D. Sadigh, and S. S. Sastry, "Towards verified artificial intelligence," *arXiv preprint arXiv:1606.08514*, 2016.