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 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 software violated the civil rights of the teachers~\cite{evaas}. 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~\cite{compas}. Amazon tried building an AI for hiring people, which was found biased against women~\cite{amazon}. 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~\cite{rte}.

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 \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 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 to verify AI systems and offer prospective solution principles~\cite{verifiedAI}. 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 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 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 safe 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.

I am Abdus Salam Azad. I am a first year Ph.D. student at the department of Electrical Engineering and Computer Sciences at 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 domain 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. There are several cases in recent times when the use of proprietary ML models have raised the question of being unjust or biased.

The Houston Independent School District, largest in Texas’, started using a proprietary system (Educational Value-Added Assessment System, or EVAAS) on 2012 that measures teacher performance based on their students' test scores. It also announced a goal of firing 85\% of the `ineffective' teachers using that system. The system did not give any explanation of its prediction and the author of the system also did not disclose its inner working procedure. The Houston ISD was sued in April 2014 and was alleged that the system curbes the teachers of their right to hear evidence about why the district has chosen to fire them. The court sided with the teachers.

In May 2017, the investigative journalism organisation ProPublica claimed that a computer program, Correctional Offender Management Profiling for Alternative Sanctions (Compas), used by a US court for risk assessment was biased against black prisoners. The report shows that it was much more prone to mistakenly label black defendants---at almost twice the rate as white people (45\% to 24\%)---as likely to reoffend. The author of the program also denied to reveal the inner decision logic of the program and refuted the study.

With the increasing use of ML in each and every sector, such issues are bound to increase. Hence, it is crucial to be able to reason about some desirable properties/questions ((e.g., does a system predicting risk of providing loan unethically discriminates people based on their race or, ethnicity?)) that an ML model should satisfy. To be more concrete the long term vision of my current research interest is: \textbf{``How can we `verify' or 'prove' that an ML model or, more generally an AI system (a system that heavily uses AI in its working) does follow a particular characteristics?''}.

To tackle this question a good starting point is to view the question 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(e.g., program/circuit) follows certain properties?". 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.

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 specification are often written as discrete formulas in SAT or, SMT. Then, tools such as, SAT/SMT solver or, model checker is 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, the current formal methods techniques may not be directly be utilized to verify AI systems. 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 methodology.

Another domain of particular interest is explainable AI (XAI), which deals with understanding how ML models think. If we can understand how an ML model `thinks' or, if we can understand the reasoning behind how it makes each individual decision, it can surely increase the element of trust for using these models, and can eventually lead us closer to formally verify AI systems.

Most of the recent success of AI has been because mainly of Deep Learning which mostly deals with different forms of large 'hard-to-understand' neural networks with millions of weights. These models are appealing as they can theoretically learn any function of arbitrary complexity \cite{universal approximator}. One direction of research use the internal weights and architecture of the models to gain an insight on how they think~\cite{from tan karuana}. On the other hand, some approaches consider the models as black-boxes and try to understand how they work, called model agnostic explainer, solely based on their predictions\cite{LIME}. There are also a line of work that focuses on designing inherently explainable/interpretable models that can learn complex functions. However, coming up with such models are hard. As the typically popular interpretable models can't represent complex function or lose their interpretability doing so.

All such research can lead us closer to verify AI systems. Particularly, model agnostic explainers can be particularly useful for the scenarios described previously, where a third-party which only has access to the predictions of the model but have no idea on what sort of model is used, its complexity, the kind of data-preprocessing, feature-representation has been used etc. One of my current projects focuses on how can we provide explanations for such scenarios.

At last, I come from Bangladesh and 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 towards automation in Government and NGOs. As BUET is the premier engineering university of Bangladesh, it was part of my responsibility to actively participate (e.g., design, develop, monitor, audit) in many of such national projects. I believe with time the use of AI will increase in Bangladesh too. And the question of bias, ethics, and trust of AI systems will also become much more important, especially when used by police, customs, justice, and so on. A research towards verifying AI 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 more applicable to the developing or underdeveloped nations.