ORCAA

Dealing with High Impact AI

Prepared by Cathy O'Neil for the AI Insight Forum

November 1, 2023

My name is Cathy O'Neil. I am a mathematician, data scientist, and algorithmic auditor. In 2016 I founded ORCAA,[1] to help develop standards for safe, fair, and trustworthy AI. In the past 7 years we have worked with clients including regulators, state and federal enforcement agencies, and private companies to safeguard algorithmic systems in industries including insurance, housing, credit, and online platforms.

Why we should be worried

AI systems are handling high-stakes decisions in nearly every area of life, and they are making mistakes.

Researchers have been chronicling this for years, from my own *Weapons of Math Destruction* in 2016, to the Gender Shades study[2] that showed how facial recognition systems work less well for darker-skinned individuals and women, to Virginia Eubanks' *Automating Inequality*.

More recent algorithmic failures in major regulated industries and other high-impact scenarios include:

- In insurance, a class action lawsuit alleges State Farm's use of AI fraud detection systems in settling homeowners claims discriminated against Black policyholders.[3]
- In the area of housing, the Department of Justice sued Meta, following a HUD investigation that found its advertising algorithms used protected characteristics to bias the display of housing ads towards certain groups, violating the Fair Housing Act.[4]
- On the National Eating Disorder Alliance website, a chatbot powered by an AI large language model was supposed to serve as a "prevention resource" for those struggling with eating disorders – but it started dispensing diet and weight loss advice instead.[5]

[1] www.orcaarisk.com
[2] http://gendershades.org/
[3] https://www.classaction.org/media/huskey-v-state-farm-fire-and-casualty-company.pdf
[4] https://www.justice.gov/media/1227811/dl?inline
[5] https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea

These have prompted responses by lawmakers, regulators, and enforcers:

- Colorado state law SB 169, passed in 2021, addresses unfair discrimination in insurance arising from big data and predictive algorithms; the Division of Insurance recently released draft regulations outlining statistical tests that will be required to demonstrate compliance.[6]
- In the area of employment, New York City Local Law 144 requires annual Bias Audits of automated decision tools used in hiring.[7]
- Under a settlement with the Department of Justice, Meta has agreed to build a system to mitigate the bias, and hit a predefined schedule of compliance targets.[8]

AI systems will always make mistakes. These will often follow historical patterns and power dynamics, which are reflected in the vast personal data – our financial, educational, and criminal histories, our digital footprints – that are the basis for these systems. And, because AI systems are deployed at scale, their mistakes have the potential to cause enormous harm.

The good news is, there is a way to police these systems for harms and, in doing so, make them better and build trust in them. It involves three ideas, which we describe below.

Ethical Matrix

At my company ORCAA, an algorithmic audit starts with the simple question: For whom could this fail? This centers the audit around stakeholders: the people whose lives could be impacted by the algorithm, for better or worse. Our audit process involves talking directly to stakeholders to understand their concerns.

The Ethical Matrix is the framework we use for audits. Its purpose is to expand the conversation around whether an algorithm "works" beyond the usual notions of accuracy, efficiency, or profit. These are often priorities for the deployer, but they may not address the concerns of external stakeholders, like job or insurance applicants. There are three steps to creating an Ethical Matrix:

1. **Fix a Use Case.** We view an algorithm not merely as lines of code, but as a deployed system in the real world. Context matters: Where, when, and how is it used, by whom and for whom, to make what decisions? Answering these questions helps define the scope of the audit and identify stakeholders.
2. **Elicit concerns from stakeholders.** With key stakeholder groups identified, we ask each how the system could fail or succeed from their perspective. How

[6] https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view
[7] https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page
[8] https://www.theverge.com/2023/1/9/23547191/meta-equitable-ads-system-settlement

exactly would they be harmed (or benefit)? Stakeholders' concerns, both positive and negative, suggest what needs to be measured and monitored.

3. **Validate and prioritize concerns.** Often we can't tell immediately whether a given concern is actually being realized, or how big a problem it is. This step involves measuring the various concerns and deciding which ones require a response most urgently.

In sum, the Ethical Matrix distills the extremely broad question "How could this AI system fail?" into a tractable set of specific stakeholder concerns – potential harms – that can be monitored and mitigated. In order to perform this audit, we narrow down a Use Case to understand the real world consequences of failure. That means there is no universal checklist for the safety of all AI systems, but there is a reliable way to assess the safety of a given AI system.

Explainable Fairness

Whether through an audit or otherwise, we often arrive at the question: Does this AI system work equally well for different kinds – races, genders, ages, languages, etc. – of people? The responsible AI community struggles to give a clear answer to this obviously valid question. On one hand, proponents of *explainability* aim to show how AI systems "think" or exactly where their predictions come from. Unfortunately, it is hard enough to explain predictions from a simple linear regression, let alone sophisticated AI; nobody (including the inventors!) can account for each word in a large language model's response to a given prompt. On the other hand, conversations about *fairness* often devolve into technical arguments over dozens of competing metrics and definitions.[9] Neither of these approaches addresses the concerns of regular people or helps win their trust.

We developed a framework called Explainable Fairness[10] that presents a way forward: we **clarify what's meant by 'fairness' for a given AI system with precise, nontechnical statements about how real people are treated by it**. For example, applying Explainable Fairness to AI-assisted underwriting of student loans, you could say (hypothetically): "The system produced fair interest rates, meaning: average APR differed by less than one percentage point across race groups, after accounting for borrowers' FICO scores." Here is a sketch of the process:

[9] A well-known session from the 2018 FAccT conference – the biggest ML fairness conference – was entitled "21 Fairness Definitions and their Politics". The debates continue today; see "A Clarification of the nuances in the fairness metrics landscape" from 2022.
[10] A preprint of our paper is available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4598305

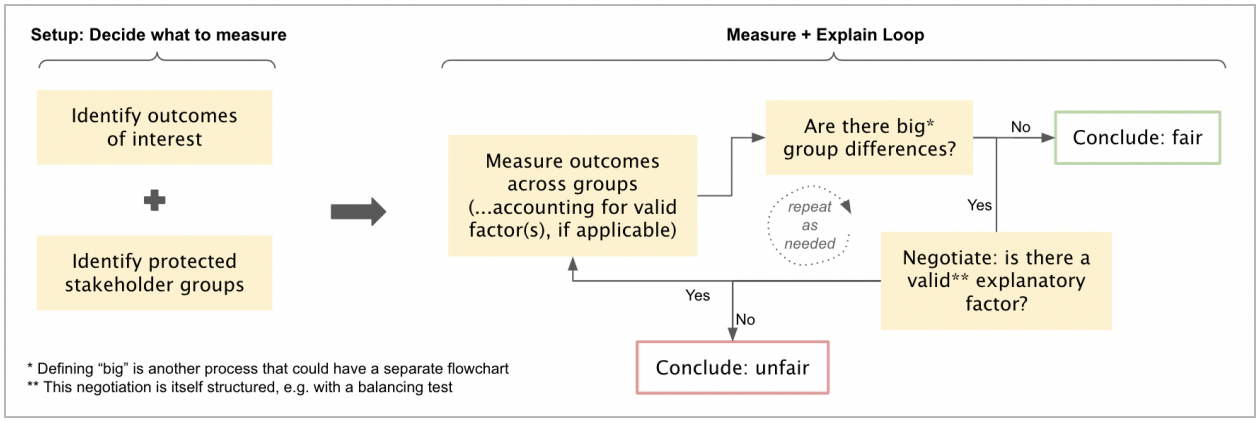

Figure 1: Explainable Fairness schematic

Explainable Fairness is a complete, end-to-end method for developing these kinds of statements about fairness. It coordinates the efforts of regulators/policymakers, AI system deployers, and data scientists in defining relevant metrics, setting thresholds of acceptability, and ultimately measuring outcomes of AI systems across groups.

Designing a cockpit

In considering the appropriate scope for monitoring/auditing AI systems and regulatory oversight, we often use the metaphor of a cockpit in an airplane. Planes are of course thoroughly tested before leaving the factory, and inspected before each flight. But we would never get on an airplane that had no cockpit. To fly safely, pilots need to monitor changing conditions, know when danger is imminent, and adjust accordingly. This is exactly what the cockpit enables.

Likewise with AI systems. Ensuring safety and fairness is an ongoing process, not a one-time effort. The major failure modes, which could be identified using the Ethical Matrix or another method, become the "dials and gauges" in the cockpit. And the thresholds of acceptability – say, the maximum allowable difference between groups in the average outcome – become the "redlines" on the various dials and gauges.

In this metaphor, we have two roles as algorithmic auditors. One is doing the pre-flight testing and inspections to catch major issues before deployment. The other is helping design the cockpits that will ensure the system operates safely and fairly when deployed.

Alignment with NIST AI Risk Management Framework

We are encouraged by the efforts of many actors in the public sector to take on this issue. NIST's AI Risk Management Framework (AI RMF)[11] is a notable project that will

[11] https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

become a central framework for how to approach high-impact AI systems across a range of industries and settings.

Therefore, I want to show how the three ideas just discussed map onto the NIST AI RMF. Here is a schematic showing the main components of the AI RMF, with our ideas overlaid in bold:

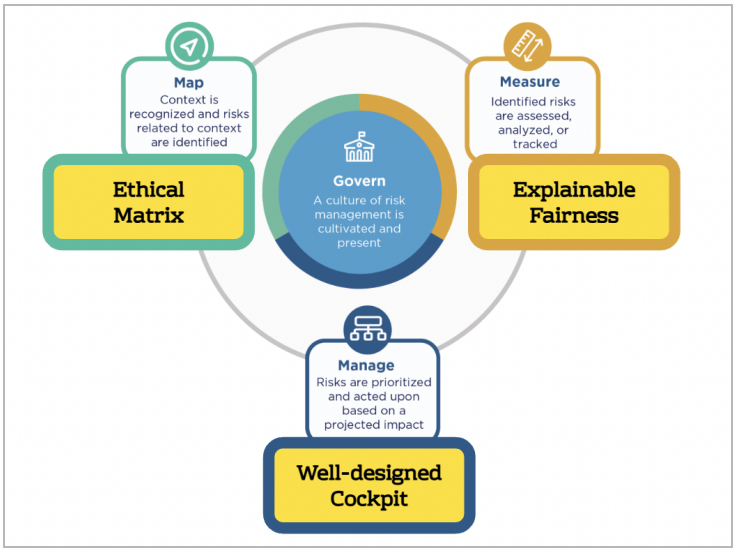

Figure 2: Alignment with NIST AI RMF Core

The AI RMF is addressed to organizations that build and deploy consequential AI. It says the risks of individual AI systems should be Mapped, Measured, and Managed; and that this ongoing process should be informed by a central backbone of Governance. We suggest that:

- The Ethical Matrix framework is a way to **map** the urgent risks of a given AI system, while deliberately taking into account the context of its deployment.
- Explainable Fairness is a meaningful, robust, and understandable way to **measure** a certain kind of risk – differential performance across groups – that is of concern in high-stakes AI systems. (This is not the only risk to worry about, of course.)
- To **manage** these risks in practice, it is helpful to think of "designing a cockpit" for a given AI system. The cockpit should deliver timely information about the status of the system and surrounding conditions, ultimately informing the pilot of any imminent danger in time to take corrective action. It should also be trusted and tested by outside auditors and performance experts.