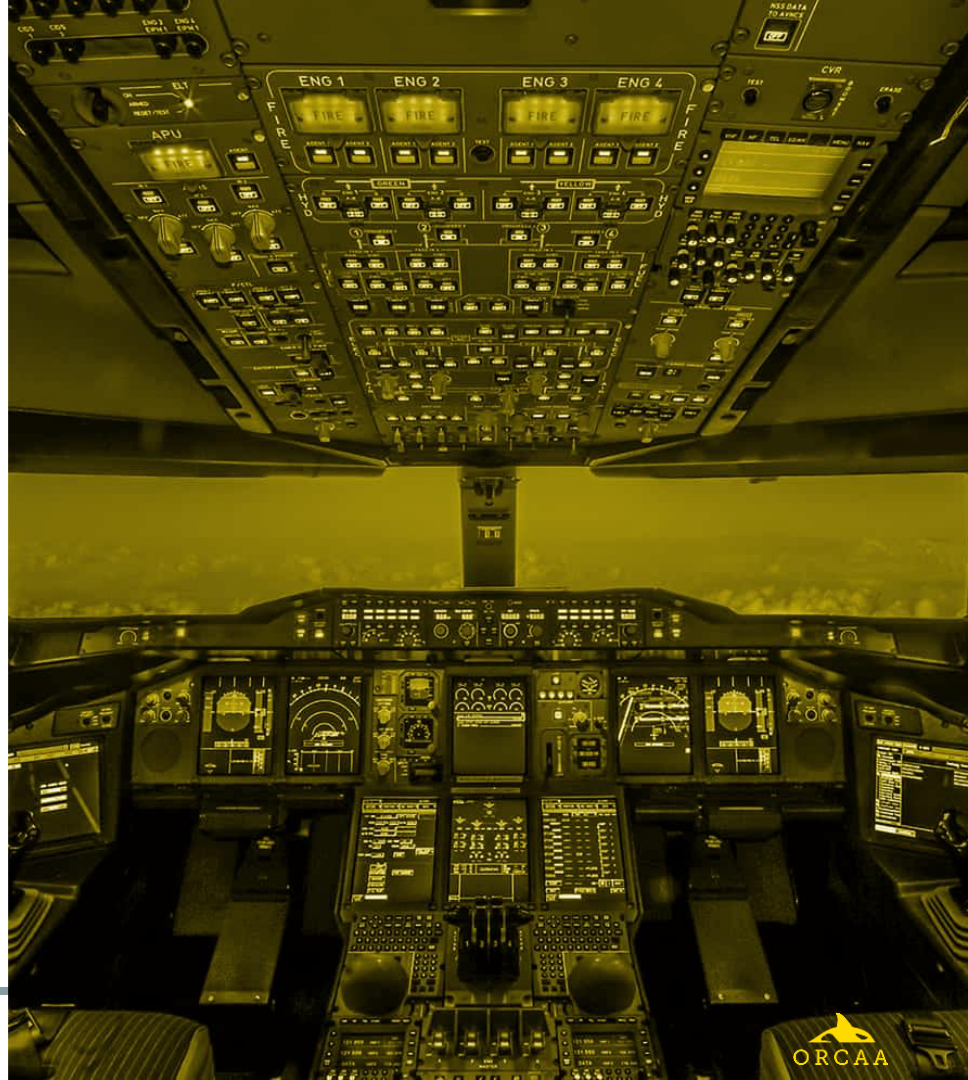
Algorithm Audits

May 27, 2025

Auditing as Cockpit Design

- Monitor potential failures
- Dials
- Redlines
- Warning vs. intervening
- Part of “system safety”



Algorithm Risk Audits

Identify harms

- What groups?
- What outcomes?

Define metrics

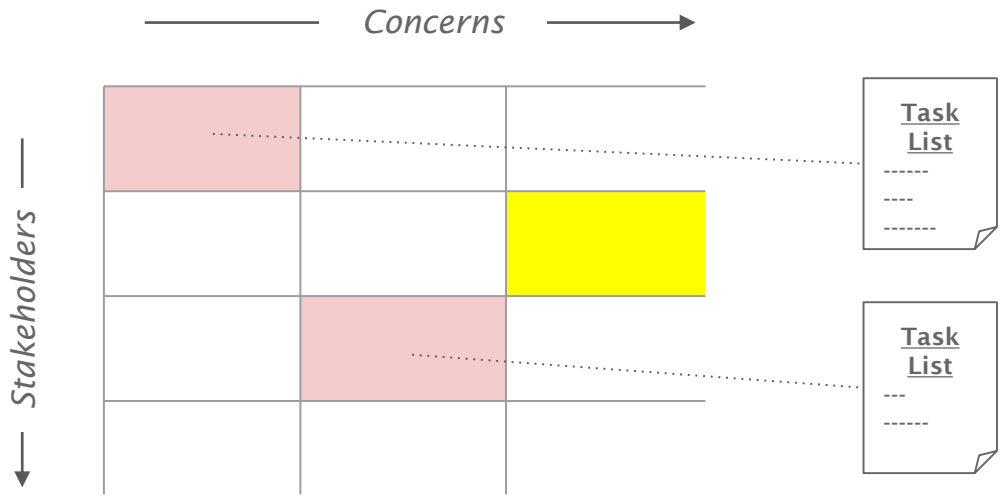
- Data needed
- Who will measure?

Set targets

- Absolute vs relative
- Can change over time

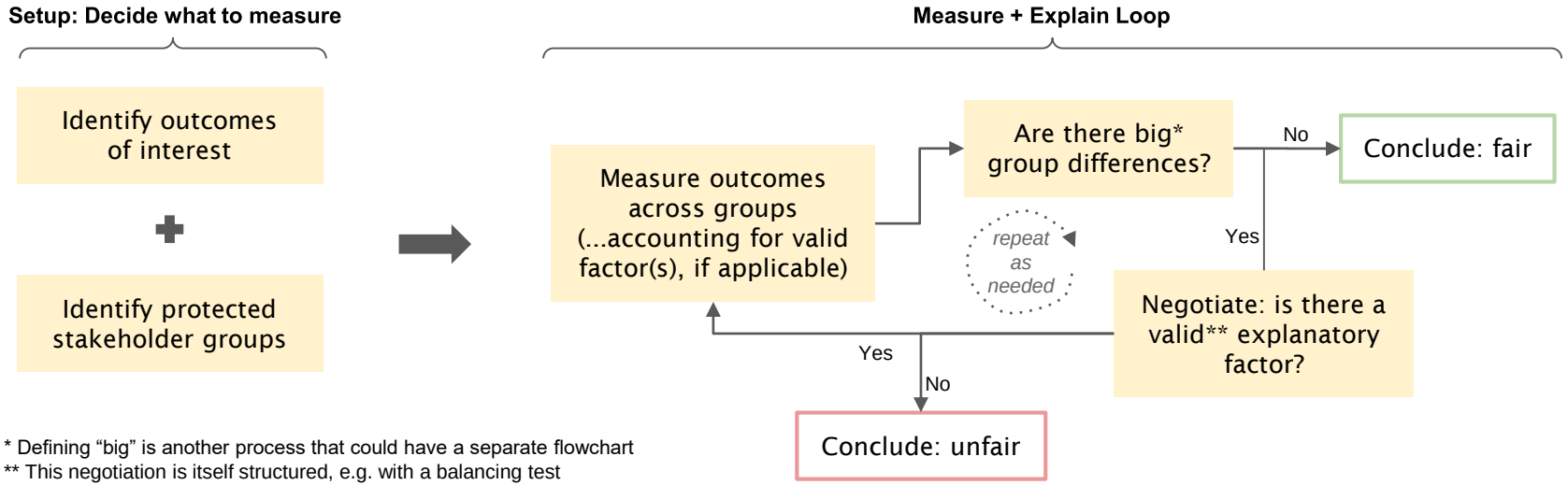
Identifying Harms: Ethical Matrix Framework

1. Identify internal and external **stakeholder groups**
2. Elicit stakeholders' concrete legal or ethical **concerns** about the model
3. **Prioritize** major risks and threats, assess tradeoffs (colors)
4. Develop **statistical tests and remediation steps** for key risks



Define Metrics: Explainable Fairness

A high-level approach for measuring bias/fairness in an AI system. Detail in WVLR article.



* Defining "big" is another process that could have a separate flowchart
** This negotiation is itself structured, e.g. with a balancing test

Case Study: Meta VRS Settlement

Identify harms

- Unequal display of housing ads
- Race and gender

Define metrics

- Eligible vs actual audience
- Earth mover distance

Set targets

- Multiple areas + metrics
- Tightens over time
- Independent monitor

https://www.justice.gov/usao-sdny/pr/united-states-attorney-resolves-groundbreaking-suit-against-meta-platforms-inc-formerly
https://www.justice.gov/d9/2023-12/meta_vrs_compliance_metrics_verification_report_october_2023-final.pdf
https://about.fb.com/wp-content/uploads/2023/07/Meta-VRS-Compliance-Metrics-Verification-Report-2023-06-29-For-Publication.pdf

ORCAA