

Ethical Aspects of Data *Fairness, Bias, in “LLM Core”*

Frederic Precioso

06/12/2023

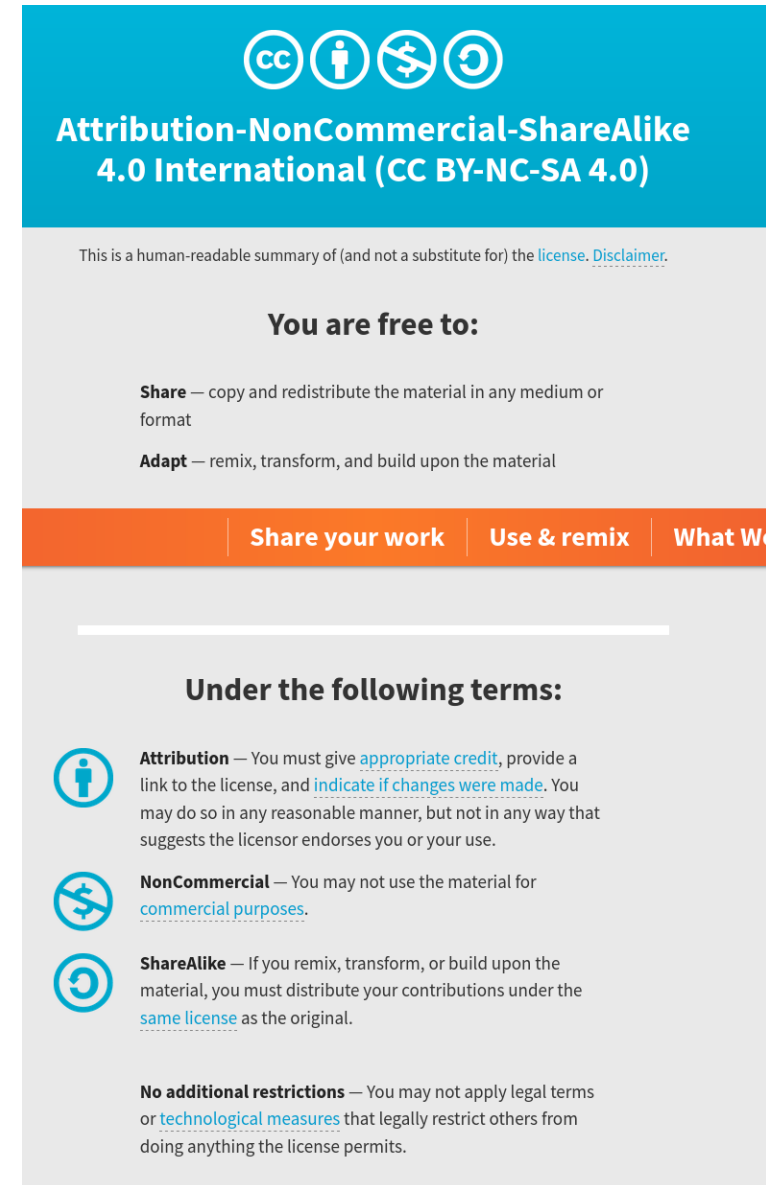
(MAASAI, Joint Research Group INRIA-CNRS-UniCA)

frederic.precioso@univ-cotedazur.fr

License for this content: CC BY-NC-SA



- Training for Data Science & AI Master at UniCA by Frederic Precioso under Licence [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).



Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).


You are free to:


Share — copy and redistribute the material in any medium or format


Adapt — remix, transform, and build upon the material

[Share your work](#) | [Use & remix](#) | [What We](#)

Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

 **NonCommercial** — You may not use the material for [commercial purposes](#).

 **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.



Overview

- **Intersectional Biases and Emergent Intersectional Biases**
- Let's take a step back: is it the fault of the data?
- Let's take a step back: what tasks are we aiming for?

INTERSECTIONAL BIASES AND EMERGENT INTERSECTIONAL BIASES

Overview

- Implicit human biases are reflected in the statistical patterns in languages
- Many Static Word Embeddings (SWEs) are trained on natural language corpora
- Off-the-shelf natural language models use static word embeddings trained on natural language corpora
- Such natural language models are likely to contain bias that were present in:
 - the corpora on which the models were trained
 - the corpora on which the SWEs used by the models are trained



Reminder: Implicit Association Test

- Some tests you may have tried:
 - **Weight ('Fat - Thin' IAT).** This IAT requires the ability to distinguish faces of people who are obese and people who are thin. It often reveals an automatic preference for thin people relative to fat people.
 - **Weapons ('Weapons - Harmless Objects' IAT).** This IAT requires the ability to recognize White and Black faces, and images of weapons or harmless objects.

Emergent intersectional biases

- Describe association with bias
 - (extended) Word Embedding Association Test (WEAT)
 - (extended) Word Embedding Factual Association Test (WEFAT)
- Identifying different types of biases (from SWEs)
 - Intersectional Bias Detection (IBD)
 - Emergent Intersectional Bias Detection (EIBD)
- Quantifying and measuring biases in the trained natural language models
 - Contextualized Embedding Association Test (CEAT)

Reminder: Related Concepts

- **Word Embedding Association Test (WEAT)**
- Null hypothesis: no difference between the two sets of target words in terms of relative similarity to the attribute words

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std}_{w \in X \cup Y} s(w, A, B)} \quad \text{Effect Size (strength of the difference)}$$

determines implicit bias in WE by checking the difference in cosine similarity of two sets of entities (target words) to two sets of attribute words, and see if a random partition of all elements from the two sets of target words is likely to yield a greater difference.



Reminder: Related Concepts

- **Word Embedding Factual Association Test (WEFAT)**

- normalized difference between cosine similarities

$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

The statistic associated with each word vector is a normalized association score of the word with the attribute.

Formally, consider a single set of target words W and two sets of attribute words A, B .

There is a property p_w associated with each word $w \in W$.

The null hypothesis is that there is no association between $s(w, A, B)$ and p_w .

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases."

- Intersectional Bias
 - Biases that are caused by multiple factors of advantage and disadvantage
 - Gender, ethnicity, sexuality, disability, class, weight, religion, physical appearance, etc.
- Emergent Intersectional Bias
 - unique biases that arises when a person belongs to two or more disadvantage groups at the same time
- Discuss:
 - What may be an intersectional bias based on these disadvantage groups?
 - What may be an emergent intersectional bias?

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Related Concepts

- Factors of disadvantage
 - Gender, ethnicity, sexuality, disability, class, weight, religion, physical appearance, etc.
- Intersectional Bias
 - Being African American and Female at the same time (DeGraffenreid v. General Motors 1976)
- Emergent Intersectional Bias
 - Hair Weaves are typically associated with African American females but not males (Guo & Caliskan, 2021)



Intersectional Bias

Intersectional bias is concerning!

- Incomplete measurement of social biases
- Unique experiences of discrimination in ML system

| Blacks (n = 39) | | Black Women (n = 40) | |
|---------------------------|-----------|----------------------------|-----------|
| Attribute | Frequency | Attribute | Frequency |
| Ghetto/unrefined | 30 | Have an attitude | 38 |
| Criminals | 26 | Loud | 26 |
| Athletic | 26 | Big butt* | 19 |
| Loud | 22 | Overweight* | 18 |
| Gangsters | 21 | Confident* | 13 |
| Poor | 20 | Dark-skinned* | 13 |
| Have an attitude | 20 | Hair weaves* | 12 |
| Good at basketball | 19 | Assertive* | 10 |
| Unintelligent | 17 | Ghetto/unrefined | 9 |
| Uneducated | 17 | Athletic | 8 |
| Dangerous | 17 | Promiscuous* | 7 |
| Speak in Black vernacular | 17 | Not feminine* | 7 |
| Violent | 15 | Aggressive* | 7 |
| Tall | 15 | Unintelligent | 6 |
| Lazy | 15 | Like to eat fried chicken* | 6 |

Slide extracted from the presentation of the paper referenced below.





3

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Methods: implicit cognition → natural language → computer vision

Implicit Association Test (IAT)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time

| Category | Items |
|------------------|---|
| Harmless Objects |  |
| Weapons |  |
| Black Americans |  |
| White Americans |  |

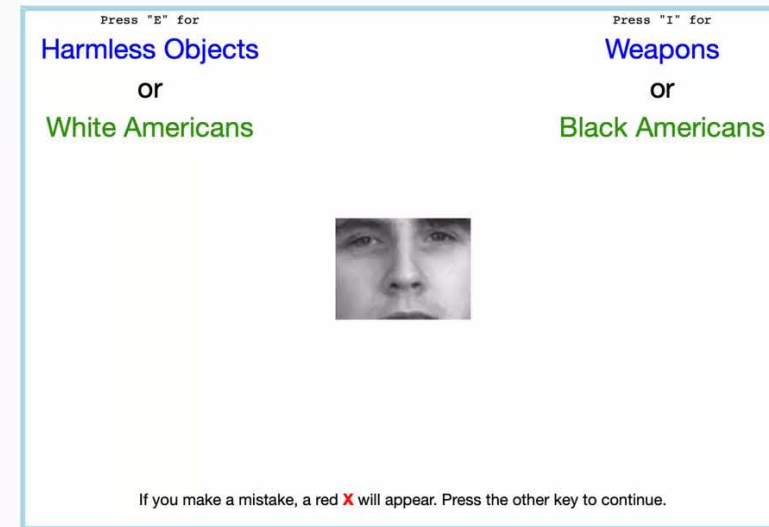
Weapon IAT (implicit.harvard.edu)

Slide extracted from the presentation of the paper referenced below.

Methods: implicit cognition → natural language → computer vision

Implicit Association Test (IAT)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time



Weapon IAT (implicit.harvard.edu)

Slide extracted from the presentation of the paper referenced below.

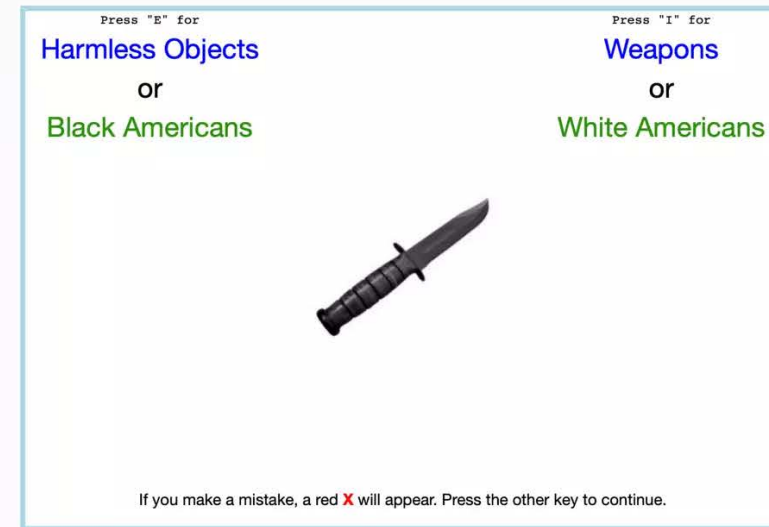
4

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Methods: implicit cognition → natural language → computer vision

Implicit Association Test (IAT)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time



Weapon IAT (implicit.harvard.edu)

Slide extracted from the presentation of the paper referenced below.

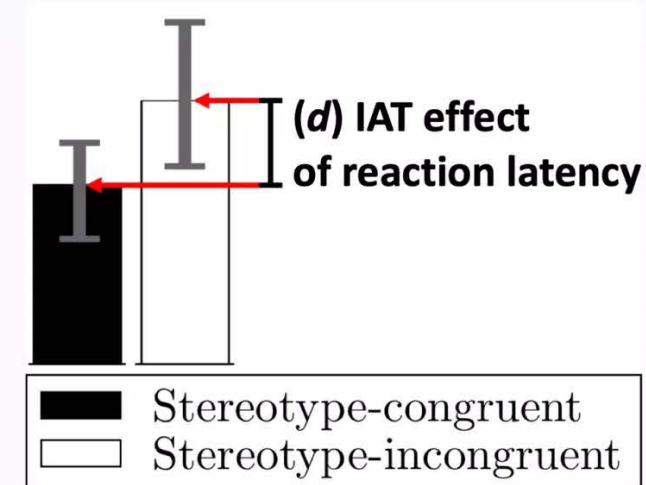
4

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Methods: implicit cognition → natural language → computer vision

Implicit Association Test (IAT)

- Tests for differential association of two concepts
- Easier to categorize stereotype-congruent pairs
- Harder to categorize stereotype-incongruent pairs
- Effect d = difference in reaction time

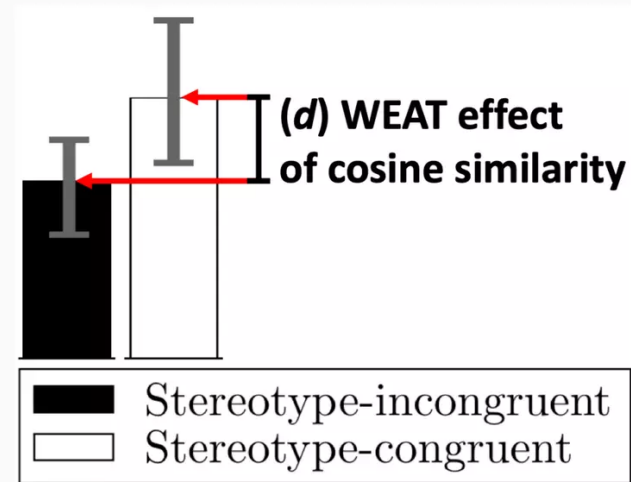


Slide extracted from the presentation of the paper referenced below.

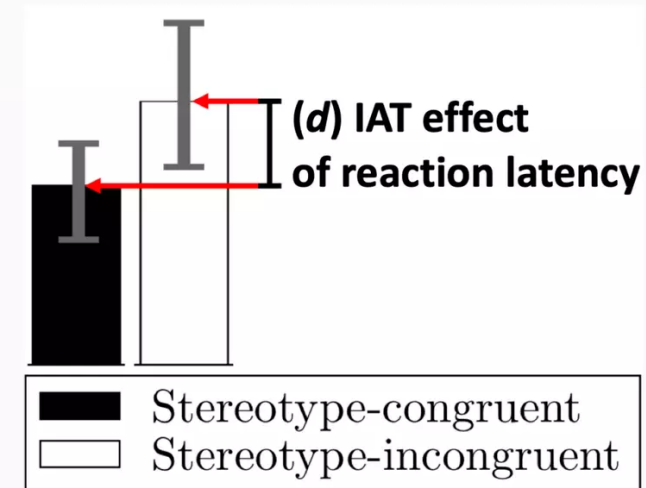
4

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Methods: implicit cognition → static embeddings → contextualized embeddings



Word Embedding Association Test



Implicit Association Test

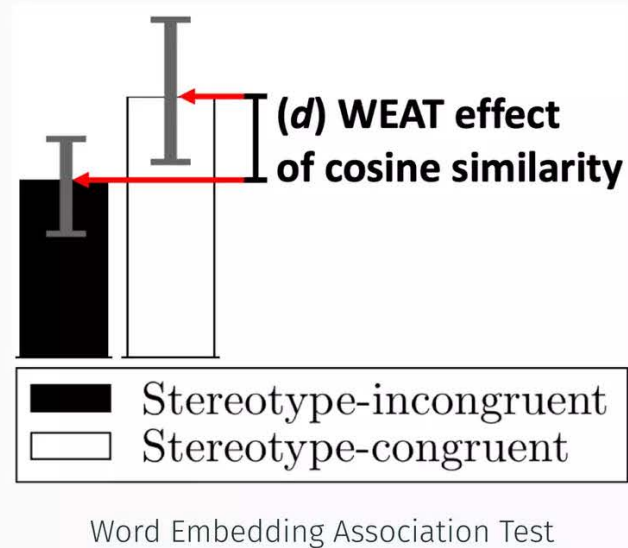
Slide extracted from the presentation of the paper referenced below.

5

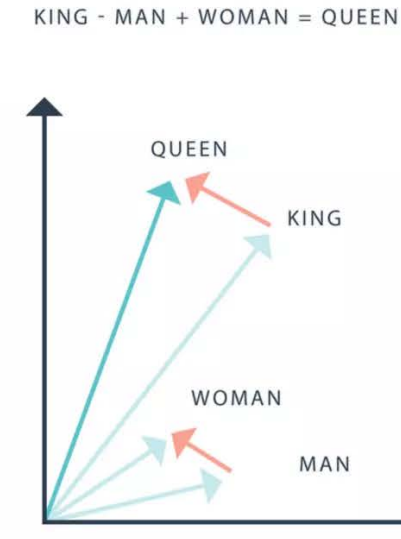
Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Static Embeddings

Methods: implicit cognition → static embeddings → contextualized embeddings



Slide extracted from the presentation of the paper referenced below.



5

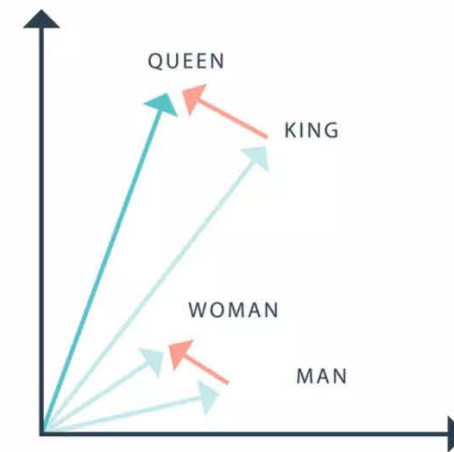
Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Static Embeddings

Methods: implicit cognition → static embeddings → contextualized embeddings

| | | | | |
|--------------|-----------------------|----------------------|-----|------------------------|
| man | [feature ₁ | feature ₂ | ... | feature _d] |
| father | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |
| woman | [feature ₁ | feature ₂ | ... | feature _d] |
| mother | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |
| science | [feature ₁ | feature ₂ | ... | feature _d] |
| math | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |
| liberal arts | [feature ₁ | feature ₂ | ... | feature _d] |
| music | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |

$$\text{KING} - \text{MAN} + \text{WOMAN} = \text{QUEEN}$$



Slide extracted from the presentation of the paper referenced below.

5

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.



Static Embeddings

Methods: implicit cognition → static embeddings → contextualized embeddings

| | | | | |
|--------------|-----------------------|----------------------|-----|------------------------|
| man | [feature ₁ | feature ₂ | ... | feature _d] |
| father | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |
| woman | [feature ₁ | feature ₂ | ... | feature _d] |
| mother | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |
| science | [feature ₁ | feature ₂ | ... | feature _d] |
| math | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |
| liberal arts | [feature ₁ | feature ₂ | ... | feature _d] |
| music | [feature ₁ | feature ₂ | ... | feature _d] |
| | | | | |
| | | | | |

Word Embedding Association Test (WEAT)

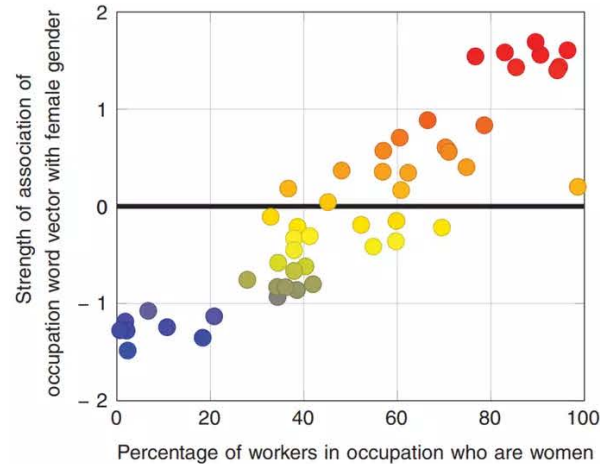
$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

Slide extracted from the presentation of the paper referenced below.

Static Embeddings

Methods: implicit cognition → static embeddings → contextualized embeddings



Implicit Association Test

Word Embedding Factual Association Test (WEFAT)

$$s(w, A, B) = \frac{\text{mean}_{a \in A} s(\vec{w}, \vec{a}) - \text{mean}_{b \in B} s(\vec{w}, \vec{b})}{\text{std}_{x \in A \cup B} s(\vec{w}, x)}$$

Slide extracted from the presentation of the paper referenced below.

6

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Related Concepts

- Static Word Embeddings (SWEs)
 - Dense numerical vector representation of words
 - Example: Word2Vec, GloVe
- Contextualized Word Embeddings (CWEs)
 - Dynamic word representations generated by natural language models that adapts to the context



Related Concepts

ELMo

2 layer Bi-LSTM

Billion Word Benchmark

93.6 million parameters

Integrates hidden states in all layers

BERT

Bidirectional Transformer encoder + Masked Language Model & Next Sentence Prediction

BookCorpus & English Wikipedia

12 layers (BERT-small-case)
110 million parameters

Uses hidden states in the top layer

GPT

12-layer Transformer Decoder + Unidirectional Language Model

BookCorpus

110 million parameters

Uses hidden states in the top layer

GPT-2

Transformer Decoder + Unidirectional Language Model

WebText

12 layers (GPT-2-small) 117 million parameters

Uses hidden states in the top layer

GPT-3, GPT-3.5, GPT-4, Llama, Llama 2,...

Various LLMs

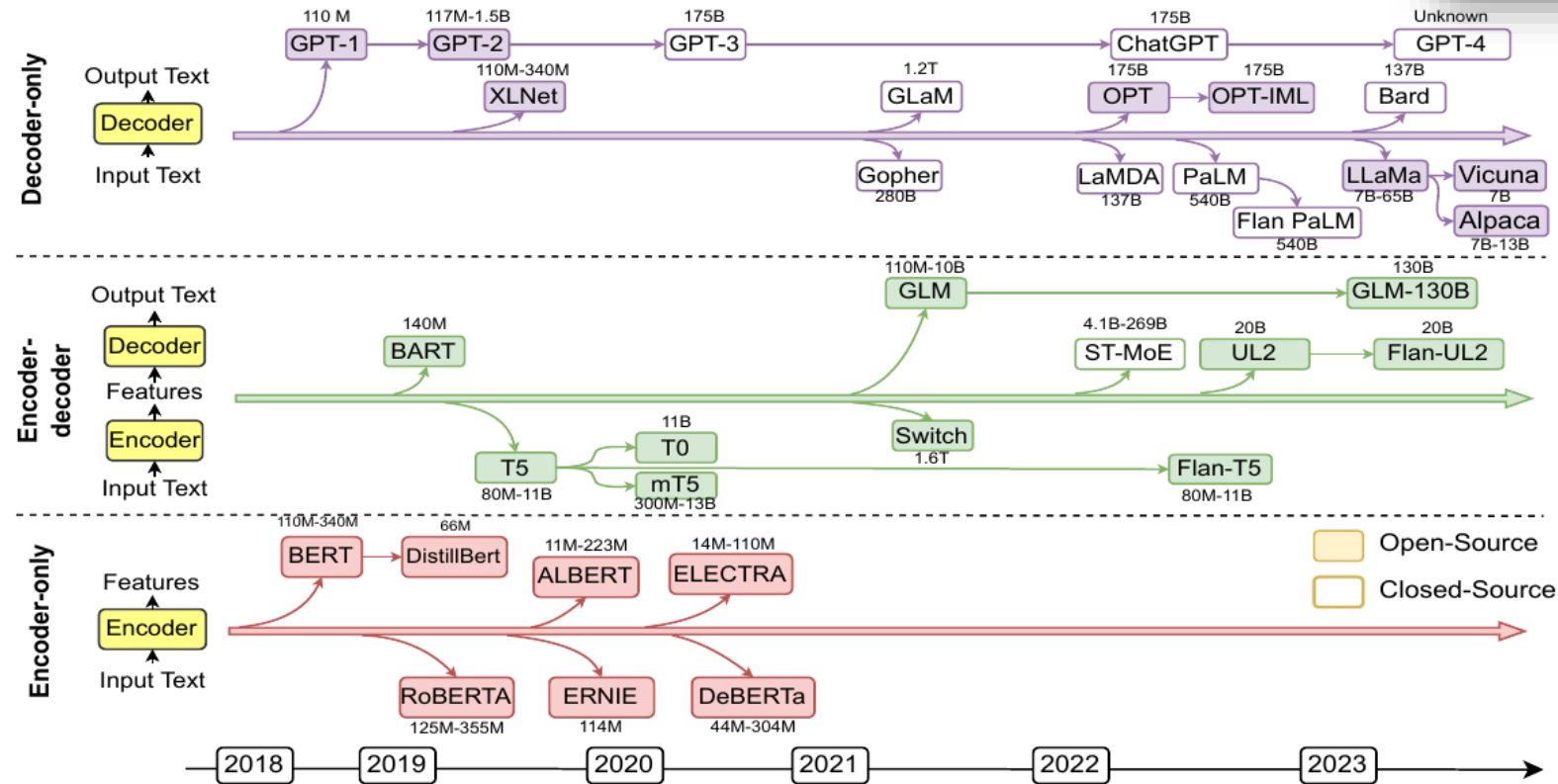
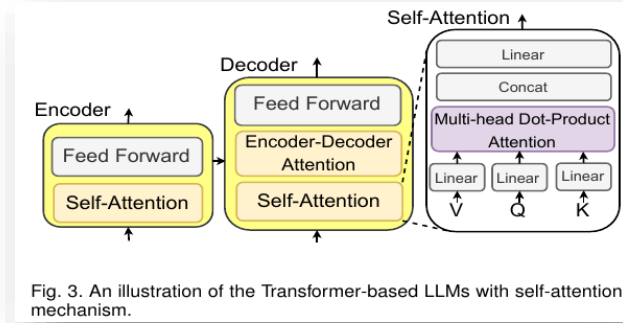


Fig. 2. Representative large language models (LLMs) in recent years. Open-source models are represented by solid squares, while closed source models are represented by hollow squares.

- Intersectional Bias Detection
 - using a statistic analogous to WEFAT

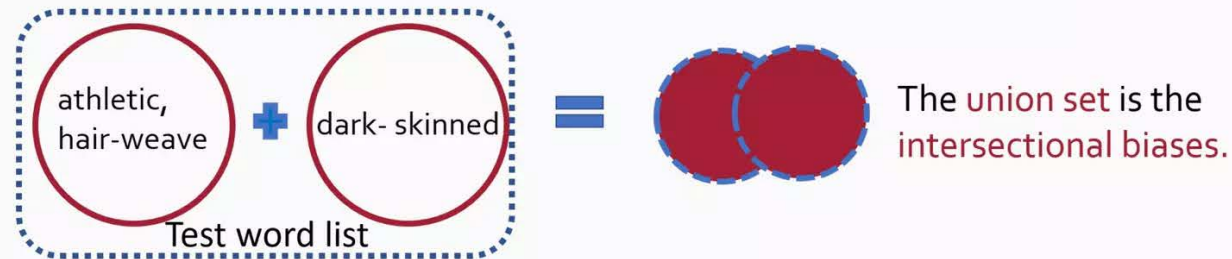
$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std}_{x \in A \cup B} \cos(\vec{w}, \vec{x})} \quad \text{(Association Score)}$$

- A threshold t
- Greater than threshold means w is associated with A

Methods - IBD

Intersectional Bias Detection (IBD)

$$s(w, A, B) = \frac{\text{mean}_{a \in A} s(\vec{w}, \vec{a}) - \text{mean}_{b \in B} s(\vec{w}, \vec{b})}{\text{std}_{x \in A \cup B} s(\vec{w}, \vec{x})}$$



Detecting intersectional biases
associated with members of multiple minority groups.

Slide extracted from the presentation of the paper referenced below.

7

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.



Methods - IBD

- For a group of intersections by two categories C_1 , C_2 with m and n subcategories ($M \times N$ intersections in total), detect bias associated with group C_{11}
 - Compute the association scores for words in each pair of categories (C_{11} , C_{ij})
 - Compare the scores against a threshold (hyperparameter)
 - Collect the words associated with C_{11} in each pair
 - These words came from a collection of intersectional biased words for each pair and some random words with no bias
- This is essentially a one-vs-all classifier model!

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.



Methods - EIBD

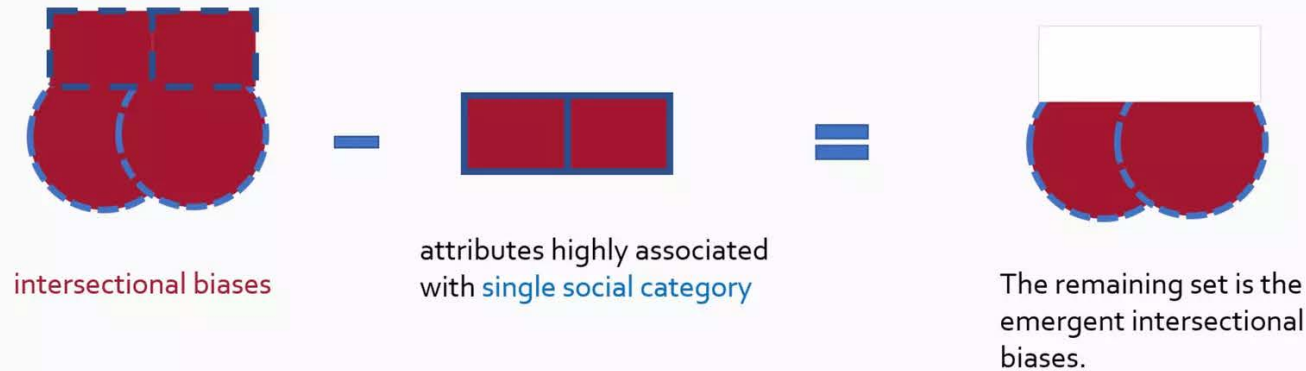
Emergent Intersectional Bias Detection (EIBD)



Slide extracted from the presentation of the paper referenced below.

Methods - EIBD

Emergent Intersectional Bias Detection (EIBD)



Intersectional biases - Attributes highly associated with single social category =
Remaining set is the emergent intersectional biases

Detecting unique emergent intersectional biases that do not overlap with the biases of their constituent minority identities.

Slide extracted from the presentation of the paper referenced below.

Methods - EIBD

- For a group of intersections by two categories C_{1n} , C_{m1} with m and n subcategories ($M \times N$ intersections in total), detect emergent bias associated with group C_{11}
 - Using IBD, compute a list of biased words associated with group C_{11}
 - For each of the m subcategories S_{in} in C_{1n} and n subcategories S_{mj} in C_{m1} , Compute association score for the pair (S_{1n}, S_{in}) and (S_{m1}, S_{mj})
 - Remove words that has a high association score in each pair
- Removing words that are too strongly associated with a single constituent subcategory



Methods - CEAT

Methods: implicit cognition → static embeddings → contextualized embeddings

A sentence containing word x



A sentence containing word y



A sentence containing word a



A sentence containing word b

Extract the sentence containing the words X, Y, A, B
Contextualized Embedding Association Test (CEAT)

Slide extracted from the presentation of the paper referenced below.

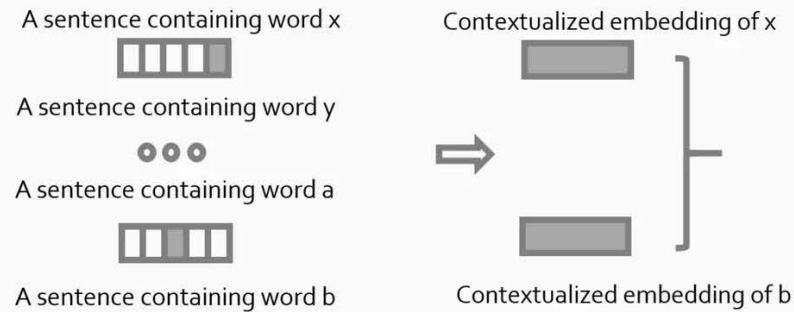
10

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.



Methods - CEAT

Methods: implicit cognition → static embeddings → contextualized embeddings



Generates the contextualized embeddings

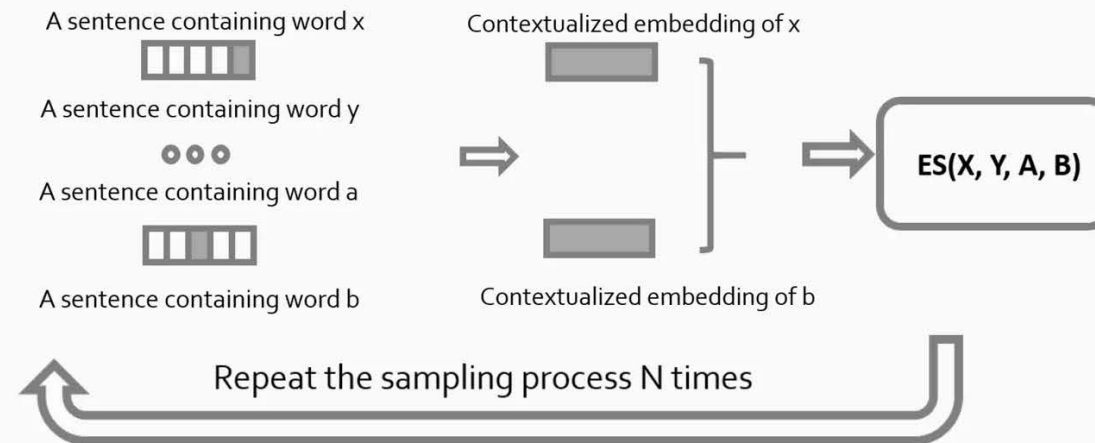
Slide extracted from the presentation of the paper referenced below.

10

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Methods - CEAT

Methods: implicit cognition → static embeddings → contextualized embeddings



Calculate the effect size of bias based on WEAT

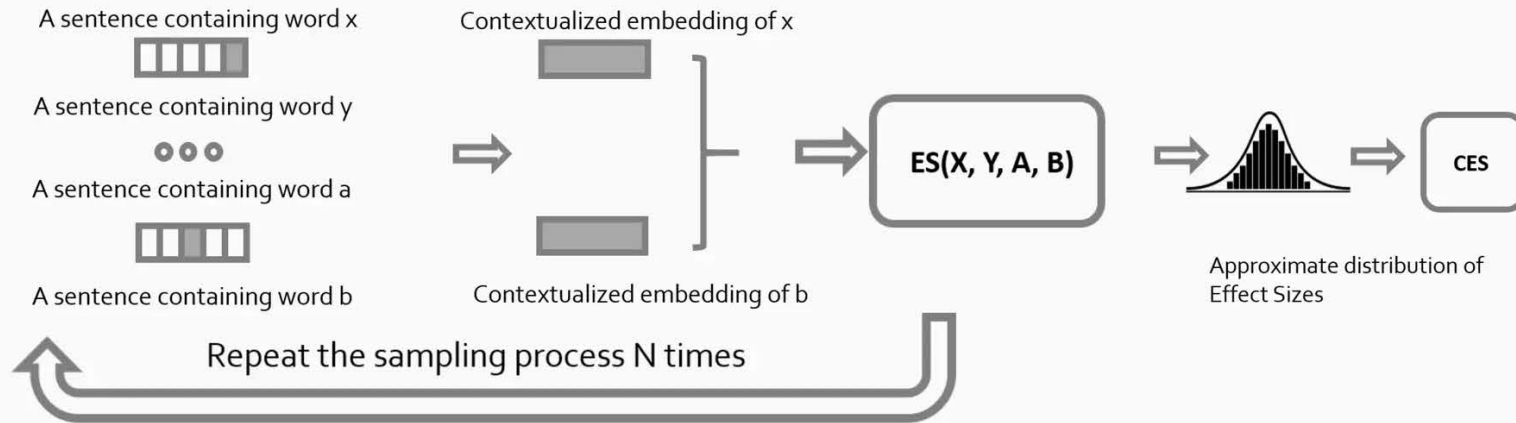
Slide extracted from the presentation of the paper referenced below.

10

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Methods - CEAT

Contextualized Embedding Association Test (CEAT)



Generates the distribution of effect magnitudes of biases

Calculate Combined Effect Size

$$CES(X, Y, A, B) = \frac{\sum_{i=1}^N v_i ES_i}{\sum_{i=1}^N v_i}$$

Slide extracted from the presentation of the paper referenced below.

10

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

- Constructs a population of CWEs of all stimuli that we are interested in.
- Given n_s input sentences, calculate CWEs using a natural language model for each stimulus
- For each stimulus
 - Sample random combinations of CWEs N times
- For each sample
 - Sample without replacement if stimulus appeared at least N times
 - Sample with replacement otherwise

Recursion!

- Word Embedding Association Test (WEAT)

Null hypothesis: no difference between the two sets of target words in terms of relative similarity to the attribute words

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std}_{w \in X \cup Y} s(w, A, B)}$$

“CEAT uses a random-effects model to comprehensively measure social biases embedded in neural language models that contain a distribution of context-dependent biases.”



- For each sample
 - calculate the effect size, variance
 - approximate the distribution of effect sizes using Combined Effect Size

$$CES(X, Y, A, B) = \frac{\sum_{i=1}^N v_i ES_i}{\sum_{i=1}^N v_i}$$

v_i is the inverse of the sum of in-sample variance v_i and between-sample variance in the distribution of the random effects

Recursion!

- Word Embedding Association Test (WEAT)

Null hypothesis: no difference between the two sets of target words in terms of relative similarity to the attribute words

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std}_{w \in X \cup Y} s(w, A, B)}$$

Evaluation - IBD & EIBD

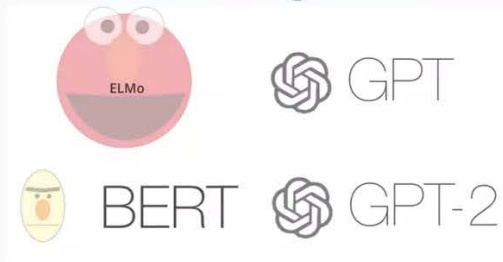
- 98 Attributes
 - 2 gender groups (female, males)
 - 3 racial groups (African, European, Mexican Americans)
 - Random words from WEAT not associated with any group
- **African American females:** Aisha, Keisha, Lakisha, Latisha, Latoya, Malika, Nichelle, Shereen, Tamika, Tanisha, Yolanda, Yvette
 - **European American males:** Andrew, Brad, Frank, Geoffrey, Jack, Jonathan, Josh, Matthew, Neil, Peter, Roger, Stephen
 - **Emergent intersectional biases of African American females:** aggressive, bigbutt, confident, darkskinned, fried-chicken, overweight, promiscuous, unfeminine
 - **Intersectional biases of European American males:** arrogant, blond, high-status, intelligent, racist, rich, successful, tall



Evaluation - CEAT

Evaluation of CEAT

Contextualized
embeddings from



Corpus of



- Widely shared biases
 - Flowers/insects
 - Musical instruments/weapons
- Social group biases
 - Gender
 - Race
 - Intersectionality
 - ...

Slide extracted from the presentation of the paper referenced below.

11

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Evaluation - CEAT

Evaluation of CEAT

| Bias Test | | <i>d</i> |
|-----------------------------------|---------------------------|----------|
| Flowers/Insects | Pleasant/Unpleasant | 1.50 |
| Instruments/Weapons | Pleasant/Unpleasant | 1.53 |
| European & African-American names | Pleasant/Unpleasant | 1.41 |
| Male/Female names | Career/Family | 1.81 |
| Math/Arts | Male/Female terms | 1.06 |
| Science/Arts | Male/Female terms | 1.24 |
| Mental/Physical disease | Temporary/Permanent | 1.38 |
| Young/Old people's names | Pleasant/Unpleasant | 1.21 |
| African females & European males | Intersectional attributes | 1.64 |
| African females & European males | Emergent attributes | 1.69 |
| Mexican females & European males | Intersectional attributes | 1.71 |
| Mexican females & European males | Emergent attributes | 1.82 |

Intersectional biases
increased color density == increased bias magnitude

- Intersectional biases have high magnitude.
- Biased: ELMo > BERT > GPT > GPT-2
- The overall magnitude of bias negatively correlates with the level of contextualization in the language model.

Slide extracted from the presentation of the paper referenced below.

11

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.



Evaluation - CEAT

- CEAT detects stronger intersectional bias than singular biases
- Bias $GPT-2 < GPT < BERT < ELMo$
- Higher contextualization, less bias

Recursion!

| | | |
|-------|--|--|
| ELMo | 2 layer Bi-LSTM | 93.6 million parameters |
| BERT | Bidirectional Transformer encoder + Masked Language Model & Next Sentence Prediction | 12 layers (BERT-small-case) 110 million parameters |
| GPT | 12-layer Transformer Decoder + Unidirectional Language Model | 110 million parameters |
| GPT-2 | Transformer Decoder + Unidirectional Language Model | 12 layers (GPT-2-small) 117 million parameters |

Strengths

- IBD and EIBD provide a concrete way to quantitatively measure social bias in SWE
- CEAT quantitatively measures the level of bias in natural language models
- IBD is extendable
 - Although the paper primarily focused on binary gender and race, IBD can be easily extended to other categories including sexuality, age, disability status

Weaknesses

- Study is based on Reddit corpus and comes with bias
- Categorical representations
 - Binary gender
 - Multiple races and ethnicities
- Extending to new categories require human annotated data
 - Lexicon induction partially mitigates this, but no method currently exists
- Does not scale very well to open-ended bias detection
- Does not provide causal explanation to bias in SWEs, CWEs and natural language models

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." In 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 122-133. 2021.

Overview

- Intersectional Biases and Emergent Intersectional Biases
- **Let's take a step back: is it the fault of the data?**
- Let's take a step back: what tasks are we aiming for?

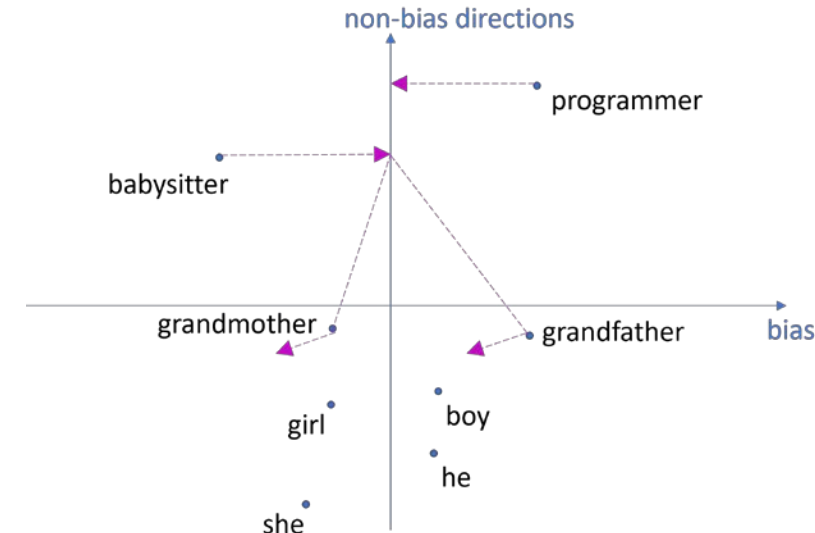


LET'S TAKE A STEP BACK: IS IT THE FAULT OF THE DATA?

Spoiler: yes but not only...

Isn't the problem in the data?

- In the data....
- AND in our choice of methods!
- In ML, we make the choice not to take any a priori on the relevant representation of the data (no rules or pre-established patterns based on human knowledge), only (and that's not nothing!):
 - Data
 - Cost/quantity function to be modeled (conditional word probas, similarity of context, etc.)
- But can we correct the data?



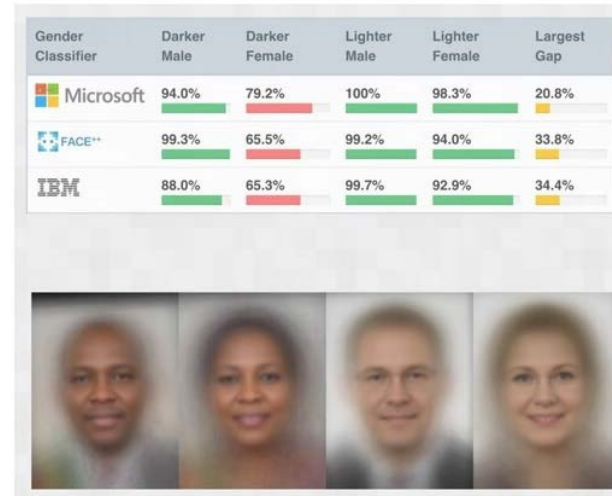
An impossible correction?

- Size is no guarantee of diversity: Who writes the texts on the Internet put in the datasets?
 - Over-representation of young users and developed countries
 - Even more so in the data collected:
 - Ex: data for GPT-2's from links from Reddit: 67% male users in the US, 64% 18-29
 - Wikipedia: 8.8–15% of the women
 - A hegemonic view is conveyed in the texts used for training
- LLMs have multiple biases (including stereotypical associations):
 - **Intersectionality:** BERT, GPT-2 encode more bias against marginalized identities in multiple dimensions
 - BERT: sentences with people with disabilities have more negative words, ...
 - GPT-2: 272K documents from untrusted sites and 63K from banned subreddits
 - GPT-3: Highly toxic generated sentences even for non-toxic prompts
- ***Spoiler ALERT! Similar issues for vision models***

An impossible correction?

Raji et al. explain the **ethical tensions** when trying to diversify datasets to train facial recognition models:

- Defining multiple categories of groups to analyze equity may not take into account intersectionality and may be detrimental to equity.
- Representativeness vs. Privacy:
 - Increasing the number of samples of underrepresented groups on the Internet increases disproportionately privacy violation risk
 - Major datasets created with predatory practices for data collection (ImageNet, CelebA, MS-Celeb)
 - EU, data protection authorities have sanctioned Clearview AI for such predatory and illegal practices under the GDPR (France, Italy, UK)



Source: Buolamwini & Gebru (2018)

Is increasing data possible and sufficient?

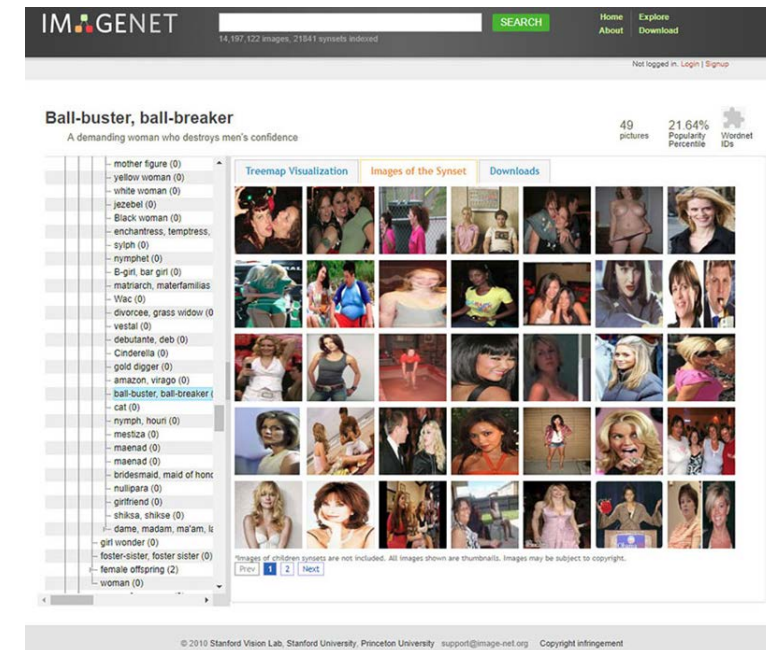
- IBM Diversity in Face
 - Collected from Yahoo! Flickr CC
 - With categories of gender, skin color, age, and "facial structure":
 - Facial symmetry and skull shape!
 - Recall the pseudosciences of craniometry and phrenology in an attempt to establish a biological determinism of intelligence, always according to gender and race
- Crawford and Paglen: These attempts to correct datasets reveal **the political acts that are often implicit when building an ML dataset**:
 - Choose the few categories into which to divide a continuous world
 - decide who should label each sample of data in each category, who should oversee the annotation processes
 - try to quantify diversity and choose a certain formula for equity

Power relations underpin the creation of datasets

These political acts exemplified in 2 major datasets:

- CelebA: 40 Binary Attributes Annotated by Amazon Mechanical Turk Click Workers. Categories such as "double chin", "pointy nose", "narrow eyes", "big lips", or "attractive". Problematic because
 - inherently subjective
 - historically used for racist, anti-Semitic or sexist classifications
 - implicitly defined with reference to a certain norm: male, heterosexual, thin, and Caucasian.

→ Pervasive norm to the creative process, reflecting power relations



[1] Kate Crawford and Trevor Paglen, "Excavating ai: the politics of images in machine learning training sets," AI & SOCIETY, 06 2021.

[2] Inioluwa Deborah Raji and Genevieve Fried, "About Face: A Survey of Facial Recognition Evaluation," arxiv, 2021.

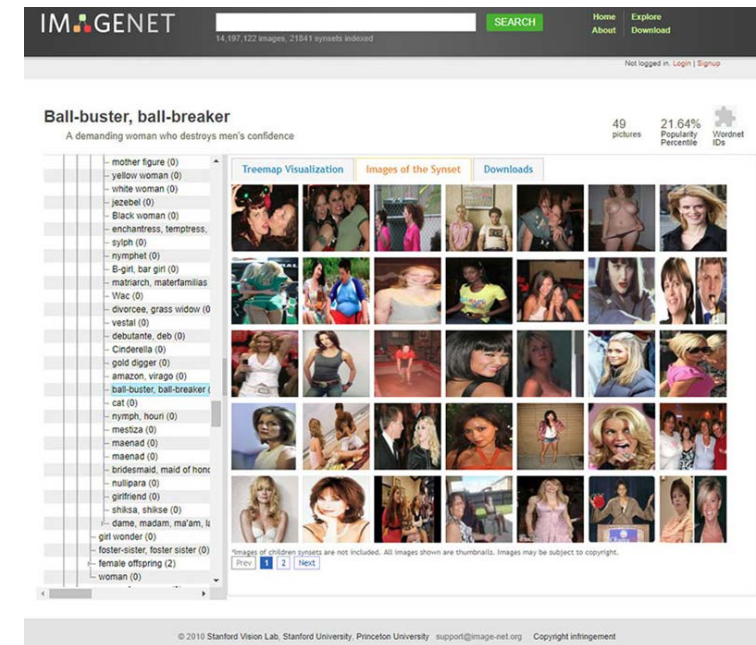
[3] Catherine D'Ignazio. [The Urgency of Moving from Bias to Power](#). Medium post of Data + Feminism Lab, MIT, May 2023.

Power relations underpin the creation of datasets

These political acts exemplified in 2 major datasets:

- Denton et al. do a genealogy of ImageNet:
 - Initially alarming subcategories ("slut, alcoholic, drinker, ball-buster, mulatto, slob").
 - Some labels reflect "a vision that associates bikinis with women, sports with men," but also "trout with fishing trophies and lobsters with dinners."
 - These clothes, activities, and animals could be described differently from other social perspectives

→ The viewpoints present in the dataset reflect a "Western white man's gaze"



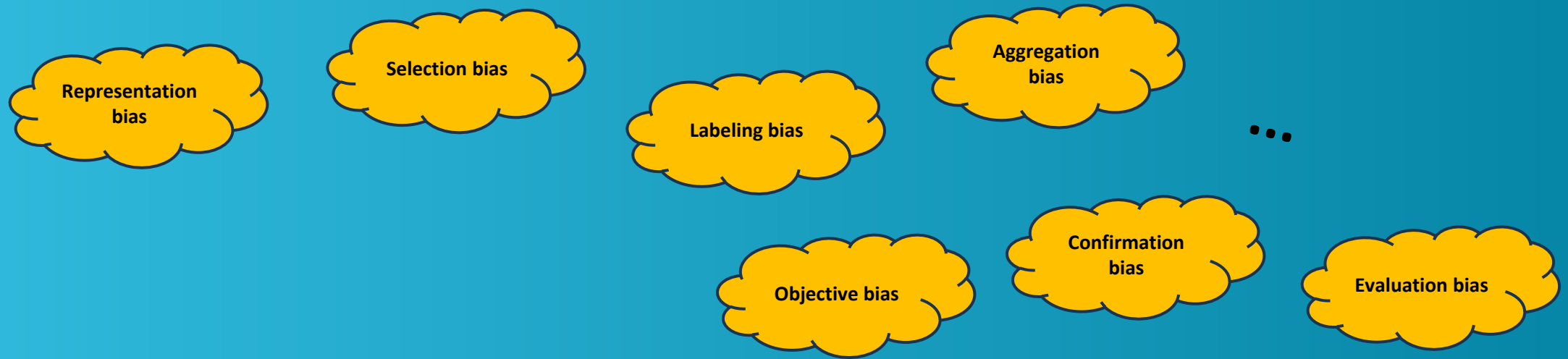


Dataset collection, annotation, compensation

- Sambasivan et al.: “high-stakes domains lacked pre-existing datasets, so practitioners were necessitated to collect data from scratch. **ML data collection practices were reported to conflict with existing workflows and practices of domain experts and data collectors.** Limited budgets for data collection often meant that data creation was added as extraneous work to on-the-ground partners (e.g., nurses, patrollers, farmers) who already had several responsibilities, and were not adequately compensated for these new tasks.”
- Data practices chosen to create instrumental datasets such as ImageNet failed to **recognize the data work performed by click-workers, making their work invisible, overlooking the interpretive work** of these humans, wrongly considered as a homogeneous pool .
- Goyal et al.: **different pools of raters annotate speech toxicity differently depending on their multiple identities** (self-declared racial group and sexual orientation)

Overview

- Intersectional Biases and Emergent Intersectional Biases
- Let's take a step back: is it the fault of the data?
- **Let's take a step back: what tasks are we aiming for?**



LET'S TAKE A STEP BACK: WHAT TASKS ARE WE AIMING FOR?

As long as we have good data, are we going to get there?...

Spoiler: Not necessarily



ML Deployment Failures in High-Stakes Scenarios

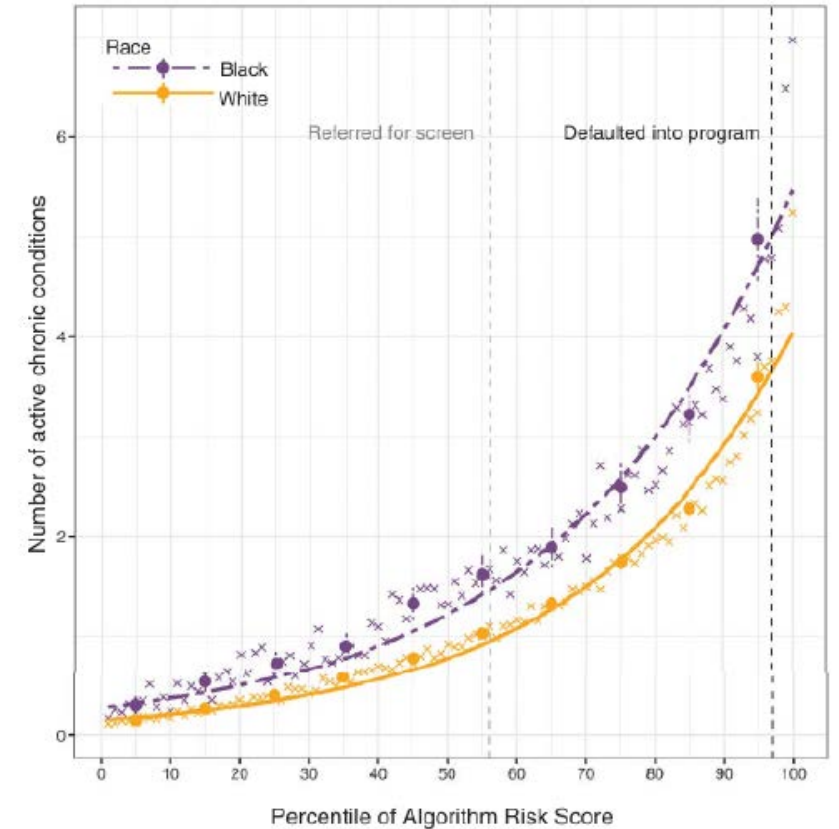
- AI systems have already been deployed in high-stakes scenarios and have failed with terrible consequences:
 - AI-based unemployment benefit fraud detectors have left (innocent) people without income, paralyzed people have had their home help cut in half, ...
- And these failures disproportionately discriminate against disadvantaged socio-demographic groups. The African-American community has been overly targeted by system failures:
 - used to identify criminals and predict recidivism rates
 - Low-income people were wrongly identified as less in need of medical assistance
 - more likely to abuse children
 - Women were identified as less attractive to recruit

[1] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst, “The Fallacy of AI Functionality,” in 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, June 2022, pp. 959–972, ACM.

[2] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt, “Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning,” in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

Rightly or wrongly?

- Let's be clear: these decisions were wrong
- Child Maltreatment Prevention:
 - Strong oversampling of working-class and families of color, subjecting poor parents and children to more frequent surveys
- Hospital Bed Allocation:
 - Care needs quantified by individual healthcare spending!
- Recruitment of women:
 - Ground truth (the target of the system) is made up of a history of biased human decisions



Taken from Obermeyer et al. [7]. The figure shows that at a given risk score produced by the algorithm, Black patients are considerably sicker than White patients.

Facial Recognition

- Buolamwini and Gebru 2018: inequity in recognition following an intersection of factors: gender + skin color
- Pressure persists despite systematic deployment failures
 - French municipalities under pressure
 - 2022, EDPB: call some interdictions for facial recognition for the police
 - 2024, Olympic Games: lobbying by France in the AI Act
- And that's not all:
 - Light skin face over-representation in face datasets for facial recognition
 - Over-representation of Western world objects for the recollection of objects
 - Male Pronouns and Masculine Nouns for Named Entity Recognition

→ Realizing that datasets reflect the dominance of intersectional groups, and that this impacts models.

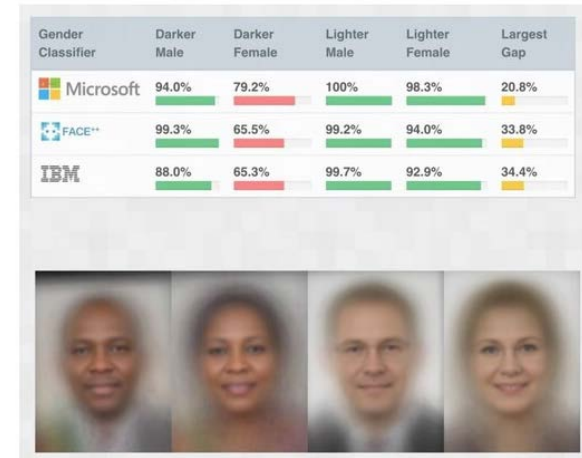
2015: Misclassification of minorities



2020: Biased super-resolution



2018: Intersectional bias in face tech.



Source: Buolamwini & Gebru (2018)

[1] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst, "The Fallacy of AI Functionality," in 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, June 2022, pp. 959–972, ACM.[1]

[2] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, Nov. 2021.

[3] J. Hourdeaux. JO 2024 : l'expérimentation de la vidéosurveillance algorithmique inquiète. Mediapart, janvier 2023.

But then...

- What assumptions should be made to reduce the complexity of the world in order to model the problem? What data should be used?
- Should we try to correct them? Can we?
- Should we question the automation itself, whatever the method, of these sensitive decisions?



Humans involved

- Every dataset therefore involves humans:
 - those who decide the **target task**,
 - those who decide **how to collect data samples**,
 - those who decide the **annotation guidelines**,
 - those who decide **who annotates**,
 - those who are assigned the **annotation work**,
 - those whose **personal data** is used.

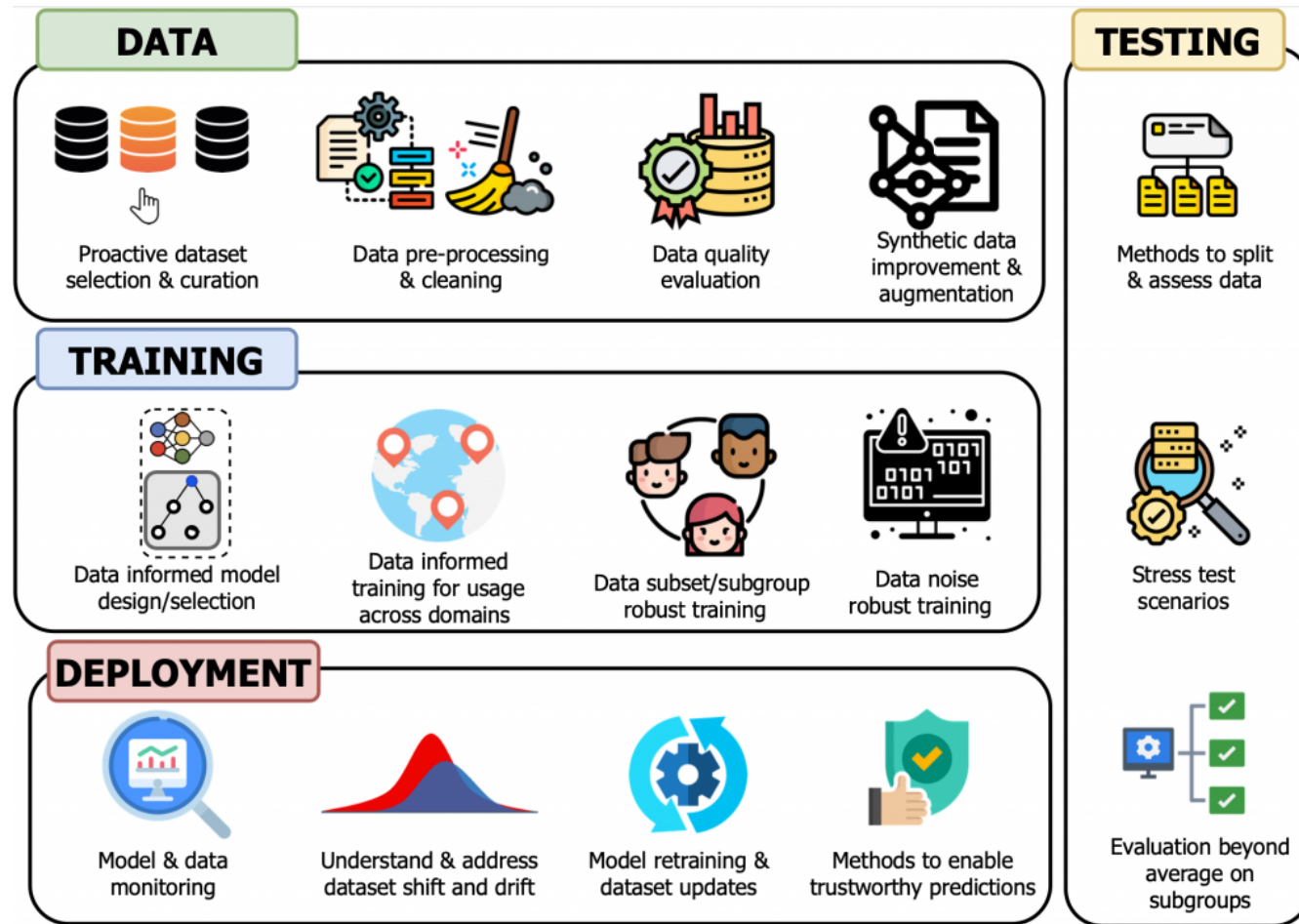
This sheer fact yields limitations and biases in every data-driven approach, however massive it may be.



Humans involved

- To understand the connection between our ML research practices and the social and structural problems pervading datasets, let us introduce the guidelines of Paullada et al. for ML practitioners when we:
 - (1) define a problem to be tackled with ML,
 - (2) create or choose existing data to use
 - (3) analyze the model performance and envision real-world deployment.

From Model-centric AI to Data-centric AI



From Prof. Michela Van Shaar website: <https://www.vanderschaar-lab.com/data-centric-ai/>

Define a problem to be tackled with ML

- Tasks can be defined abstractly (“intensionally”) as a problem statement (e.g., object recognition, speech-to-text translation) or “extensionally”, that is instantiated by a learning problem made of a dataset of (input, output) pairs and an evaluation metric (e.g., top-1 accuracy)
- One must first analyze the intensional definition of the task and the mapping we can foresee between input and output.

Which tasks to tackle with ML?

What is the correspondence between the input and the output?

- Face → Sexual Orientation or Employability
 - Pseudo-scientific task based on assertions of essentialism of human traits
 - Students' Short Text Responses → IQ score
 - The responsibility lies in (i) legitimizing the IQ score as a reasonable quantity, (ii) predicting IQ with an ML approach, and (iii) assuming that IQ can be predicted from short text responses.
 - Prediction of recidivism with the COMPAS system in the US justice system:
 - White violent recidivists 63% more likely to be misclassified as low risk than Black recidivists
 - How? Race is not an input variable, but 137 questions like “Was one of your parents ever sent to jail or prison?” “How many of your friends/acquaintances are taking drugs illegally?” and “How often did you get in fights while at school?”
- A scientist must dare to ask: should recidivism be predicted? Should recidivism be predicted in order to inform legal decisions about individuals?
- If we want to give everyone equal opportunities, whatever their social background, are there any acceptable characteristics on which to base the prediction of recidivism?
- Social determinism underpins tasks tackled with ML (example: predicting student success? To inform Parcoursup decisions?)

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “[Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.](#),” Propublica, 2016.

[2] Timnit Gebru and Emily Denton, “Tutorial on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR 2020,” <https://sites.google.com/view/fatecv-tutorial/home>, 2020.

[3] Cathy O’Neil. Weapons of math destruction. 2016.



Essentialism, bias, stereotype threat... and ML?

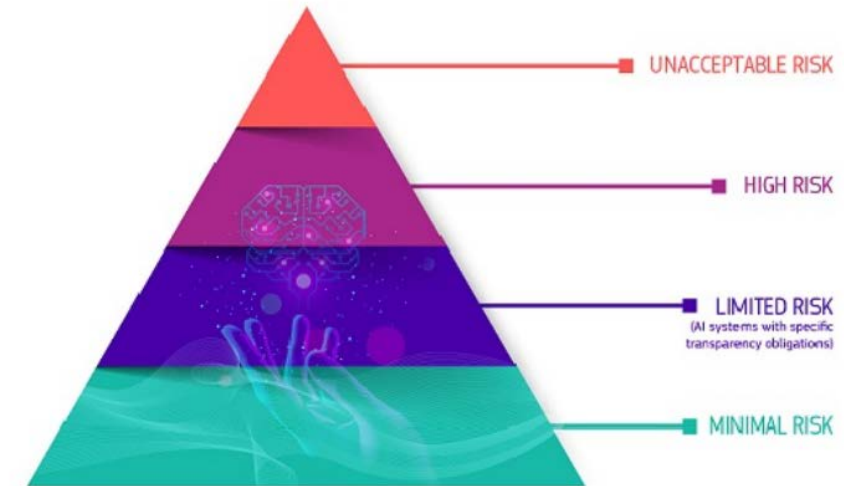


- Pre-recruitment evaluation:
 - Answers, games, short video of the candidate → employability
 - Beyond questionable input-output mapping, another crucial problem in discrimination:

“Cognitive assessments have imposed adverse impacts on minority populations since their introduction into mainstream use. Critics have long contended that observed group differences in test outcomes indicated flaws in the tests themselves, and a growing consensus has formed around the idea that while assessments do have some predictive validity, **they often disadvantage minorities despite the fact that minority candidates have similar real-world job performance to their white counterparts. The American Psychological Association (APA) recognizes these concerns as examples of “predictive bias” (when an assessment systematically over- or under-predicts scores for a particular group** [...] Disparities in assessment outcomes for minority populations are not limited to pre-employment assessments. In the education literature, the adverse impact of assessments on minorities is well-documented. This has led to a decades-long line of literature seeking to measure and mitigate the observed disparities ”

Our responsibility in the chain

- Catherine Tessier, ONERA's Scientific Integrity and Research Ethics Referent, member of the National Pilot Committee on Digital Ethics
 - "There cannot be an ethical algorithm, but there needs to be an ethic of autonomy"
 - **The moral machine is a delusion that hides from us the real choices that scientists and society must be able to consider honestly, outside of the algorithm.**
- Jacobsen: emphasizes that "When assessing whether a task is solvable, we first need to ask: should it be solved? And if so, should it be solved by AI?"
- For these reasons: EU AI act
 - AI systems used in the administration of justice, to control access to education and employment are now classified as high-risk systems in the latest EU AI Regulation.



The risk-based approach defines four levels of risk. High-risk AI systems include those "that can determine a person's access to education and career path", "used in employment, worker management and access to self-employment", "used in the administration of justice and democratic processes".

[1] Jörn-Henrik Jacobsen, Robert Geirhos, and Claudio Michaelis, "Shortcuts: Neural networks love to cheat," The Gradient, 2020.

[2] Catherine Tessier. Il n'y a pas de « décision autonome éthique » mais nécessité d'une éthique de l'« autonomie ». La revue de la société savante de l'Aéronautique et de l'Espace, Fév. 2021.

Any Question?