For Workshop on Trustworthy Algorithmic Decision-Making

A framework for incremental progress to assure fairness in consequential machine learning

Chuck Howell, the MITRE Corporation, howell@mitre.org, December 4 and 5, 2017

Concerns about fairness in AI-based systems are expressed in best-selling books (e.g., *Weapons of Math Destruction*), focused workshops (e.g., www.fatml.org), in the 2016 White House report *Preparing for the Future of Artificial Intelligence* [1], etc.). As public, end user, legal, and government concerns about AI fairness grow, failure to adequately address the concerns is likely to be a barrier to the adoption and use of consequential AI systems, especially those that rely on machine learning. Across the research community, technical solutions are being explored (e.g., explainable AI [2], audit logs, interpretable models [3], [4], [5]) to enable increased transparency and understanding of machine learning systems. However, these solutions tend to provide insight only late in the ML development cycle – during model training, testing, and deployment. At MITRE, we are exploring how concepts from the safety critical systems community can be adapted to support the calibration, mitigation, and informed acceptance of fairness risks in consequential ML systems.

The development of safety-critical systems in domains such as avionics, transportation systems, medical devices, and weapons systems is subject to extensive scrutiny for obvious reasons. Over the years, a variety of system engineering **t**ools, **t**echniques, and **p**ractices (TTPs) have evolved to facilitate safety-critical software development and to support the communication and review of reasons why the developers assert that the system is adequately safe for use. Consequently, we hypothesize that overall assurance regarding characteristics such as fairness for an ML-based system could benefit from adapting TTPs from the safety community. As Admiral Rickover recognized when leading the original development of naval nuclear propulsion, transformational potential is enabled only if it can be used safely and with confidence [6].

The four topics we are exploring are:

1. Adapting structured assurance and dependability cases [7] [8] [9] to produce a *fairness case*. An assurance case is a documented body of evidence that provides a compelling case that the system satisfies certain critical properties for specific contexts. There are TTPs to facilitate the development and communication of claims, arguments, and evidence in a rigorous manner to support critical developments. A structured framework to communicate engineering and operational tradeoffs and decisions is essential for early agreement with various stakeholders, and can reduce the engineering churn and rework that increases system costs and delays. We are exploring adapting the development and review of assurance cases specifically to focus on fairness for ML systems.

2. Hazard or risk analysis as applied to subtle and unexpected potential causes of mishaps [10]. As an example, Systems-Theoretic Process Analysis (STPA), developed at MIT, is a hazard analysis approach that is part of a broader framework for safety called STAMP (System-Theoretic Accident Model and Processes) [11]. Industry uptake illustrates the value of STPA to provide disciplined exploration of potential hazards (any kind of undesirable outcome from system performance). We are exploring adapting these TTPs to the risks of unfair ML.

3. Instrumentation and monitoring of complex systems for runtime verification, anomaly detection, and enforcement of defined operational constraints (e.g., policy enforcement).

4. Tools and notations for incident investigation to expose subtle contributing causes to mishaps and to reduce the consequences of confirmation bias in the investigation [12] [13].

It is important to distinguish between two different causes of perceptions of unfairness: *bias* and *capricious behavior*. Two widely known illustrations of unfair behavior by judges illustrate the difference.

"In looking at decisions handed down by judges in Louisiana's juvenile courts between 1996 and 2012, the pair found that when LSU lost football games it was expected to win, judges—specifically those who had earned their bachelor's degrees from the school—issued harsher sentences in the week following the loss. When the team was ranked in the top 10 before the losing game, kids wound up behind bars for about two months longer, on average. When the team was not as highly ranked, it was a little more than a month. **The pair found that the harsher sentences disproportionately affected black defendants**." [14]

"**A prisoner's chance of parole depends on when the judge hearing the case last took a break**, say researchers who have studied decisions in Israeli courts. As judges tire and get hungry, they slip towards the easy option of denying parole, say the researchers." [15]

The first example illustrates (unintentional) **bias** in the decision process, and would widely be viewed as an unfair process. The second example would also be viewed as unfair, but in this case because sentencing severity was influenced by an arbitrary factor unrelated to the case; however, there is no indication that the change in sentencing severity disproportionately affected any specific cohort. This process is unfair because it is **capricious**. Mitigating risks to fair decision and recommendation processes involving machine learning requires addressing both sources of unfairness, biased and capricious behavior. The tools and techniques to calibrate and mitigate these risks may well overlap but in some important ways they are distinct.

References

[1] obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
[2] www.darpa.mil/program/explainable-artificial-intelligence
[3] github.com/marcotcr/lime
[4] www.csail.mit.edu/making_computers_explain_themselves
[5] sites.google.com/site/2016whi/
[6] www.navy.mil/navydata/testimony/safety/bowman031029.txt
[7] www.csl.sri.com/users/rushby/papers/sri-csl-15-1-assurance-cases.pdf
[8] ntrs.nasa.gov/search.jsp?R=20150002819
[9] ISO/IEC 15026-2:2011, Systems and software engineering -- Systems and software assurance – Part 2: Assurance case
[10] MIL-STD-882E, 11 May 2012, Department of Defense Standard Practice System Safety
[11] psas.scripts.mit.edu/home/wp-content/uploads/2015/06/STPA-Primer-v1.pdf
[12] sunnyday.mit.edu/safer-world/Arnold-Thesis.pdf
[13] www.dcs.gla.ac.uk/~johnson/book/
[14] www.theatlantic.com/education/archive/2016/09/judges-issue-longer-sentences-when-their-college-football-team-loses/498980/
[15] www.nature.com/news/2011/110411/full/news.2011.227.html