NSF PI Meeting Breakout: Issues Surrounding Trustworthy Machine Learning

Somesh Jha, Aleksander Madry, Percy Liang and Patrick McDaniel

October 28 and 29, 2019

1- *What is the topic? Why is it important to society? To a secure and trustworthy cyberspace? In other ways?*

Our topic is issues surrounding trustworthy Machine Learning. ML will impact all areas of society in the future- medicine, education, transportation, etc. Because of the cumulative impact, it is of paramount importance to insure security, reliability and safety on ML in cyberspace.

2- *Is there an existing body of research and/or practice? What are some highlights or pointers to it?*

- NSF **Center for Trustworthy Machine Learning** (https://ctml.psu.edu/)
- **gradsci.org/intro_adversarial/**
- **cleverhans.io** (library + blog) and **github.com/madrylab/robustness** (library)
- Adversarial robustness tutorial at NeurIPS 2018 (with Z. Kolter):
- **adversarial-ml-tutorial.org**
- Robust ML unit of Science of Deep Learning class (co-taught with C. Daskalakis):**people.csail.mit.edu/madry/6.883/**
- Workshops at NeurIPS 2018 & ICML 2019

3- *What are important challenges that remain? Are there new challenges that have arisen base on new models, new knowledge, new technologies, new uses, etc?*

- To avoid redundancies, design vehicles for sharing education, outreach and research
- The security of ML- how to protect data and models
- Data sourcing, reliability
- Formalizing methodologies and processes
- Explainability of results
- Making the ML community proactive instead of reactive
- Understanding the trade-offs between robustness/accuracy
- Developing formalized verification techniques
- Incorporating lessons learned from other learning models/communities (federated learning models, crypto, game theory, info theory, etc)
- Formalizing threat models for domains

3- *Cont.*

- Develop End-to-End system-level adversaries
- Benchmark & expand robustness beyond just attacks
  - include connections between privacy, fairness, etc
  - Include more then just attacks
- How to measure safety
- Industrial outreach
- Proof systems to create reductions comparing defenses
- Level playing field of different sized groups access to resources
  - Data
  - Computational
  - Human
- Incentivizing data sharing and kindness within community
  - not attacking others when models fail

4- *Are there promising directions to addressing them? What kinds of expertise and collaboration is needed (disciplines and subdisciplines)?*

- Build a better institutional memory of what works/fails in ML
- Can we utilize methods/processes from other disciplines (software design)
- Better understanding of the nexus of human/ML interactions
- Collection, maintenance and sharing of data
- Principled way for data introspection
- Develop tools to help explore data sets (poisoning, insufficient characteristics)
- Increase transparence/visibility into ML models to enable research
- Assuming ML is not robust, how can we build robustness into models?
- Create a stronger nexus between various communities (explainability, interpretability, security, etc)