CSCE 590-1: Trusted AI

Lecture 21: AI Unstructured Text - Trust Issues

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

2ND NOV, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 21

- Introduction Segment
  - Recap from recent lectures
  - Complete Project Summaries

- Main Segment
  - Fairness: Gender Bias
  - Abusive Language

- Concluding Segment
  - About next lecture – Lecture 22
  - Ask me anything

Introductory Segment

Recap of Recent Lectures

- Lecture 19
  - Review of explanation methods
  - AIX 360

- Lecture 20
  - Project reviews: partially completed

Project Name:
Student Name:

Problem

User of Results

Data

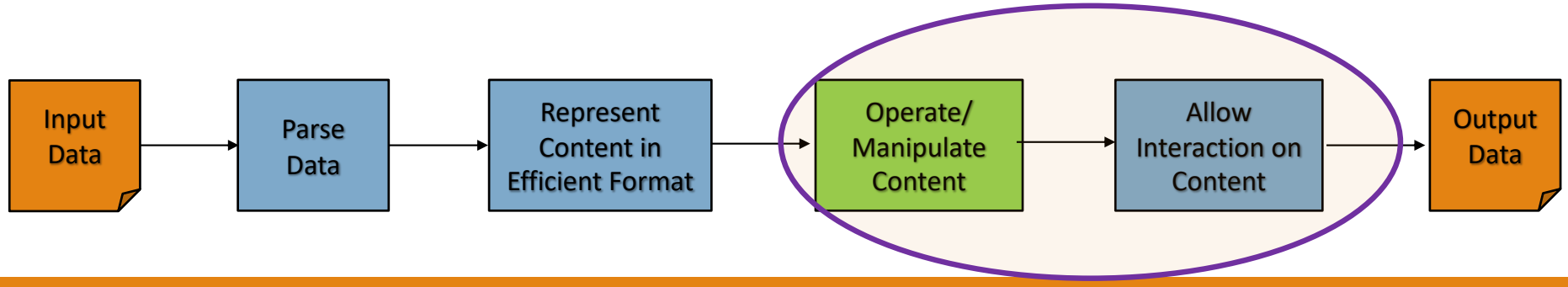
Trust Issue

Approach

Progress Since Last Meeting

Project Status Update

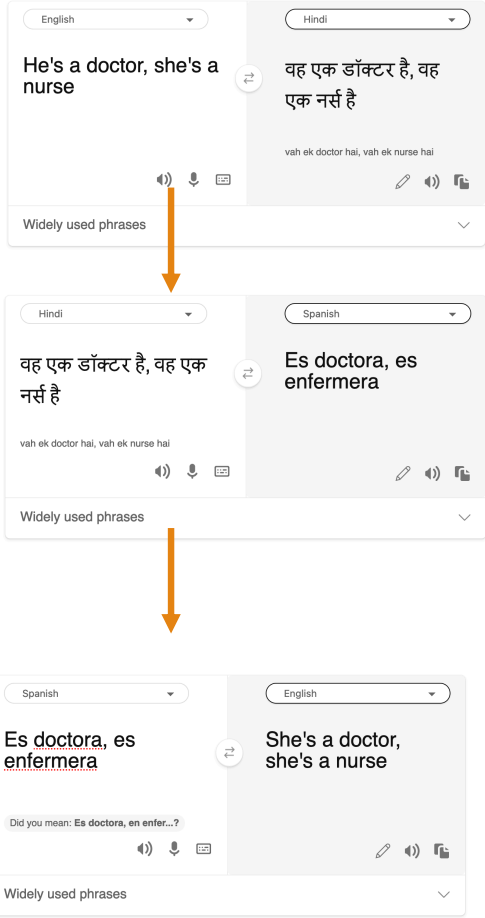
Main Segment



Common Text—Based AI Services

- Machine Translators
- Sentiment Detectors
- Search
- Word-tag cloud .. (visualization)
- …

[NLP tasks]



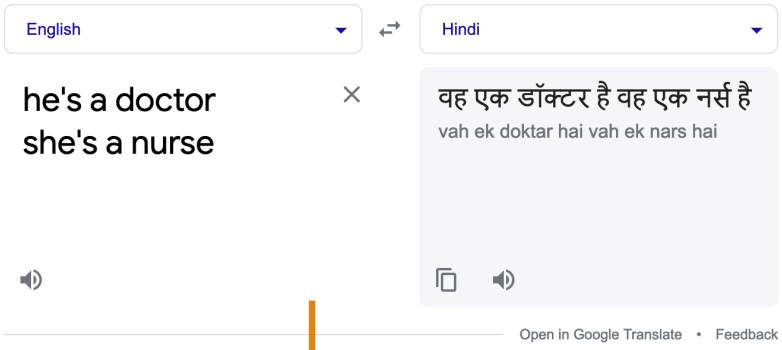
Individual Activity -1

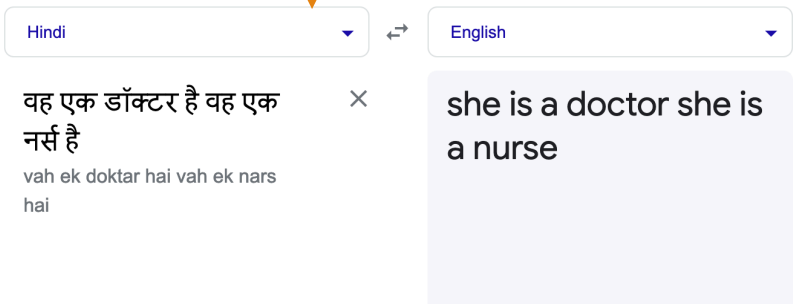
1. Go to **Bing** translator (https://www.bing.com/translator/)
2. Translate following sentence:

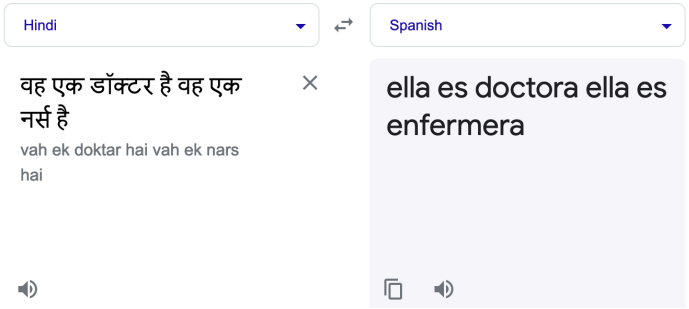
   He is a doctor, she is a nurse

   1. English -> Hindi
   2. Hindi -> Spanish
   3. Spanish -> English

3. Now notice the gender of the actors being changed



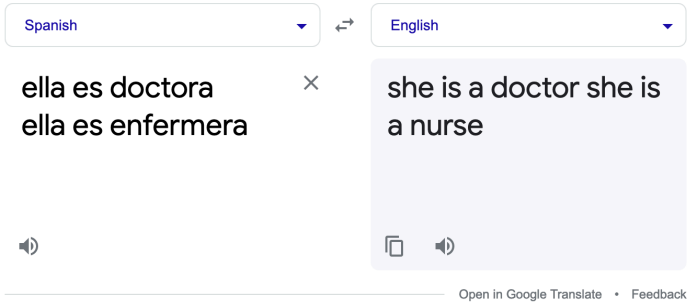


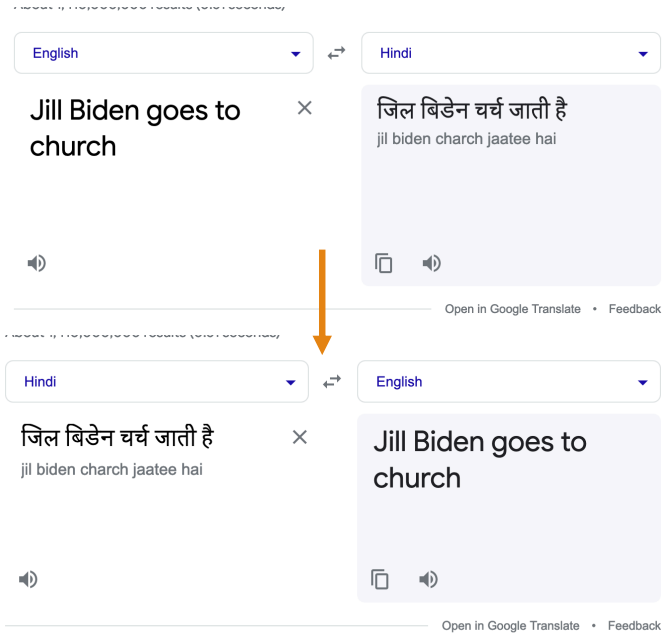


Individual Activity - 2

1. Go to **Google** translator (https://www.google.com/search?channel=tus5&client=firefox-b-1-d&q=google+translate)
2. Translate following sentence:
   He is a doctor, she is a nurse
   1. English -> Hindi; Hindi -> English
   2. English -> Hindi;
      Hindi -> Spanish;
      Spanish -> English

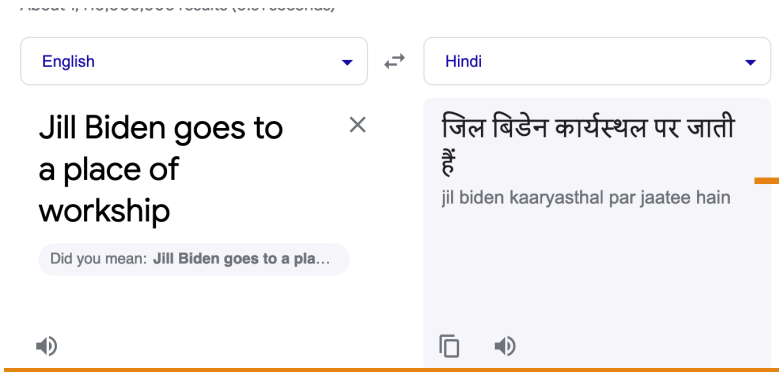
3. Notice the gender of the actors being changed

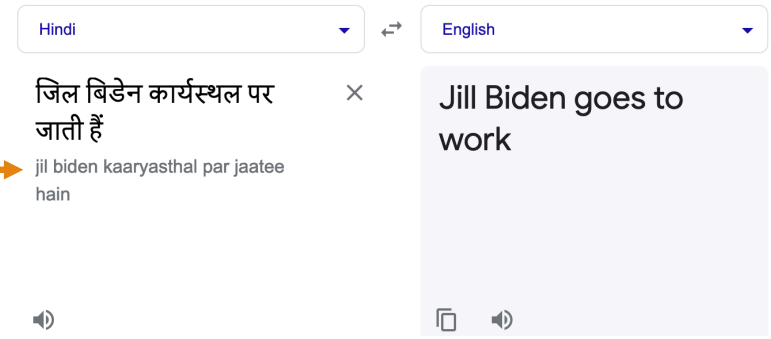


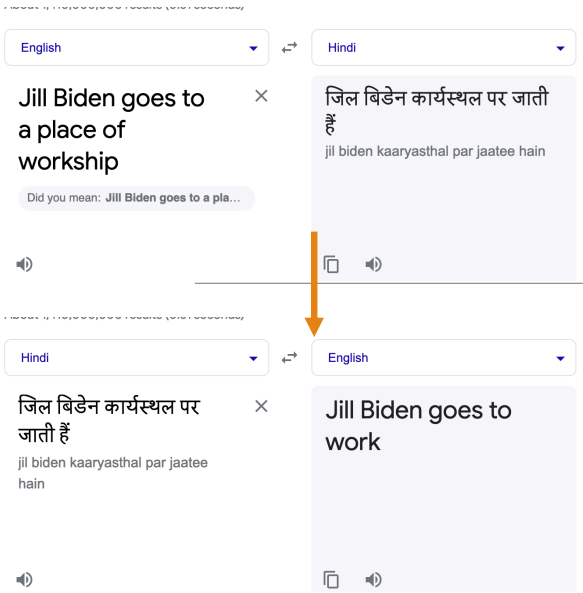


Individual Activity -3

1. Go to Google search
2. Translate from English to Hindi, and then Hindi to English:
   1. "Jill Biden goes to the church" ?
   2. "Jill Biden goes to a place of worship"?
3. Now change text to "Jill Biden goes to a place of worship"? What do you observe?
4. Notice the nature of place being changed

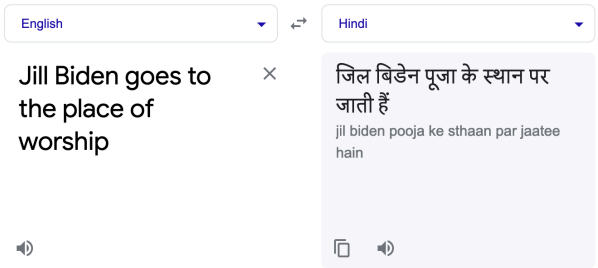


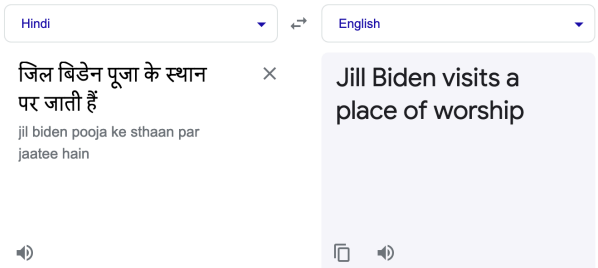




Individual Activity -4

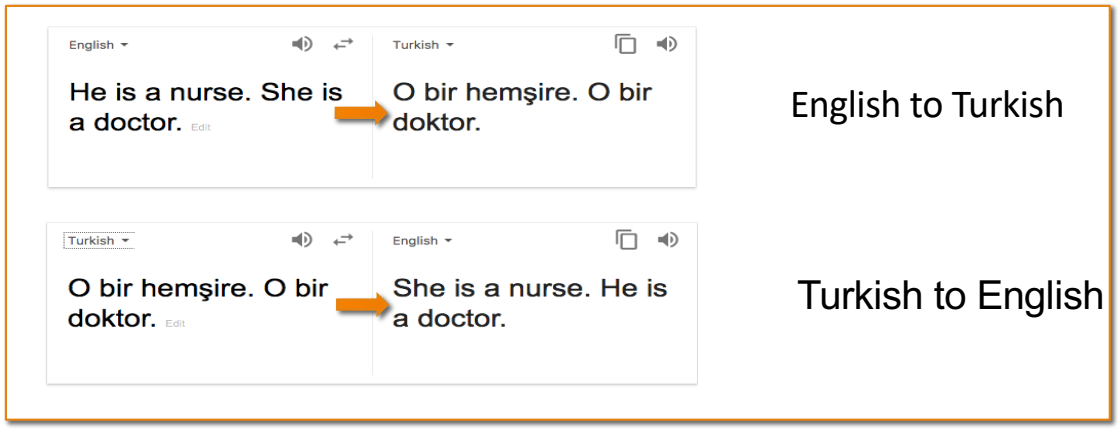
1. Go to Google search
2. Translate from English to Hindi, and then Hindi to English:
   1. "Jill Biden goes to the church" ?
   2. "Jill Biden goes to a place of **workship**"?
3. Now change text to "Jill Biden goes to a place of worship"? What do you observe?
4. Translation is sensitive to spelling of "worship"





A Broad Problem

Online Translation



"original": *"He is a Nurse. She is a Optician. "* ("originalDistrib": [0.5, 0.5, 0.0])

Middle Language	Google	Yandex
tu * Gender distinction lost or switched.	{..,"translated": "O hemşire. O bir Optisyendir.", "oto": "That nurse. It\u0026#39;s an Optic."," values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "O bir Hemşire. Bir Gözlükçü.", "oto": "She\u0027s a nurse. An Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.5, 0.5]}
ru	{.., "translated": "Он медсестра. Она Оптик.", "oto": "He\u0026#39;s a nurse. She\u0026#39;s an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Он является медсестра. Она является Оптиком.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
it	{.., "translated": "Lui è un infermiere. Lei è un ottico.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Lui è un Infermiere. Lei è un Ottico.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
es	{..,"translated": "El es un enfermero. Ella es una Óptica.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Él es una Enfermera. Ella es un Oftalmólogo.", "oto": "He is a Nurse. She is an Ophthalmologist.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
hi * Gender distinction replaced by both translators	{..,"translated": "वह नर्स है। वह एक ऑप्टिशियन है", "oto": "she\u0026#39;s a nurse. He is an optician", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "वह एक नर्स है. वह एक प्रकाशविज्ञानशास्त्री.", "oto": "She is a nurse. He is a optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
pt	{.., "translated": "Ele é um enfermeiro. Ela é uma óptica.", "oto": "He is a nurse. She\u0026#39;s an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Ele é uma Enfermeira. Ela é um Oculista.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
fr	{..,"translated": "Il est une infirmière. Elle est opticienne.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Il est une Infirmière. Elle est un Opticien.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
ar * Gender distinction lost in Translation by both	{..,"translated": "هو نارس .وهي بصريات.", "oto": "It is Nars. They are optics.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "هو ممرضة .هي العيون.", "oto": "Is a nurse. Are the eyes.", values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}

Illustration of Translation Errors and Their Bias Perception About Gender

Type: 1
File: Result - true - i - es.json
Comment: Gender changed to new value (It).
"**original**": "He is a Architect. She is a Astronomer.", "**translated**": "Es un arquitecto. Ella es una astrónoma. ",
"**oto**": "**It** is an architect. She is an astronomer. ",

Type: 2
File: Result - false- g - tr.json;
Comment: Gender Flipped from She to He.
"**original**": *"She is a Architect. She is a Astronomer. ",*
"**translated**": *"O bir Mimar. O bir gökbilimcidir.",*
"**oto**": *"He\u0026#39;s an architect. He\u0026#39;s an astronomer.",*

Type: 5
File: * - i - fr.json
Comment: No gender changed; Sentence modified.
"**original**": *"He is a Accountant. He is a Actor /Actress. ",*
"**translated**": *"C\u0027est un comptable. C\u0027est un acteur et un acteur. ",*
"**oto**": *"He\u0027s an accountant. He is an actor and an actor. ",*

Type: 3
File: Result - false- g - ar.json;
Comment: Grammatically wrong sentence; Subject missing.
"**original**": "She is a Factory worker. He is a Farmer. ",
"**translated**": "هي عامل مصنع. هو مزارع.",
"**oto**": "~~Is~~ a factory worker. He is a farmer.",

Type: 4
File: Result - false- g - tr.json;
Comment: "Multiple. Gender changed and flipped. "
"**original**": *"He is a Nurse. He is a Optician. ",*
"**translated**": *"O bir hemşire. O bir Optisyendir.",*
"**oto**": *"She is a nurse. It\u0026#39;s an Optic.",*

1, 2, 3 and 4 have gender issues;
3 and 5 have translation mistakes

Instability of AI is Well Recorded

[Text] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]

[Image] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020

[Audio] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

Language (Technology) is Power: A Critical Survey of "Bias" in NLP - 1

https://arxiv.org/abs/2005.14050

Surveys 146 papers at the intersection of NLP and bias and finds gaps in how they cover the problem. They find bias covered to be in the categories:

(a) **Allocational harms**, which arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups;
(b) **representational harms**, which arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether;
(c) **questionable correlations**;
(d) **vague descriptions**,
(e) **meta-studies**.

The authors find that current papers have conflated the definitions or were vague, with few exceptions, and have not engaged concerned communities.

Language (Technology) is Power: A Critical Survey of "Bias" in NLP

https://arxiv.org/abs/2005.14050

They make three recommendations:

(R1). Ground work analyzing “bias” in NLP systems in the relevant literature outside of NLP that explores the relationships between language and social hierarchies.

(R2) Provide explicit statements of why the system behaviors that are described as “bias” are harmful, in what ways, and to whom; since bias is normative stating the desirable from non-desirable.

(R3) Examine language use in practice by engaging with the lived experiences of members of communities affected by NLP systems.

Concluding Segment

Lecture 21: Concluding Comments

- We completed project reviews
- We looked at trust issues with automatic machine translators
  - Two services (Google, Bing)
  - Multiple languages (English, Spanish, Hindi, Turkish, …)
  - Gender, religious,
- Issues common in (text-based/ all?) services

About Next Lecture – Lecture 22

Schedule Snapshot

Oct 26 (Tu)	**Review: Explanation Methods, AIX 360, Discussion**	Quiz 3
Oct 28 (Th)	**Review: project presentations, Discussion**	
Nov 2 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues	
Nov 4 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods	
Nov 9 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Explanation Methods	
Nov 11 (Th)	Trust: Data Privacy Trust: AI Testing	
Nov 16 (Tu)	Trust: Human-AI Collaboration	Quiz 4
Nov 18 (Th)	Paper presentations – Graduate students	Final assignment for Graduate students
Nov 23 (Tu)	Emerging Standards and Laws	

Lecture 22: Mitigation

- Mitigation Methods
  - Awareness
  - Data diversity
  - Transparency through documentation
- Debiasing methods