MODERATE

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

10–14 years | 35–50 min | 8–30 participants | Card game, investigation and discussion

OVERVIEW

Participants investigate a fictional AI system that selects candidates for a space-rescue team. The system has learned from a limited set of previous examples and begins treating irrelevant characteristics as signs of ability. By testing and improving the system, participants discover how incomplete or unbalanced training data can produce biased and unfair results.

LEARNING OBJECTIVES

Participants will:

- understand that AI systems learn patterns from training data
- recognize that a pattern is not necessarily a meaningful cause
- observe how incomplete or unbalanced data can distort predictions
- identify the difference between relevant and irrelevant characteristics
- understand how biased systems can affect people unfairly
- explore ways humans can test and improve an AI system

MATERIALS

- A set of fictional candidate cards
- A set of previous team-member cards labelled SELECTED
- A set of previous candidate cards labelled NOT SELECTED
- A board, flip chart or large sheet of paper
- Markers
- Two signs: SELECT and DO NOT SELECT
- Optional highlighters or coloured stickers
- Optional machine rule cards

Each fictional candidate card should include several characteristics, such as:

- name or number
- special skill
- training completed
- ability to work in a team
- type or colour of equipment
- favourite food
- home planet or region
- number of missions completed
- whether the candidate carries a backpack
- whether the candidate wears a helmet or hat

Use fictional characters, robots or space explorers so participants can examine bias without assigning stereotypes to real classmates or social groups.

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

PREPARATION

Create six cards representing members of previous successful space-rescue teams.
Give them a mixture of relevant and irrelevant characteristics.

For example, most previously selected candidates might:

- **have completed rescue training**
- **demonstrate teamwork**
- **have navigation or medical skills**
- **carry blue backpacks**

Create four previous NOT SELECTED cards.
These candidates might lack relevant training, but none carries a blue backpack.

This creates a misleading pattern: the machine may learn that carrying a blue backpack is connected to success, even though the backpack colour has no effect on rescue ability.

Prepare six to eight new candidate cards, including:

- **highly qualified candidates without blue backpacks**
- **less-qualified candidates with blue backpacks**
- **candidates with different combinations of skills and equipment**
- **at least two candidates who are difficult to compare**

Do not tell participants which characteristics are intended to be relevant.

INTRODUCTION

Ask the participants:

- **What skills would someone need to join a rescue team?**
- **Should the colour of their clothing or equipment affect the decision?**
- **How might an AI system learn whom to select?**
- **What could happen if it learns from previous decisions that were incomplete or unfair?**

Explain:
An AI system can find patterns in previous examples. However, it does not automatically know which patterns are meaningful. Today, we will test a fictional machine that has learned how to select members of a space-rescue team.

Clarify that the machine is not deliberately trying to be unfair. It is following patterns in the information it received.

MODERATE

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

INSTRUCTIONS

1) EXAMINE THE PREVIOUS SELECTIONS

Divide participants into small groups.
Give each group the cards showing previous candidates who were selected and not selected.

Ask groups to compare the cards and identify possible patterns:

- What do the selected candidates have in common?
- What do the non-selected candidates have in common?
- Which characteristics might the machine consider important?
- Which characteristics actually seem relevant to rescue work?

Groups write down two or three possible selection rules.

2) BUILD THE MACHINE'S RULE

Invite groups to share their proposed rules.

They may suggest:

- Select candidates with rescue training.
- Select candidates with teamwork experience.
- Select candidates with a blue backpack.
- Select candidates who have completed at least two missions.
- Select candidates with medical or navigation skills.

Explain that an AI may detect all these patterns without understanding which ones contribute to successful rescue work.

Select one misleading machine rule based largely on backpack colour:

Candidates with blue backpacks are likely to be good rescue-team members.

Do not describe this as a good rule. It represents what the machine may have learned from the limited examples.

3) TEST NEW CANDIDATES

Show the new candidate cards one at a time.

The groups must apply the machine's learned rule and place each candidate under:

- SELECT
- DO NOT SELECT

Include a highly qualified candidate without a blue backpack and a poorly qualified candidate with one.

Ask after each decision:

- Did the machine follow its rule?
- Does the decision appear reasonable?
- What important information did the machine ignore?

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

INSTRUCTIONS

4) COMPARE PREDICTION AND SUITABILITY

After all candidates have been classified, ask groups to reassess them using only relevant rescue criteria:

- appropriate training
- useful skills
- experience
- teamwork
- ability to respond safely

Place the candidates into the two categories again.

Compare the machine's predictions with the group's decisions.

Discuss who was unfairly excluded and who was selected for the wrong reason.

5) INVESTIGATE THE TRAINING DATA

Return to the original training cards.

Ask:

- Did carrying a blue backpack cause previous candidates to succeed?
- Could it have been a coincidence?
- Was the backpack simply common among the examples?
- Were enough different kinds of successful candidates included?
- What examples were missing?

Explain that the machine identified a correlation—two things appearing together—but this did not prove that one caused the other.

6) IMPROVE THE TRAINING SET

Give groups additional examples of successful rescue workers with:

- different backpack colours
- different equipment
- different appearances
- varied combinations of relevant skills

Also include some unsuccessful candidates who carry blue backpacks.

Ask groups to develop improved selection criteria based on relevant abilities rather than accidental patterns.

Possible criteria include:

- completion of basic rescue training
- at least one relevant specialist skill
- evidence of teamwork
- ability to follow safety procedures

7) AUDIT THE IMPROVED MACHINE

Test the revised rules using another group of fictional candidates.

Ask participants to check:

- Does the system evaluate relevant skills?
- Does it exclude qualified candidates for irrelevant reasons?
- Are the criteria applied consistently?
- Are there unusual cases that require human review?
- What information should the machine not use?

Explain that checking a system for errors and unfair patterns can be called an audit.

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

DISCUSSION AND REFLECTION

Ask the participants:

- Why did the machine associate blue backpacks with successful candidates?
- Was the machine intentionally unfair?
- What was missing from its training data?
- What is the difference between finding a pattern and understanding a cause?
- Who was affected by the machine's mistake?
- Could adding more examples improve the system?
- Is more data always enough, or must the data also be appropriate?
- Who should decide which characteristics are relevant?
- Should a person check important AI-assisted decisions?
- What should happen if an AI system repeatedly disadvantages certain people?

Conclude with the key idea: AI systems learn from examples and previous patterns. If those examples are incomplete, unbalanced or based on unfair decisions, the system may reproduce or strengthen the same problems.

CONNECTION TO ARTIFICIAL INTELLIGENCE

AI systems often make predictions by identifying patterns in training data.
These patterns can be useful, but they can also be misleading.

Bias may enter an AI system when:

- some groups or situations are missing from the training data
- one group appears much more frequently than another
- examples are labelled inaccurately
- previous human decisions were already unfair
- the system uses an irrelevant characteristic as a substitute for something else
- the system is tested only under limited conditions

For example, an image-recognition system trained mainly on photographs taken in bright daylight may perform poorly at night. A medical system trained primarily using data from one population may be less accurate for others. A selection system trained on previous decisions may repeat the inequalities contained in those decisions.

Reducing bias can involve:

- using more representative training data
- checking data and labels carefully
- testing results across different groups and situations
- removing irrelevant information
- involving people with different perspectives
- providing human review and a way to challenge important decisions

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

EASIER VARIATION

Use a simple animal-recognition example.

Tell participants that a machine has learned to recognize birds using pictures in which every bird is shown flying in a blue sky. The machine may incorrectly learn that “blue background” is an important feature of a bird.

Test it with:

- a bird sitting in a tree
- a bird photographed at night
- an aeroplane in a blue sky
- a butterfly in a blue sky
- a bird standing on the ground

Ask which examples the machine might misclassify and what new training pictures it needs.

This version focuses on unbalanced training data without introducing decisions about candidates.

MORE CHALLENGING VARIATION

Give groups one of the following fictional AI systems to audit:

- a school system recommending students for a science club
- a health app predicting who may need an examination
- a wildlife camera identifying endangered animals
- a recruitment system selecting candidates for an interview
- a voice assistant recognizing spoken instructions

Groups identify:

1. What decision the system makes
2. What data it uses
3. Which groups or situations might be missing
4. What irrelevant patterns it might learn
5. Who could be harmed by a mistake
6. How the system should be tested
7. When human review is necessary

Groups present an “AI fairness checklist” for their system.

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

FACILITATOR NOTES

- Use fictional examples during the main game. Do not classify participants by appearance, identity, ability or background.
- Avoid implying that bias is simply a technical error. It may also reflect human decisions and existing inequalities.
- Explain that biased results can occur without a programmer deliberately creating an unfair rule.
- Distinguish between an imperfect prediction and unfair impact. Some errors have more serious consequences than others.
- Do not suggest that removing one characteristic automatically makes a system fair. Other information can sometimes act as a substitute for it.
- Emphasize that larger datasets are not necessarily better if they reproduce the same imbalance.
- Allow participants to question the categories and the purpose of the system itself.
- Avoid presenting AI as the final decision-maker in sensitive situations.
- Reinforce the importance of testing, transparency, human oversight and opportunities to challenge decisions.
- Keep real-world examples age-appropriate and avoid asking participants to share experiences of discrimination.

MODERATE

THE BIASED MACHINE

BIAS AND INCOMPLETE DATA

Science Film Festival

KEY WORDS

BIAS	A systematic tendency that can produce distorted or unfair results.
TRAINING DATA	Examples or information used to teach an AI system.
CORRELATION	A relationship in which two things occur together, without necessarily proving that one causes the other.
CAUSE	Something that directly contributes to a particular result.
REPRESENTATIVE DATA	Data that adequately include the different groups or situations in which a system will be used.
FAIRNESS	The principle of treating people appropriately and avoiding unjust disadvantage.
AUDIT	A structured examination of how a system works and what effects it produces.
HUMAN OVERSIGHT	Meaningful review or control of an automated system by people.