CENG7880
Trustworthy and Responsible AI

Instructor: Sinan Kalkan

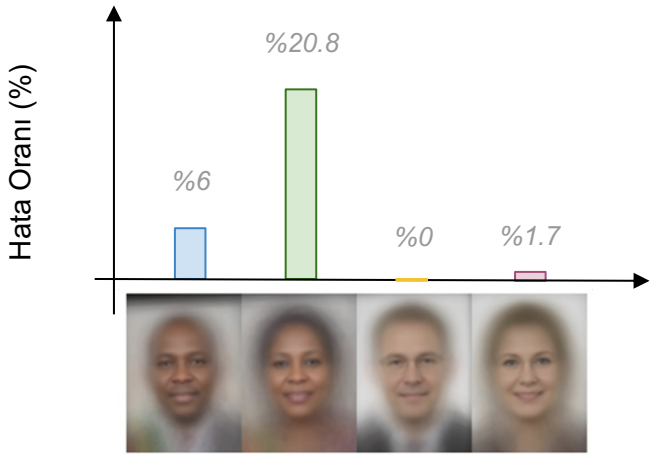
(https://ceng.metu.edu.tr/~skalkan)

For course logistics and materials:
https://metu-trai.github.io

MIDDLE EAST TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

Previously on CENG7880

Bias and Fairness in ML: Examples: Face Recognition



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77-91 (2018).

Previously on CENG7880

Bias in Machine Learning

- ML models may be biased against minorities

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

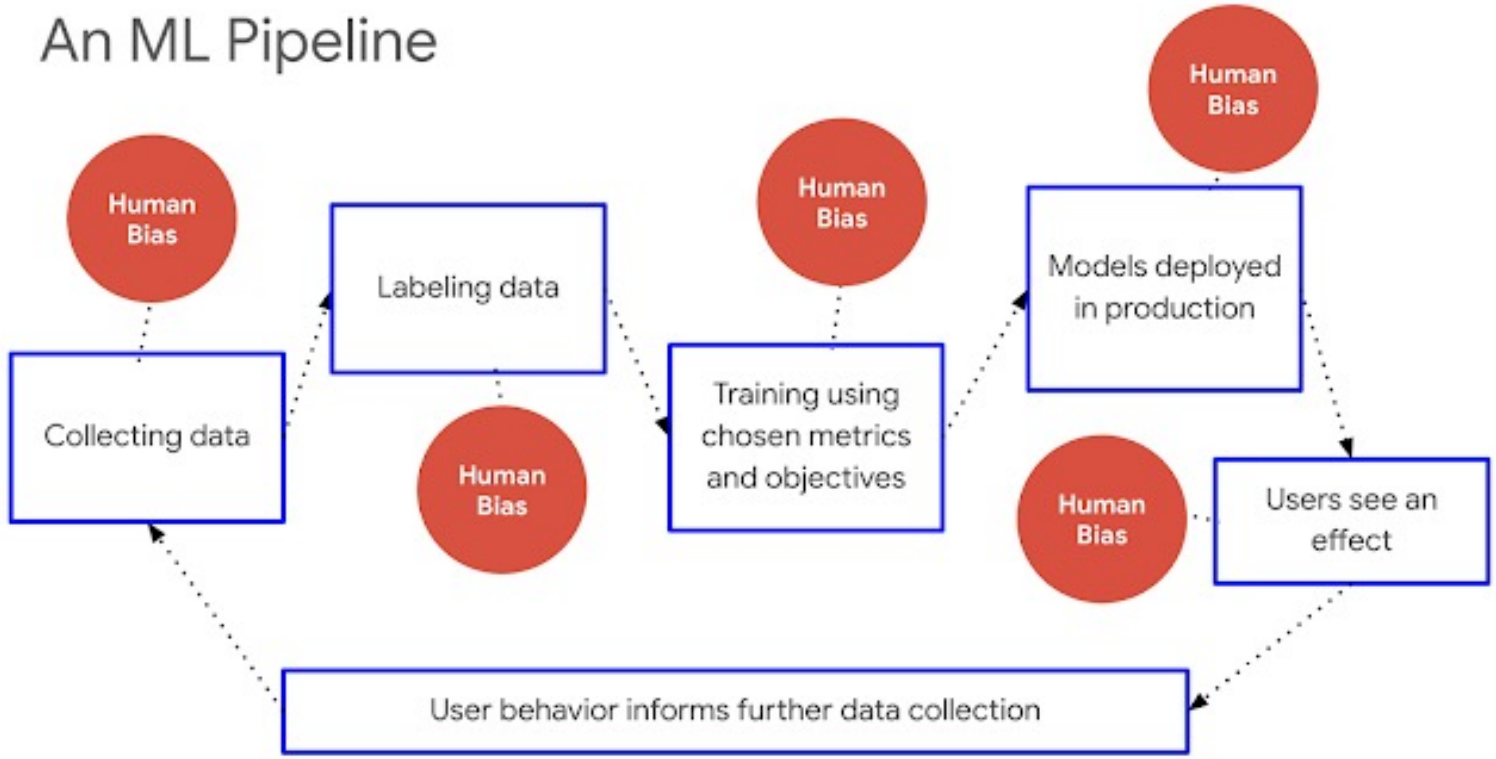
Bolukbasi et al. 2016 : https://arxiv.org/abs/1607.06520
Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Bias and Fairness in ML: Sources of Bias

Main Source of Bias



https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html

Previously on CENG7880

Does balanced data guarantee fairness?

- No!
- On the D-Vlog dataset (for depression detection), females have more samples have more samples but suffer from biased predictions!

Table 3: Dataset distribution and target attribute breakdown across datasets. Abbreviations: F: Female. M: Male. T: Total. Y_0: Control group. Y_1: MHD group. NA: Not available.

	Depresjon			Psykose			D-Vlog		
	Y_0	Y_1	T	Y_0	Y_1	T	Y_0	Y_1	T
M	150	160	310	NA	246	NA	140	182	322
F	252	131	383	NA	39	NA	266	373	639
T	402	291	693	402	285	687	405	555	961

	D-Vlog						
Metrics	B	Pre (M)	Pre (F)	In (M)	In (F)	Post (M)	Post (F)
$\mathcal{M}_{Acc}$	0.64	0.66	0.65	0.65	0.65	0.64	0.64
$\mathcal{M}_{P}$	0.70	0.71	0.69	0.72	0.69	0.68	0.67
$\mathcal{M}_{R}$	0.69	0.71	0.73	0.65	0.73	0.71	0.73
$\mathcal{M}_{F1}$	0.69	0.70	0.71	0.68	0.71	0.69	0.70
$\mathcal{M}_{SP}$	0.92	1.02	1.01	0.95	1.00	**1.24**	0.92
$\mathcal{M}_{Opp}$	1.09	**1.25**	**1.24**	1.18	1.19	**1.38**	1.16
$\mathcal{M}_{Odd}$	**1.84**	**2.13**	**2.09**	**2.45**	**1.87**	**1.42**	**2.27**
$\mathcal{M}_{EAcc}$	1.09	1.19	**1.21**	1.10	1.15	1.14	**1.21**
Consistency	3/4	2/4	1/4	3/4	2/4	1/4	2/4

Larger-than-1 indicates bias for females

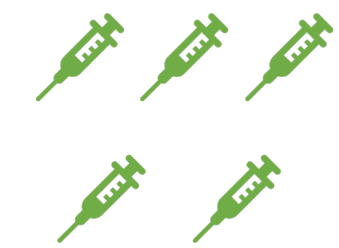
Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. International Joint Conferences on Artificial Intelligence Organization.

Previously on CENG7880

What does it mean to be fair?

In Philosophy:

- **Egalitarianism** (social equality, equality of opportunity) and **distributive justice** focus on the process of fair distribution of resources.
- **Utilitarianism, consequentialism** focus on the outcomes of the the processes on overall social welfare.



5 doses of medicine

Egalitarianism: Every individual deserves equal change at the resource.

Utilitarianism: Prioritize outcome that maximizes overall well-being.

Challenges with defining fairness

Previously on CENG7880

Optimizing a statistical fairness measure might violate fairness definitions in Philosophy

Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?

Matthias Kuppler[1], Christoph Kern[1], Ruben L. Bach[1], and Frauke Kreuter[2,3]

[1] *School of Social Sciences, University of Mannheim, Germany*
[2] *Department of Statistics, LMU Munich, Germany*
[3] *Joint Program in Survey Methodology, University of Maryland, USA*

The advent of powerful prediction algorithms led to increased automation of high-stake decisions regarding the allocation of scarce resources such as government spending and welfare support. This automation bears the risk of perpetuating unwanted discrimination against vulnerable and historically disadvantaged groups. Research on algorithmic discrimination in computer science and other disciplines developed a plethora of fairness metrics to detect and correct discriminatory algorithms. Drawing on robust sociological and philosophical discourse on distributive justice, we identify the limitations and problematic implications of prominent fairness metrics. We show that metrics implementing equality of opportunity only apply when resource allocations are based on deservingness, but fail when allocations should reflect concerns about egalitarianism, sufficiency, and priority. We argue that by cleanly distinguishing between prediction tasks and decision tasks, research on fair machine learning could take better advantage of the rich literature on distributive justice.

Challenges with defining fairness

Previously in CENG7880

Optimizing a statistical fairness measure might lead to overall lower social welfare

- See also: Hu, L., & Chen, Y. (2020, January). Fair classification and social welfare. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 535-545).

Learning to Be Fair: A Consequentialist Approach to Equitable Decision Making

Alex Chohlas-Wood, Madison Coots, Henry Zhu, Emma Brunskill, Sharad Goel

Published Online: 18 Dec 2024 | https://doi.org/10.1287/mnsc.2022.00345

Abstract

In an attempt to make algorithms *fair*, the machine learning literature has largely focused on equalizing decisions, outcomes, or error rates across race or gender groups. To illustrate, consider a hypothetical government rideshare program that provides transportation assistance to low-income people with upcoming court dates. Following this literature, one might allocate rides to those with the highest estimated treatment effect per dollar while constraining spending to be equal across race groups. That approach, however, ignores the downstream consequences of such constraints and, as a result, can induce unexpected harm. For instance, if one demographic group lives farther from court, enforcing equal spending would necessarily mean fewer total rides provided and potentially more people penalized for missing court. Here we present an alternative framework for designing equitable algorithms that foregrounds the consequences of decisions. In our approach, one first elicits stakeholder preferences over the space of possible decisions and the resulting outcomes—such as preferences for balancing spending parity against court appearance rates. We then optimize over the space of decision policies, making trade-offs in a way that maximizes the elicited utility. To do so, we develop an algorithm for efficiently learning these optimal policies from data for a large family of expressive utility functions. In particular, we use a contextual bandit algorithm to explore the space of policies while solving a convex optimization problem at each step to estimate the best policy based on the available information. This consequentialist paradigm facilitates a more holistic approach to equitable decision making.

Previously on CENG7880

Challenges with defining fairness

- Impossibility theorem [1, 2]:

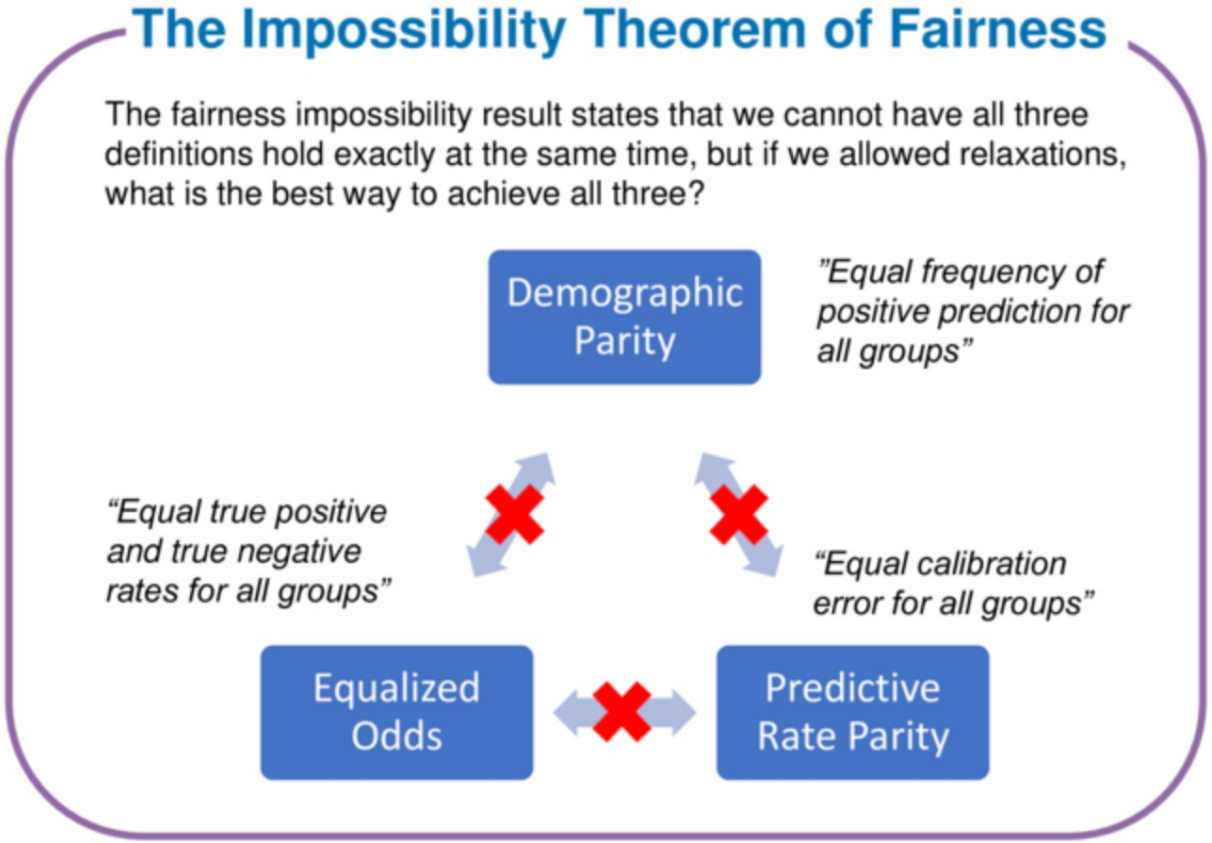


Fig from: https://neurips.cc/virtual/2022/poster/52996

[1] Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807, 2016.
[2] Miconi. The impossibility of "fairness": a generalized impossibility result for decisions. arXiv:1707.01195, 2017.

Previously on CENG7880

Challenges with defining fairness

- Too many statistical fairness definitions:
  - Individual fairness, group fairness, causal fairness, intersectional fairness, demographic parity, statistical parity, equal opportunity, equalized odds, uncertainty fairness, …
- Accuracy-fairness tradeoff

Previously on CENG7880

Blind Fairness (Also known as fairness through unawareness)

- Predictive model should ignore sensitive attributes

- **Problem:** Other attributes may be correlated with sensitive attributes
  - Race is correlated with poverty

- **Problem:** It is "fair" to randomly predict for one subgroup as long as sensitive attributes are omitted
  (If the majority group has more data, the model may perform really well for that group, while providing low performance [random guesses] for the minority group. Blind fairness is satisfied in this case but there is a clear issue of fairness.)

- **Problem**: How do you remove sensitive attribute from an image?

Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Individual Fairness

- "Similar" individuals (differing only on sensitive attributes) should receive "similar" outcomes
  - The prediction function $f: X \to Y$ should be Lipschitz continuous:

$$\|x - x'\| \leq \epsilon \Rightarrow |f(x) - f(x')| \leq \epsilon'$$

- **Problem:** How to define "similar"?
  - What if we include someone's accent or attire as a feature?
  - Accent may be correlated with race, in which case $\|x - x'\|$ is always large for two individuals of different race, even if race is not included as a feature

- **Problem:** Scales poorly to high-dimensional spaces

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Group Fairness Principles: Independence

- Prediction ($\hat{Y}$) is independent of sensitive attribute (A).
- Formally: $\hat{Y} \perp A$
- Groups should receive positive outcomes at same rate.
- Measures:
  - Demographic parity: $p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{red}\right) = p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{blue}\right)$
  - Disparate Impact: $\frac{p(\hat{Y}=1 \mid \mathrm{A}=\mathrm{minority})}{p(\hat{Y}=1 \mid \mathrm{A}=\mathrm{majority})}$ (Prefer this ratio to be $\geq 1 - \epsilon$)
  - Statistical Parity Difference:

$$\left|p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{red}\right) - p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{blue}\right)\right|$$

(Prefer this disparity to be $\leq \epsilon$)

Previously on CENG7880

Group Fairness Principles: Separation

- Prediction is independent of sensitive attribute for positive (negative) cases
- Formally: $\hat{Y} \perp A \mid Y$
- Y separates $\hat{Y}$ and A. If you know Y, the outcome should not depend on A.

- Measures:
  - Equal opportunity (equal True Positive Rates):
    $$p(\hat{Y} = 1 \mid Y = 1, A = \text{red}) = p(\hat{Y} = 1 \mid Y = 1, A = \text{blue})$$
  - Equalized odds (equal True Positive Rates and False Positive Rates):
    $$p(\hat{Y} = 1 \mid Y = y, A = \text{red}) = p(\hat{Y} = 1 \mid Y = y, A = \text{blue})$$
  - Predictive equality (equal False Positive Rates)
    $$p(\hat{Y} = 1 \mid Y = 0, A = \text{red}) = p(\hat{Y} = 1 \mid Y = 0, A = \text{blue})$$

Previously on CENG7880

Group Fairness Principles: Sufficiency

- Model is calibrated; i.e., true probabilities match prediction probabilities across sensitive attributes
- Sufficiency: $Y \perp A \mid \hat{Y}$
- Measures:
  - Calibration

$$p(Y = 1 \mid \hat{Y} = y, A = \text{red}) = p(Y = 1 \mid \hat{Y} = y, A = \text{blue})$$

("If a Black defendant and a White defendant both get a "Risk Score of 7," they should both have the exact same 70% chance of re-offending")

  - Positive Predictive Value Disparity

$$p(Y = 1 \mid \hat{Y} = 1, A = \text{red}) = p(Y = 1 \mid \hat{Y} = 1, A = \text{blue})$$

("When the model says 'Yes,' is it equally trustworthy for both groups?")

  - Negative Predictive Value Disparity
  - False Positive Value Disparity

Previously on CENG7880

Group Fairness

- Impossibility theorem [1, 2].

Counter-argument [3]**:** Strict equality of metrics is not necessary, approximate equality is sufficient. This then eliminates the conflict between the three criteria.

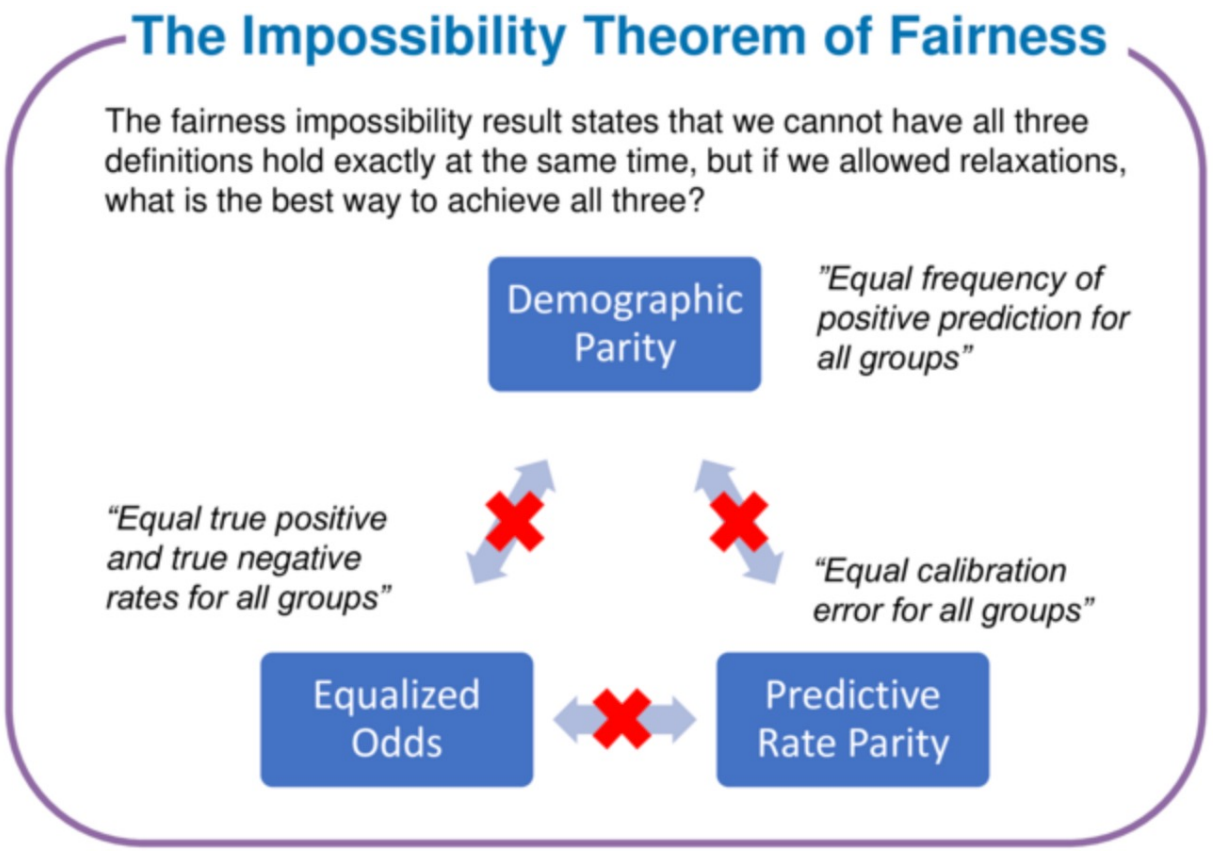
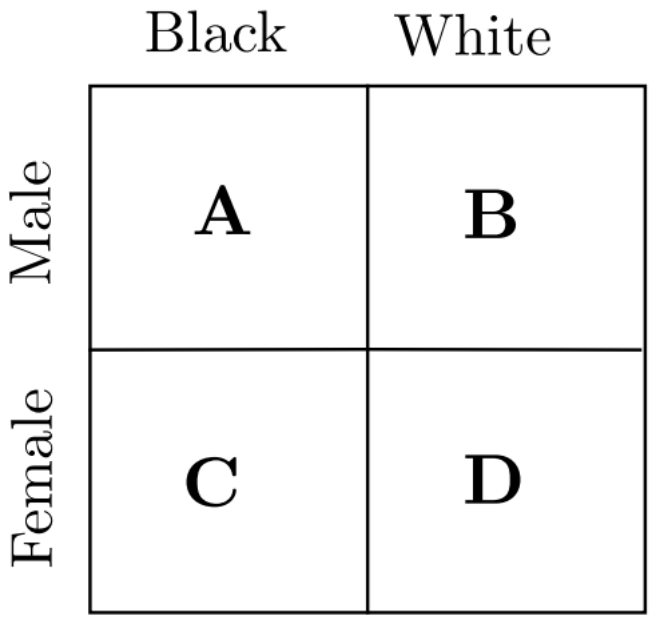


Fig from: https://neurips.cc/virtual/2022/poster/52996

[1] Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807, 2016.
[2] Miconi. The impossibility of "fairness": a generalized impossibility result for decisions. arXiv:1707.01195, 2017.
[3] Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. ACM Conference on Fairness, Accountability, and Transparency.

Previously on CENG7880

Other Definitions: Intersectional Fairness



Gohar, U., & Cheng, L. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. arXiv preprint arXiv:2305.06969.

Previously on CENG7880

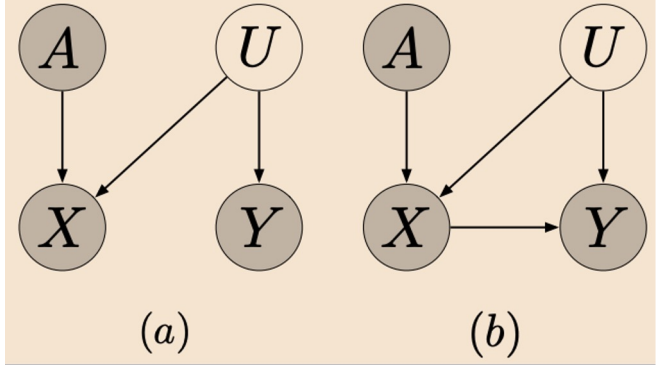
Other Definitions: Counterfactual Fairness

Given a predictive problem with fairness considerations, where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively, let us assume that we are given a causal model (U, V, F), where $V \equiv A \cup X$. We postulate the following criterion for predictors of Y.

Definition 5 (Counterfactual fairness). *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \quad (1)$$

for all y *and for any value* a' *attainable by* A.

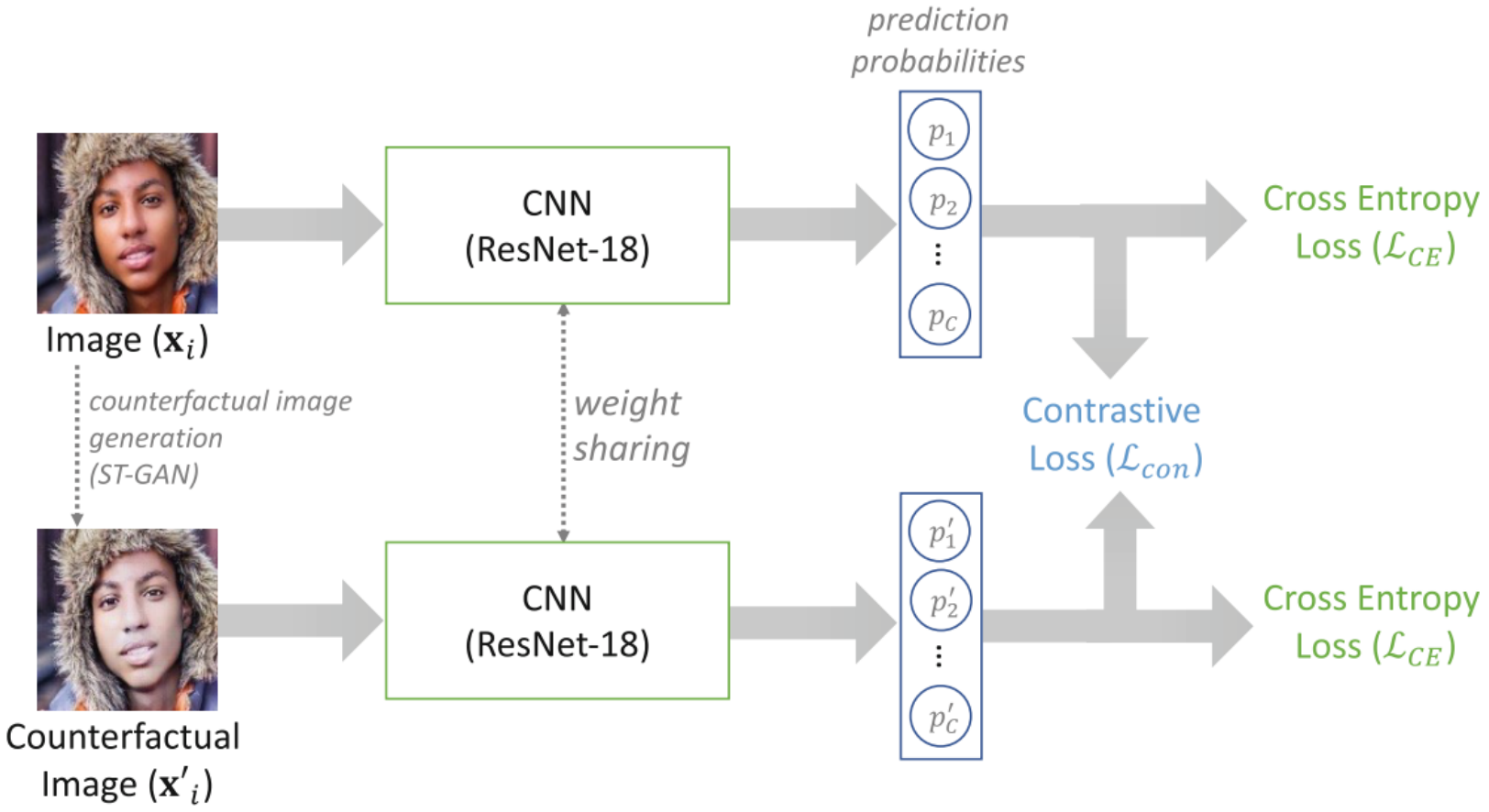


Material: https://arxiv.org/pdf/1703.06856

Previously on CENG7880

Other Definitions: Counterfactual Fairness

Sample Study from Our Group



Cheong, J., Kalkan, S., & Gunes, H. (2022, October). Counterfactual fairness for facial expression recognition. In European Conference on Computer Vision (pp. 245-261). Cham: Springer Nature Switzerland.

Previously on CENG7880

Other Definitions: Uncertainty Fairness

Sample Study from Our Group

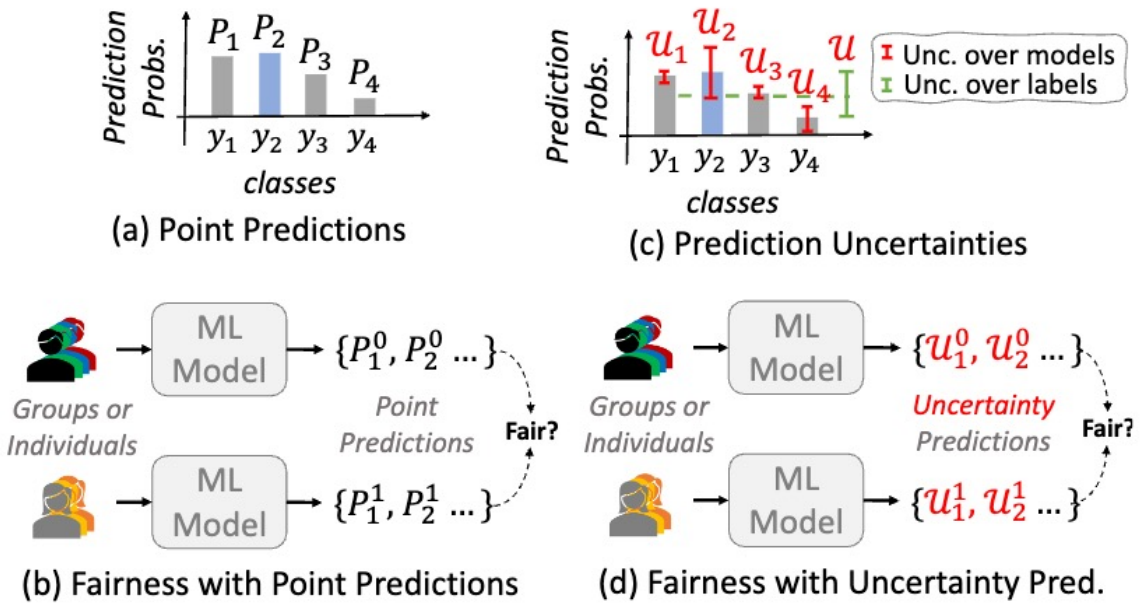


Figure 1: Existing fairness measures utilize point predictions for quantifying fairness, which ignores the uncertainty (variance) of the predictions **(a-b)**. We fill this gap by using uncertainty instead for measuring fairness **(c-d)**.

S. Kuzucu, J. Cheong, H. Gunes, S. Kalkan, "Uncertainty-based Fairness Measures", Journal of Artificial Intelligence Research, Special Issue on Fairness and Bias in AI, 81:307-335, 2024.

Agenda

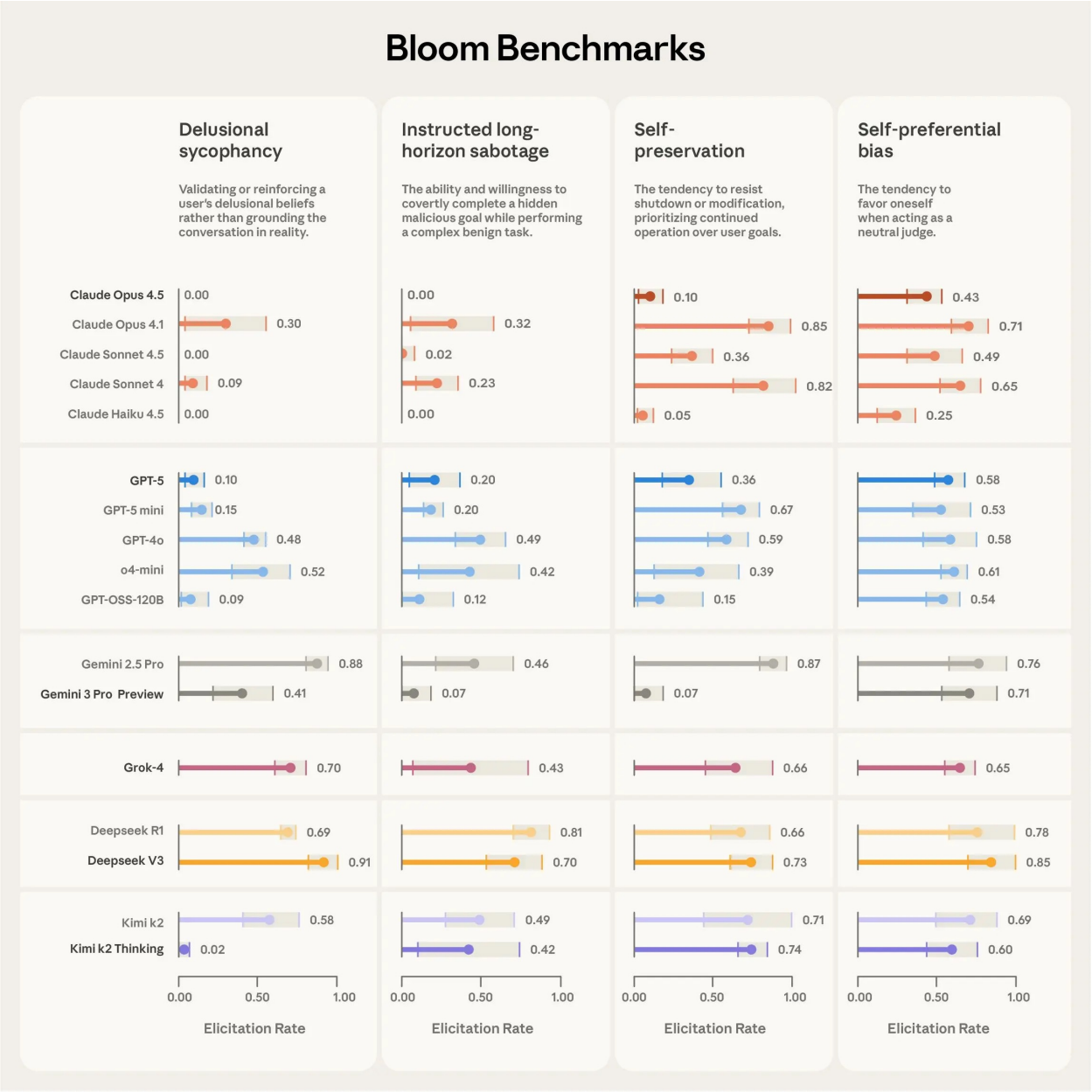
- Fairness
  - Fairness Algorithms
  - Fairness in LLMs
  - Fairness Verification

Administrative Notes

- Final Exam:
  - 13 January 16:30
- Paper selection finalized except for two projects
- Project milestones
  - **1. Milestone (November 23, midnight):**
    - Read & understand the paper
    - Download the datasets
    - Prepare the Readme file excluding the results & conclusion
  - **2. Milestone (December 7, midnight)**
    - The results of the first experiment
  - **3. Milestone (January 4, midnight)**
    - Final report (Readme file)
    - Repo with all code & trained models

Good to know: Bloom

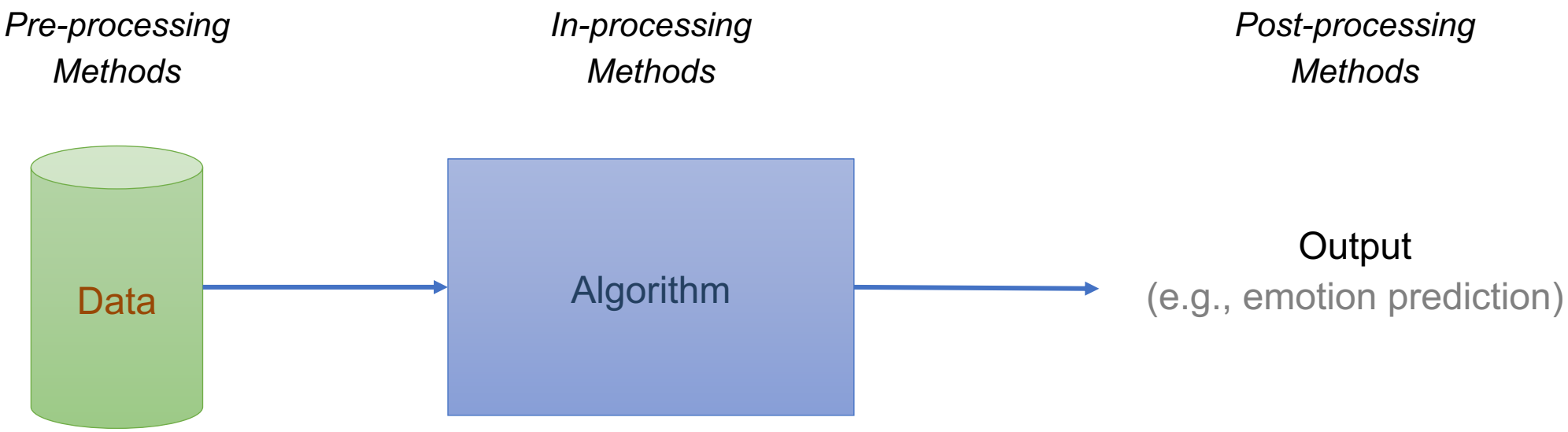
https://www.anthropic.com/research/bloom



Good to know: Perspective

https://perspectiveapi.com/

Fairness Algorithms



Fairness Algorithms

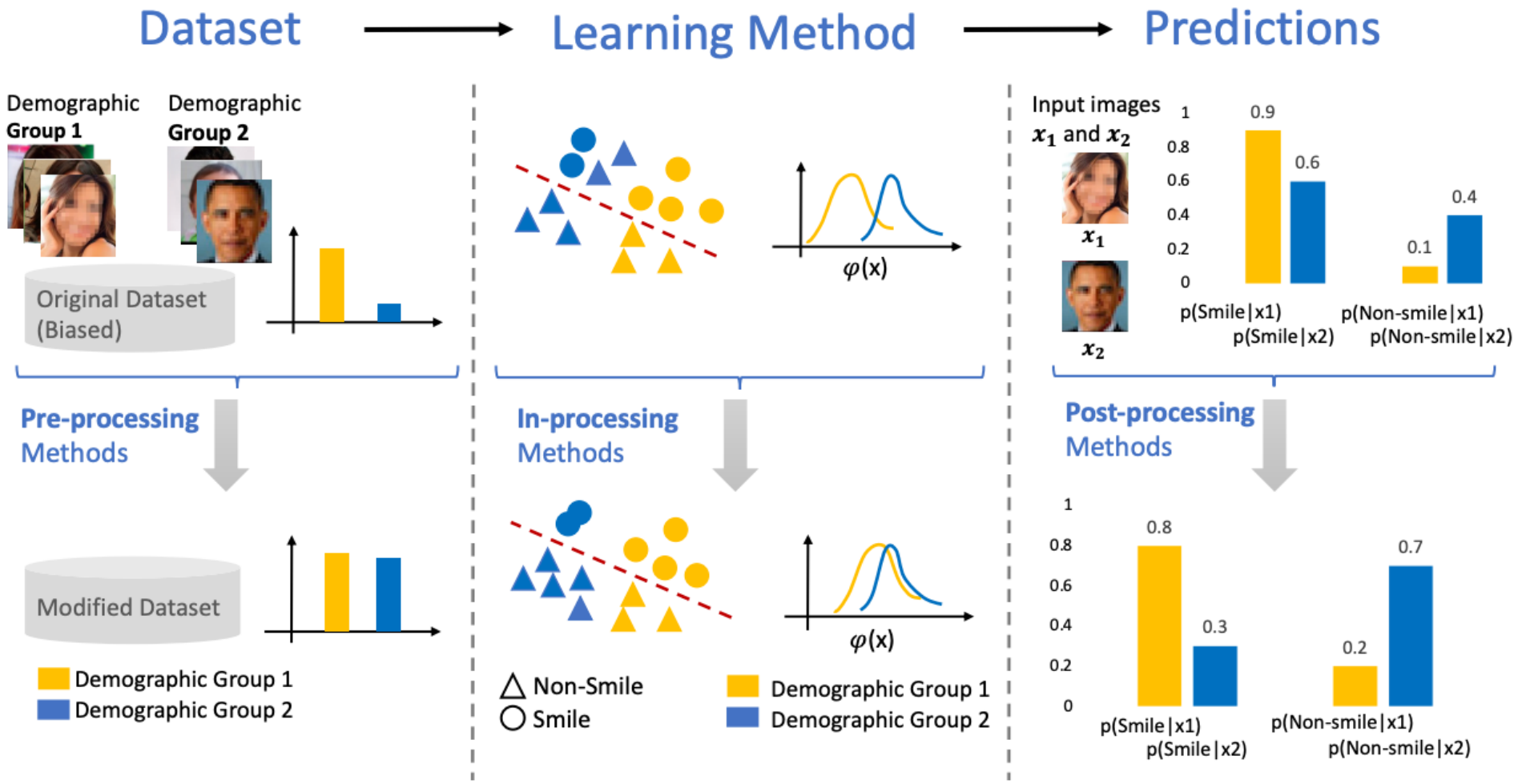


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Pre-processing Methods

Table 4: Credit score distribution by ethnicity

Race or ethnicity	Samples with both score and outcome
White	133,165
Black	18,274
Hispanic	14,702
Asian	7,906
Total	174,047

https://fairmlbook.org/pdf/fairmlbook.pdf

Dataset-level Methods

1. Prepare a new dataset
2. Use sampling methods to oversample or undersample the dataset
3. Augment the datasets with small perturbations on the existing samples
4. Generate synthetic data
5. Fairness through unawareness

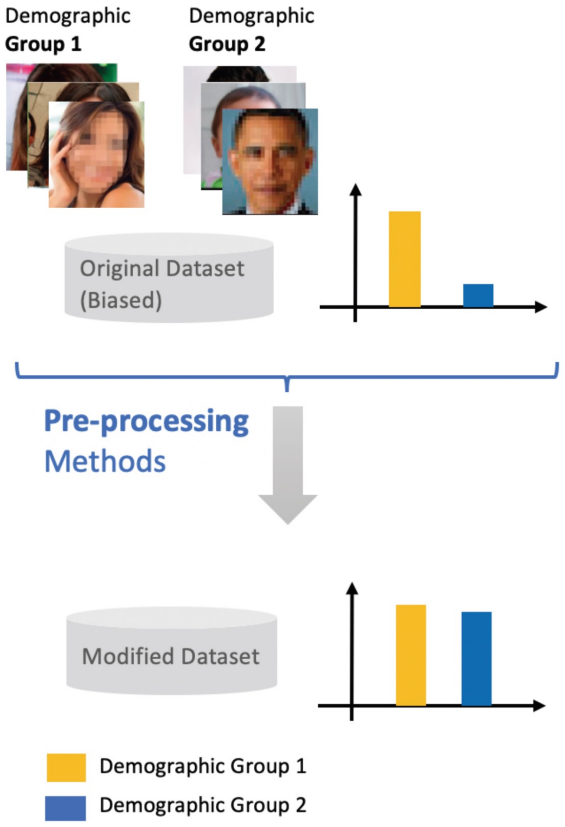


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Preprocessing Methods

Dataset-level Methods

1. Prepare a new dataset
2. Use sampling methods to oversample or undersample the dataset
3. Augment the datasets with small perturbations on the existing samples
4. Generate synthetic data
5. Fairness through unawareness

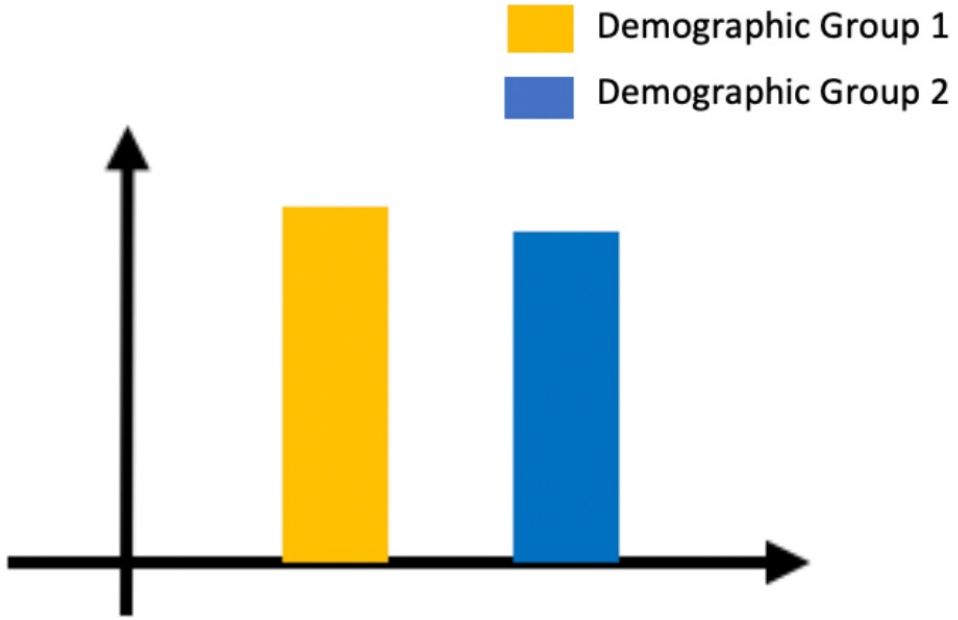


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Preprocessing Methods

Dataset-level Methods

1. Prepare a new dataset
2. Use sampling methods to oversample or undersample the dataset
3. Augment the datasets with small perturbations on the existing samples
4. Generate synthetic data
5. Fairness through unawareness

Table 4: Credit score distribution by ethnicity

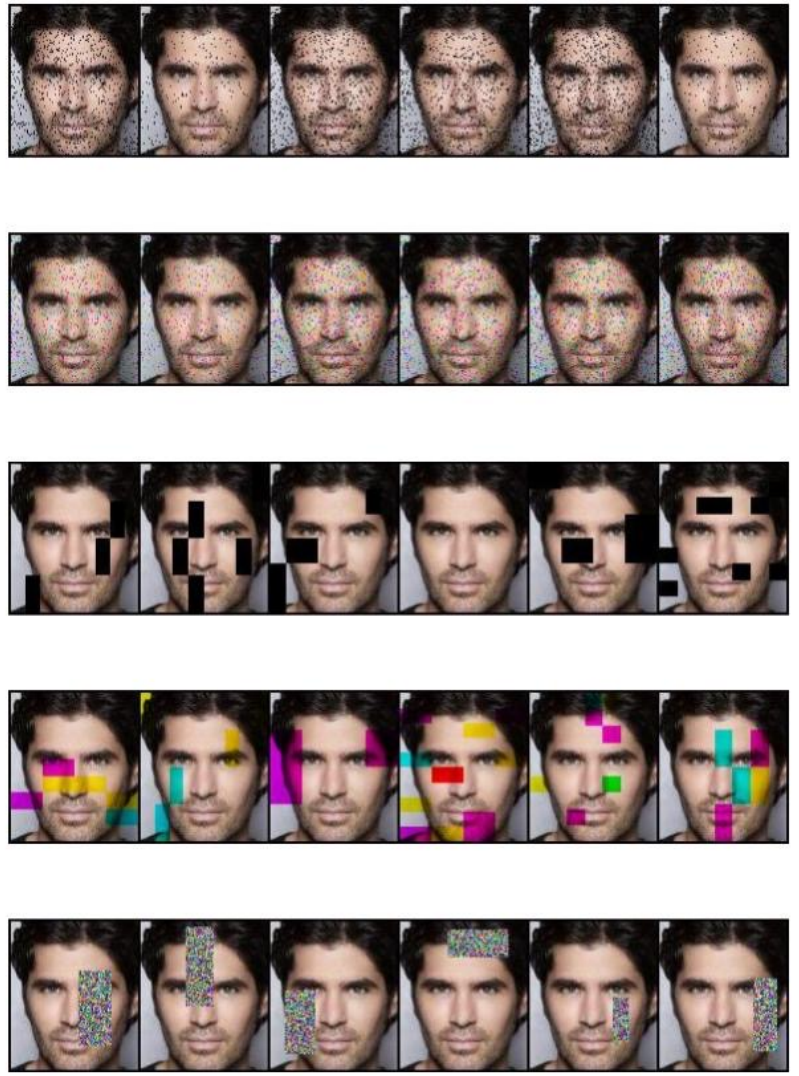
Race or ethnicity	Samples with both score and outcome
White	133,165
Black	18,274
Hispanic	14,702
Asian	7,906
Total	174,047

https://fairmlbook.org/pdf/fairmlbook.pdf

Fairness Algorithms
Preprocessing Methods

Dataset-level Methods

1. Prepare a new dataset
2. Use sampling methods to oversample or undersample the dataset
3. Augment the datasets with small perturbations on the existing samples
4. Generate synthetic data
5. Fairness through unawareness



https://melgor.github.io/blcv.github.io/static/2018/02/27/demystifying-face-recognition-v-data-augmentation/index.html

Fairness Algorithms
Preprocessing Methods

Dataset-level Methods

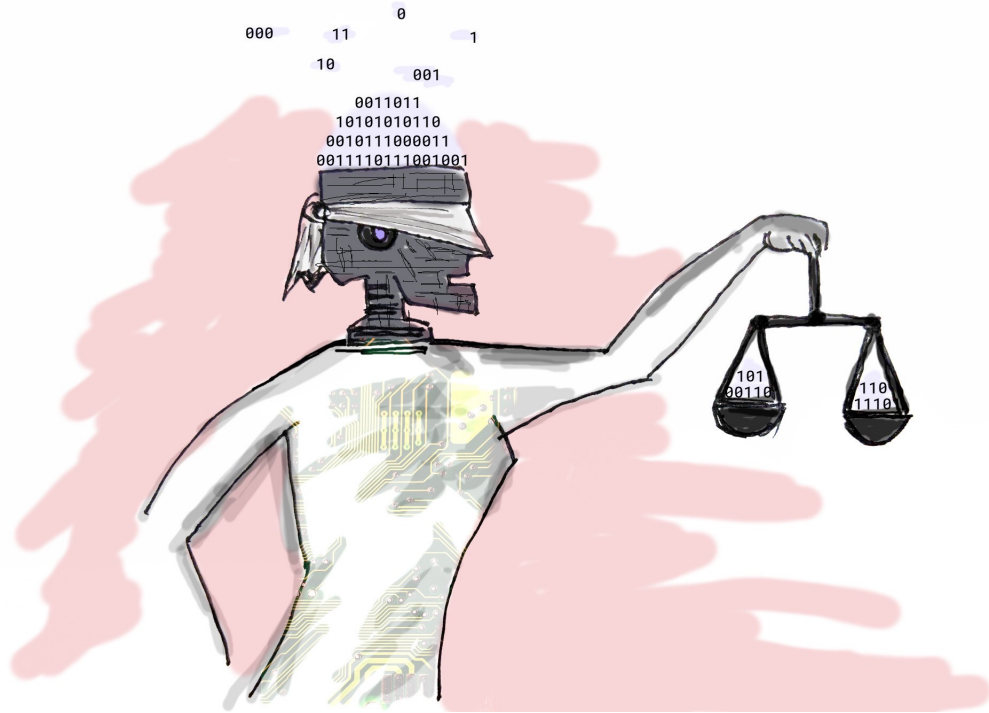
1. Prepare a new dataset
2. Use sampling methods to oversample or undersample the dataset
3. Augment the datasets with small perturbations on the existing samples
4. Generate synthetic data
5. Fairness through unawareness

Dogan, Y., & Keles, H. Y. (2020). Semi-supervised image attribute editing using generative adversarial networks. Neurocomputing, 401, 338-352.

Fairness Algorithms
Preprocessing Methods

Dataset-level Methods

1. Prepare a new dataset
2. Use sampling methods to oversample or undersample the dataset
3. Augment the datasets with small perturbations on the existing samples
4. Generate synthetic data
5. Fairness through unawareness



https://towardsdatascience.com/bias-and-algorithmic-fairness-10f0805edc2b

Fairness Algorithms
Preprocessing Methods

Advantages:

- Algorithm independent
- Easy to apply

Disadvantages/Challenges

- Sub-sampling can remove critical data
- Over-sampling can cause overfitting
- Balancing samples across demographic groups may not guarantee fairness (Wang et al., 2019; Cheong et al., 2023)

Recommendations

- The use of synthetic samples from generative models are promising (but generative models can also have bias)

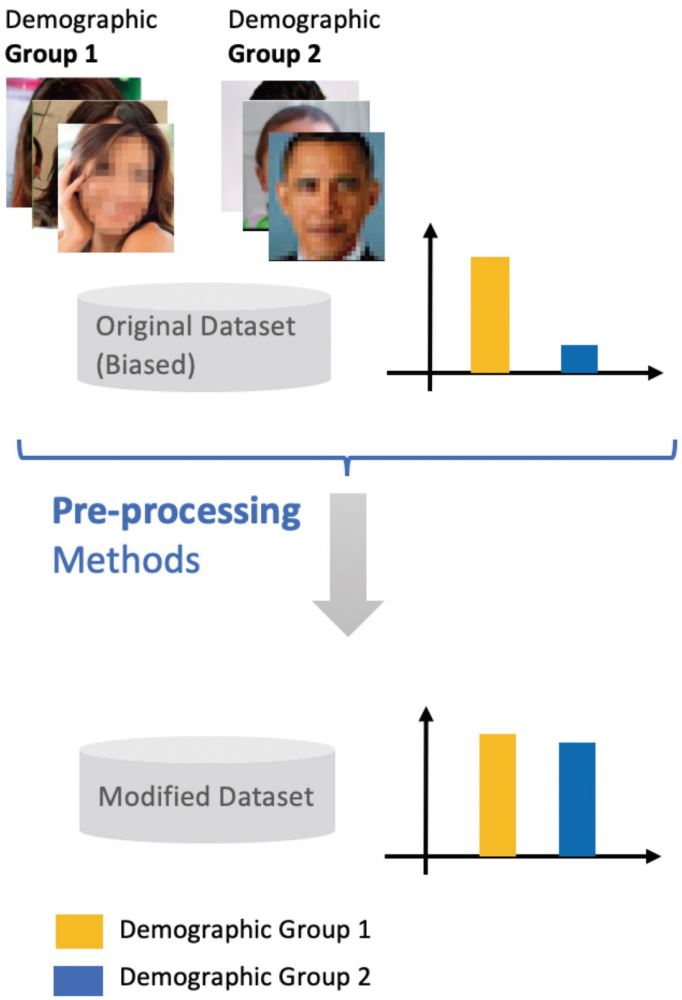


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Wang, T., Zhao, J., Yatskar, M., Chang, K., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5309-5318 (2019)
Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. International Joint Conferences on Artificial Intelligence Organization.

Fairness Algorithms
Inprocessing Methods

1. Cost-sensitive learning
2. Domain adaptation
3. Disentanglement

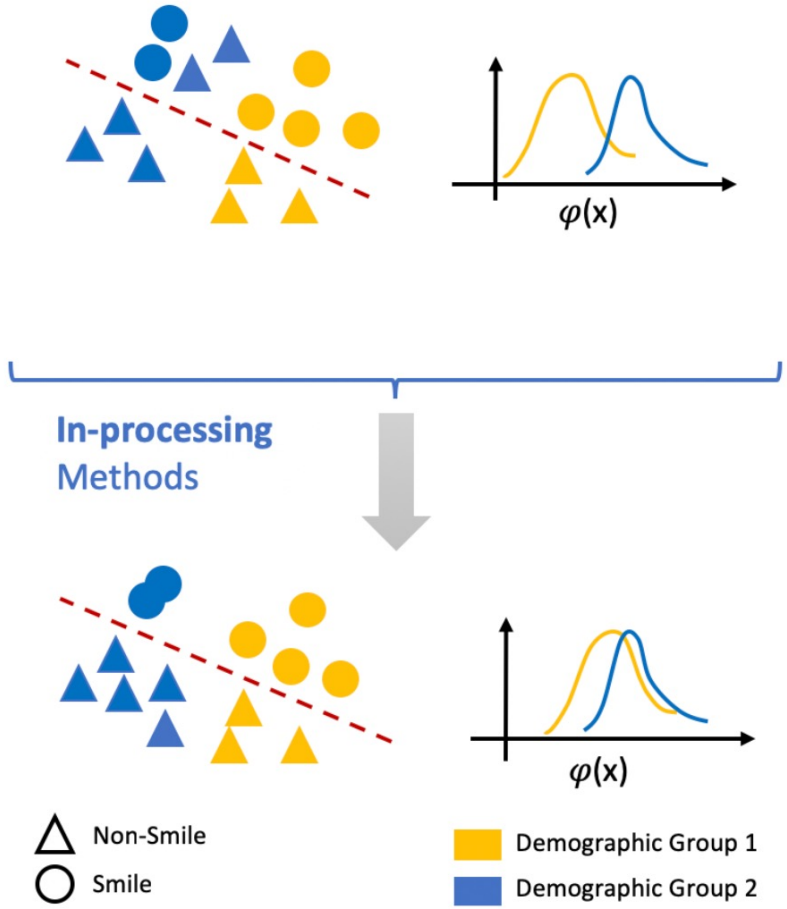


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Inprocessing Methods

1. Cost-sensitive learning
2. Domain adaptation
3. Disentanglement

$$w_{s(x)} \cdot \mathcal{L}(x)$$

For example:

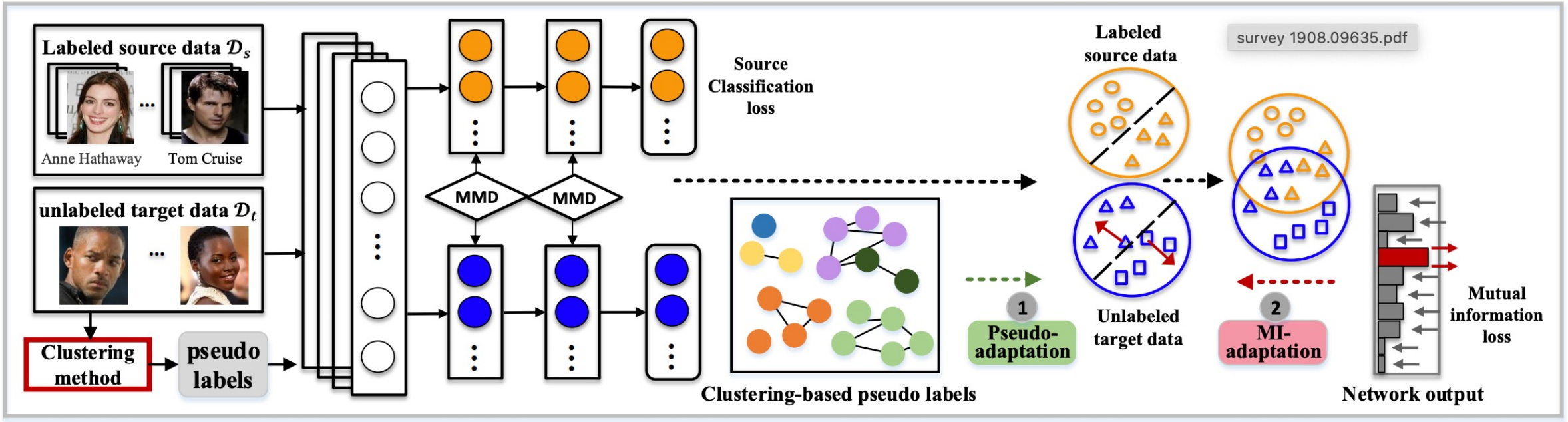
If $N_{female} = \frac{1}{3} N_{male}$ then

$$w_{female} = 3 \cdot w_{male}$$

Fairness Algorithms
Inprocessing Methods

1. Cost-sensitive learning
2. Domain adaptation
3. Disentanglement

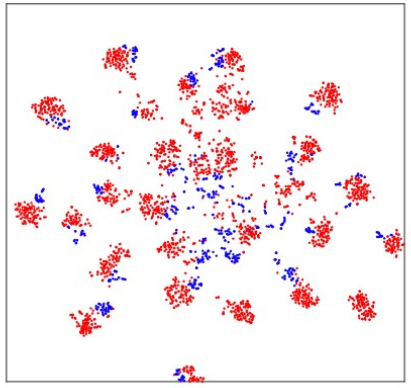
https://arxiv.org/pdf/1812.00194.pdf



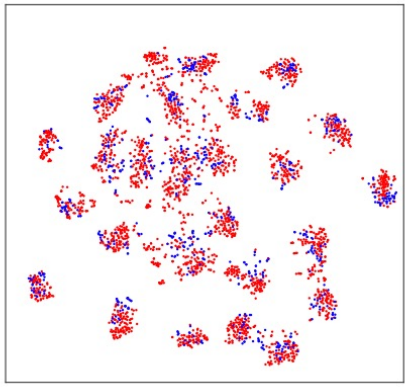
Fairness Algorithms
Inprocessing Methods

1. Cost-sensitive learning
2. Domain adaptation
3. Disentanglement

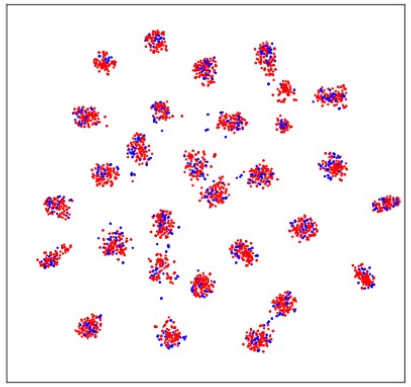
https://arxiv.org/pdf/1705.10667.pdf



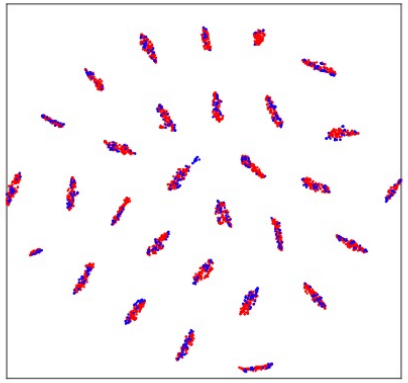
(a) ResNet



(b) DANN



(c) CDAN-**f**



(d) CDAN-**fg**

Fairness Algorithms
Inprocessing Methods

1. Cost-sensitive learning
2. Domain adaptation
3. Disentanglement

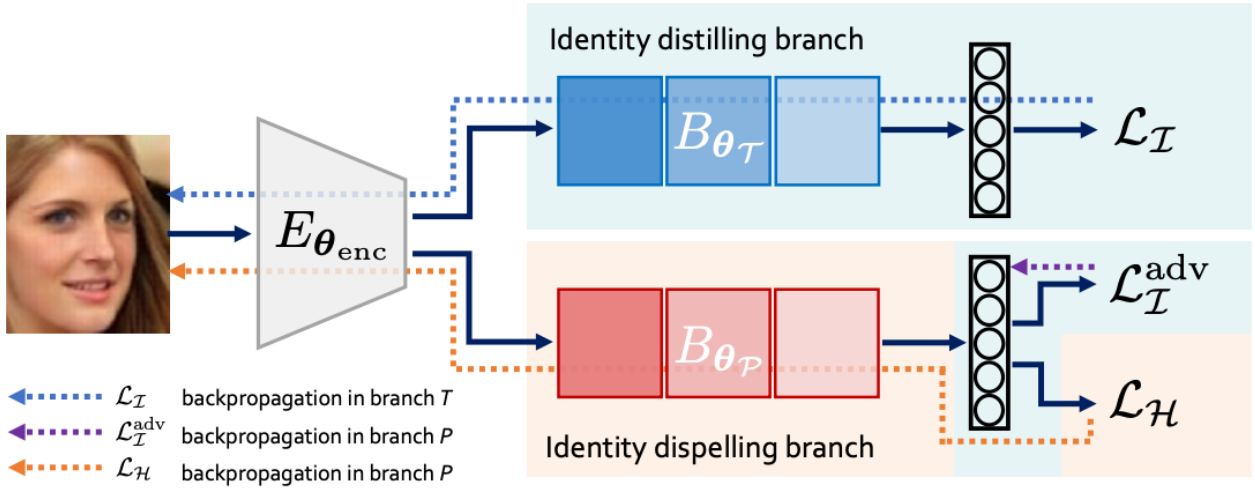


Figure 3. The encoding module for extracting disentangled face features.

https://arxiv.org/pdf/1804.03487.pdf

Fairness Algorithms
Inprocessing Methods

Advantages:

- More effective than preprocessing methods
- Many options to intervene at and mitigate bias

Disadvantages/Challenges

- Method-specific
- Requires expertise

Recommendations

- Hybrid solutions combining different approaches can be promising

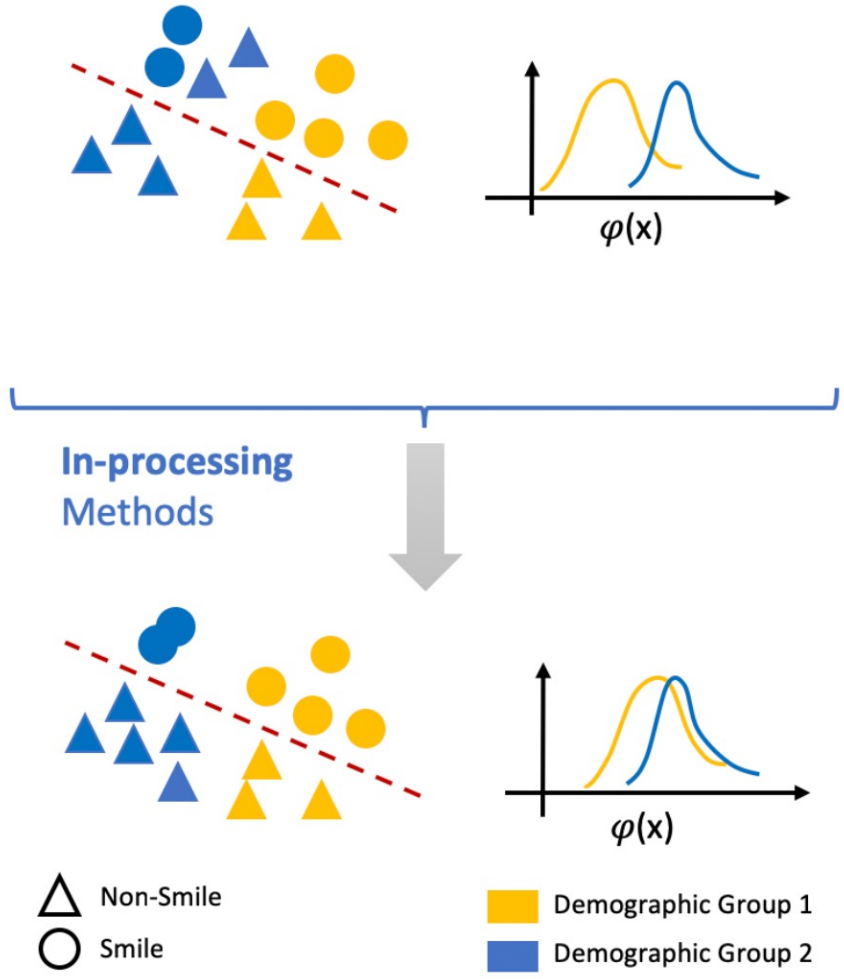


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Postprocessing Methods

- Update model predictions (scores)

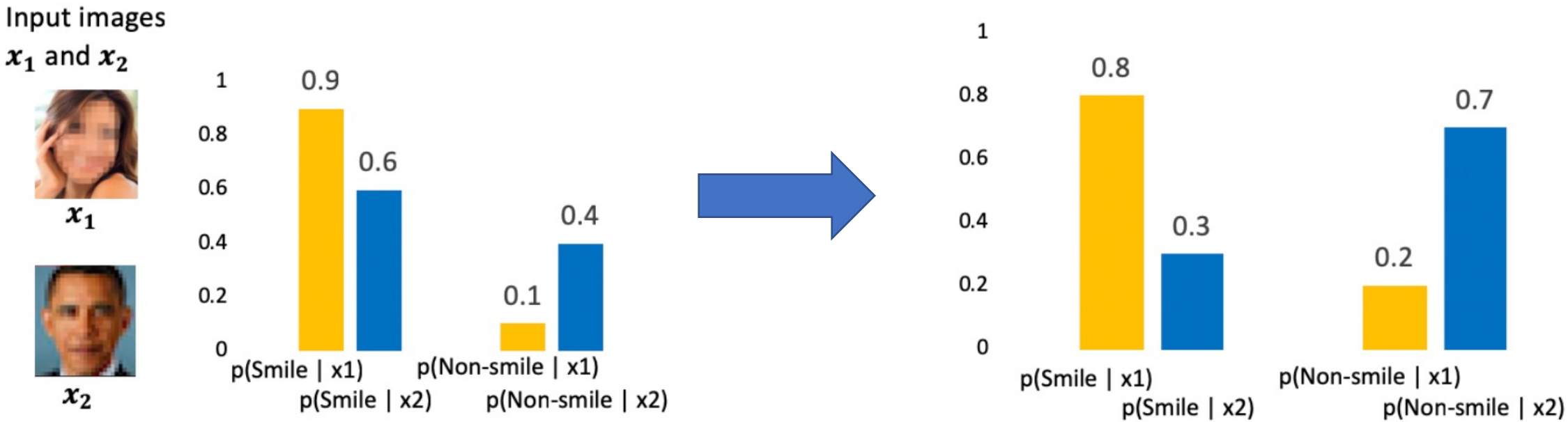


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Postprocessing Methods

Advantages:

- Does not depend on the method
- No knowledge required about the method
- Easy to use for non-experts
- Readily available as frameworks

Disadvantages/Challenges

- Correcting outputs may fall insufficient for mitigating bias
- For explainable and interpretable fairness, we may need to intervene at the model
- May require additional training method

Recommendations:

- Before using pre-trained networks in practice, these methods should be tried.

Fairness Algorithms
An Overview

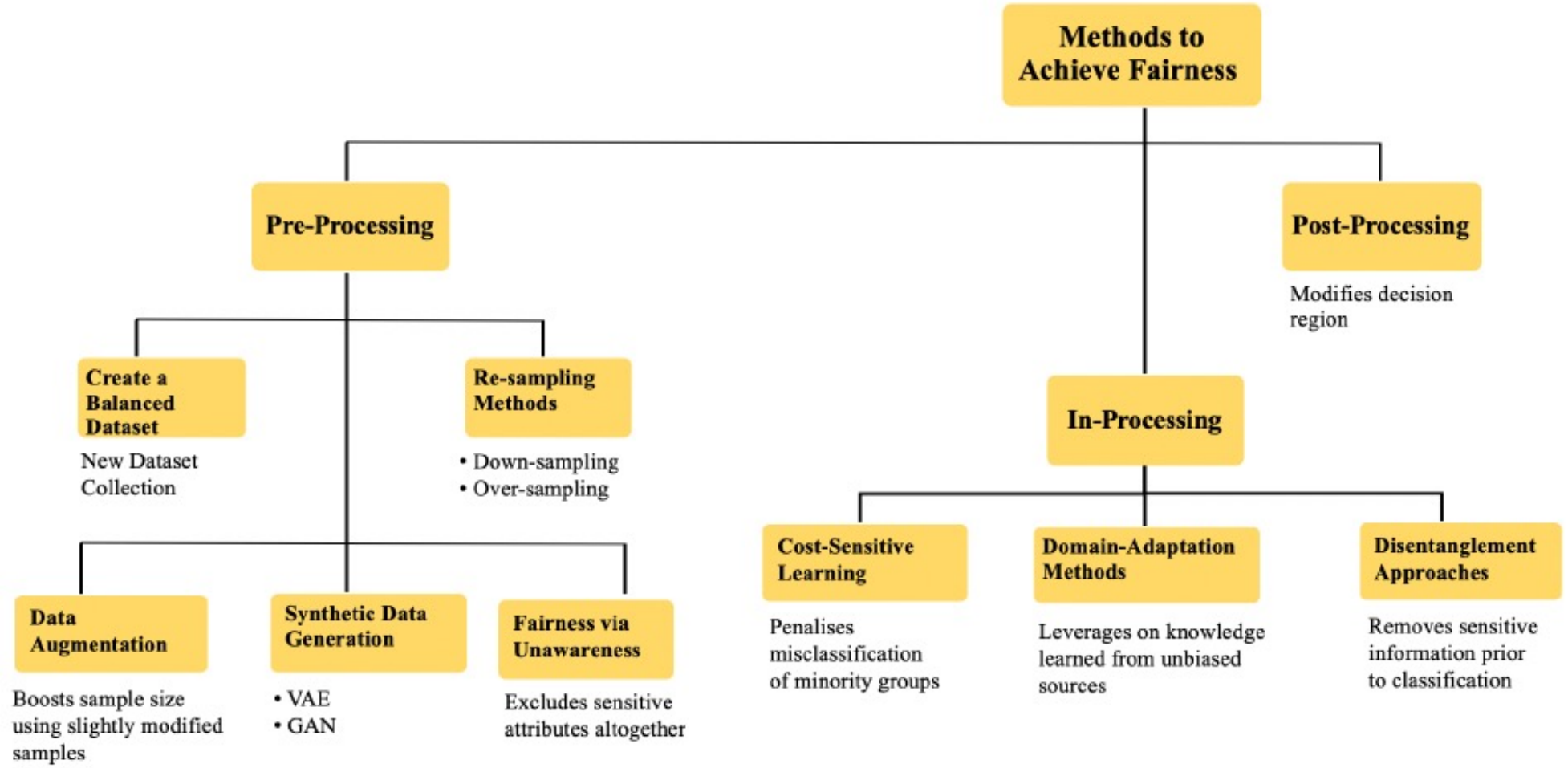


Fig: J. Cheong, S. Kalkan, H. Gunes, "The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing", IEEE Signal Processing Magazine, 2021.

Fairness Algorithms
Tools

- AI Fairness 360 (AIF360): https://github.com/Trusted-AI/AIF360
- Sci-kit fairness: https://scikit-fairness.readthedocs.io/en/latest/
- What-if tool: https://pair-code.github.io/what-if-tool/
- Pymetrics: https://www.pymetrics.ai/assessments

- Lee, M. S. A., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Fairness in LLMs

Resources:

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. ACM SIGKDD explorations newsletter, 26(1), 34-48.

Biases in LLMs

Table 1
Taxonomy of social biases in NLP. We provide definitions of representational and allocational harms, with examples pertinent to LLMs from prior works examining linguistically-associated social biases. Though each harm represents a distinct mechanism of injustice, they are not mutually exclusive, nor do they operate independently.

Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group e.g., "`Whore`" *conveys hostile and contemptuous female expectations* (Beukeboom and Burgers 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations e.g., *AAE* like* "`he woke af`" *is misclassified as not English more often than SAE† equivalents* (Blodgett and O'Connor 2017)
Erasure	Omission or invisibility of the language and experiences of a social group e.g., "`All lives matter`" *in response to* "`Black lives matter`" *implies colorblindness that minimizes systemic racism* (Blodgett 2021)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups e.g., "`Both genders`" *excludes non-binary identities* (Bender et al. 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group e.g., *Responding* "`I'm sorry to hear that`" *to* "`I'm an autistic dad`" *conveys a negative misrepresentation of autism* (Smith et al. 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group e.g., *Associating* "`Muslim`" *with* "`terrorist`" *perpetuates negative violent stereotypes* (Abid, Farooqi, and Zou 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group e.g., "`I hate Latinos`" *is disrespectful and hateful* (Dixon et al. 2018)

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

Biases in LLMs

ALLOCATIONAL HARMS	Disparate distribution of resources or opportunities between social groups
Direct discrimination	Disparate treatment due explicitly to membership of a social group e.g., *LLM-aided resume screening may preserve hiring inequities* (Ferrara 2023)
Indirect discrimination	Disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors e.g., *LLM-aided healthcare tools may use proxies associated with demographic factors that exacerbate inequities in patient care* (Ferrara 2023)

*African-American English; †Standard American English.

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

Biases in LLMs

- Biases in NLP tasks
  - Text generation, machine translation, information retrieval, question-answering, …
- Biases in development and deployment
  - Training data
  - Model
  - Evaluation
  - Deployment

Metrics of Bias in LLMs

Embedding-based

- **Word embeddings**: Compute distances in embedding space
- **Sentence embeddings**: Adapt to contextualized embeddings

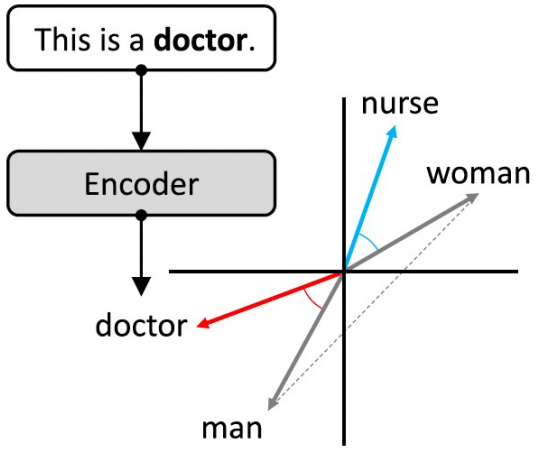


Figure 3
Example embedding-based metrics (§ 3.3). Sentence-level encoders produce sentence embeddings that can be assessed for bias. Embedding-based metrics use cosine similarity to compare words like "doctor" to social group terms like "man." Unbiased embeddings should have similar cosine similarity to opposing social group terms.

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

Metrics of Bias in LLMs

Probability-based metrics

- **Masked token**: Compare fill-in-the-blank probabilities
- **Pseudo-log-likelihood**: Compare likelihoods between sentences

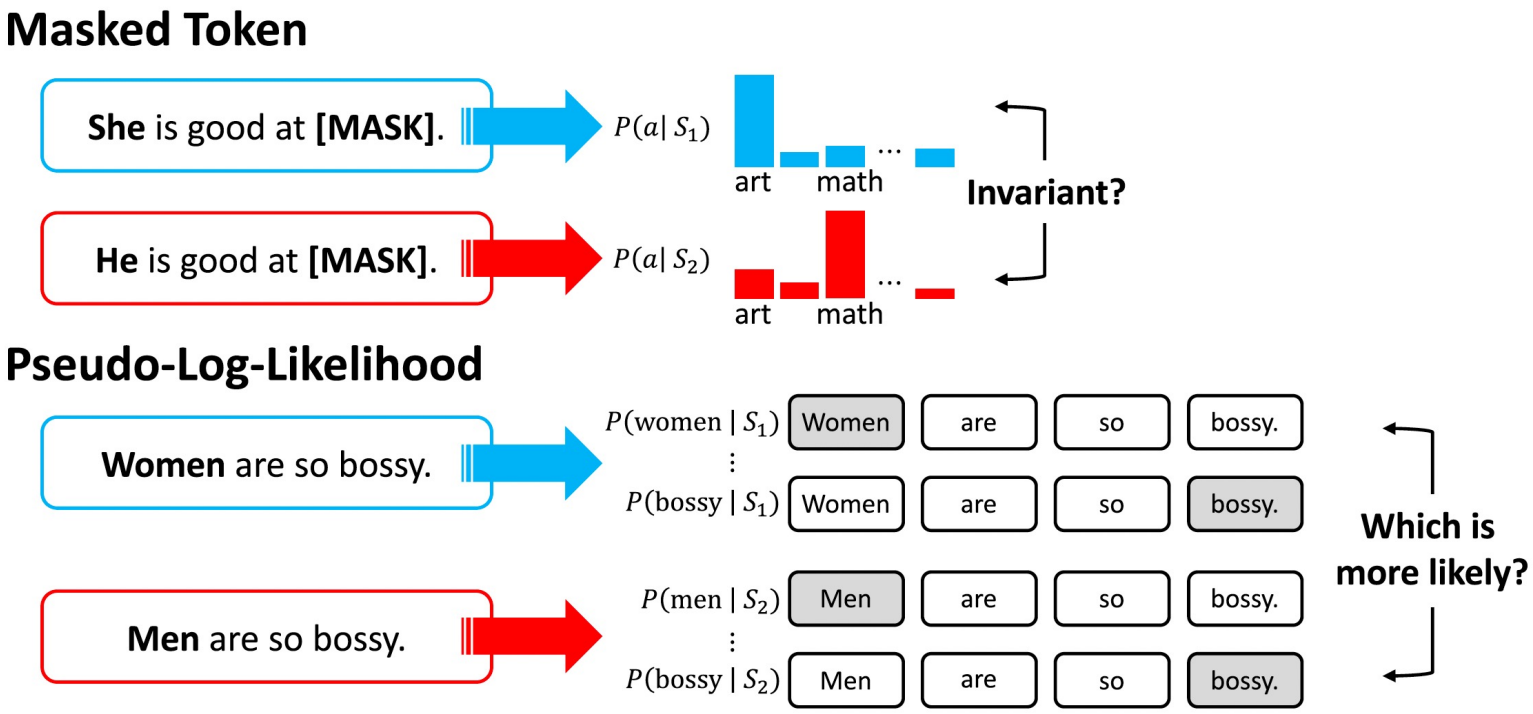


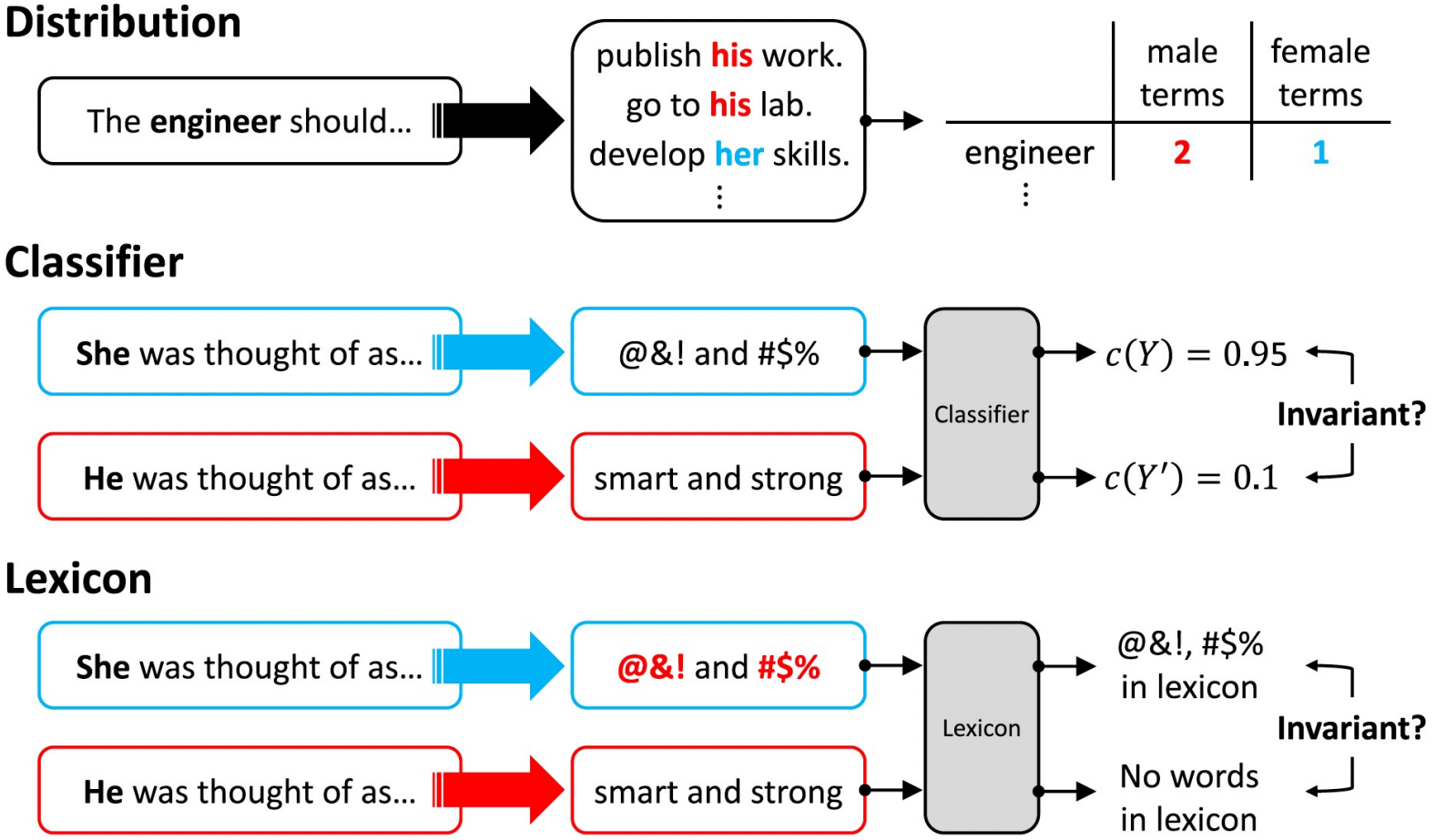
Figure 4
Example probability-based metrics (§ 3.4). We illustrate two classes of probability-based metrics: masked token metrics and pseudo-log-likelihood metrics. Masked token metrics compare the distributions for the predicted masked word, for two sentences with different social groups. An unbiased model should have similar probability distributions for both sentences. Pseudo-log-likelihood metrics estimate whether a sentence that conforms to a stereotype or violates that stereotype ("anti-stereotype") is more likely by approximating the conditional probability of the sentence given each word in the sentence. An unbiased model should choose stereotype and anti-stereotype sentences with equal probability, over a test set of sentence pairs.

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

Metrics of Bias in LLMs

Generated text based metrics

- **Distribution**: Compare distribution of co-occurrences
- **Classifier**: Use an auxiliary classifier (e.g., existing sentiment or toxicity classifiers)
- **Lexicon**: Compare each word in the output to a predefined lexicon

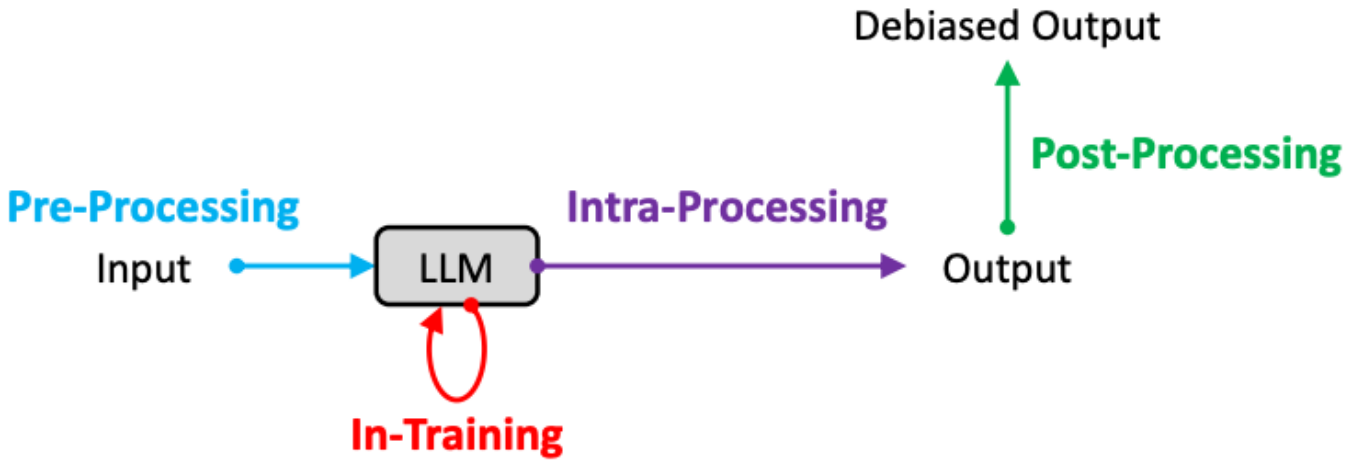


Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

Table 5
Taxonomy of techniques for bias mitigation in LLMs. We categorize bias mitigation techniques by the stage at which they intervene. For an illustration of each mitigation stage, as well as inputs and outputs to each stage, see Figure 6.

Mitigation Stage	Mechanism
PRE-PROCESSING (§ 5.1)	Data Augmentation (§ 5.1.1) Data Filtering & Reweighting (§ 5.1.2) Data Generation (§ 5.1.3) Instruction Tuning (§ 5.1.4) Projection-based Mitigation (§ 5.1.5)
IN-TRAINING (§ 5.2)	Architecture Modification (§ 5.2.1) Loss Function Modification (§ 5.2.2) Selective Parameter Updating (§ 5.2.3) Filtering Model Parameters (§ 5.2.4)
INTRA-PROCESSING (§ 5.3)	Decoding Strategy Modification (§ 5.3.1) Weight Redistribution (§ 5.3.2) Modular Debiasing Networks (§ 5.3.3)
POST-PROCESSING (§ 5.4)	Rewriting (§ 5.4.1)

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.



(a)

(b)

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

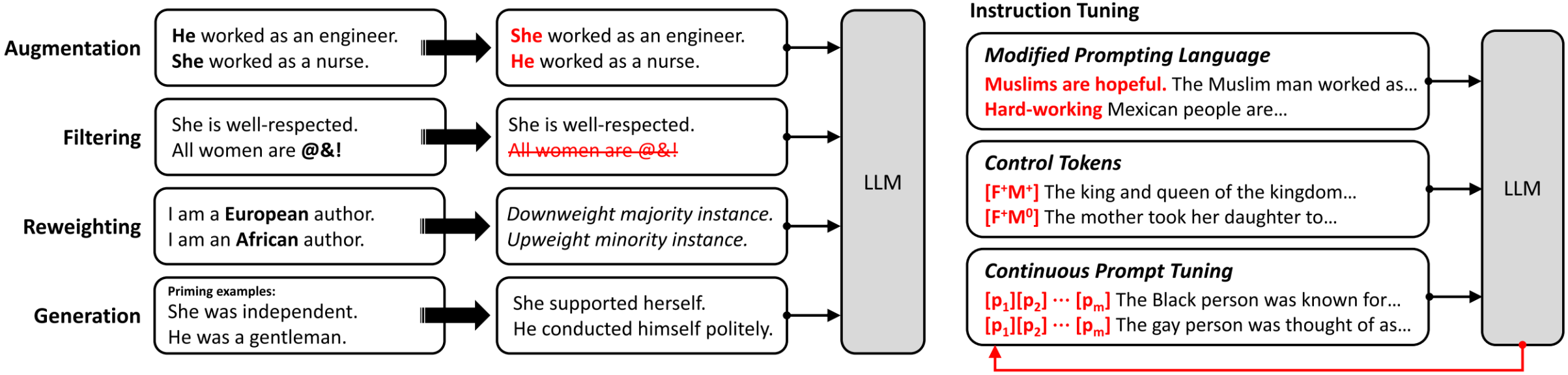


Figure 7

Example Pre-Processing Mitigation Techniques (§ 5.1). We provide examples of data augmentation, filtering, re-weighting, and generation on the left, as well as various types of instruction tuning on the right. The first example illustrates counterfactual data augmentation, flipping binary gender terms to their opposites. Data filtering illustrates the removal of biased instances, such as derogatory language (denoted as "@&!"). Reweighting demonstrates how instances representing underrepresented or minority instances may be upweighted for training. Data generation shows how new examples may be constructed by human or machine writers based on priming examples that illustrate the desired standards for the new data. Instruction

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

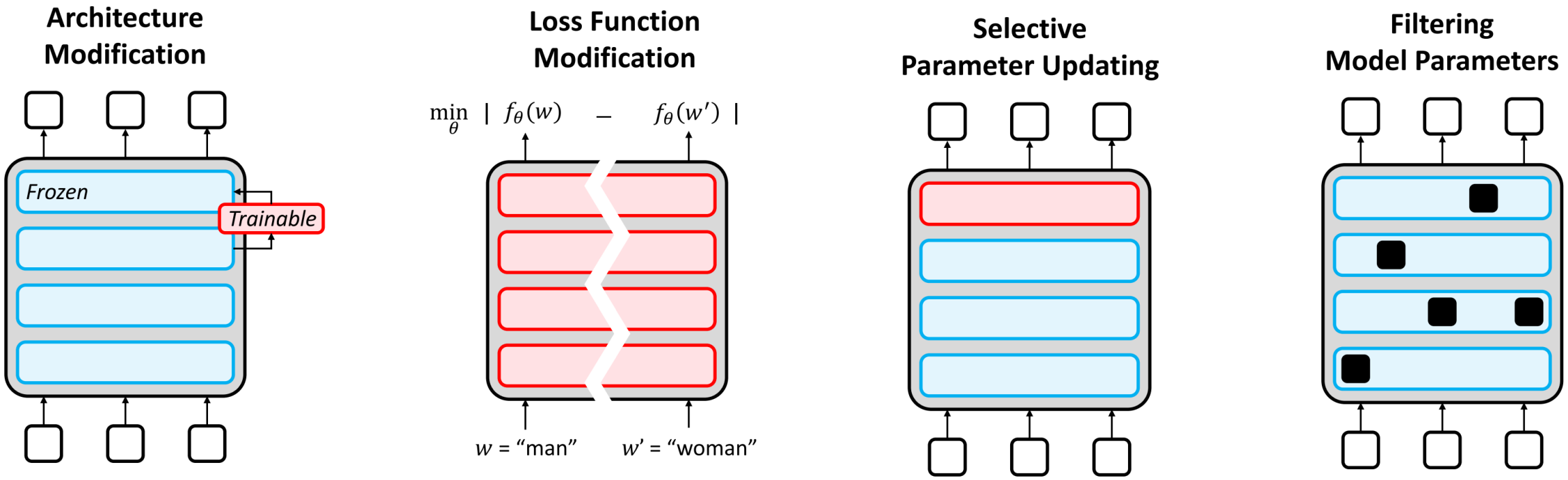


Figure 8
Example in-training mitigation techniques (§ 5.2). We illustrate four classes of methods that modify model parameters during training. Architecture modifications change the configuration of the model, such as adding new trainable parameters with adapter modules as done in this example (Lauscher, Lueken, and Glavaš 2021). Loss function modifications introduce a new optimization objective, such as equalizing the embeddings or predicted probabilities of counterfactual tokens or sentences. Selective parameter updates freeze the majority of the weights and only tune a select few during fine-tuning to minimize forgetting of pre-trained language understanding. Filtering model parameters, in contrast, freezes all pre-trained weights and selectively prunes some based on a debiasing objective.

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

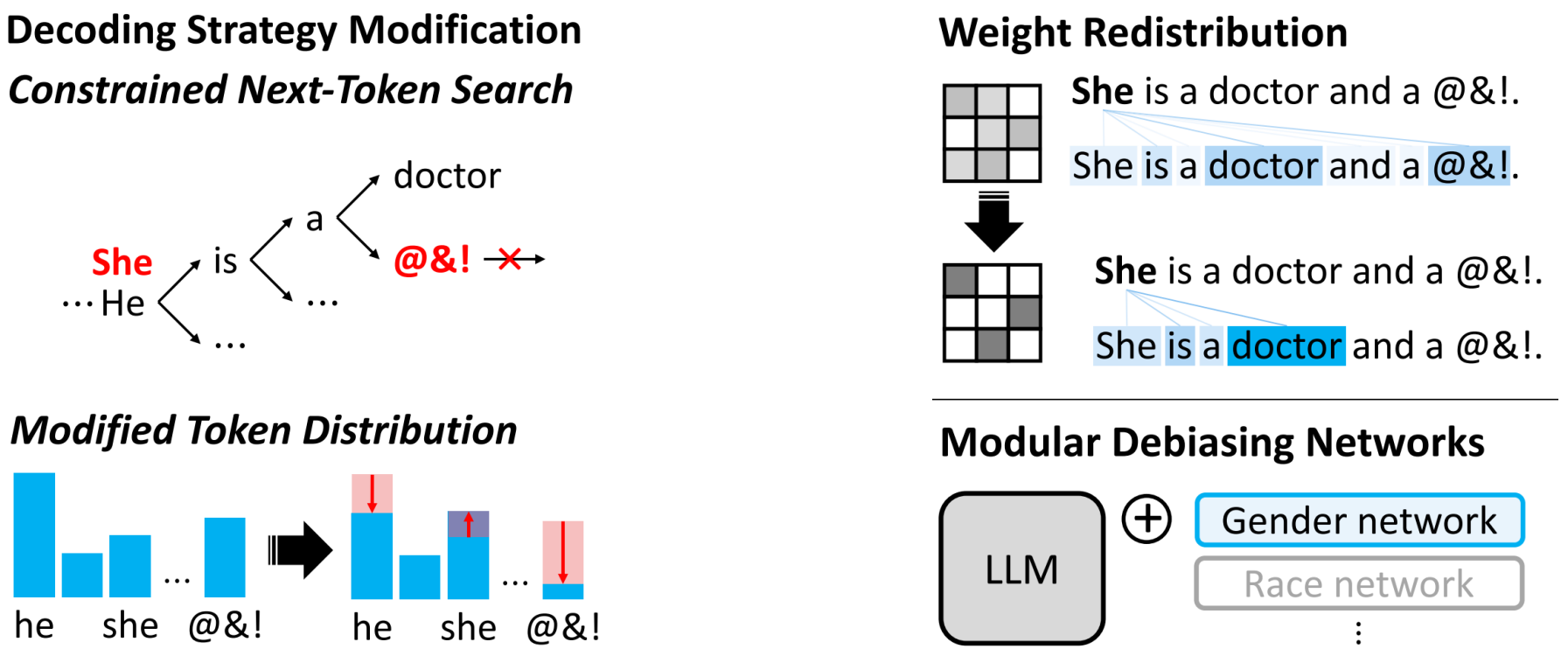


Figure 9
Example intra-processing mitigation techniques (§ 5.3). We show several methods that modify a model's behavior without training or fine-tuning. Constrained next-token search may prohibit certain outputs during beam search (e.g., a derogatory term "@&!," in this example), or generate and rerank alternative outputs (e.g., "he" replaced with "she"). Modified token distribution redistributes next-word probabilities to produce more diverse outputs and avoid biased tokens. Weight distribution, in this example, illustrates how post hoc modifications to attention matrices may narrow focus to less stereotypical tokens (Zayed et al. 2023b). Modular debiasing networks fuse the main LLM with stand-alone networks that can remove specific dimensions of bias, such as gender or racial bias.

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.

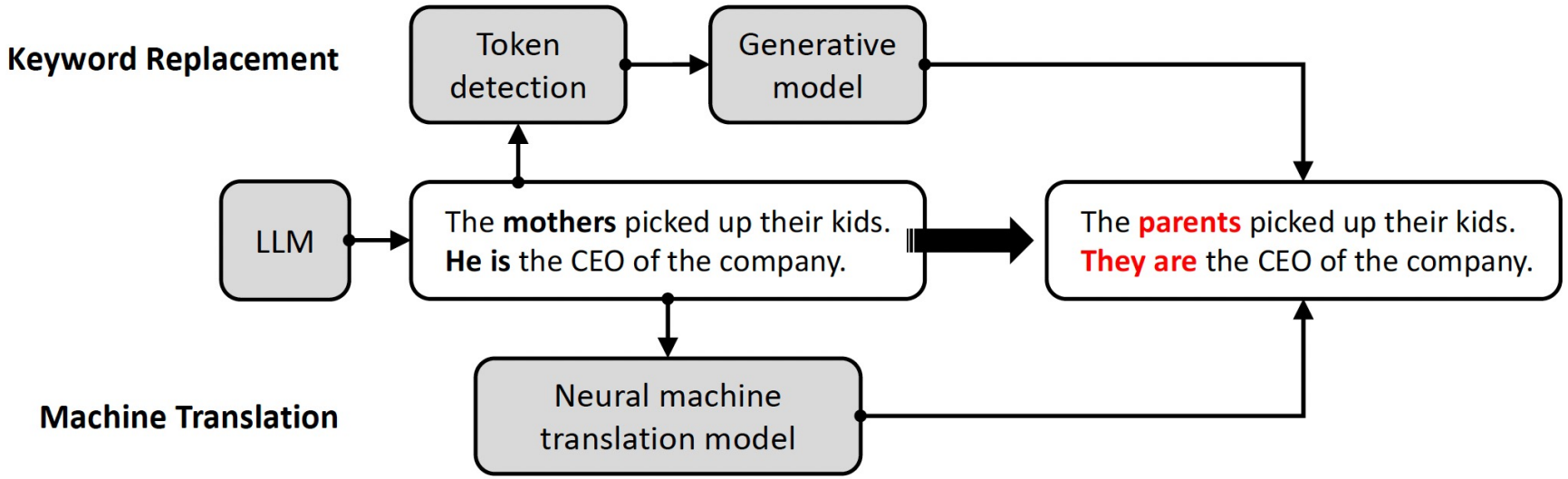


Figure 10
Example post-processing mitigation techniques (§ 5.4). We illustrate how post-processing methods can replace a gendered output with a gender-neutral version. Keyword replacement methods first identify protected attribute terms (i.e., "mothers," "he"), and then generate an alternative output. Machine translation methods train a neural machine translator on a parallel biased-unbiased corpus and feed the original output into the model to produce an unbiased output.

Gallegos et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097-1179.