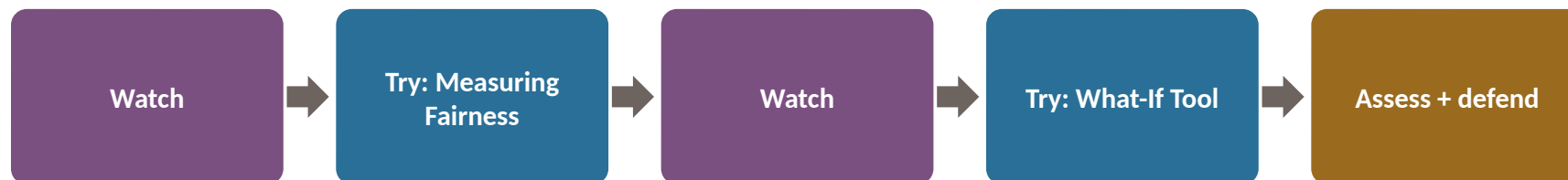


MODULE 7

Data and AI ethics and governance

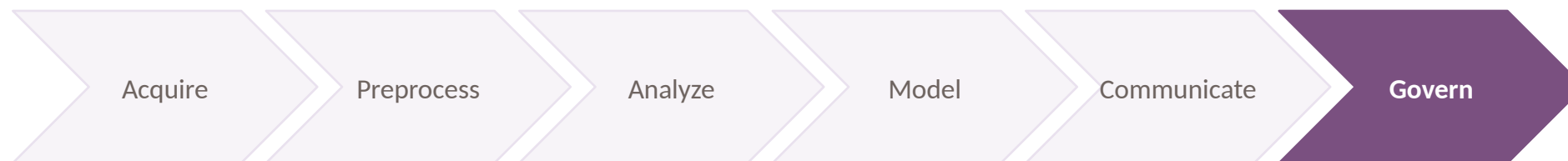
Harms get specified, measured, and controlled with the same rigor as latency. “We didn’t intend it” is not a control.

Watch a little, then do a little



Each recorded segment is short. You pause, practice the concept hands-on, then return for the next piece.

Where this module sits in the lifecycle



This is the govern layer: culture and accountability, contextualizing the system, quantifying risk, and acting on it. Governance is ongoing work, because models drift and a one-time audit is not governance.

PART 1 OF 2

Measuring harm and the impossibility results

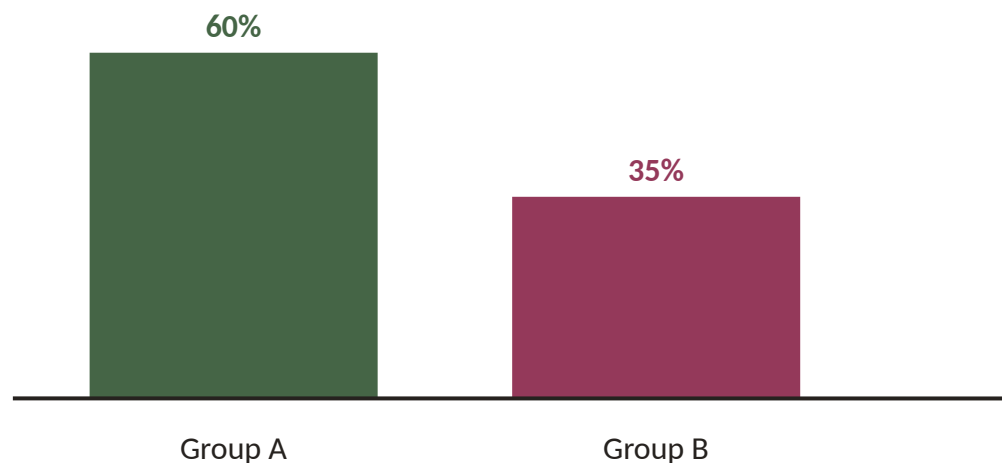
About 12 minutes on camera

Disparate impact, measured

- Compare subgroup base rates, selection rates, and error rates rather than trusting intent.
- The four-fifths rule is a screening heuristic, not a legal safe harbor.
- A flag warrants a closer audit; it does not by itself prove or clear discrimination.

A worked disparate-impact check

Selection rate by group



The four-fifths screen

Ratio = $0.35 \div 0.60 = 0.58$. Below the 0.80 threshold, so this is a screening flag that the classifier warrants a closer disparate-impact audit, not proof and not a safe harbor.

Fairness definitions conflict

Calibration

Equalized odds

Demographic parity

Pick at most two

When base rates differ across groups, these three cannot generally hold at once. Choosing a fairness definition is a value judgment you justify, not a quantity you compute away.

NOW TRY IT

PAIR Measuring Fairness + Equality of Odds

- In PAIR's Measuring Fairness, move the threshold and watch improving one group's outcome degrade a competing metric.
- In MLU-Explain's Equality of Odds, see how equalized true- and false-positive rates differ from parity and calibration.

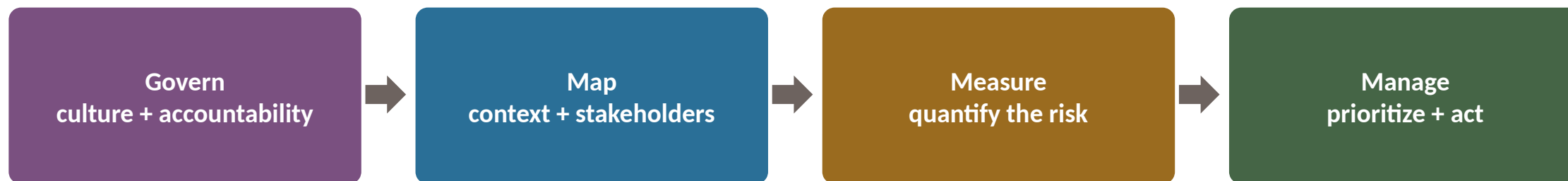
About 15 minutes, then come back for Part 2.

PART 2 OF 2

The NIST AI RMF and governance

About 13 minutes on camera

The NIST AI Risk Management Framework



The four functions are connective tissue: Govern sets accountability, Map contextualizes the system, Measure quantifies risk, and Manage acts on it.

Document what the model is and where the data came from

Model Card

States behavior, intended use, and limitations, including subgroup performance.

Datasheet

Records provenance and consent: where the data came from and whether subjects agreed to this use.

Consent + privacy

De-identification is fragile under re-identification and linkage attacks.

Monitoring

Populations shift and models drift; documentation, thresholds, and a review cadence make governance durable.

NOW TRY IT

Aequitas and the What-If Tool

- Audit a model with Aequitas to see subgroup disparities reported against a reference group.
- Probe predictions with Google's What-If Tool to inspect counterfactuals and per-subgroup performance without new code.
- Play Survival of the Best Fit to see how historical data launders past discrimination into an automated system.

About 20 minutes, then run the formative check.

Formative deliverable (completion points)

- Run a disparate-impact check on the module's case classifier and its dataset.
- Choose a sensitive attribute, compute one fairness metric (selection rate or false-positive rate) per subgroup, and report the disparate-impact ratio of the least- to most-favored group.
- In two or three sentences, name which competing metric your chosen one trades against, and in which direction.

How it counts. Submit the work with a short reflection or evidence you ran it. Credited on completion, not scored for correctness. This is the practice that makes the graded simulation go well.

GRADED ASSESSMENT

The Model-Governance Review Board

- You convene a NIST AI RMF review of a classifier that makes decisions about people, built on records repurposed without consent, whose subgroup table shows real disparate impact.
- You assemble a governance dossier (Model Card, Datasheet, measured harms, proposed controls) and recommend approve, reject, or condition.
- You defend it before a rotating three-person board that asks unscripted questions.

Includes a live or recorded oral defense: rehearse it first with the on-device defense-practice tool.