Module 6

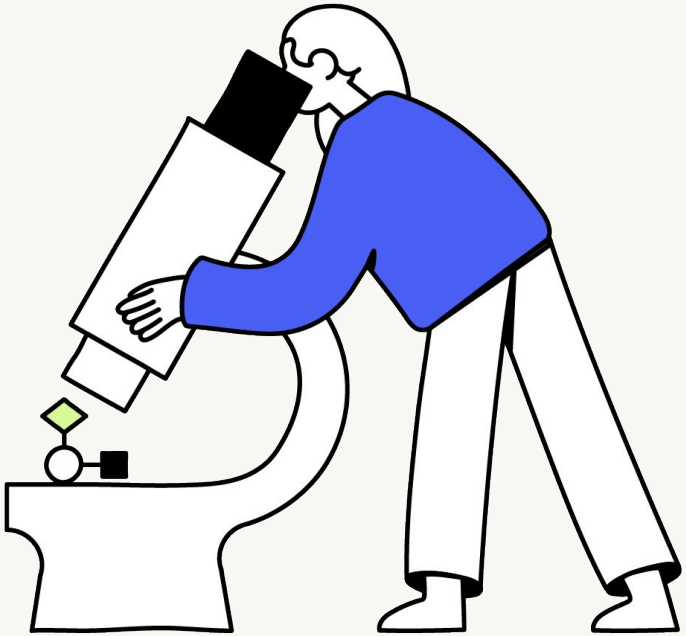
# Workshop 2: Auditing Fairness and Bias in ML Models

# Fair or Flawed?

Which do you think is harder to detect — performance issues or fairness issues in a model?

What makes one trickier to uncover than the other?

**Type your answer in the chat!**



Module 6

# Workshop 2: Auditing Fairness and Bias in ML Models

# Agenda

- **Review:** Recap key concepts.
- **Practical exercise:** Bias under the microscope.
- **Closing:** Wrap up and reflection.

# Learning objectives

- Detect algorithmic bias in model outputs using fairness metrics and explainable AI tools.
- Interpret subgroup disparities using fairness visualisations.
- Recommend mitigation strategies and documentation practices for ethical AI deployment.

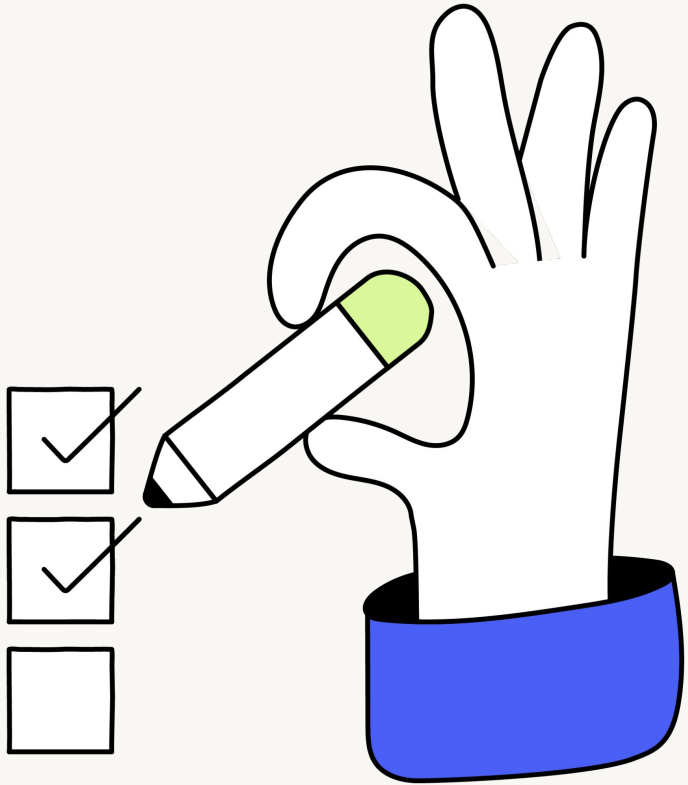
# Async review: Unit 3

Go to My Multiverse, Module 6
Workshop 2

Navigate to the **“Async review”** page

Review the summary of Unit 3, then complete the following Poll:

**How fair is your model?**



# Poll debrief

- Fairness auditing can break down at different stages—detection, evaluation, or long-term monitoring.

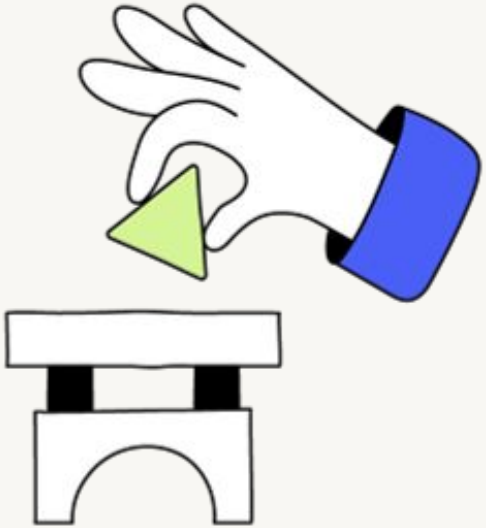
**Where do you see the biggest opportunity to strengthen fairness in your own ML projects?**

# What is fairness bias in ML?

- Fairness bias happens when model outcomes differ unfairly across subgroups.
- It can emerge at any stage of the ML lifecycle:
  - **Data:** Underrepresentation or skewed labels.
  - **Modeling:** Algorithms that favor majority patterns.
  - **Evaluation:** Metrics that miss subgroup differences.
- Consequences include ethical risks, compliance issues, and loss of trust.

# Detecting bias – looking beyond accuracy

- High accuracy doesn’t guarantee fairness.
- Use fairness metrics such as:
  - **Demographic parity** – equal outcomes.
  - **Equal opportunity** – equal true positive rates.
  - **Predictive parity** – equal precision.
- Tools like Fairlearn, Aequitas, SHAP, and LIME help reveal subgroup disparities.

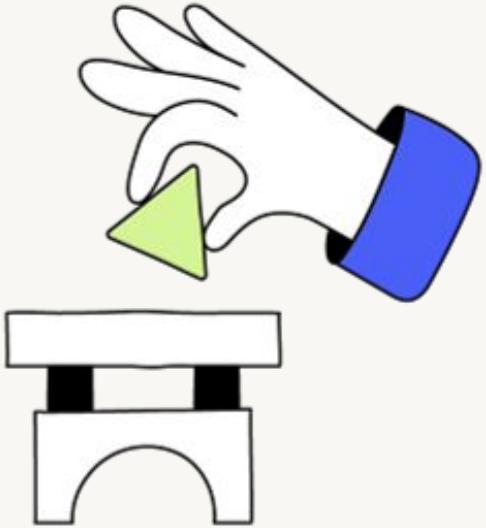


# Mitigating and monitoring bias

Once bias is detected, the next step is to reduce its impact while maintaining model performance.

- **Pre-processing:** Rebalance or reweight data.
- **In-processing:** Apply fairness constraints or regularisation.
- **Post-processing:** Adjust thresholds or outputs.

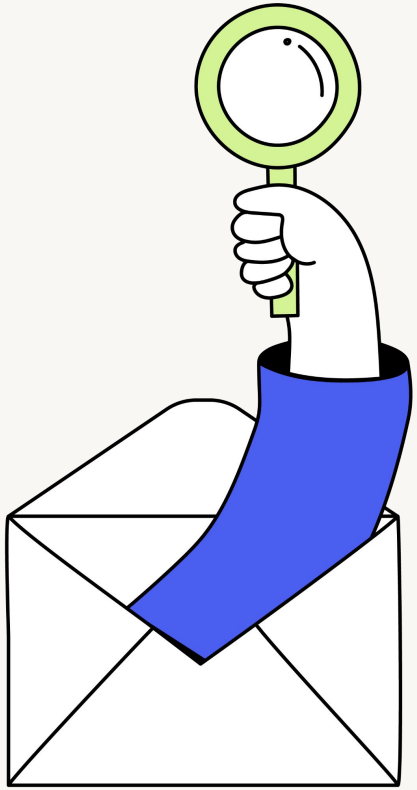
Document findings in model audit cards and monitor fairness continuously as data changes.



# Practical exercise: Bias under the microscope

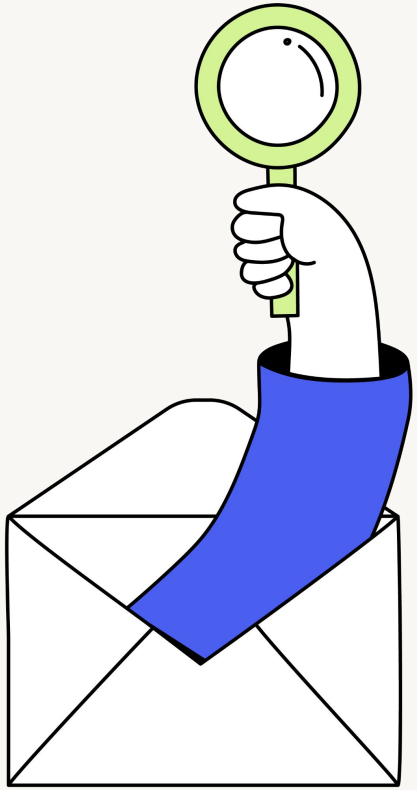
# Practical exercise: Context

- You’re part of a data science team **conducting a fairness audit for a loan approval model** trained on five years of applicant data.
- The model performs well overall but shows **lower approval rates for women and minority groups.**
- **Your task** is to use fairness metrics and explainability tools to detect, interpret, and mitigate potential bias.
- **The goal** is to ensure the model is fair, transparent, and compliant with ethical AI standards before deployment.



# Practical exercise instructions

- Work in breakout rooms in small groups.
- Step into the role of the data science fairness audit team.
- Find all instructions and resources on Ariel.



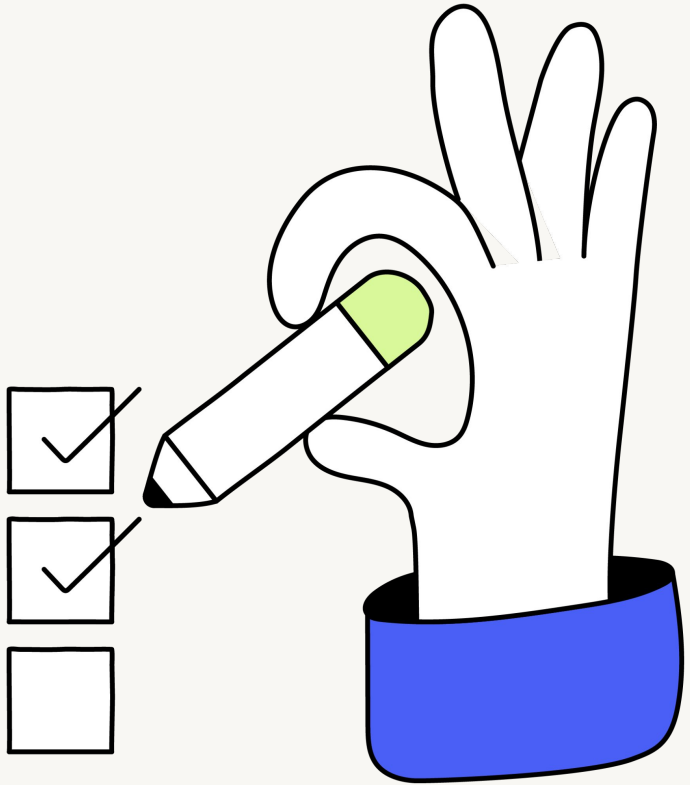
# Access the skills application

Go to My Multiverse, Module 6 Workshop 2

Navigate to the **“Practical exercise”** page

Start working on the activity tasks

Be prepared to share in 20 minutes when we come back together



# Group discussion debrief

# Practical exercise debrief

1. Which subgroup showed the most significant disparity across fairness metrics such as demographic parity or equal opportunity?

2. How did your threshold adjustments affect the balance between model accuracy and fairness?

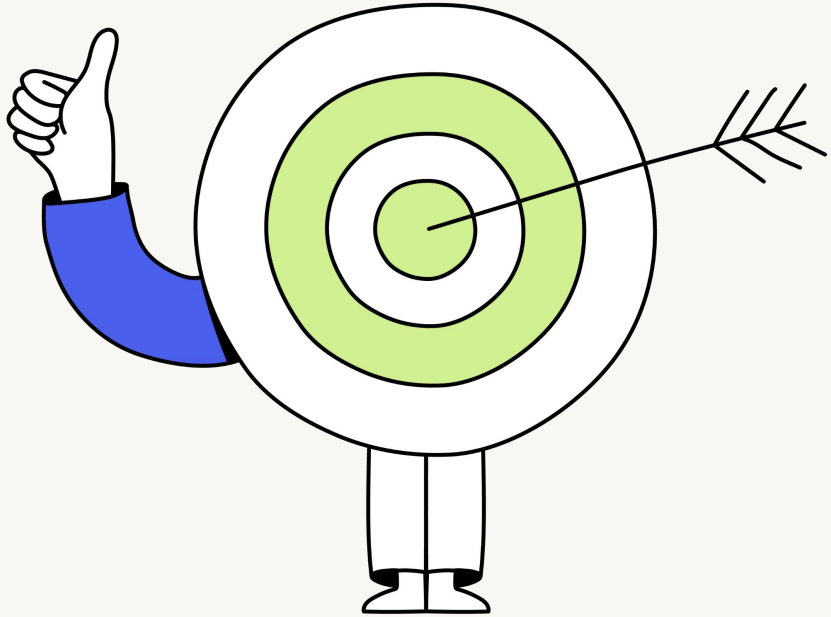
3. What mitigation strategy would you prioritise next, and how would you document it to support transparent and responsible deployment?

# Key takeaways

1. Fairness audits reveal how model performance can vary across demographic groups.

2. Adjusting thresholds or applying mitigation strategies helps balance accuracy and equity.

3. Transparent documentation and continuous monitoring are essential for responsible AI deployment.



# Applying key takeaways to your organisation

# Applying key takeaways

1. How does your organisation ensure fairness and transparency in ML model development and deployment?

2. Where do bias risks or fairness gaps most often appear in your workflows or decision processes?

3. What actions could your team take to make fairness auditing and documentation more consistent, collaborative, and accountable across projects?

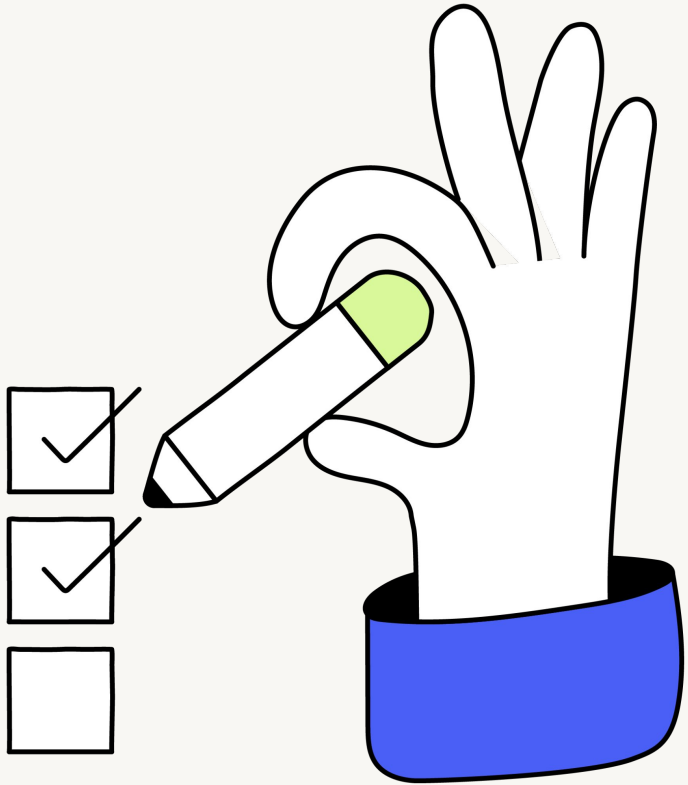
# Share your plan to apply new skills

Go to My Multiverse, Module 6
Workshop 2

Navigate to the discussion activity page **“Apply new skills to your role”**

Take 5 minutes to share how you will apply this skill in your role

Take time this week to read what others have written and get new ideas!



# Closing out

- **Next up:** Module Wrap-Up Workshop
- **If you have questions, you can:**
  - Chat with Atlas
  - Reach out to your Cohort Coach
  - Sign up for Tutoring
- **Log your OTJ hours**
- **Take the Workshop Survey** on My Multiverse

multiverse