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 calculated by a software. The software did not accompany any explanations behind the scores it assigns, nor its authors revealed the inner working of the software. The court ruled against Houston ISD as the software violated the civil rights of the teachers. 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. Amazon AI for hiring people was biased against women. 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 trustable?

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 trust. 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.

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 proof or, quantified measure of trust. We need to design methods that can mathematically or at least empirically, ‘proof’ 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 \textbf{``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), 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 directly solve the problem of AI verification. Nonetheless, the vast literature of formal methods can identify which challenges are to be overcome to verify AI systems~\cite{verifiedAI} and offer prospective solution methodologies. Hence, one interesting research direction I would like to explore is to leverage the ideas, methods, and tools from the domain of formal methods/verification to verify AI systems. Another domain of particular interest is explainable AI (XAI). Research on interpretable models or, explaining predictions of ML models will 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 more and more automation 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, audit) in some of such projects. I believe with time the use of AI will increase in Bangladesh too. And the question trusting AI systems will also become much more important, especially when used by police, customs, justice, and so on. A research towards my intended direction will allow me to serve my country for a safe use of AI and this prestigious fellowship can allow me the freedom to work in problems that are unique to the developing or underdeveloped nations.

Sometimes the authors of the ML models may not reveal its working (e.g., the algorithm may be proprietary and considered a trade secret).

Several challenges will arise in that direction. Typically the current methods deal with precisely known systems such as a program written in C or, a circuit described in a hardware specification language. Typically a formal specification is written for the system with respect to the properties of interest. These specifications are often written as discrete formulas in SAT or SMT. Then, tools such as SAT/SMT solver or model checker are used to mathematically prove whether the system satisfies the property or not. Now, even in the case of precisely known systems (e.g. programs or circuits), coming up with such formal specifications can be very hard. Hence, the current formal specification languages will come short to describe the desired/undesired properties for AI systems, which involve stochastic complex ML models often involving millions of real valued weights. Hence, we will need new ways to specify the desired behaviors of AI systems. New computation engines will also be needed to verify whether an AI system satisfies a property or not. Thus,

Such methods will allow the authors (i.e., systems/softwares that internally use AI/ML), their users, and the concerned third parties to verify the AI systems independently to be certain

These methods can be used by the ML system authors, users.