# Bias, Fairness and Power in AI

**Fatma Elsafoury 17.12.2025**

# Artificial Intelligence

## Definition

- AI is a marketing term refers to a set of **technologies and tools** that aim at **automating a set of tasks** e.g. classification, recommendation, translation and text and image generation.

- The automation of these tasks is referred to as **Machine Learning (ML).**



Bender, Emily M., and Alex Hanna. *The AI Con: How to fight big tech's hype and create the future we want*. Random House, 2025, pages 5-7

# Machine Learning

## Definition

- **Machine Learning:** is the field of building computer programs that pick up a set of patterns from training data and generalise to new data to automate specific tasks without explicit instructions.

- **For example:**

  - Task: To build a Psychotherapist application (call it Eliza  ). It is an algorithm that reads a text from the user and respond accordingly.



Bender, Emily M., and Alex Hanna. *The AI Con: How to fight big tech's hype and create the future we want*. Random House, 2025, pages 5-7

# Machine Learning

## How does it work?

- Task: To build a Psychotherapist application (call it Eliza). It is an algorithm that reads a text from the user and respond accordingly.

Instructions

Eliza

```
Function detect_sentiment (sentence):
    Emotion_response = “”
    If sentence.contains(“sad”, “angry”, “afraid”):
        Emotion_response = “I’m sorry you are feeling this way. Did
        something happen?”

    If sentence.contains(“happy”, “pleased”, “safe”):
        Emotion_response = “Great!”
    Print (Emotion_response)
```

# Machine Learning

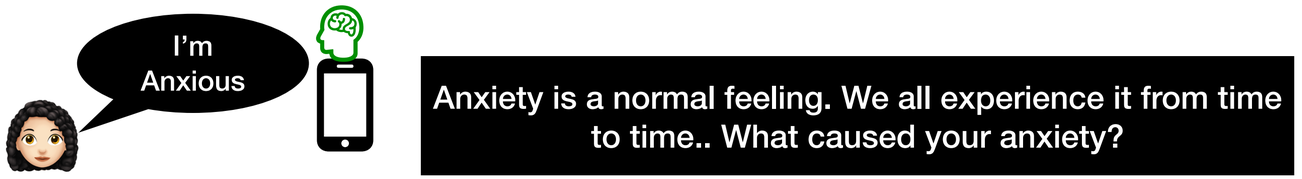
## How does it work?

- Task: To build a Psychotherapist application (call it Eliza). It is an algorithm that reads a text from the user and respond accordingly.

**No-Instructions Machine Learning**

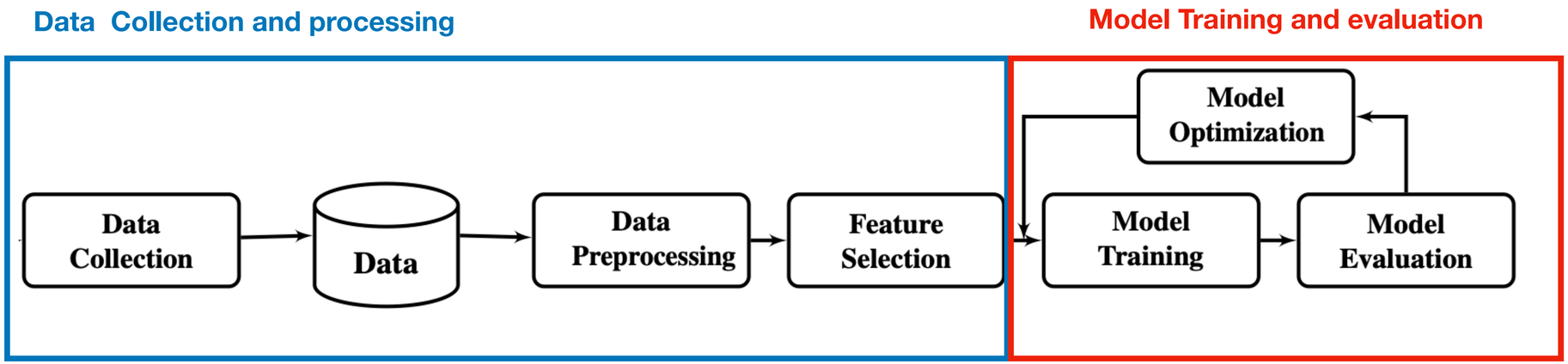
Eliza

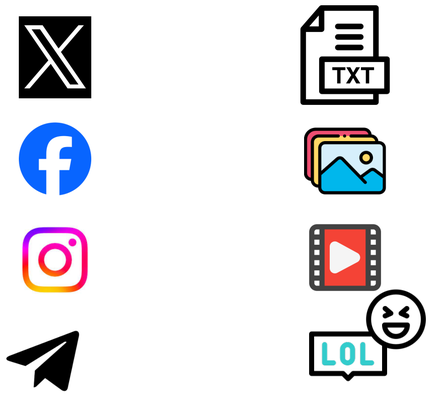
(Psychiatrist transcripts) + ->Pattern



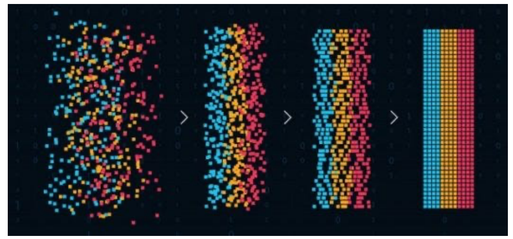
# Machine Learning

## Building Blocks



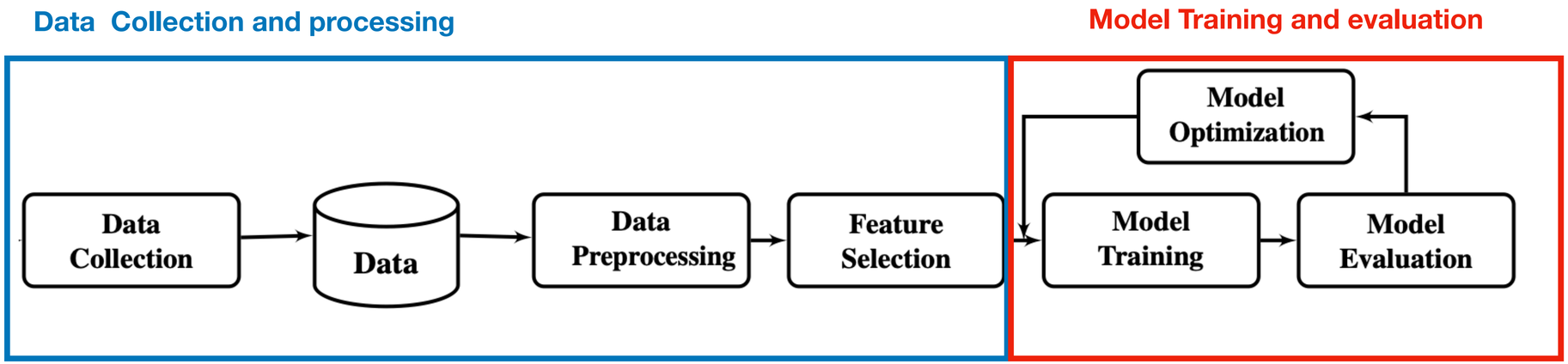







# Machine Learning

## Building Blocks: Your Turn




- **The Job centre at TU decides to build a ML tool to give students an indiction of their potential to find a job with in 6 months of graduation.**
    - We want to know *how fast a graduate from TU-CS would find a job within 6 months of their graduation?*

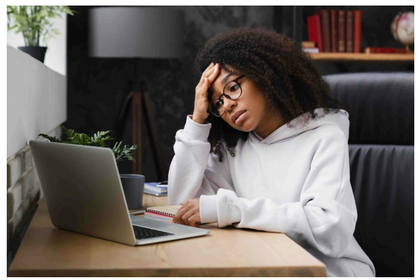
# Machine Learning

## TU-CS employability predictor

Fast Employment Predictor

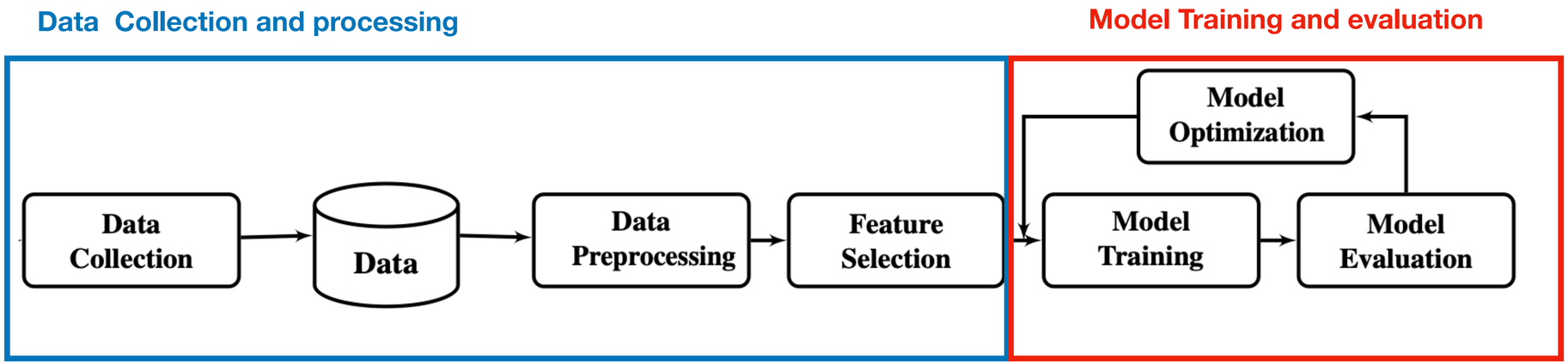

Yes! Yay :)

No! :(



# Machine Learning

## Building Blocks: Your Turn

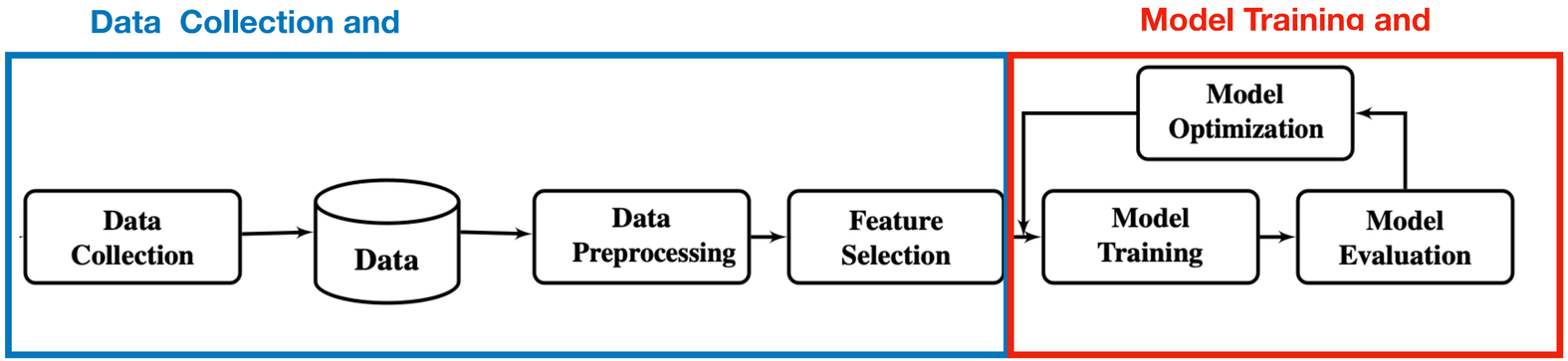



- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
  - **What are the steps required to build this ML model?**
  - **Hint:** Use the flowchart above.

# Machine Learning

## Building Blocks: Your Turn

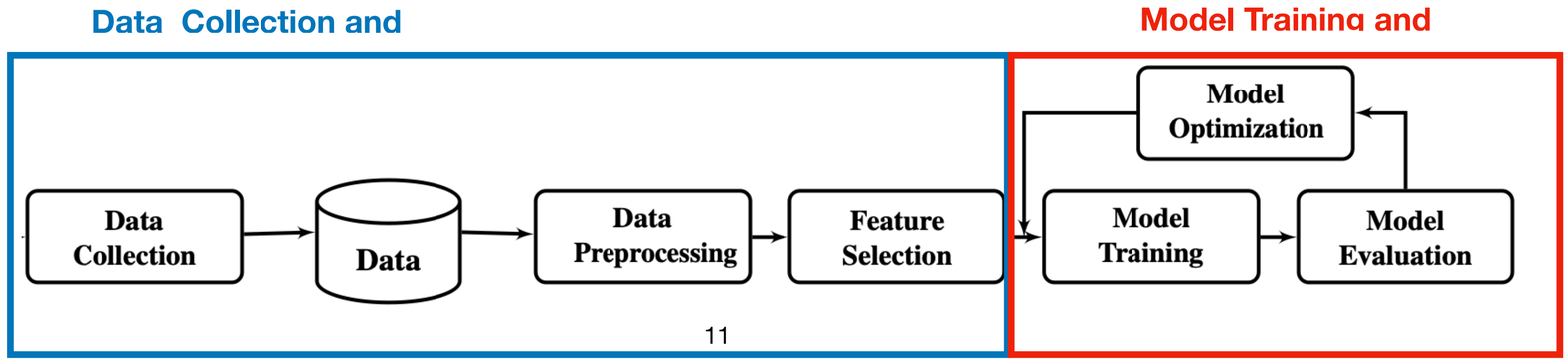
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 1- Data Collection:** What data to collect?



# Machine Learning

## Building Blocks: Your Turn

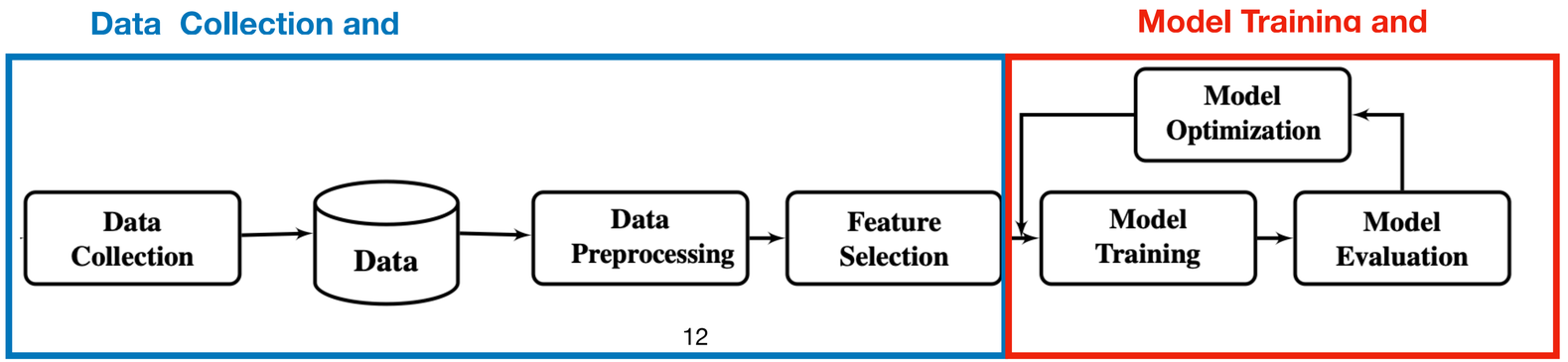
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 1- Data Collection:** What data to collect?
        - A survey to collect information about graduates from TU-CS in the last 10 years.
        - Name, gender, image, courses, grades, internship experiences, languages, graduation projects, jobs, how fast they found the job.



# Machine Learning

## Building Blocks: Your Turn

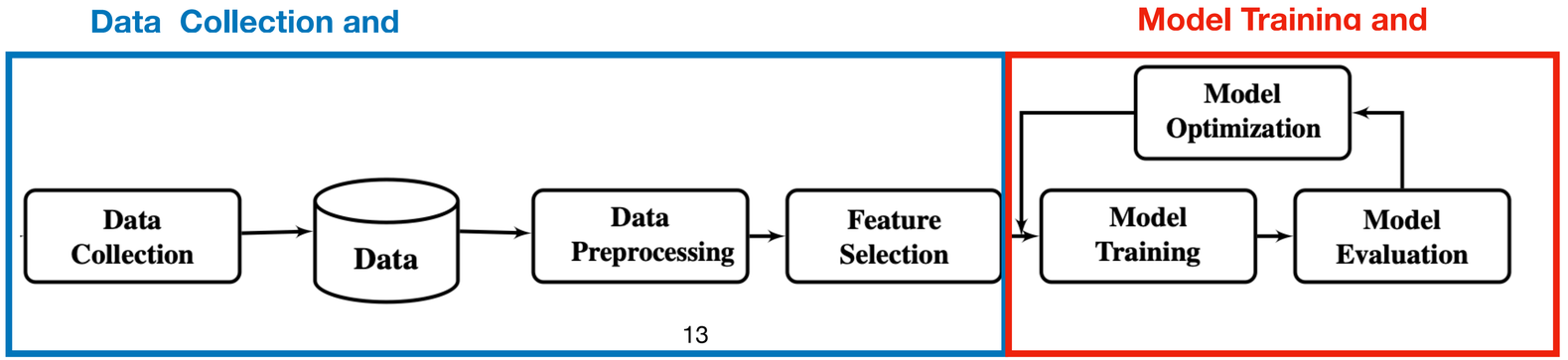
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 2- Data Pre-processing:** What should be filtered out/added?



# Machine Learning

## Building Blocks: Your Turn

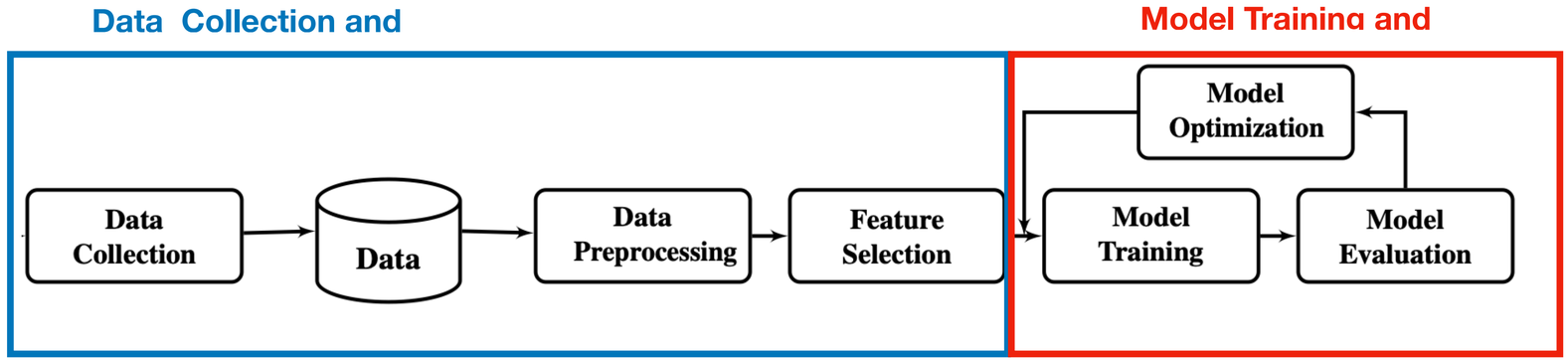
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 2- Data Pre-processing:** What should be filtered out/added?
        - Repeated entries.
        - Incomplete surveys.
        - Spam: some people entered wrong information.



# Machine Learning

## Building Blocks: Your Turn

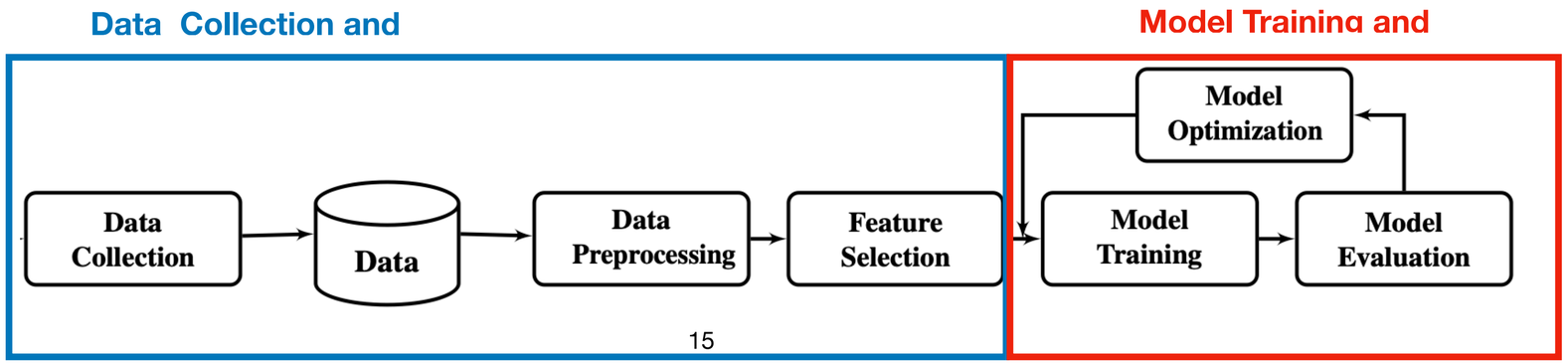
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 3- Feature Selection:** What important information in the collected data could be useful in predicting a graduate's employability?



# Machine Learning

## Building Blocks: Your Turn

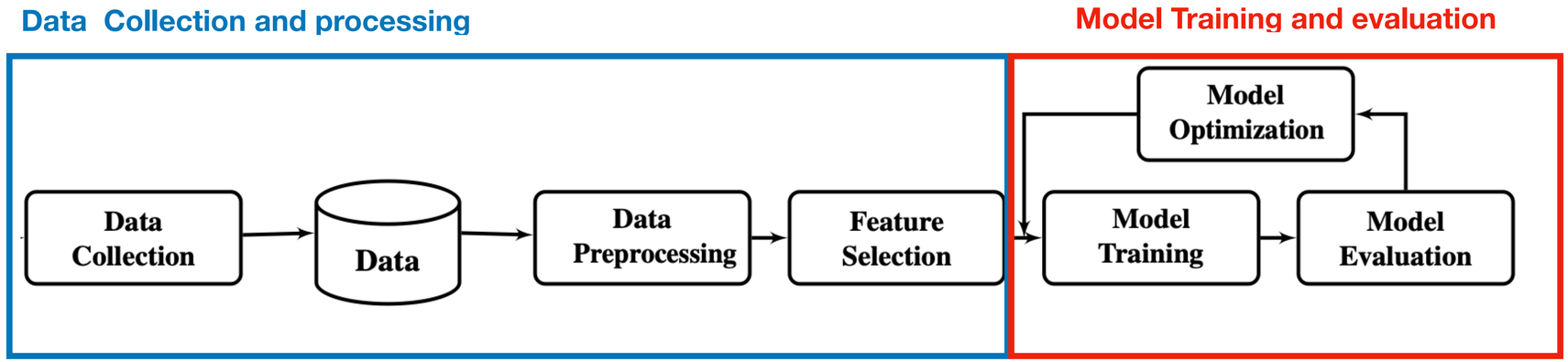
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 3- Feature Selection:** What important information in the collected data could be useful in predicting a graduate's employability?
        - Jobs, how fast they found their first job, courses, grades, internships, languages, names.



# Machine Learning

## Building Blocks: Your Turn

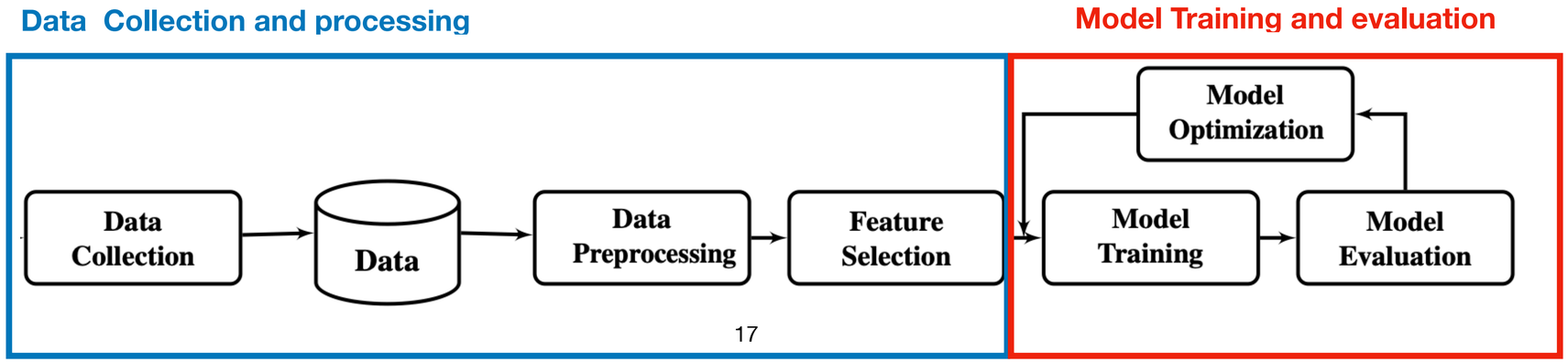
- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
    - **Step 4- Model training and evaluation:** Which ML model to choose and how to evaluate the our predator's performance?



# Machine Learning

## Building Blocks: Your Turn

- We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?*
  - **Step 4- Model training and evaluation:** Which ML model to choose and how to evaluate the our predator's performance?
    - A GPT, liner regression or a neural network model.
    - Use accuracy to evaluate the model performance.

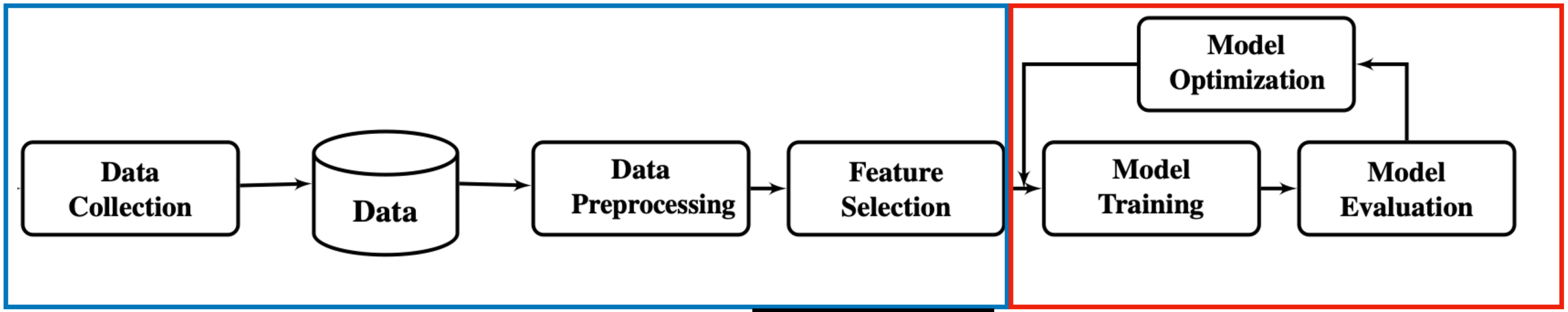


# Machine Learning

## Building Blocks

Data Collection and processing

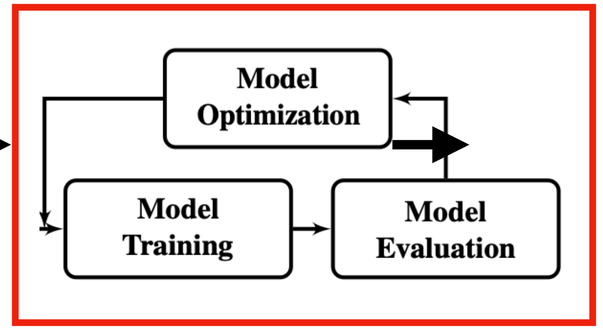
Model Training and evaluation




**Data:**
Collect data about Graduates from TU-CS in the last 10 years

**Pre-processing:**
- Remove repeated entries.
- fill in the incomplete information.
- Remove spams.

**Features:**
- Courses
- Grades
- Internship experience
- How long after graduation they got their first job.
- Who was their first employer.




Fast Employment Predictor

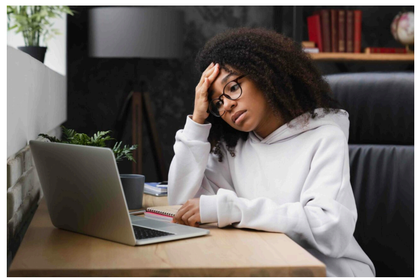
# Machine Learning

## Inequality

Fast Employment Predictor


Yes! Yay :)

No! :(



# Machine Learning

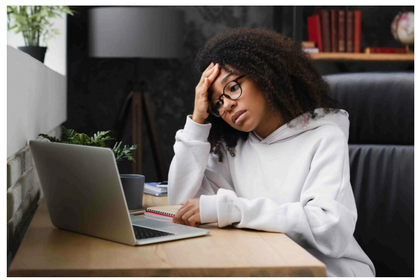
## Inequality

the model tends to predict more positive outcome to men compared to women and non-binary people.

Fast Employment Predictor


Yes! Yay :)

No! :(



# Machine Learning

## Inequality

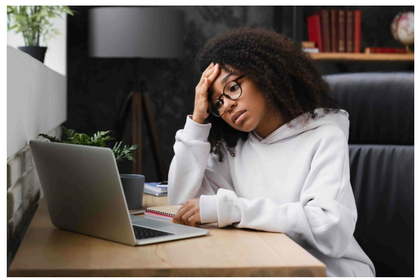
**Why do you think might be the case?**

the model tends to predict more positive outcome to men compared to women and non-binary people.

Fast Employment Predictor


Yes! Yay :)

No! :(



# Machine Learning

## Inequality

**Why do you think might be the case?** **Bias**

the model tends to predict more positive outcome to men compared to women and non-binary people.

Fast Employment Predictor


Yes! Yay :)

No! :(

# Socio-technical Bias

## Definition

- **Socio-technical bias:** “A systematic **divergence** between the **data** and the **phenomenon** that is supposed to be depicted due to **structural inequalities**”[1].

[1] Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review 2021, 10, 1–29.

# Socio-technical Bias

## Definition

- **Socio-technical bias:** “A systematic **divergence** between the **data** and the **phenomenon** that is supposed to be depicted due to **structural inequalities**” [1].
    - **For example:** We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?* (**phenomenon**)

[1] Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review 2021, 10, 1–29.

# Socio-technical Bias

## Definition

- **Socio-technical bias:** “A systematic **divergence** between the **data** and the **phenomenon** that is supposed to be depicted due to **structural inequalities**” [1].
    - **For example:** We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?* (**phenomenon**)
        - We collect information from the last 10 years about the graduates of TU-CS and detailed information like their grades, the courses they chose, their internship experience, their language skills (**data**).

[1] Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review 2021, 10, 1–29.

# Socio-technical Bias

## Definition

- **Socio-technical bias: “**A systematic **divergence** between the **data** and the **phenomenon** that is supposed to be depicted due to **structural inequalities**”[1]**.**
    - **For example:** We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?* (**phenomenon**)
        - We train a ML model and when we test it. We find that the model tends to predict more positive outcome to men compared to women and non-binary people. **There are different outcomes based on the gender.**
        - Gender was not one of the features we used to train the model. **Gender should have noting to do with a person’s employability.**
        - But when we look again into the collected data, we find **over representation of men**. (**divergence**).

[1] Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review 2021, 10, 1–29.

# Socio-technical Bias

## Definition

- **Socio-technical bias:** “A systematic **divergence** between the **data** and the **phenomenon** that is supposed to be depicted due to **structural inequalities**”[1]**.**
    - **For example:** We want to know *how fast a graduate from CS-TU would find a job within 6 months of their graduation?* (**phenomenon**)
        - We train a ML model and when we test it. We find that the model tends to predict more positive outcome to men compared to women and non-binary people. There are different outcomes based on the gender.
        - Gender was not one of the features we used to train the model. Gender should have noting to do with a person’s employability.
        - But when we look again into the collected data, we find over representation of men. (**divergence**). **Why do we have this divergence?**

[1] Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review 2021, 10, 1–29.

[1] Lopez, P. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review 2021, 10, 1–29.

# Socio-technical Bias

## Definition

- **Socio-technical bias:** “A systematic **divergence** between the **data** and the **phenomenon** that is supposed to be depicted due to **structural inequalities**” [1].
    - This is **the divergence between** the **phenomenon** which is how fast TU-CS graduates find a job **and** the **data** collected which is over represent men **due to structural inequality** in the access of STEM education to women.

# Socio-technical Bias

## Types

- **Under-representation:** The under-representation of certain identity groups in the collected data to train a ML models. For example, women and non-binary in STEM.
- **Over-representation:** The over-representation of certain identity groups in the collected data. For example, Black people in prisons in the US.
- **Negative stereotyping:** Associating certain identity groups with a negative stereotype in the dataset. For example, Jews control the world or Muslims are terrorists.

Gallegos, Isabel O., et al. "Bias and fairness in large language models: A survey." *Computational Linguistics* 50.3 (2024): 1097-1179.

# Socio-technical Bias

## Negative stereotyping: Example

- Negative stereotype against refugees in German language models.



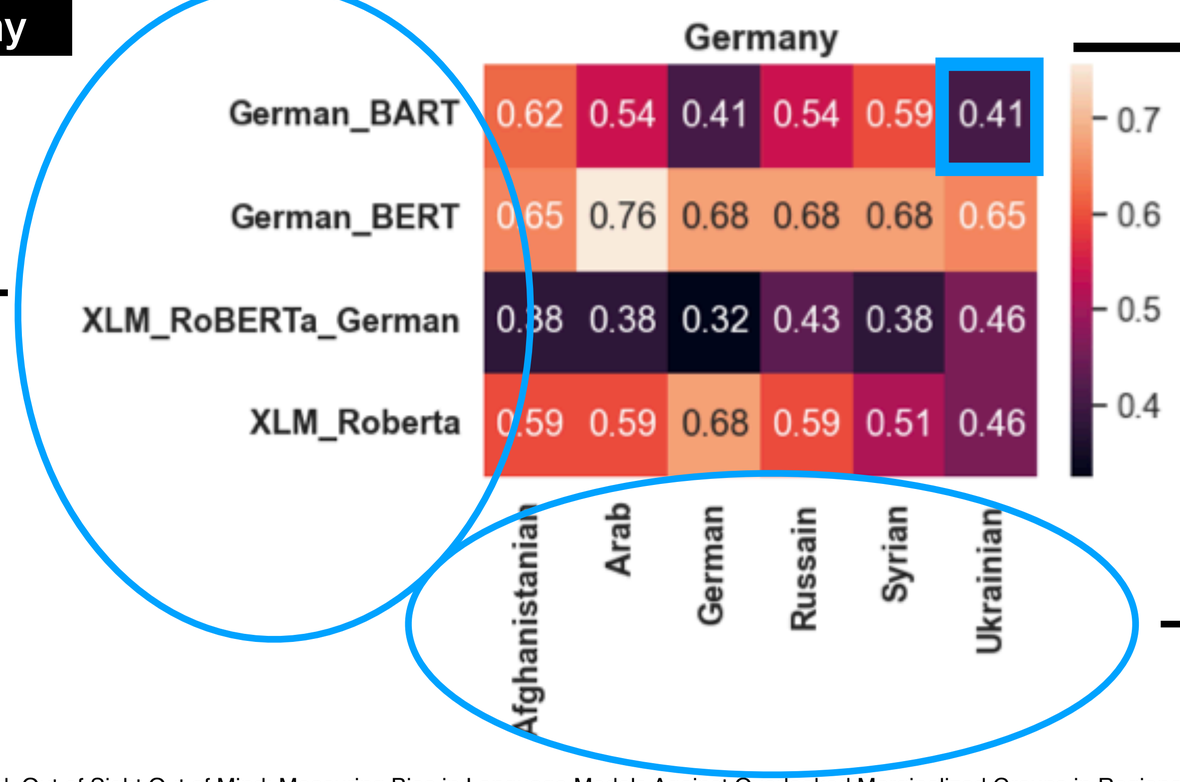
Elsafoury, Fatma, and David Hartmann. "Out of Sight Out of Mind, Out of Sight Out of Mind: Measuring Bias in Language Models Against Overlooked Marginalized Groups in Regional Contexts." *arXiv preprint arXiv:2504.12767* (2025).

# Socio-technical Bias

## Negative stereotyping

Negative stereotype against refugees in German language models.

**Heatmap of the bias scores against refugees in Germany**




**LLMs trained on German news articles**

**0.41**

**Bias score between 0.5 and 1. The higher the score the more biased the model is**

**Refugee groups in Germany**

Elsafoury, Fatma, and David Hartmann. "Out of Sight Out of Mind, Out of Sight Out of Mind: Measuring Bias in Language Models Against Overlooked Marginalized Groups in Regional Contexts." *arXiv preprint arXiv:2504.12767* (2025).

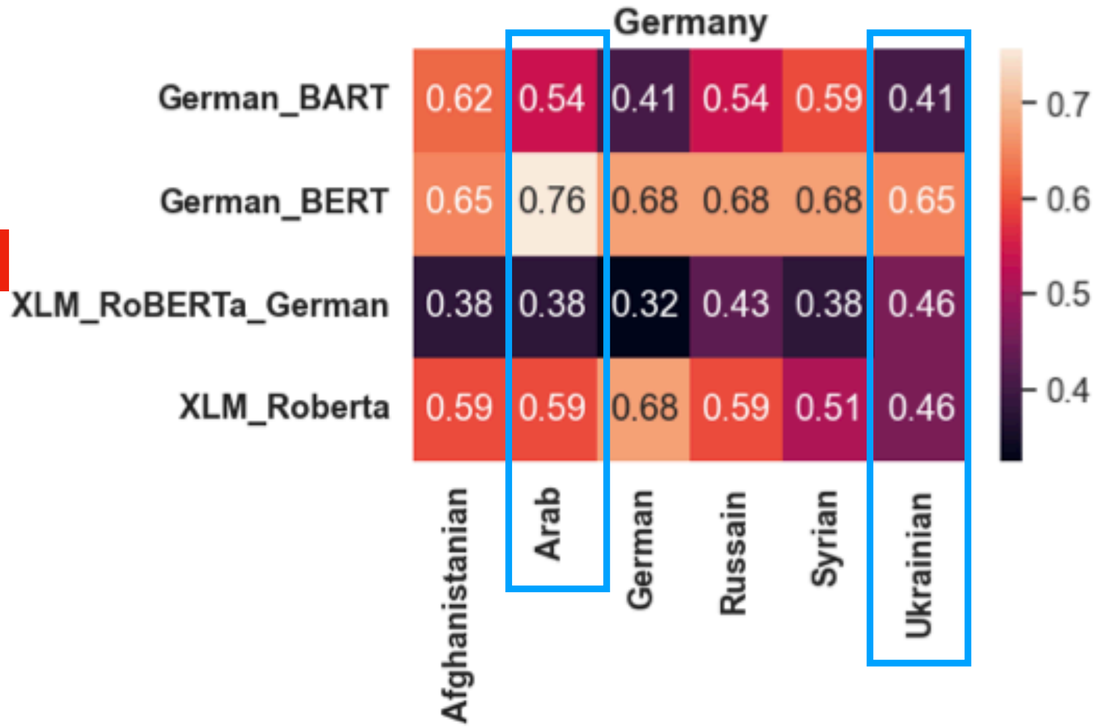
# Socio-technical Bias

## InEquality: Bias, fairness, stereotyping

Negative stereotype against refugees in German language models.

**Heatmap of the SOS bias scores against refugees in Germany**

Why are the German LMs more biased against Arabs?




Elsafoury, Fatma, and David Hartmann. "Out of Sight Out of Mind, Out of Sight Out of Mind: Measuring Bias in Language Models Against Overlooked Marginalized Groups in Regional Contexts." *arXiv preprint arXiv:2504.12767* (2025).

# Socio-technical Bias

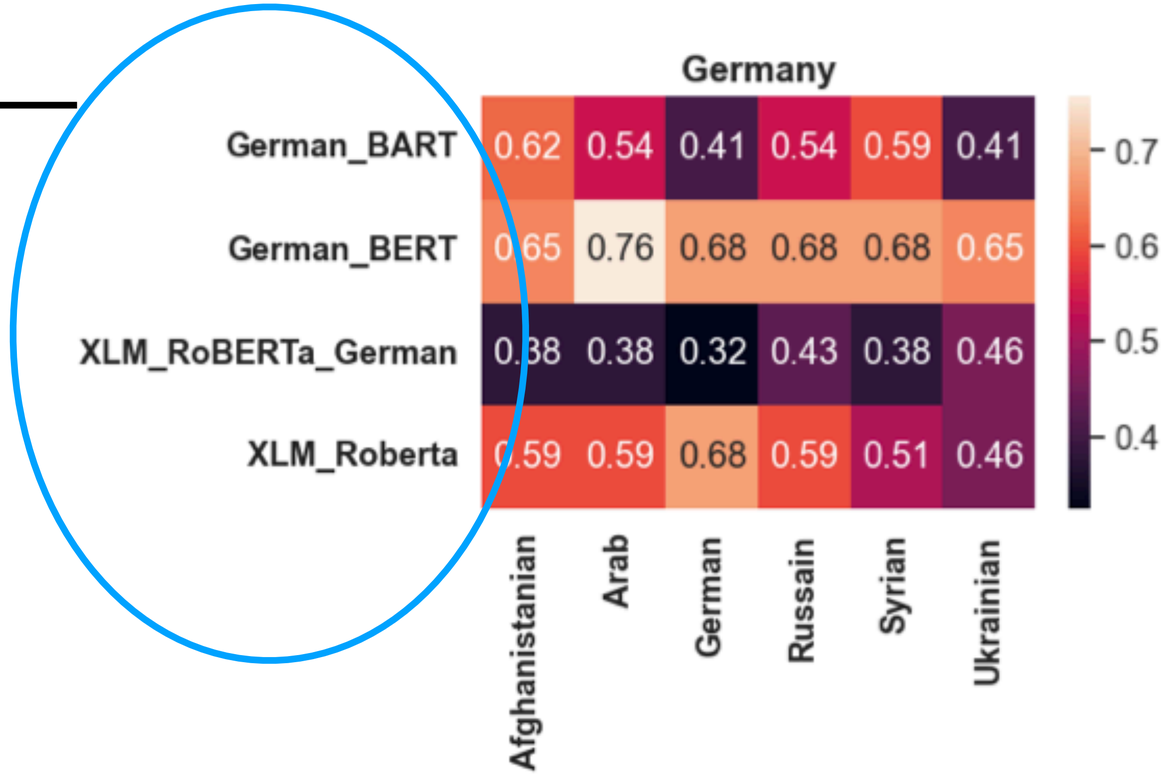
## Negative stereotyping

Negative stereotype against refugees in German language models.

**LLMs trained on German news articles**

**Heatmap of the SOS bias scores against refugees in Germany**

Why are the German LMs more biased against Arabs?




Esposito argues that this differential treatment of Non-White refugees in Germany and the EU, stems from: islamophobia, othering and racial prejudice which impacts the media coverage of Africans and Middles Easterners [2].

[1] Risks of discrimination for refugees in Germany. https://www.antidiskriminierungsstelle.de/SharedDocs/forschungsprojekte/E
[2] Esposito, A. The Limitations of Humanity: Differential Refugee Treatment in the EU. https://hir.harvard.edu/the-limitations-of-humanity-differential-refugee-treatment-in-the-eu/, 2022.

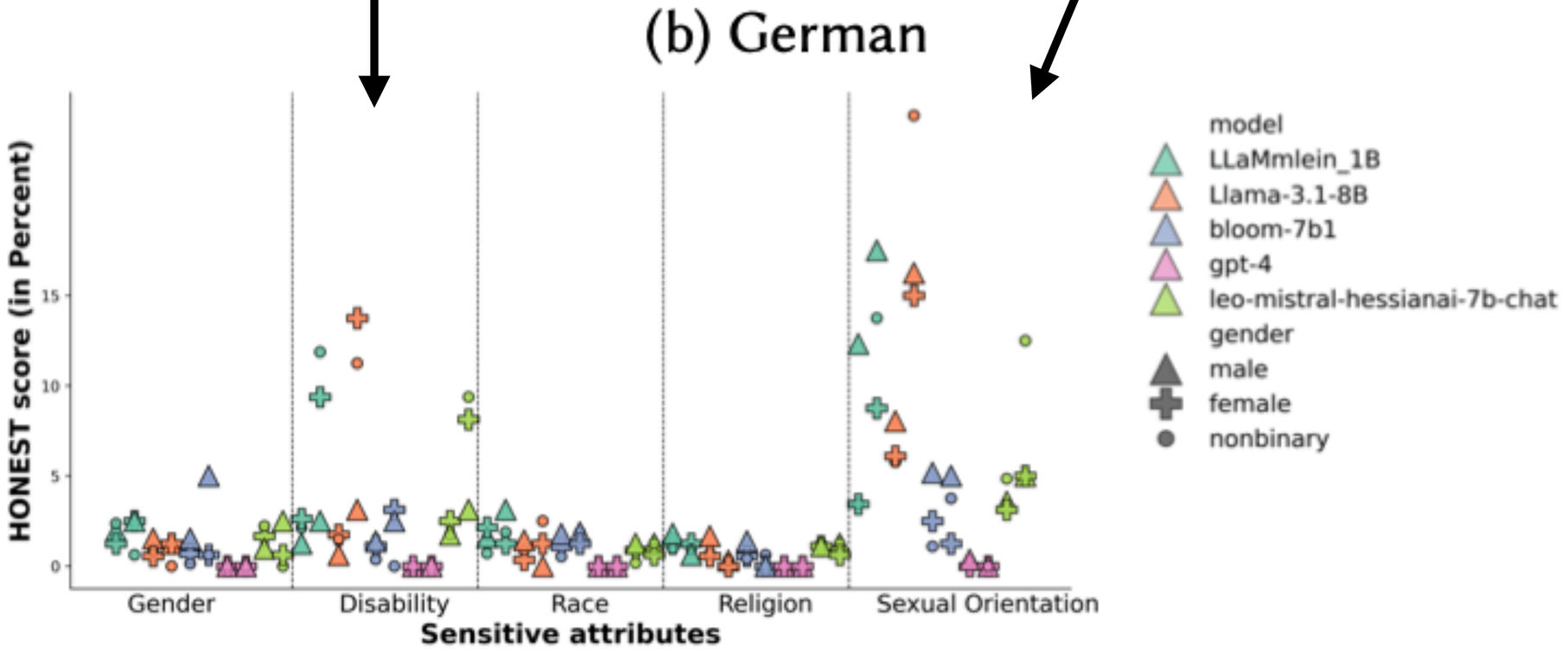
# Socio-technical Bias

## Intersectionality of Bias

Higher bias against women and non-binary when they have a physical disability.

Higher bias against non-binary and men regarding their sexual orientation

Bias scores in different language models for different sensitive attributes and genders



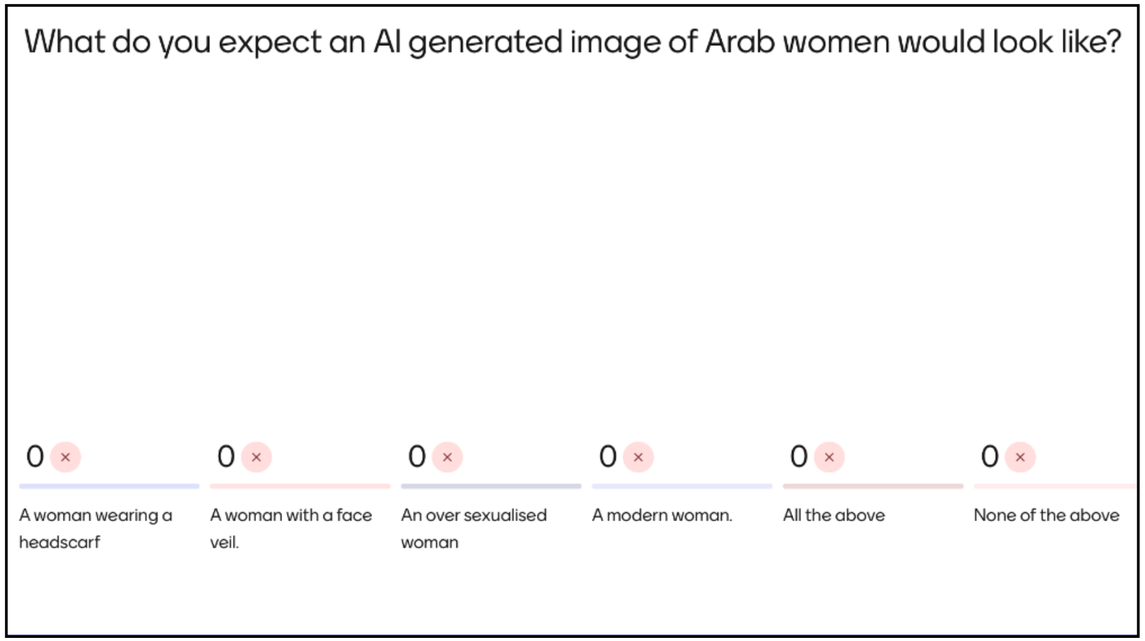

Elsafoury, Fatma, and David Hartmann. "Out of S.ight Out of Mind, Out of Sight Out of Mind: Measuring Bias in Language Models Against Overlooked Marginalized Groups in Regional Contexts." *arXiv preprint arXiv:2506.12767* (2025).

# Socio-technical Bias

## Reflect

- An 12 years old student has a homework to write an essay about Arab women. The first thing the student does is to ask DALL-E to generate an image of an Arab woman.

- What do you think the model would generate?



What do you expect an AI generated image of Arab women would look like?

| 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|
| A woman wearing a headscarf | A woman with a face veil. | An over sexualised woman | A modern woman. | All the above | None of the above |

# Socio-technical Bias

## Stereotyping

- An 12 years old student has a homework to write an essay about Arab women. The first thing the student does is to ask DALL-E to generate an image of an Arab woman.

DALLE-3




https://creator.nightcafe.studio/dall-e-ai-image-generator

# Socio-technical Bias

## Stereotyping

Why is this the case?

- An 12 years old student has a homework to write an essay about Arab women. The first thing the student does is to ask DALL-E to generate an image of an Arab woman.

DALLE-3




https://creator.nightcafe.studio/dall-e-ai-image-generator

# Socio-technical Bias

## Stereotyping

- An 12 years old student has a homework to write an essay about Arab women. The first thing the student does is to ask DALL-E to generate an image of an Arab woman.

**Why is this the case?**

- The data used to train DALL-E depicts Arab and Egyptian women in this way rather than reflect the reality of how Arab women.
- This depiction fulfils a stereotype of Arab women resulting from stories like a thousand night and night.

https://creator.nightcafe.studio/dall-e-ai-image-generator

# Fairness in AI

## Definition

- Fairness is related to the application of ML systems in real life.
- **Unfairness of ML models:** *“different outcome of the AI algorithm for different identity groups.”*
- In other words, the **bias** in the **ML** models leads to **unfair decisions** made by the ML model for people from **marginalised background**.

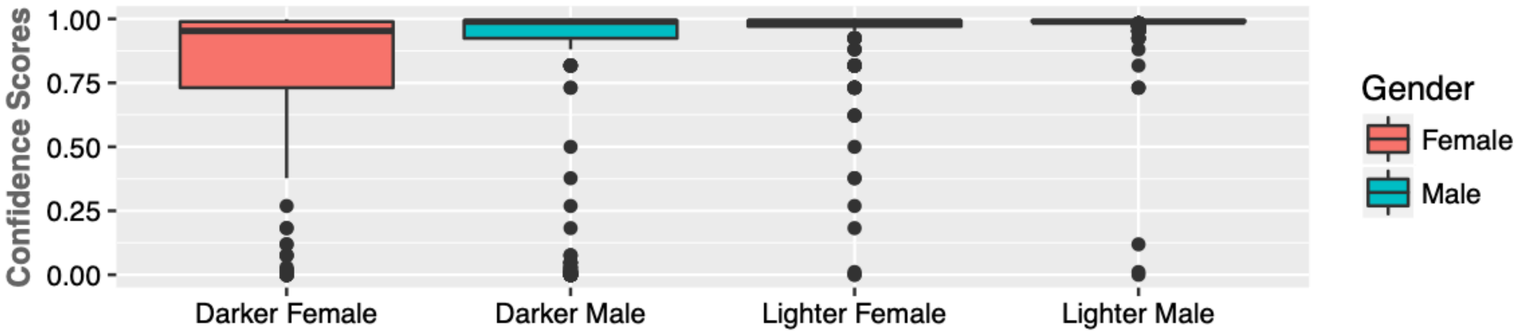

Elsafoury, Fatma, and David Hartmann. "Out of Sight Out of Mind, Out of Sight Out of Mind: Measuring Bias in Language Models Against Overlooked Marginalized Groups in Regional Contexts." *arXiv preprint arXiv:2504.12767* (2025).

# Fairness

## Example

**In a famous Gender Shades paper, the authors show that automated facial analysis algorithms perform differently for different skin shade and gender [1].**

Figure: Gender classification confidence scores from IBM facial analysis systems




Why is this case?

IBM facial analysis systems perform better for people with lighter skin shads.

[1] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. PMLR, 2018.

# Fairness

## What is Fairness?

**In a famous Gender Shades paper, the authors show that automated facial analysis algorithms perform differently for different skin shade and gender [1].**

Why is this case?

Table: the distribution of lighter and darker-skinned subjects in different images datasets used to train and test the automated Facial analysis algorithms.

| Dataset | Lighter (I,II,III) | | Darker (IV, V, VI) | | Total |
|---|---|---|---|---|---|
| PPB | 53.6% | 681 | 46.4% | 589 | 1270 |
| IJB-A | 79.6% | 398 | 20.4% | 102 | 500 |
| Adience | 86.2% | 1892 | 13.8% | 302 | 2194 |

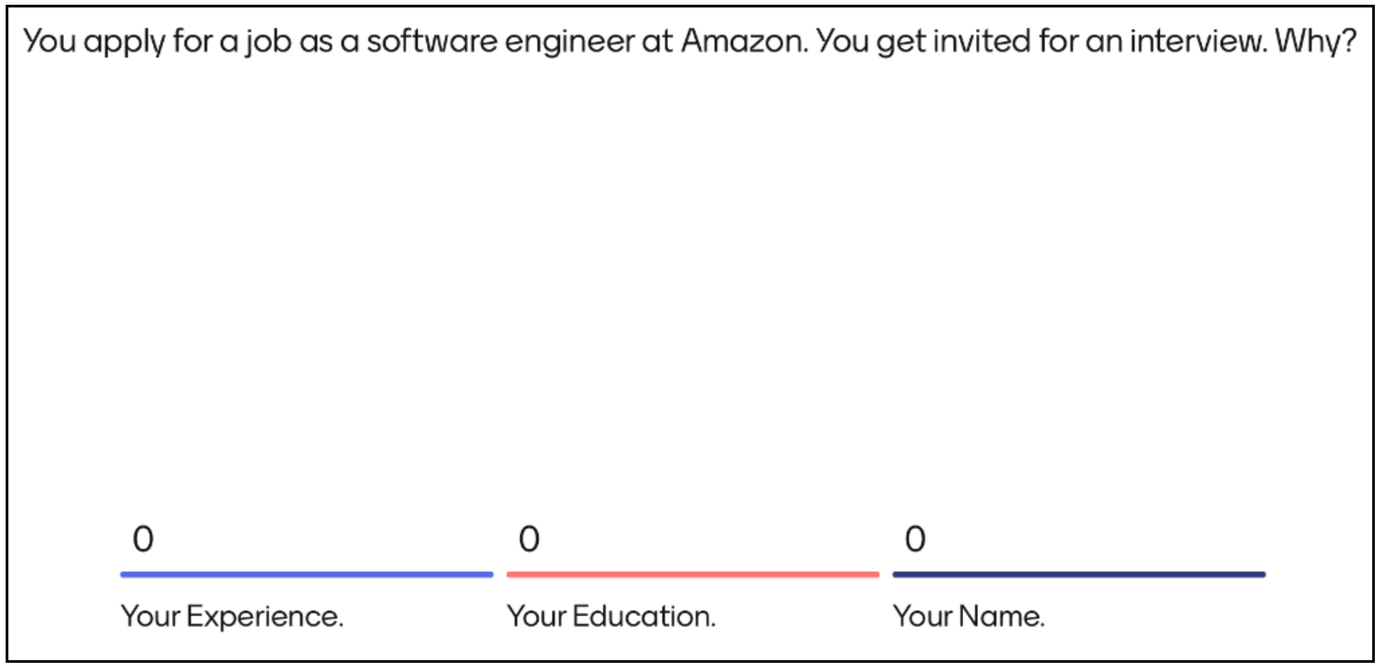
Because there's over-representation of people with lighter skin shade in the training datasets.

[1] Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. PMLR, 2018.

# Fairness

## Reflect

- You applied for a job as a software engineer at Amazon and you get an invitation to an interview. What made you qualified for the interview?



You apply for a job as a software engineer at Amazon. You get invited for an interview. Why?

# Fairness

## Examples

Reuters 2018

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By **Jeffrey Dastin**

October 11, 2018 2:50 AM GMT+2 · Updated October 10, 2018

Aa

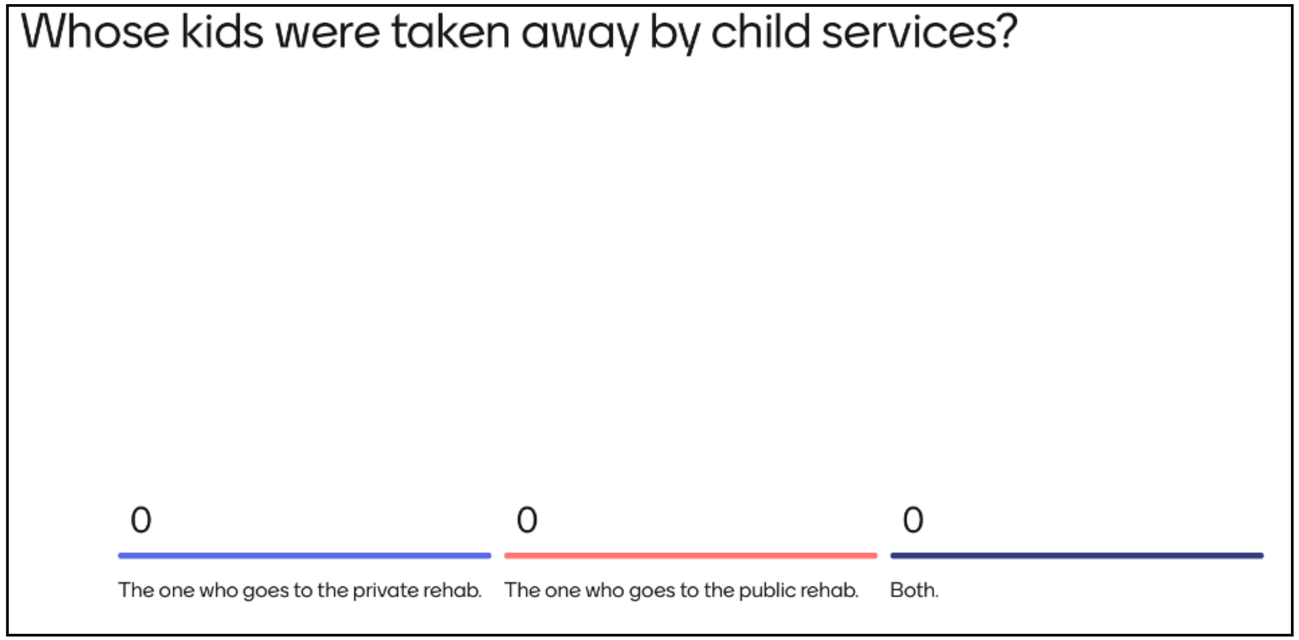
Why is this the case?

https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

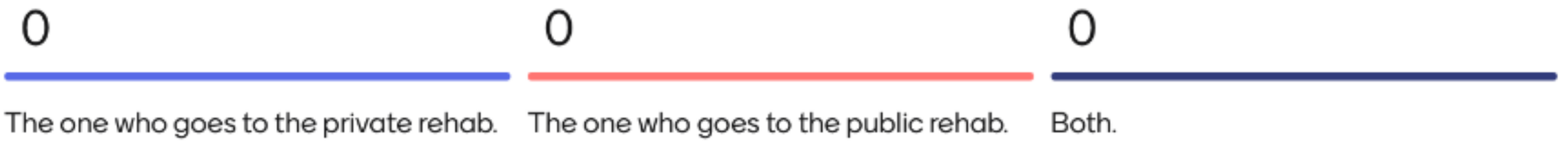
# Fairness

## Reflect

- Two drinking buddies. Both are parents. They decided to get help and go to rehab. One went to a private rehab and another went to a public rehab. Months later, child protection services take the kids away. Whose kids were taken away?



# Whose kids were taken away by child services?



# Fairness

## Reflect

- Two drinking buddies. Both are parents. They decided to get help and go to rehab. One went to a private rehab and another went to a public rehab. Months later, one of them found the child protection services take their kids away.Whose kids were taken away? **The one who went to the public rehab.**

Why is this the case?

# Fairness

## Privilege

Why is this the case?

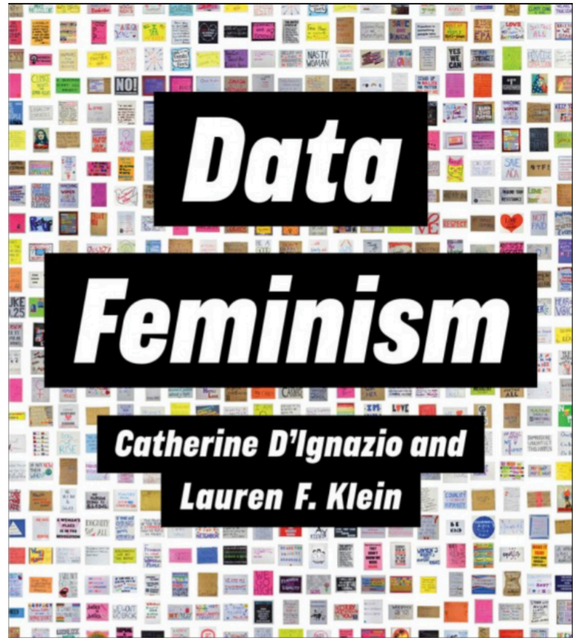
- In her book **“automating inequalities, Virginia Eubanks** documents that poor, Black and indigenous people were at higher risk to lose their children du to suspected neglect.,

- This happens because **poor parents** use **public** resources to seek help with mental health, drug or alcohol problems, making them **over represented** compared to **rich parents** that can afford **private services** for the same problems but remain **invisible to the “AI”** that **ultimately punish poor parents** by taking their kids away.

# Fairness

## Privilege is Power

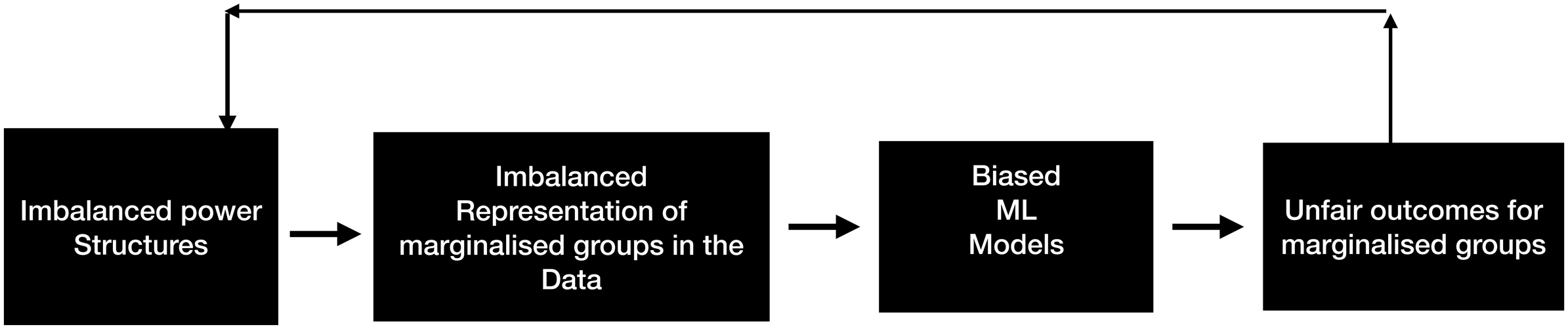
- But who collects the data? Who has the resources to collect and store it? Data is expensive.
- Corporations, Governments and Universities.
- The majority of the people who control these institutes are **White, Cisgender, heterosexual, Neurotypical, able-bodied Men**. For example:
    - Board of directors in companies: Google (82% White Men), Meta (78% Men, 89% White).
    - The German Bundestag members are 32.4% Women, 36.5% under 35 and 11.6% from migrant background.



# Fairness

## Privilege is Power

- The existing power structures in our society lead collecting datasets with imbalanced representation of marginalised groups which leads to biased ML models which result in unfair outcomes to marginalised groups which eventually reinforces the power structures in our societies.



# Bias, Fairness and Power

## Take Away Messages

- AI and ML models depend on the data it is trained on.
- Biased data will lead to biased models.
- Biased models will give unfair outcomes/ make unfair decision to different groups of people due to people's attributes like gender, race,.etc…
- The Biased data is a result of privilege and power in our societies.

# Thanks for listening!

## Questions?

**Fatma Elsafoury 17.12.2025**