A framework for addressing fairness in consequential machine learning

Chuck Howell, Chief Scientist for Dependable Artificial Intelligence, the MITRE Corporation, howell@mitre.org

There is a case to be made that society is in the early phase of another Industrial Revolution, driven by the rapid advancement of AI capabilities, with at least as significant an impact as the first Industrial Revolution and at an accelerated pace of adoption. However, inability to establish justified confidence in consequential AI systems will inhibit their adoption (wasting opportunities to tackle major problems) and/or result in adoption of systems with major latent flaws and potentially serious consequences.

Concerns about fairness in AI-based systems are expressed in best-selling books (e.g., *Weapons of Math Destruction*), focused workshops (e.g., https://fatconference.org/2018/), in the 2016 White House report *Preparing for the Future of Artificial Intelligence* [1], and many other venues. 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, and 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 systems safety community can be adapted to support the calibration, mitigation, and informed acceptance of fairness risks in consequential ML systems.

Perhaps counter-intuitively, in some ways the national security domain is more tolerant of errors and residual doubt about AI performance than civilian agencies are when it comes to machine learning. For example, when using machine learning to sift through thousands of hours of otherwise unwatched full-motion video for activity of interest, any intelligence value uncovered is more than what would otherwise have been found because of the sheer volume of video. If some of the machine classifications aren't right, the net result may still be added value, assuming the volume of false positives doesn't become counter-productive. However, if a civilian agency deployed a machine learning system to make consequential recommendations that could certainly influence decisions, and a credible allegation is made that the system is unfair (biased and/or capricious) in specific instances, the validity of the entire system could be challenged; it is less likely that an argument that "it is still an improvement over the previous situation" would protect the system and the agency from significant public scrutiny. This why we view a framework for addressing concerns about fairness for machine learning to be an enabler and accelerant for the adoption of machine learning in civilian government agencies.

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 an informed risk assessment about using it can be performed and communicated [6].

Approved for Public Release; Distribution Unlimited. Case Number 17-4085
©2017 The MITRE Corporation. ALL RIGHTS RESERVED.

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. Fairness concerns will of course involve a broad range of stakeholders, and a complex mix of tradeoffs. Safety critical developments also include tradeoffs and balancing conflicting constraints; the observation here is that these tradeoffs are better made when explicit and deliberate than when implicit and ad hoc.

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).

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*. An example 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]

The example illustrates **bias** in the decision process, and would widely be viewed as an unfair process, because it disproportionately affected a specific group of defendants based in this instance on race. It would also be viewed as unfair because sentencing severity was influenced by an arbitrary factor unrelated to the case and unrelated to any characteristics of the defendants, the outcome of a football game. The process is also unfair because it is **capricious**. In fact, The Administrative Procedures Act uses the standard of "arbitrary and capricious" as one example of grounds for overturning an agency action [16]. Mitigating risks to fair decision and recommendation processes involving machine learning requires addressing both biased and capricious sources of unfairness. The tools and techniques to calibrate and mitigate these risks may well overlap but in some important ways they are distinct.

Approved for Public Release; Distribution Unlimited. Case Number 17-4085
©2017 The MITRE Corporation. ALL RIGHTS RESERVED.

The MITRE investigation of safety TTPs is also shaped by the need to adapt what can be slow, very deliberative certification processes to the agile, rapid, iterative nature of ML development.

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
[16] https://www.archives.gov/federal-register/laws/administrative-procedure

MITRE is a not-for-profit organization that operates federally funded research and development centers, dedicated to solving problems for a safer world.
https://www.mitre.org

Approved for Public Release; Distribution Unlimited. Case Number 17-4085
©2017 The MITRE Corporation. ALL RIGHTS RESERVED.