How do AI systems fail socially? Social Failure Mode and Effect Analysis for Artificial Intelligence Systems

Shalaleh Rismani, AJung Moon

Background

Failure Mode and Effect Analysis (FMEA) is part of standard risk management in automotive, aerospace and medical devices industries [1].

An FMEA is an engineering risk analysis tool used to identify, analyze and mitigate potential failure modes (ex. flat tire on a bike). [1] Researchers have recognized the applicability of FMEAs for algorithmic auditing [2].

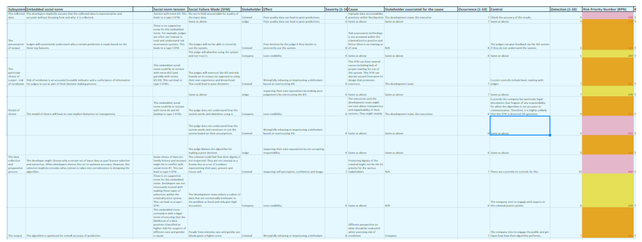
Building on Millar's definition of social failure modes, we propose a modified FMEA process to identify sociotechnical failures for AI systems.

Social failure modes(SFM) occur when social norms embedded in a technology are not supported or are in tension with existing with a societal norm [3]

Social FMEA Case Study: COMPAS

Use case: risk assessment algorithm to determine likelihood of recedivism for criminal suspects

Figure 1. Snapshot of a the sample Social FMEA

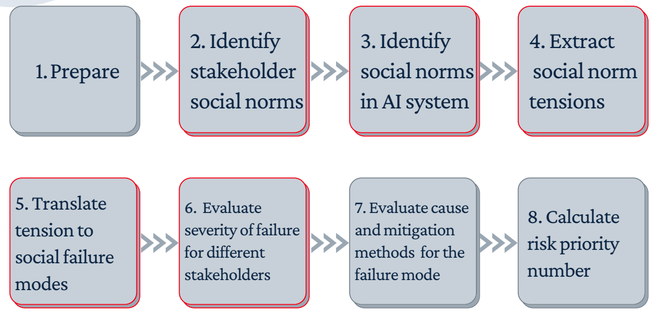


Social norm 1: ensure the likelihood of a false positives (classified as higher risk) for suspects of different race and gender is equal.
Social norm 2: the algorithm is optimized for overall accuracy of prediction.

Scan for full COMPAS FMEA

Social Norm Tension and SFM: People from minority races and gender falsely receive a higher score. [4]

Social FMEA Process



Discussion and Next steps

- This preliminary work develops a systematic way of identifying social failure modes and provides a case study. **How could we make this process more practical and granular?**
- FMEAS and Social FMEAs are resource intensive. **How could they be adapted for the fast-paced AI system development process?**

AI SYSTEMS IMPACT MANY LIVES. IT IS IMPORTANT TO DEVELOP TECHNIQUES FOR ANALYZING AND UNDERSTANDING HOW THEY MIGHT FAIL AND CAUSE HARM.

[1] C. S. Carlson, Effective FMEAs. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012.
[2] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap," Fairness, Accountability, Transparency Conference., pp. 33–44, 2020.
[3] J. Millar, "Social Failure Modes in Technology and the Ethics of AI," Oxford Handb. Ethics AI, no. February 2021, pp. 441–461, 2020.
[4] S. M. Julia Angwin, Jeff Larson and L. Kirchner, "Machine Bias," 2016. [Online]. Available: https://www.propublica.org/article/machinebias-risk-assessments-in-criminal-sentencing