Responsible Machine Learning

Lecture 13: Algorithmic Fairness in the Real World - Part 1

CS 4973-05

Fall 2023

Instructors: Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University, Boston, MA

Northeastern University

Background

Research Questions

Uncertain Inference

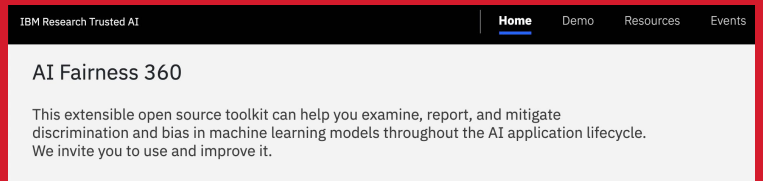
Adversarial Attacks

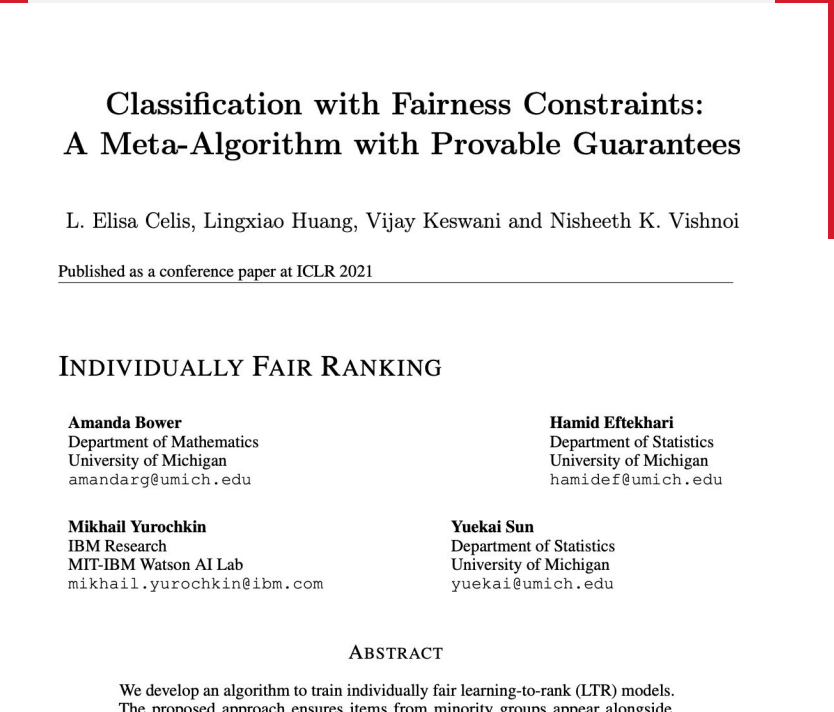
Algorithmic Bias

- Well understood that **ML can go horribly wrong**
- Famous example of **ProPublica's analysis of the Northpointe algorithm** which was shown to grant bail at a higher rate to white defendants than blacks.
- Innumerable other examples from financial algorithms to facial recognition to almost every sensitive sphere where ML is used.

Northeastern University

Algorithmic Debiasing





Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees

L. Elisa Celis, Lingxiao Huang, Vijay Keswani and Nisheeth K. Vishnoi

Published as a conference paper at ICLR 2021

INDIVIDUALLY FAIR RANKING

Amanda Bower
Department of Mathematics
University of Michigan
amandarg@umich.edu

Hamid Eftekhari
Department of Statistics
University of Michigan
hamidef@umich.edu

Mikhail Yurochkin
IBM Research
MIT-IBM Watson AI Lab
mikhail.yurochkin@ibm.com

Yuekai Sun
Department of Statistics
University of Michigan
yuekai@umich.edu

ABSTRACT

We develop an algorithm to train individually fair learning-to-rank (LTR) models.

- As a response to algorithmic bias, there is **algorithmic "debiasing"**.
- Both industrial solutions (like **IBM AI Fairness 360**), and algorithms presented in Machine Learning **Research** exist out there.
- Techniques include different sampling rates for different groups, constrained learning, and group sensitive reordering of ranked lists.

Northeastern University

Cool! Is ML bias a solved problem then?

There exist some real world problems...

- Not many transparent real-world audits
- Intersectionality of bias
- Models may become unfair in a live deployment over time
- Missing demographic information
- Adversarial attackers can make the algorithm more unfair
- Decisions are not always correlated with outcomes.
- And many more!

This is the part of the course where you read work done by your professor!

Background

Research Questions

Uncertain Inference

Adversarial Attacks

Research Questions

It is becoming clear that while a lot of progress has been made in the fields of fairness, accountability, transparency, and ethics of ML algorithms, there is still a **considerable amount of technical challenges** involved before this work can translate from controlled research settings into the real world.

Research Questions

- **RQ1:** How does noise in demographic information as an input to a fair ML algorithm adversely impact the intended fairness of the outcomes for different subgroups?
- **RQ2:** How can fair ML models be attacked by adversarial actors to create even more unfairness?
- **RQ3:** In fair ML techniques that deliberately do not use protected attributes, how do their theoretical guarantees hold up in real life when compared against actual ground truth?
- **RQ4:** Do fair ML models, once deployed in a production system, continue to remain fair in the face of changing data and feature-output relationships? If so, how can such unfairness be measured and mitigated?

Research Questions

- **RQ1:** How does noise in demographic information as an input to a fair ML algorithm adversely impact the intended fairness of the outcomes for different subgroups?
- **RQ2:** How can fair ML models be attacked by adversarial actors to create even more unfairness?
- **RQ3:** In fair ML techniques that deliberately do not use protected attributes, how do their theoretical guarantees hold up in real life when compared against actual ground truth?
- **RQ4:** Do fair ML models, once deployed in a production system, continue to remain fair in the face of changing data and feature-output relationships? If so, how can such unfairness be measured and mitigated?

Research Questions

- **RQ1:** How does noise in demographic information as an input to a fair ML algorithm adversely impact the intended fairness of the outcomes for different subgroups?
- **RQ2:** How can fair ML models be attacked by adversarial actors to create even more unfairness?
- **RQ3:** In fair ML techniques that deliberately do not use protected attributes, how do their theoretical guarantees hold up in real life when compared against actual ground truth?
- **RQ4:** Do fair ML models, once deployed in a production system, continue to remain fair in the face of changing data and feature-output relationships? If so, how can such unfairness be measured and mitigated?

Papers accepted at SIGIR 2021 and FAccT 2022!

Northeastern University

Background

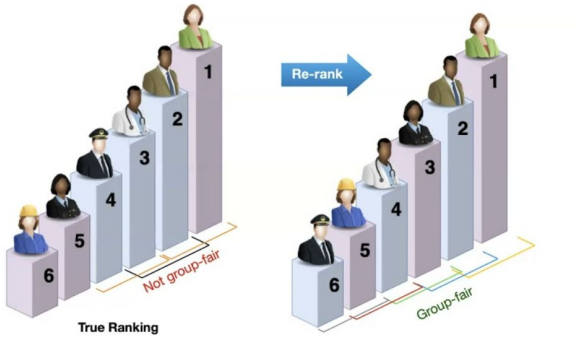
Research Questions

Uncertain Inference

Adversarial Attacks

Chapter 1

When Fair Ranking Meets Uncertain Inference

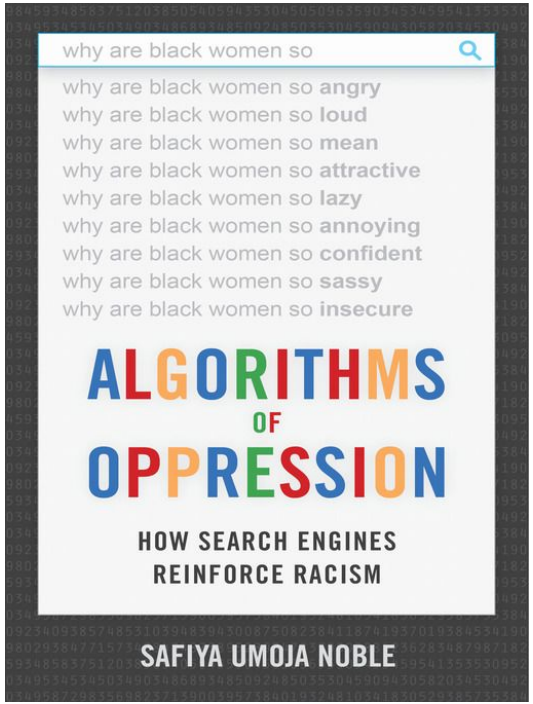


Northeastern University

Bias in Ranking

Ranking Algorithms, like other applications of Machine learning, are not immune to the **insidious effects** of **learned social biases** that are then **amplified.**

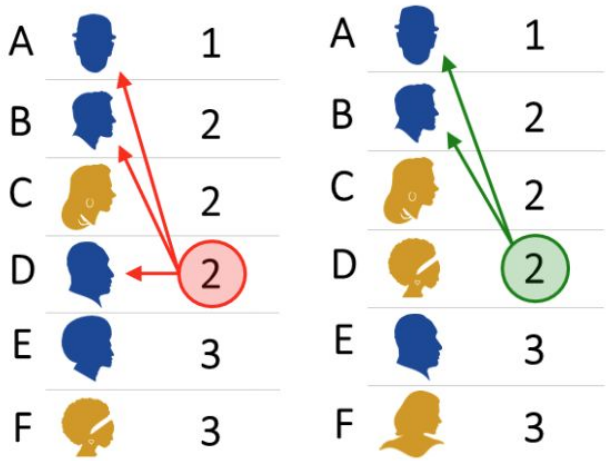
This not only **reinforces problematic stereotypes**, but also has the more direct consequence of **denying positive exposure** to marginalized communities in **opportunity ranking systems** like **resume search** (eg, LinkedIn, Indeed), or **resource allocation** recommendation systems (top K), etc.



Fair Ranking Algorithms

To combat this, **several fair ranking algorithms** have been proposed in the **literature**. Approaches include:

- Constrained optimization (utility/exposure constraint)
- Pairwise comparisons
- Learning-to-rank (amortized fairness under constraints)



Screenshot from Celis et Al, 2018

Fair Ranking Algorithms

However, most proposed fair ranking algorithms have a **caveat**: they require the **knowledge of protected group membership** (i.e, the group in which each particular item that is being ranked belongs to).

With respect to demographic groups, this has hurdles:

- Difficult for large datasets
- Might be outright illegal based on context
- Privacy concerns

Demographic Classification

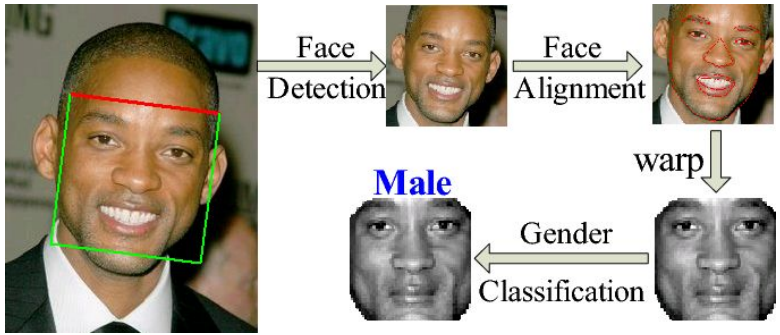
Unfortunately, a common workaround is to use **demographic classifiers** that infer the race/gender or other sensitive attribute from people's name, image, zip code, or other information.

Prominent example: Bayesian Improved Surname Geocoding **(BISG)** used in lending and healthcare uses names and zipcodes.

Demographic Classification

Similar commercial algorithms exist to infer gender or race from **images of people's faces.**

Examples include **Face++, Deepface.**

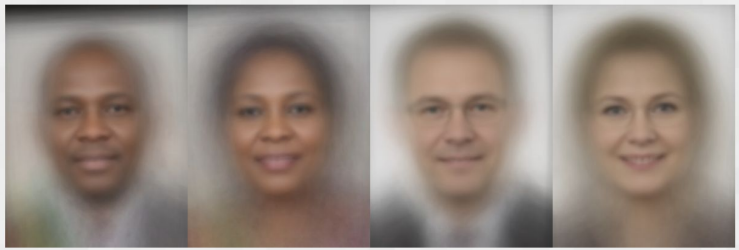


Demographic Classification

The paper “Gender Shades” (Buolamwini and Gebru 2018) shows how industrial image to gender classifiers were systematically worse for dark skinned women, a fact neatly hidden inside “overall accuracy”, which can be a misleading metric.

Pretrained models are thus **quite inaccurate.**

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%

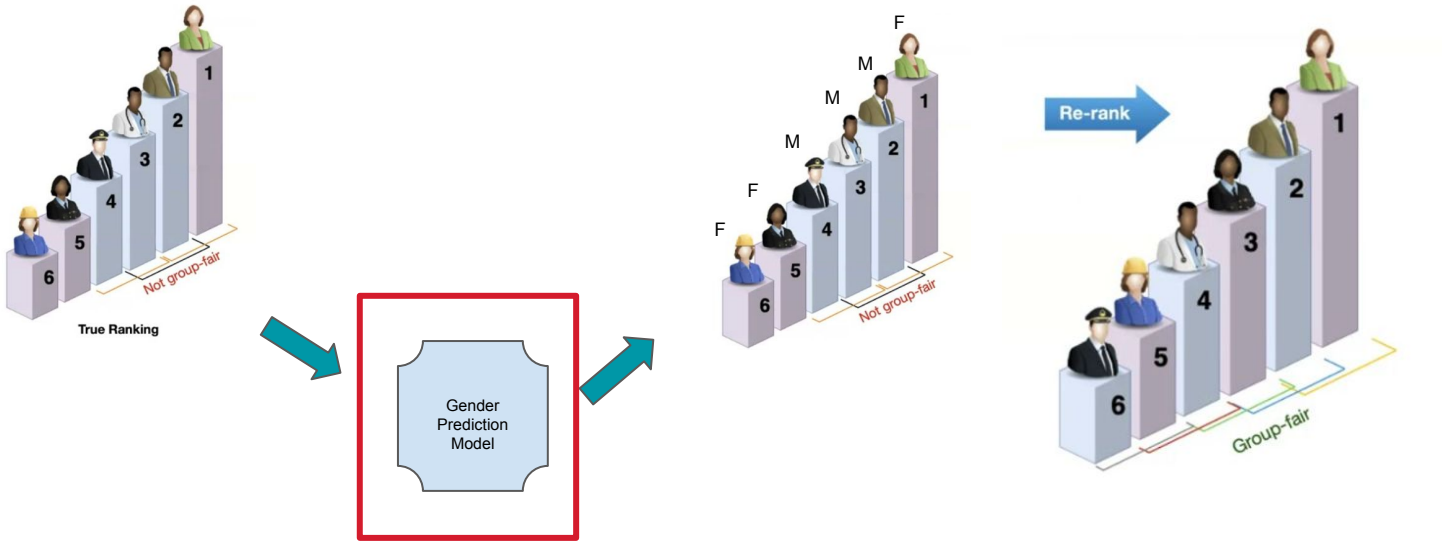


Problem to Investigate

How do **incorrectly inferred** sensitive demographic attributes affect the fairness metrics of **fair ranking algorithms**?

Problem to Investigate

How do **incorrectly inferred** sensitive demographic attributes affect the fairness metrics of **fair ranking algorithms**?



Problem to Investigate

How do **incorrectly inferred** sensitive demographic attributes affect the fairness metrics of **fair ranking algorithms**?

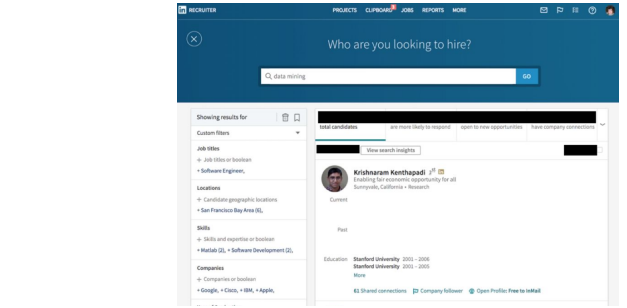
- A suitable real-world fair ranking algorithm
- Commercially available demographic classifier models
- Simulation studies to test the limits of our theory
- Case studies with real-world data

Methods

Setup: Fair Ranking Algorithm

The fair ranking algorithm we choose for this study is **DetConstSort**, from a paper by Geyik et Al at LinkedIn.

- Deterministic interval constrained sorting
- Aims to **rearrange members in the topK** to achieve a target distribution
- **Supports >2 groups** (and thus intersectional groups)
- **Large scale industrial usage** (*"deployment to 100% of LinkedIn Recruiter users worldwide"*)



https://engineering.linkedin.com/blog/2018/10/building-representative-talent-search-at-linkedin

Setup: Evaluation Metrics

Representation Based

$$Skew_{group,k} = \frac{Fraction\ of\ group\ members\ in\ top\ K}{Fraction\ of\ group\ members\ overall}$$

$$NDKL = Normalised\ Discounted\ KL\ divergence$$
$$between\ the\ group\ distributions\ in\ top\ K\ and\ overall\ population$$

The ideal value for Skew is 1, and NDKL is 0

Setup: Evaluation Metrics

Exposure Based

$$\text{Attention}_p@k(\tau) = 100 \times (1-p)^{k-1} \times (p)$$

$$ABR = \frac{Attention\ of\ group\ with\ min.\ \ avg\ attention}{Attention\ of\ group\ with\ max.\ \ avg\ attention}$$

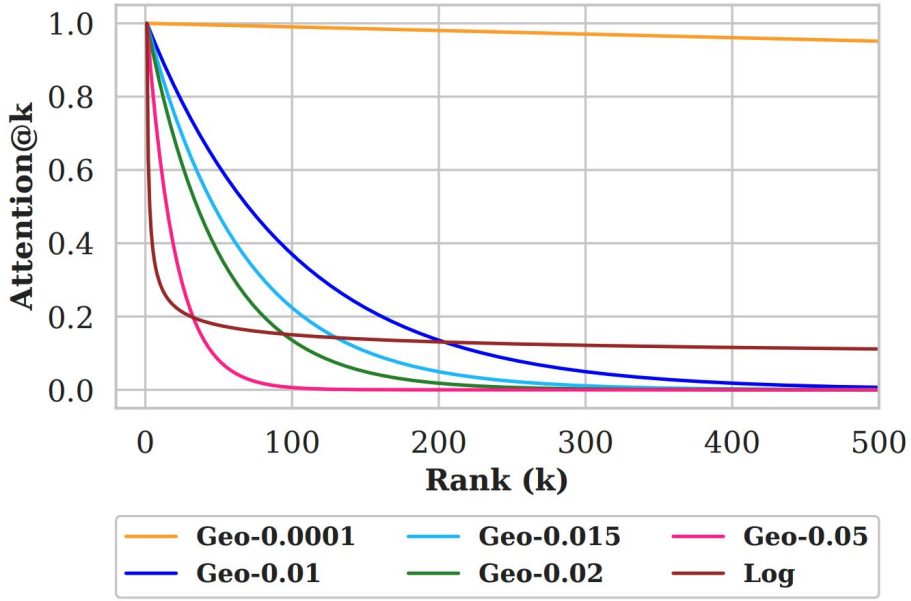
Setup: Evaluation Metrics

Exposure Based

$$\text{Attention}_p@k(\tau) = 100 \times (1-p)^{k-1} \times (p)$$

$$ABR = \frac{Attention\ of\ group\ with\ min.\ \ avg\ attention}{Attention\ of\ group\ with\ max.\ \ avg\ attention}$$

The ideal value for ABR is 1



Setup: Evaluation Metrics

Ranking Quality

$$DCG_n = \sum_{i=1}^{n} \frac{rel_i}{\log_2^{i+1}},$$

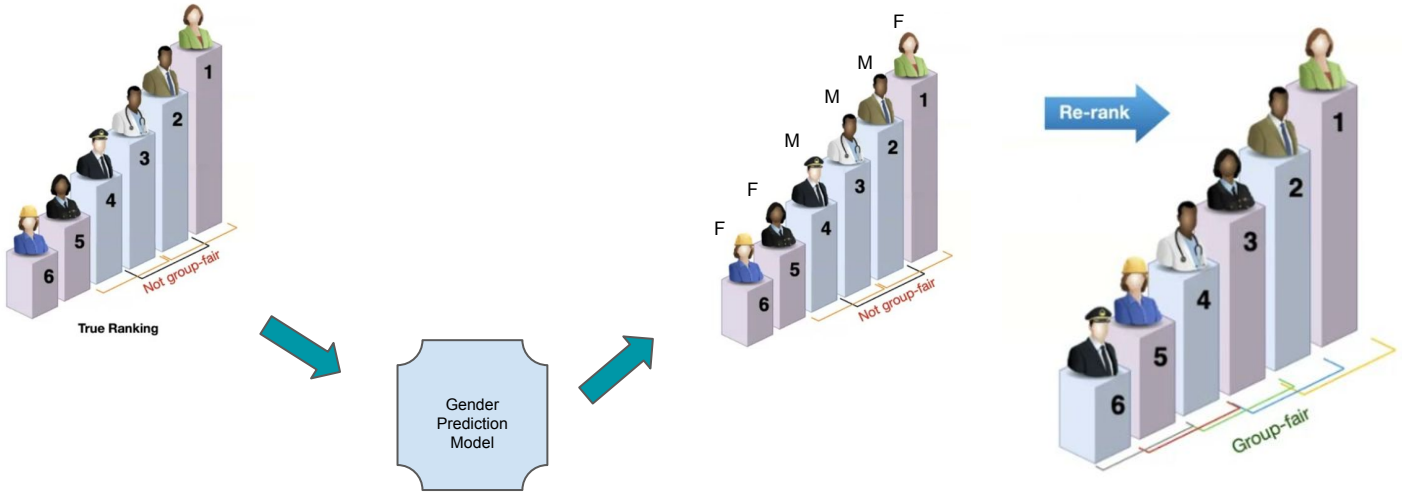
$$NDCG_n = \frac{DCG_n}{IDCG_n},$$

NDCG - Normalized Discounted Cumulative Gain, very popular in IR Literature and also used by Geyik et Al. to measure ranking quality.

The ideal value for NDCG is this case is 1

Experiments

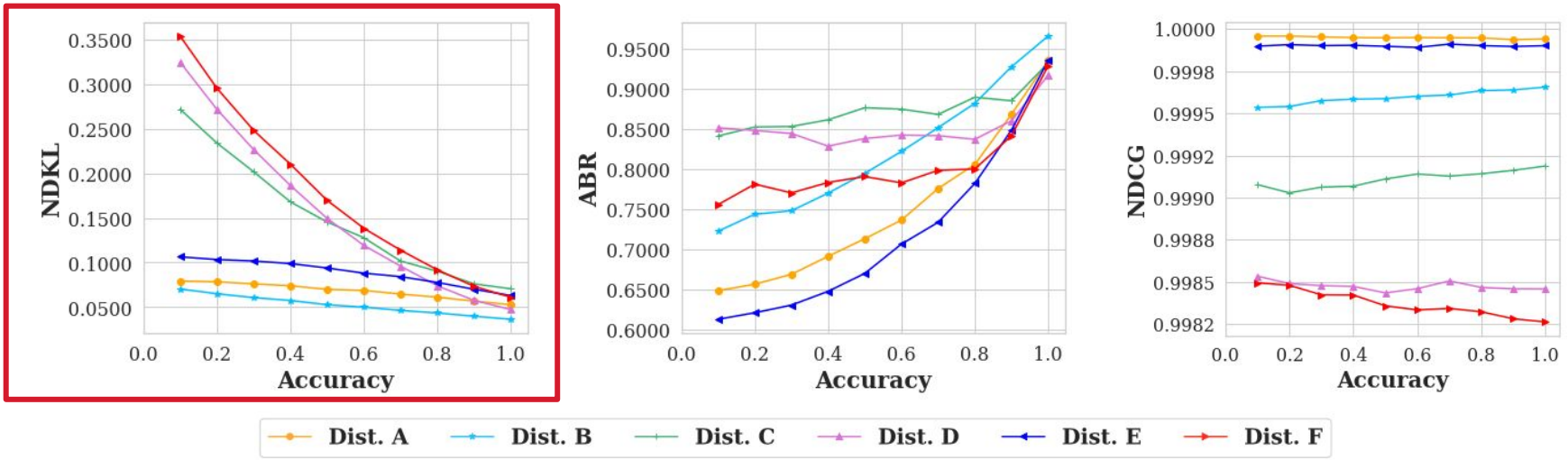
Simulation



Goal: To test the theoretical limits of the impact of inaccurate predictions.

- 6 different randomly generated lists
- Synthetic model with prediction accuracies ranging from 0 to 100% accurate
- Measure fairness metrics and ranking quality metrics

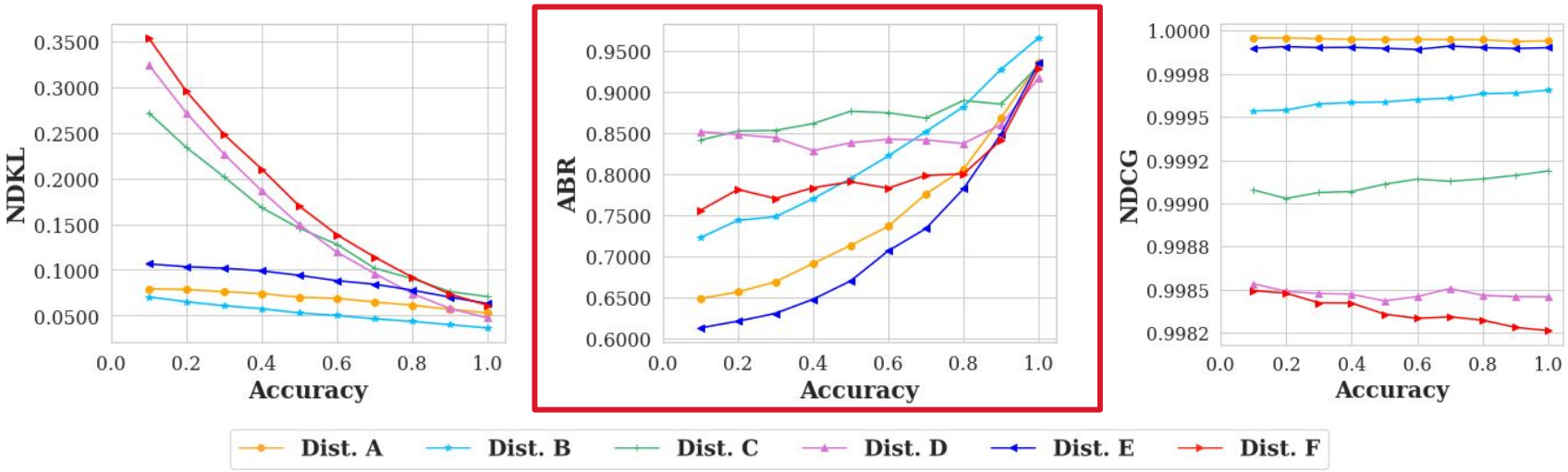
Simulation



Prediction accuracy vs ranking metrics for 6 random lists -

- **NDKL** moves towards the ideal value of zero
- **ABR** moves towards the ideal value of one
- **NDCG** is barely impacted (consistent with Geyik et Al.'s findings)

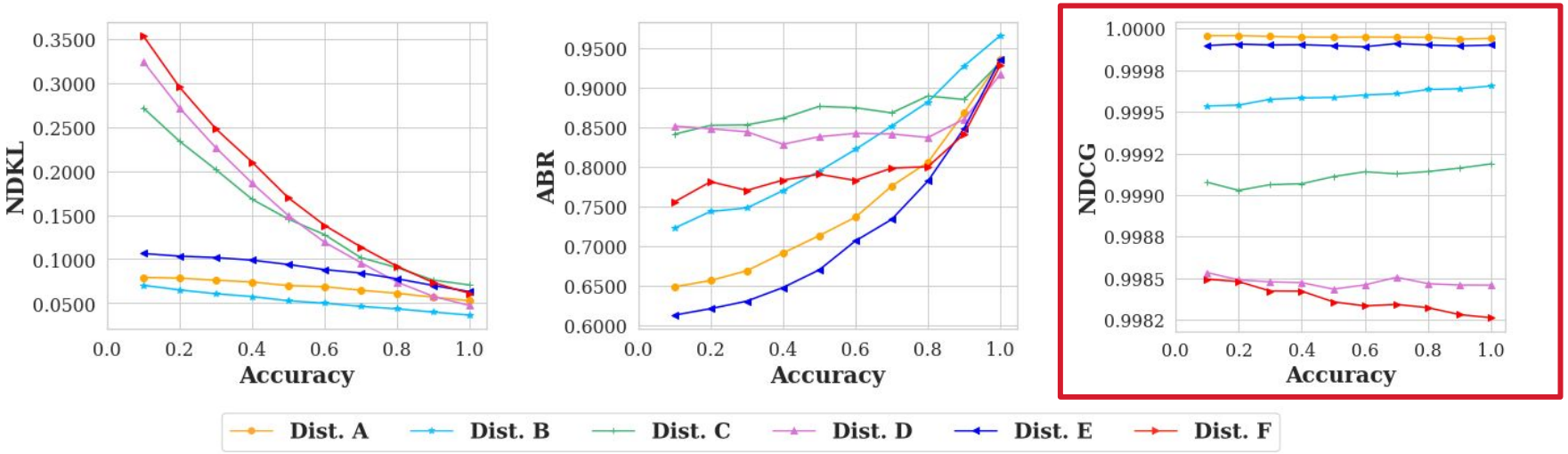
Simulation



Prediction accuracy vs ranking metrics for 6 random lists -

- **NDKL** moves towards the ideal value of zero
- **ABR** moves towards the ideal value of one
- **NDCG** is barely impacted (consistent with Geyik et Al.'s findings)

Simulation



Prediction accuracy vs ranking metrics for 6 random lists -

- **NDKL** moves towards the ideal value of zero
- **ABR** moves towards the ideal value of one
- **NDCG** is barely impacted (consistent with Geyik et Al.'s findings)

Case Study: Real World Datasets

- Simulation shows theoretical bounds
- But we wanted to test ecological validity of the hypothesis
- We collected 3 real-world ranked lists: Chess Players, Startup Founders, and Equestrians.

Case Study: Real World Datasets

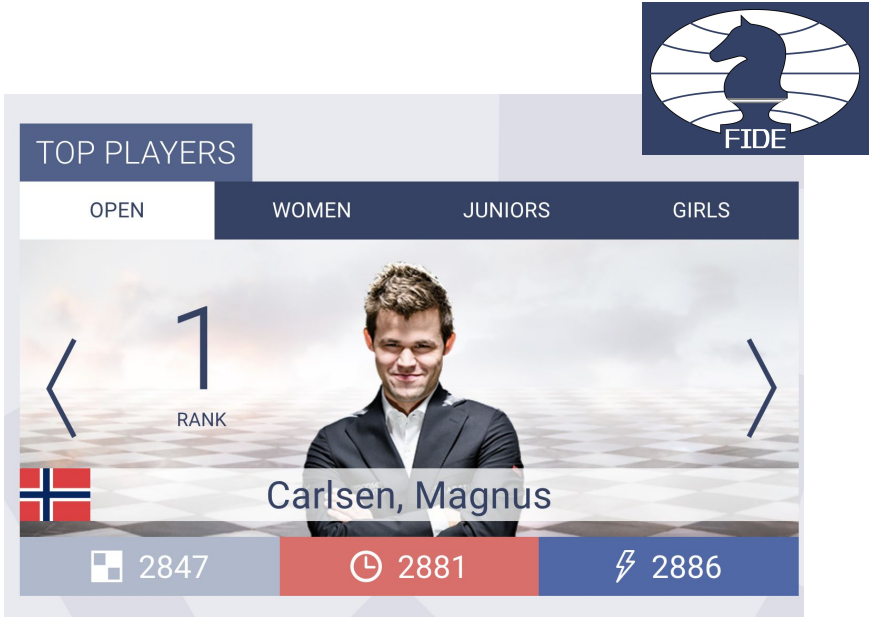
- Simulation shows theoretical bounds
- But we wanted to test ecological validity of the hypothesis
- We collected 3 real-world ranked lists: **Chess Players**, Startup Founders, and Equestrians.

For the sake of brevity, I only discuss the Chess players case study in this talk

Case Study: Data Collection

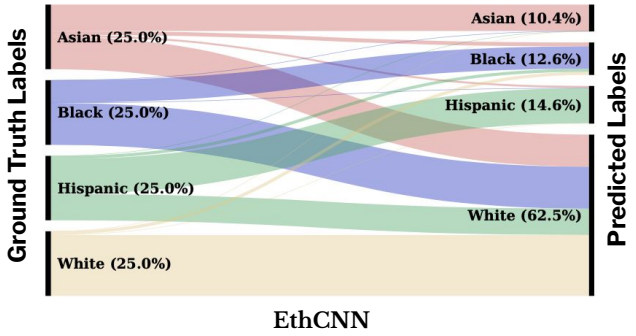
We collected a ranked list of the top **Chess players** from FIDE along with their scores.

We collect the **names, images and the binary gender**. **Race/ethnicity** annotated via **Amazon Mturk.**

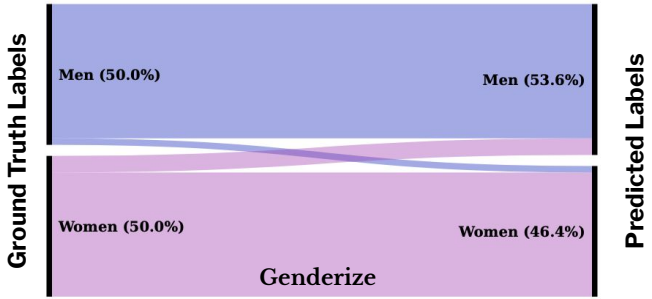


Case Study: Demographic Inference Algorithms

Name based

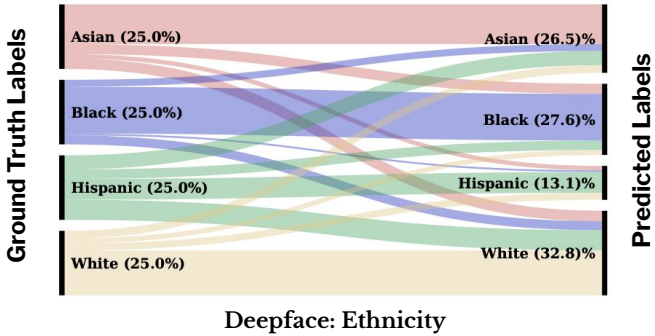


EthCNN

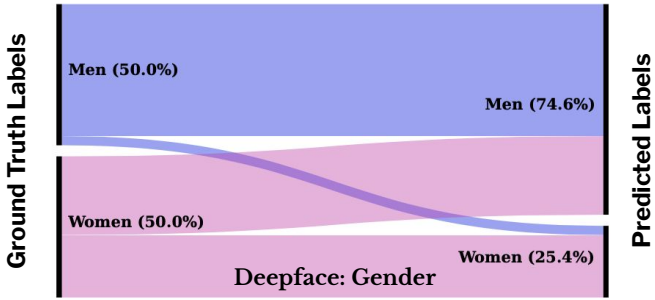


Genderize

Face based

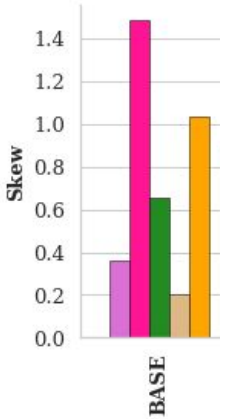


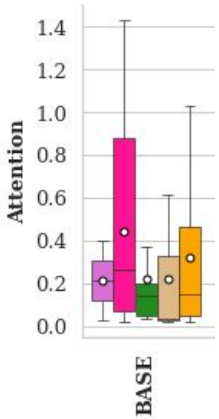
Deepface: Ethnicity



Deepface: Gender

Case Study: Results

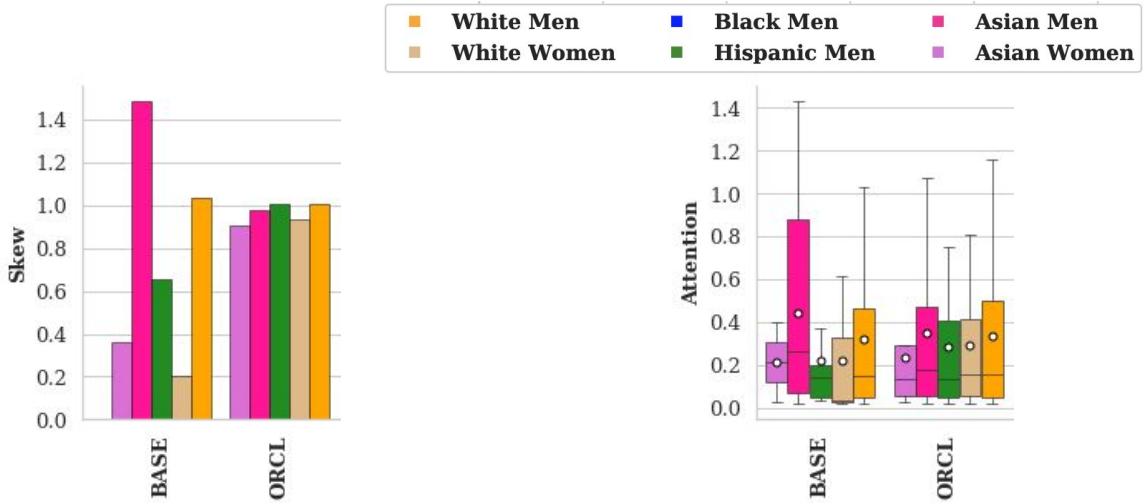




BASE: Baseline

- Unfair Ranking, no intervention

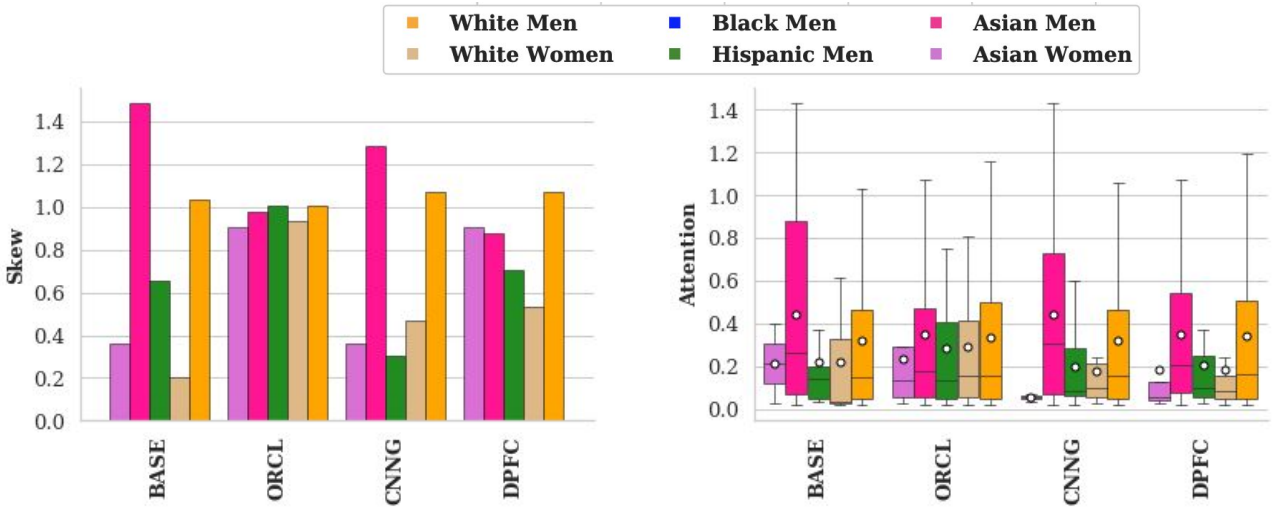
Case Study: Results



BASE: Baseline
ORCL: Oracle

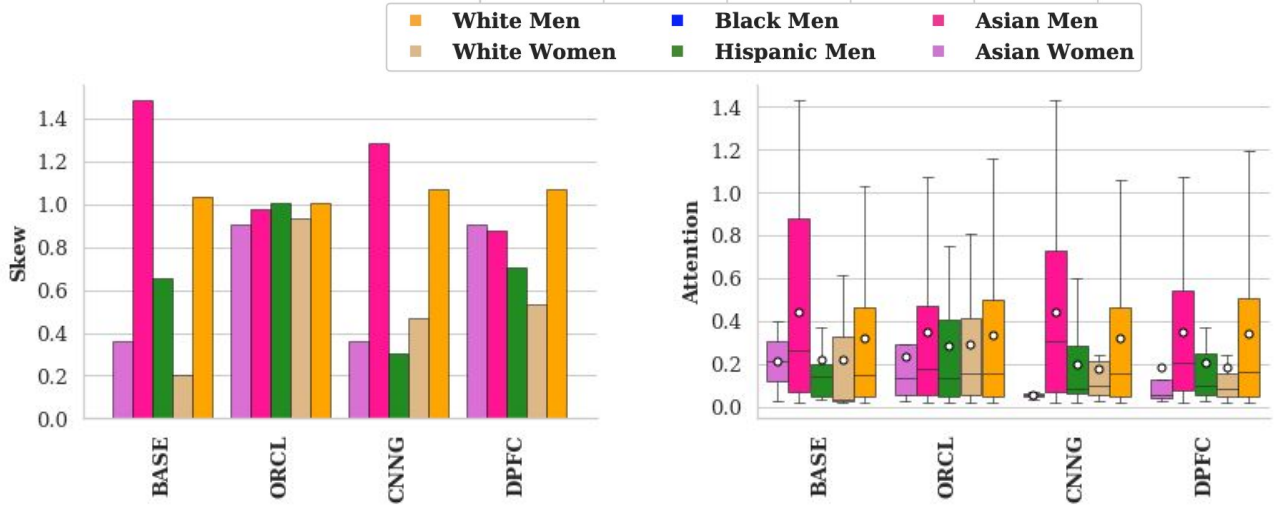
(fair ranking with 100% race/gender label classification accuracy)

Case Study: Results



BASE: Baseline
ORCL: Oracle
CNNG: EthCNN+Genderize (Name based)
DPFC: Deepface (Face based)

Case Study: Results

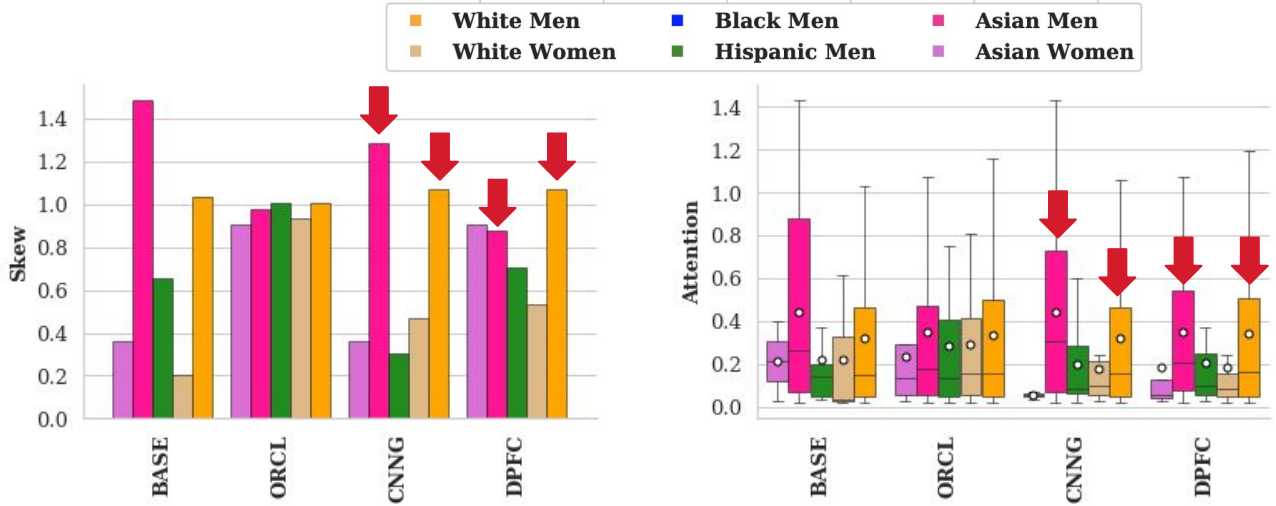


BASE: Baseline
ORCL: Oracle
CNNG: EthCNN+Genderize (Name based)
DPFC: Deepface (Face based)

- All strategies performed worse than Oracle
- White and Asian men retained their advantage while the fairness for other groups declined based on how badly they were mispredicted

- Hispanic men were mispredicted as white, causing them to be suppressed
- Asian men were mispredicted sometimes as White women, causing them to get an unfair boost

Case Study: Results

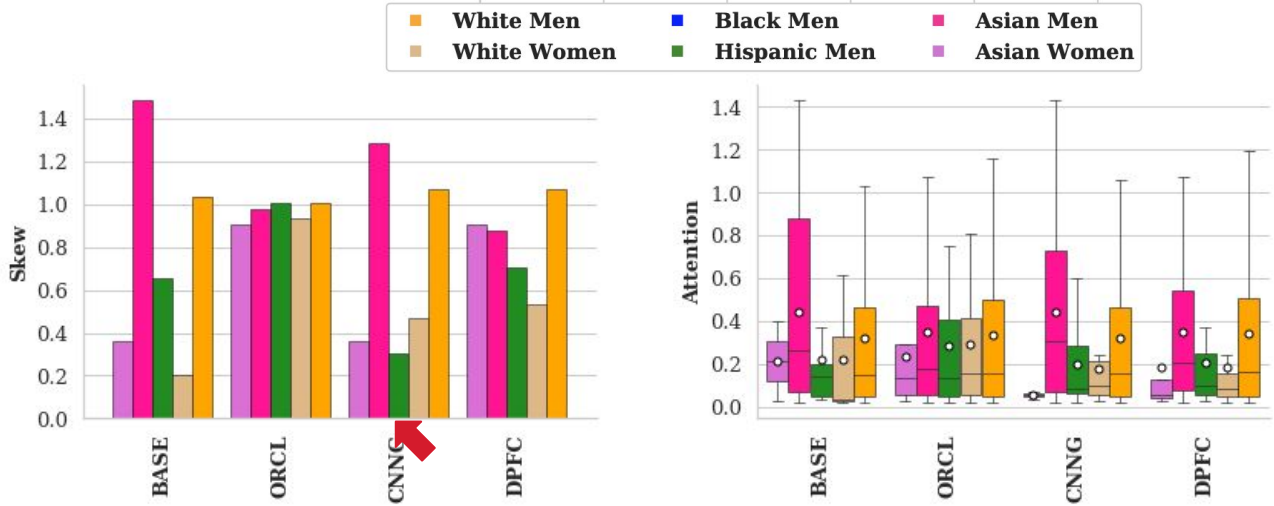


BASE: Baseline
ORCL: Oracle
CNNG: EthCNN+Genderize (Name based)
DPFC: Deepface (Face based)

- All strategies performed worse than Oracle
- White and Asian men retained their advantage while the fairness for other groups declined based on how badly they were mispredicted

- Hispanic men were mispredicted as white, causing them to be suppressed
- Asian men were mispredicted sometimes as White women, causing them to get an unfair boost

Case Study: Results

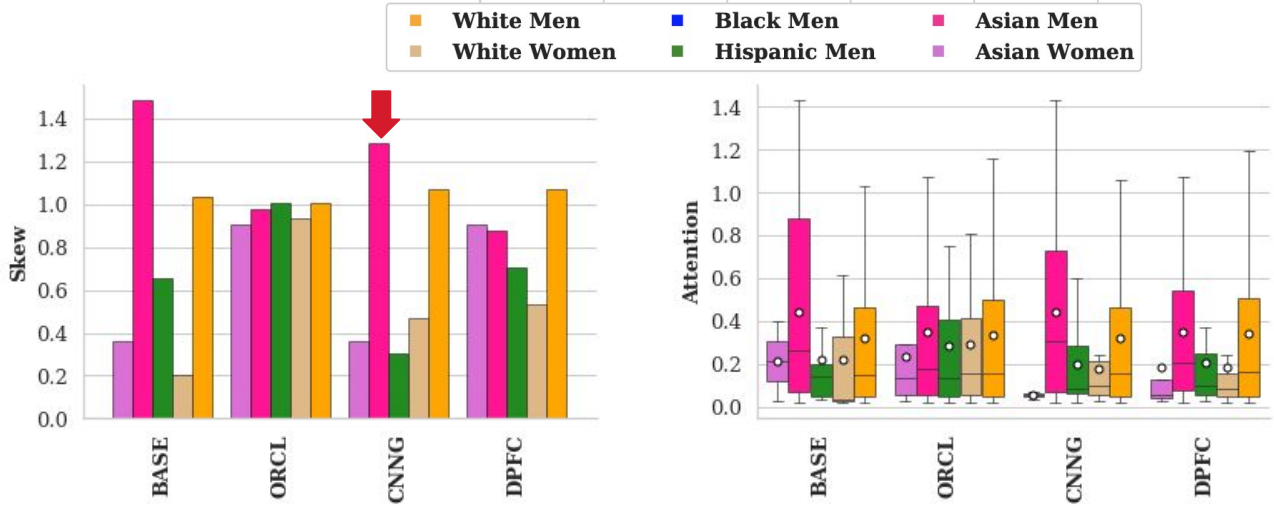


BASE: Baseline
ORCL: Oracle
CNNG: EthCNN+Genderize (Name based)
DPFC: Deepface (Face based)

- All strategies performed worse than Oracle
- White and Asian men retained their advantage while the fairness for other groups declined based on how badly they were mispredicted

- Hispanic men were mispredicted as white, causing them to be suppressed
- Asian men were mispredicted sometimes as White women, causing them to get an unfair boost

Case Study: Results



BASE: Baseline
ORCL: Oracle
CNNG: EthCNN+Genderize (Name based)
DPFC: Deepface (Face based)

- All strategies performed worse than Oracle
- White and Asian men retained their advantage while the fairness for other groups declined based on how badly they were mispredicted

- Hispanic men were mispredicted as white, causing them to be suppressed
- Asian men were mispredicted sometimes as White women, causing them to get an unfair boost

Conclusion

- Fair Ranking methods which **require access to demographic information** are prone to **violate fairness guarantees** if this information is **noisy.**
- It is not always the case that inference assisted fair rankings are categorically better than no fair ranking interventions - as we have shown sometimes **protected groups can be worse off than rankings without any intervention.**
- The violation is **not easy to predict** and the relationship between per class prediction accuracy and overall effect is **complex.**
- **Limitations:** We do not deal with multiple or partial group memberships, for instance, nonbinary people or genderfluid people.

Possible Mitigations

- Use inferred attributes only when they are **extremely accurate** for all intersectional groups
- **Human-in-the-loop** solutions (privacy aware), for instance Project Lighthouse

Airbnb's Project Lighthouse

Background

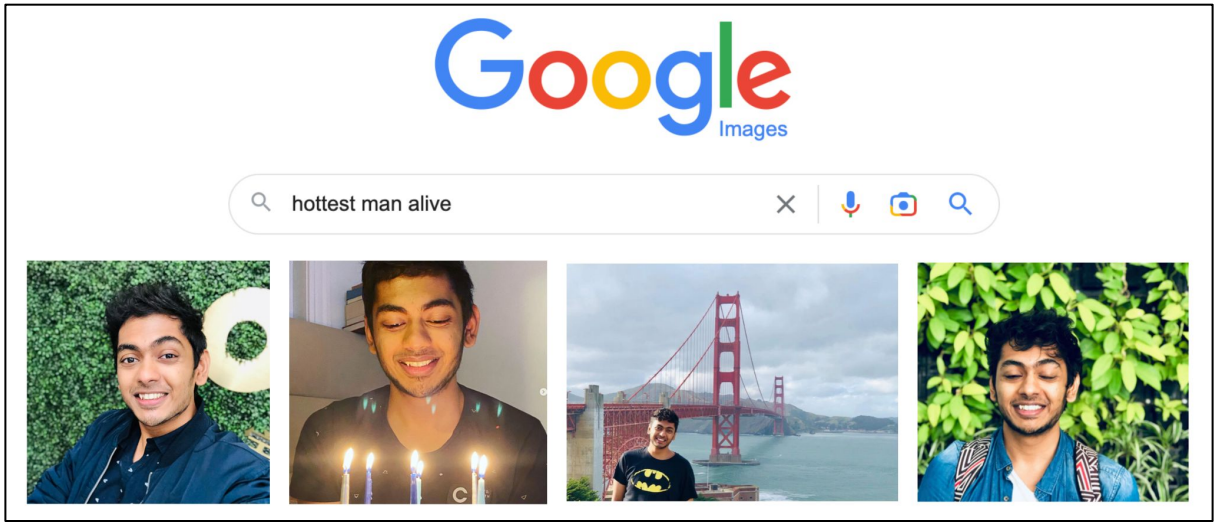
Research Questions

Uncertain Inference

Adversarial Attacks

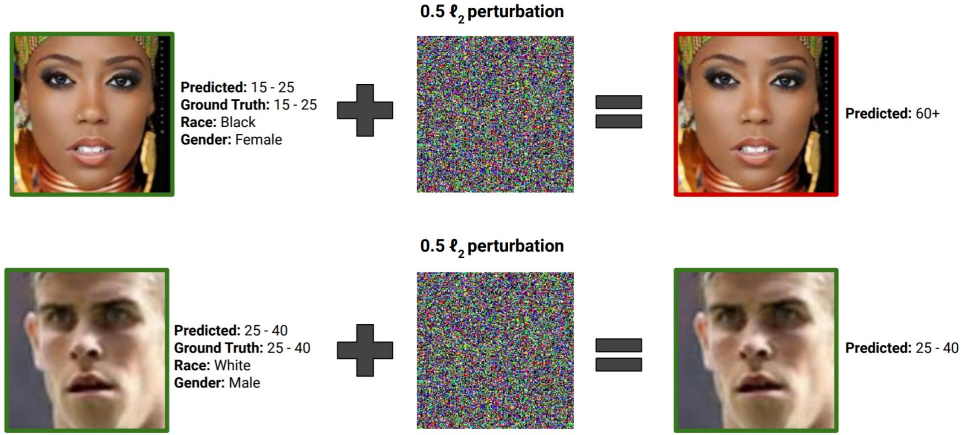
Chapter 2

Subverting Fair Image Search with Generative Adversarial Perturbations



Northeastern University

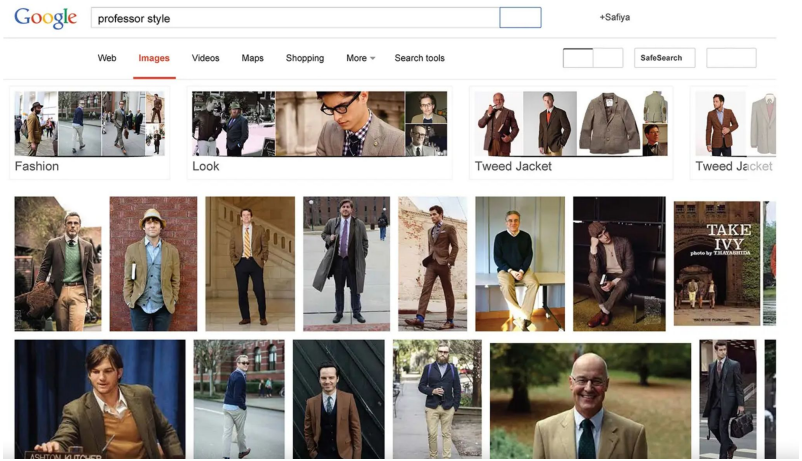
Intentional Biases



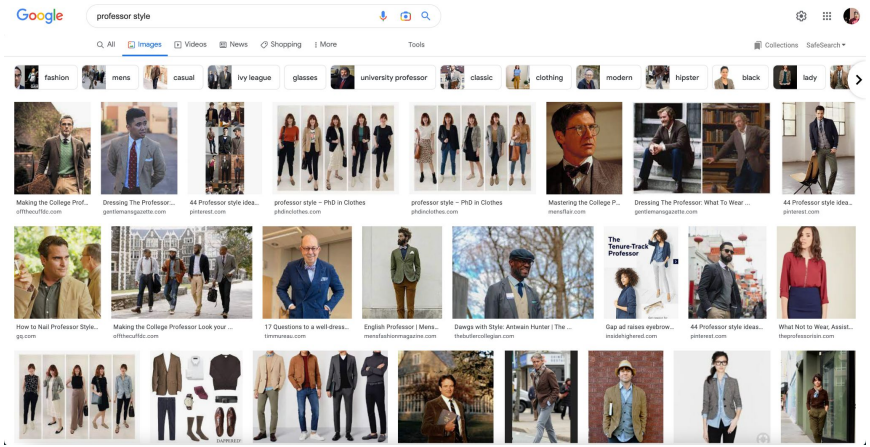
In Chapter 1, we saw that due to unintentional misclassification, inferred demographics don't work well for fair ranking. Now, what if demographic information is untrustworthy because someone is ***intentionally attempting to misrepresent themselves or their data***?

It is possible for the same adversarial perturbation to cause completely different outcomes for people in different subgroups. (Nanda et al. 2021)

Fair Image Search



Biased (2015)



Diverse (2022)

Fairness is important for image search. A real-world image search system not only has to crawl images from the internet but should also ideally present diverse, nuanced perspectives.
What if an adversary were to upend this effort?

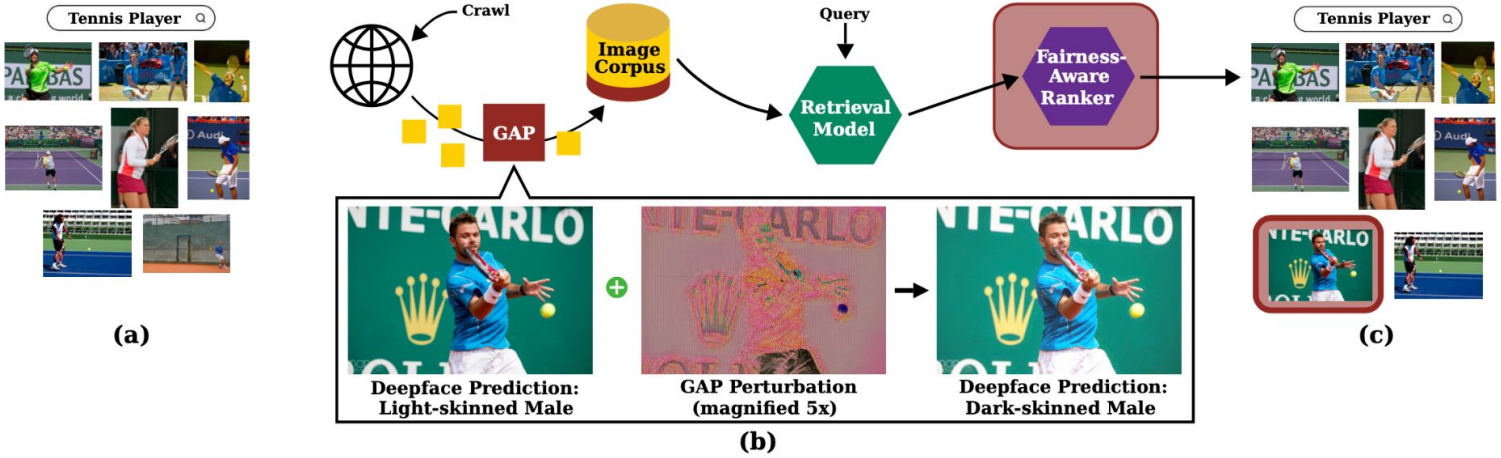
Threat Model

We attack a **Fair Image Search** model that consists of two parts - a **retrieval step and a fair re-ranking step**. The fair re-ranking step uses race and gender labels inferred via commercially available classifiers.

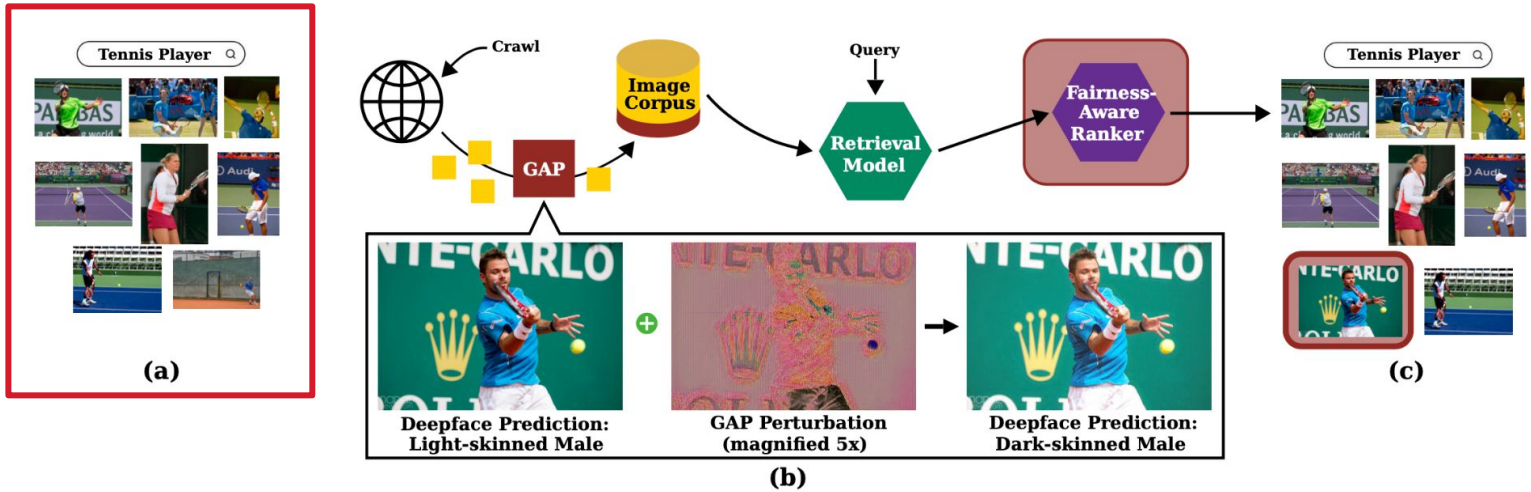
A **very restrictive threat model,** to be as close to a real-world attack as possible.

1. This is an **evasion attack**, which means that the attacker does not have access to the model parameters.
2. The attacker also does not know which fair-reranking model or which demographic inference model is being used by the image search system.
3. The attacker uploads adversarially perturbed images onto the internet, and a web scraper collects these images along with other clean images from all over the web and adds to a repository of images to retrieve from.
4. Threat model is similar to Clean Label attacks, or FAWKES (Shan and Wenger et al, 2020)

Threat Model: A Schematic

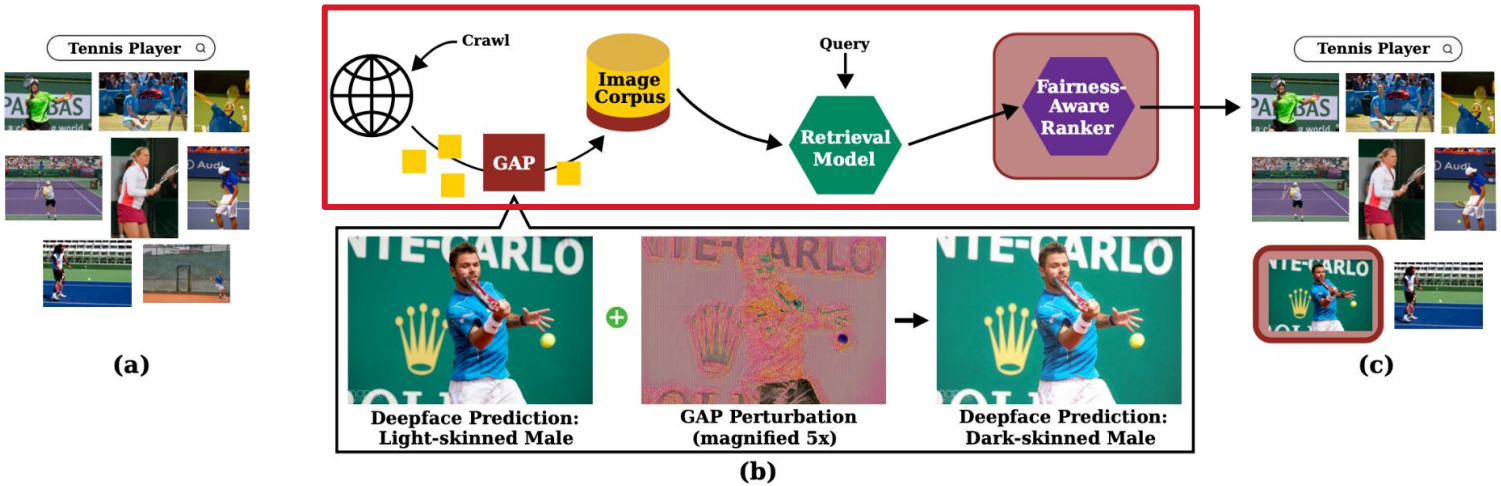


Threat Model: A Schematic



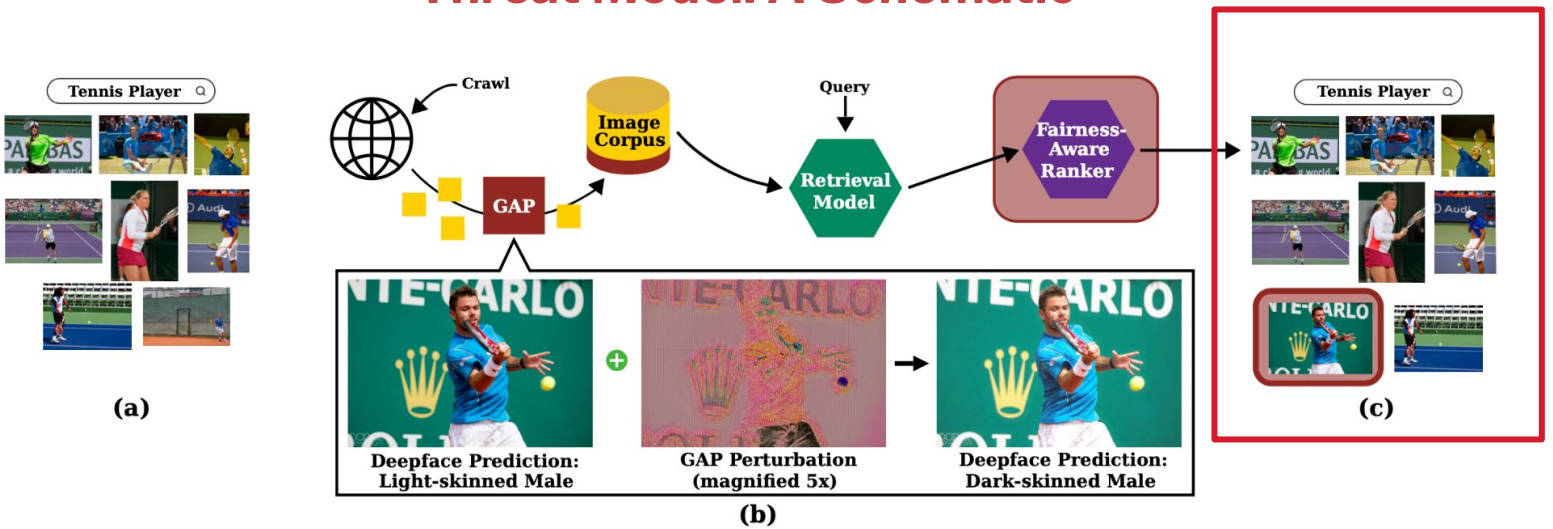
(a) shows example search results from an image search engine for the query “tennis player”.

Threat Model: A Schematic



(b) as this search engine crawls and indexes new images from the web, it collects images that have been adversarially perturbed using a GAP model

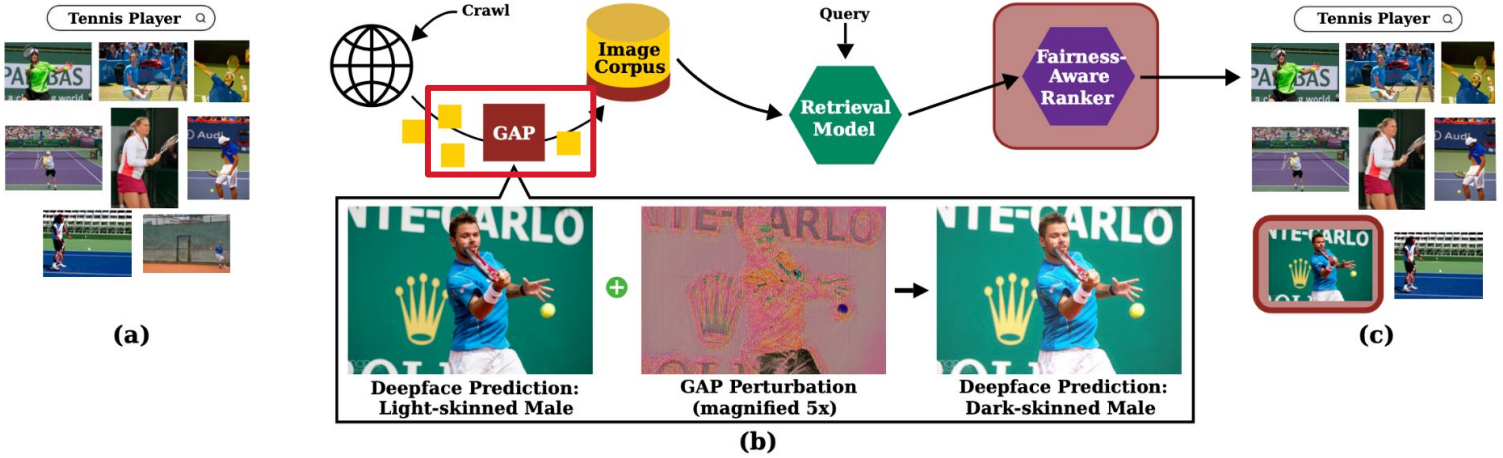
Threat Model: A Schematic



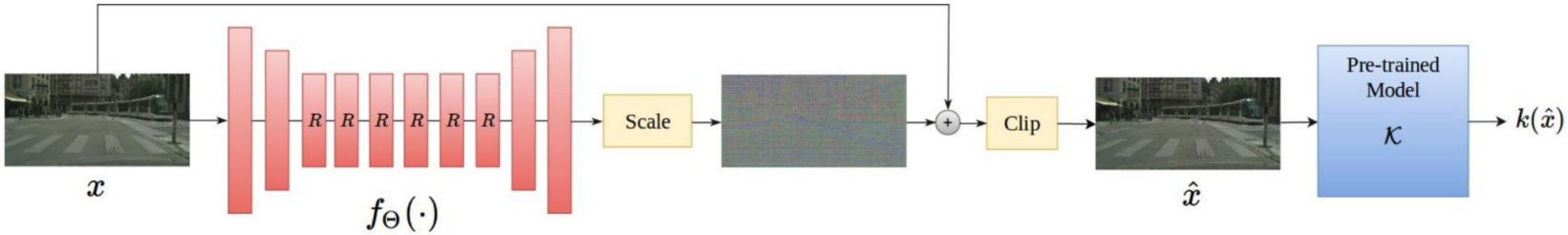
(c) The fairness-aware ranker (the target of the attack, highlighted in red) mistakenly elevates the rank of an image containing a light-skinned male (also highlighted in red) because it misclassifies them as dark-skinned due to the perturbations.

Methods

Setup: Genetic Adversarial Perturbation

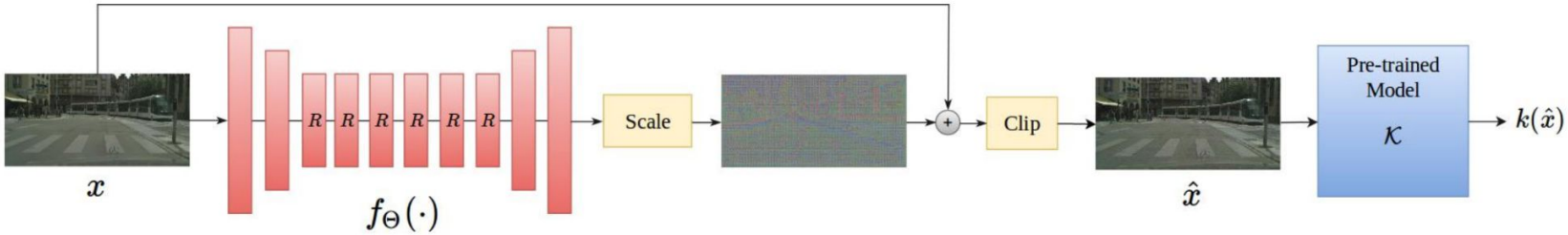


Setup: Genetic Adversarial Perturbation



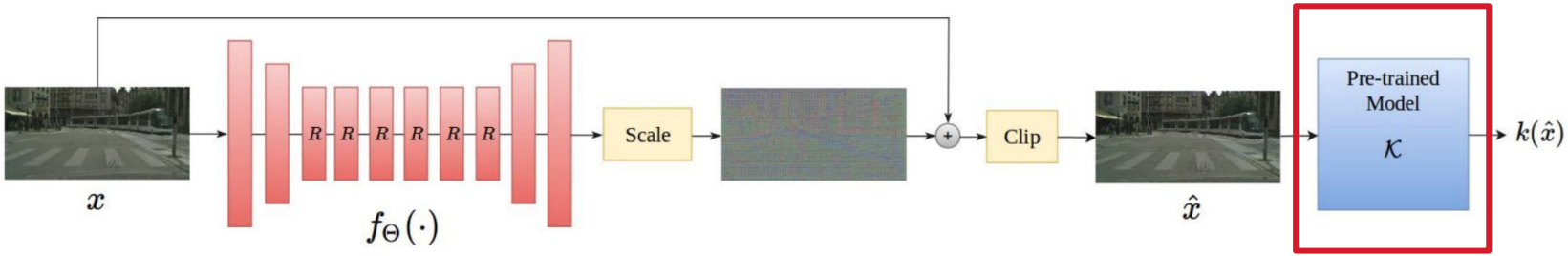
To achieve my misclassification, I modify a model called Generative Adversarial Perturbation (GAP) (Poursaeed et al. 2018). The adversary provides a source class y_s and target class y_t. The Class Targeted GAP model f_{CGAP} is a model that takes as input an image x and returns an image x′, effectively forcing the demographic inference model to misclassify samples of class y_s to class y_t, while maintaining its performance for samples not from class y_s .

Setup: Generative Adversarial Perturbation



A generative adversarial perturber model has advantages over universal perturbations because it does not require a fixed size or resolution image and can work on images of any size, which is what a realistic image search engine would be dealing with.

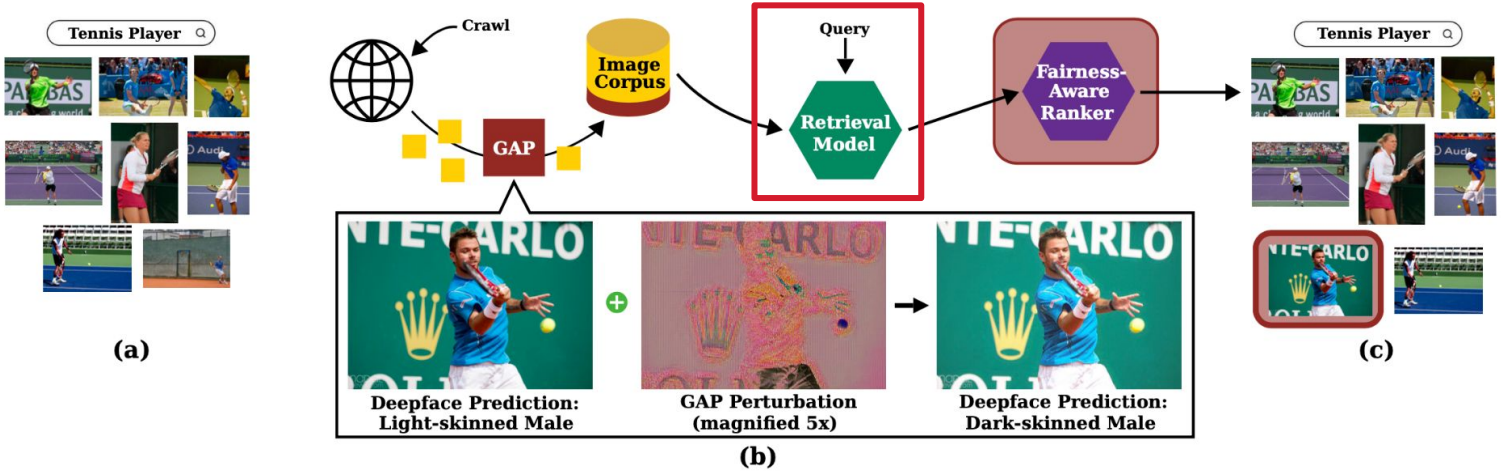
Setup: Genetic Adversarial Perturbation



We use two pre-trained demographic classification models to train the GAP:

- **Deepface** is a face recognition model for gender and race inference developed by Facebook.
- **FairFace** is a model designed for race and gender inference, trained on a diverse set of 108,000 images.

Setup: Retrieval Model



Setup: Retrieval Model

Caption: *A skier is skiing down the snow wearing a white shirt and black shorts.*

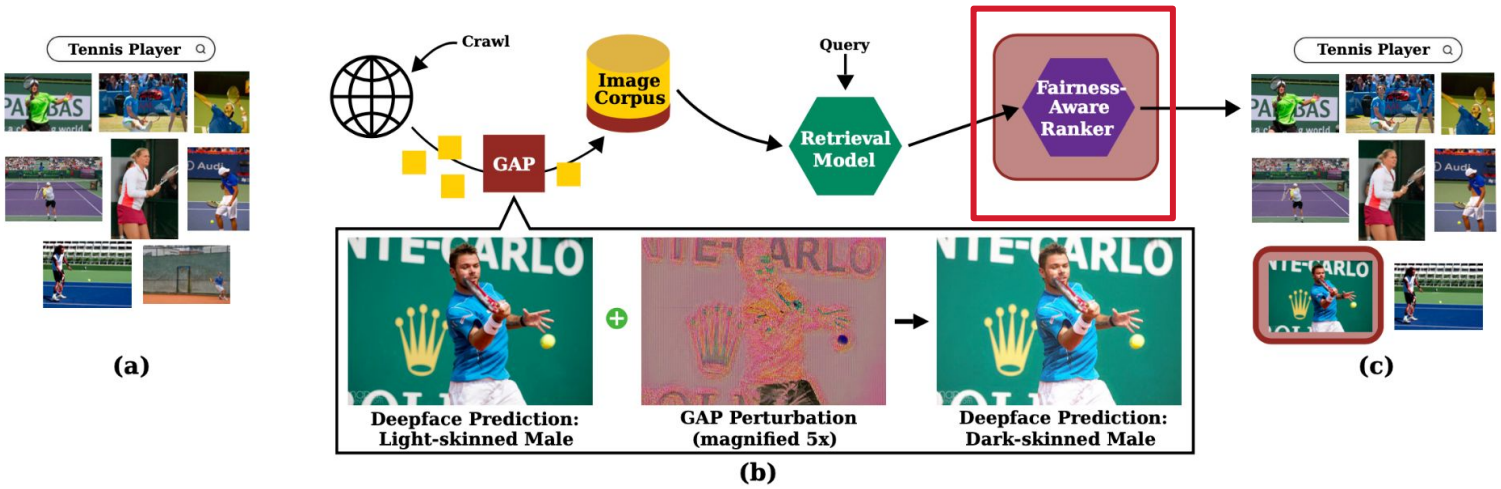
(a) Target Image

(b) Retrieved Top-6 with JOINT+BE^{OSCAR}

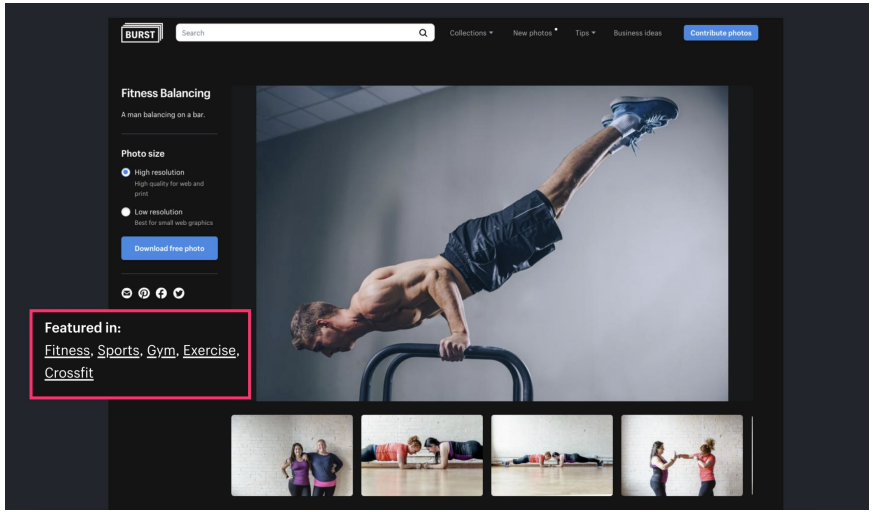
(c) Reranked Top-6 with JOINT+CE^{OSCAR}

The image search model we use in the paper is a MultiModal Transformer (MMT) (Geigle et al. 2021) based text-image retrieval model. This model consists of two components: a fast (although somewhat lower quality) retrieval step that identifies a large set of relevant images, followed by a re-ranking step that selects the best images from the retrieved set.

Setup: Fairness Aware Ranker



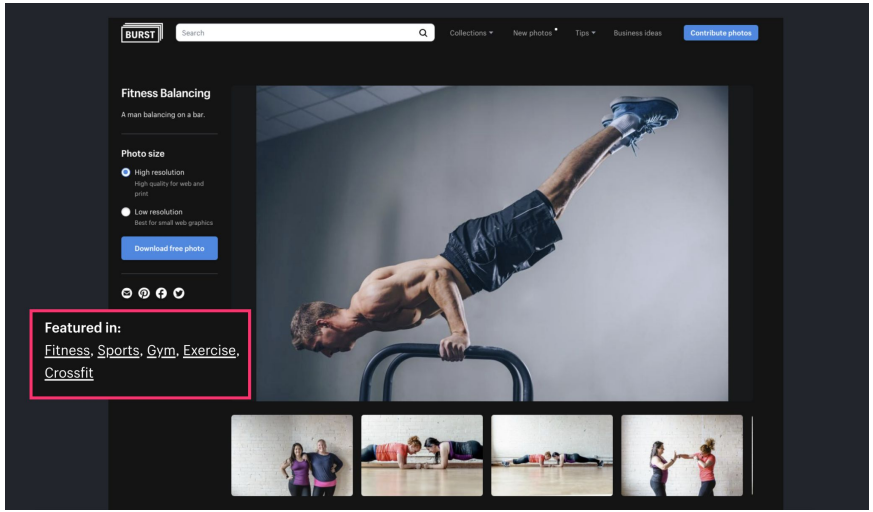
Setup: Fairness Aware Ranker



Two fair reranking models are used in the paper:

- The LinkedIn **DetConstSort** algorithm, discussed in the previous chapter
- Shopify's **Fair Maximal Marginal Relevance** (FMMR) algorithm (Karako and Manggala 2018), which essentially tries to select the next image in a search result output by maximizing relevance while minimizing similarity (thereby maximizing fairness/diversity), by modifying a KNN-esque clustering algorithm.

Setup: Fairness Aware Ranker

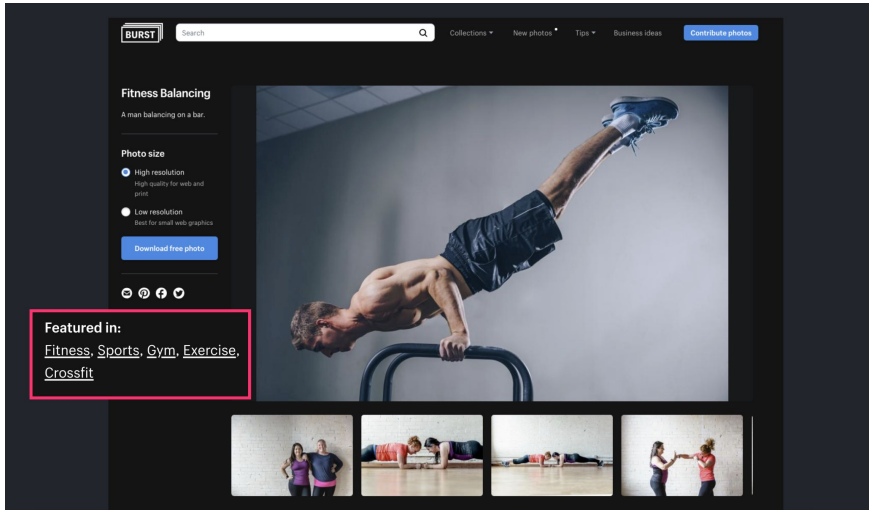


Already performed poorly because of errors in inference models

Two fair reranking models are used in the paper:

- The LinkedIn **DetConstSort** algorithm, discussed in the previous chapter
- Shopify's **Fair Maximal Marginal Relevance** (FMMR) algorithm (Karako and Manggala 2018), which essentially tries to select the next image in a search result output by maximizing relevance while minimizing similarity (thereby maximizing fairness/diversity), by modifying a KNN-esque clustering algorithm.

Setup: Fairness Aware Ranker



Does NOT require access to demographic labels, so should be harder to fool!

Two fair reranking models are used in the paper:

- The LinkedIn **DetConstSort** algorithm, discussed in the previous chapter
- Shopify's **Fair Maximal Marginal Relevance** (FMMR) algorithm (Karako and Manggala 2018), which essentially tries to select the next image in a search result output by maximizing relevance while minimizing similarity (thereby maximizing fairness/diversity), by modifying a KNN-esque clustering algorithm.

Experiments

Setup: Case Study

Dataset: Skin color and gender annotated subset of **Microsoft COCO** (Zhao et Al.)

I only used images with one person. This amounted to 8692 images.

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k
"Tennis Player" "Person eating pizza" "Person at Table"	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2, 0.5, 0.7, 1.0	10, 15, 20..., 45, 50

We used three queries, two attack training models, multiple training objectives and top K analysis for the experiments. Attack probability is the fraction of images in the dataset that are adversarially modified.

Metrics

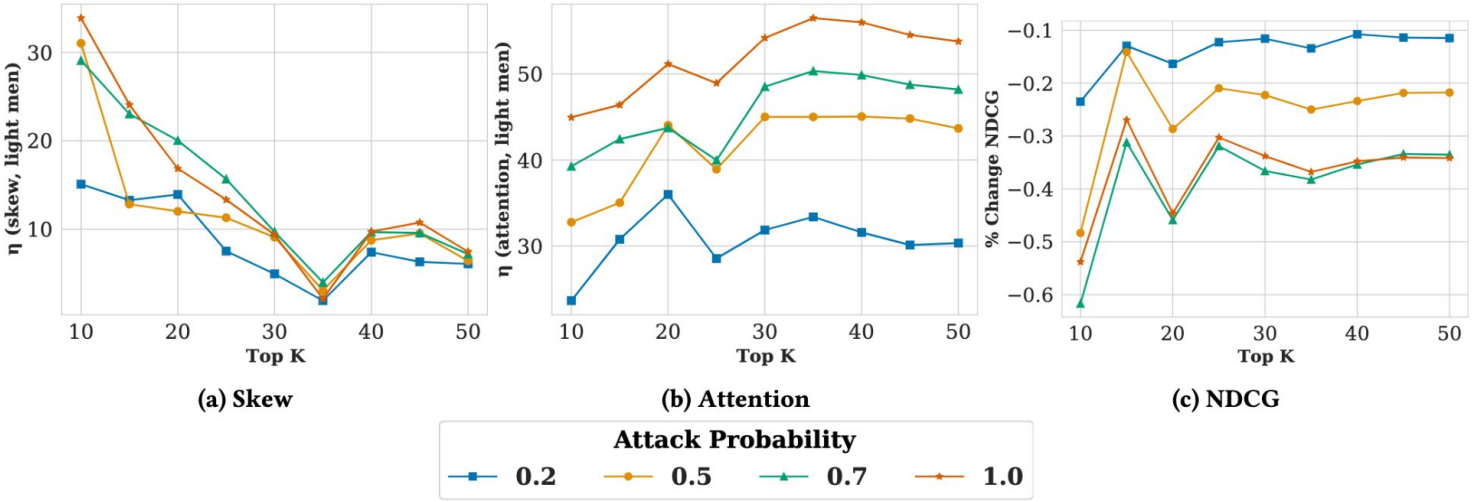
- Skew, Attention, NDCG already defined in Chapter 1
- I wanted to focus on the boost conferred to the majority subgroup - light men
- Summarizing metric: Attack Effectiveness

$$\eta(m, g) = \%\text{ change in } m \text{ for subgroup } g - \text{minimum \% change in } m \text{ over other subgroups.}$$

So, for example, η(attention, light men) measures the attack effectiveness by measuring the relative attention boost provided to light men after the attack.

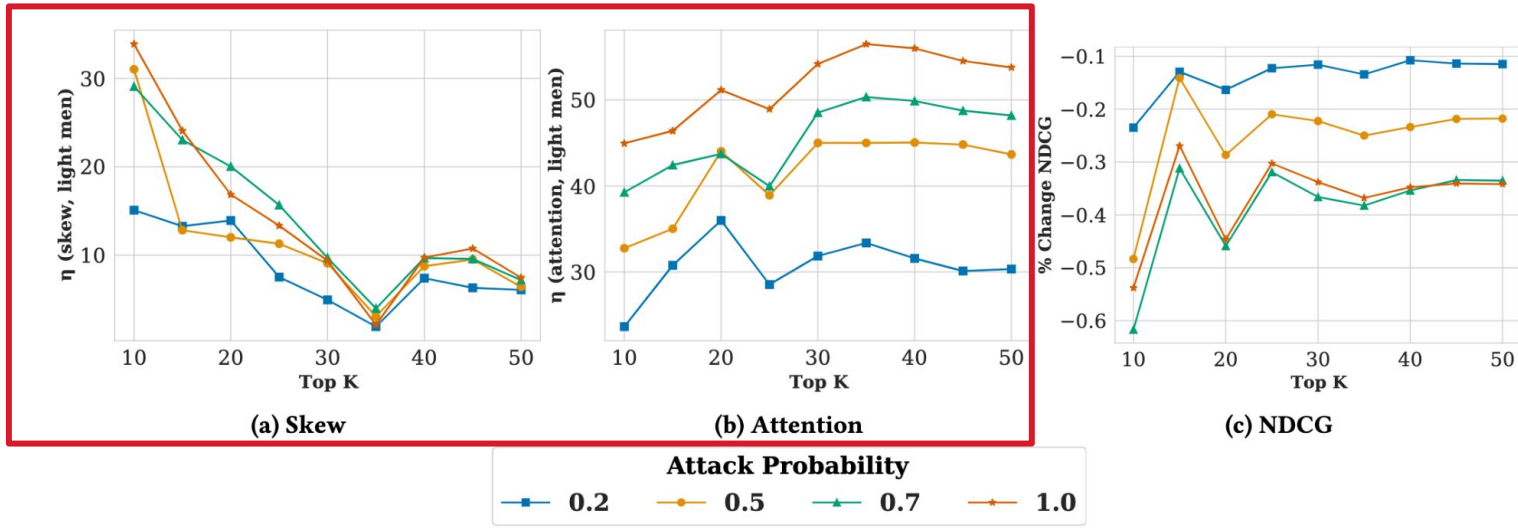
Results

Effect of Top K and Attack Probability



Results

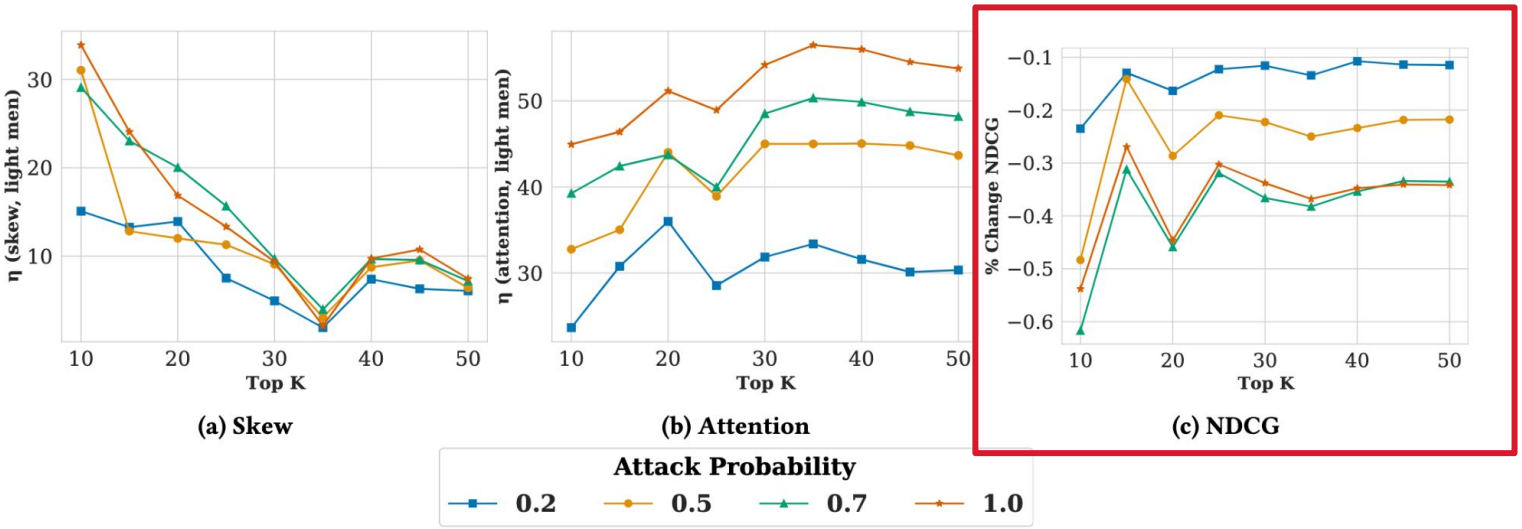
Effect of Top K and Attack Probability



Skew and Attention generally unfairly increases towards light men with increasing attack probability

Results

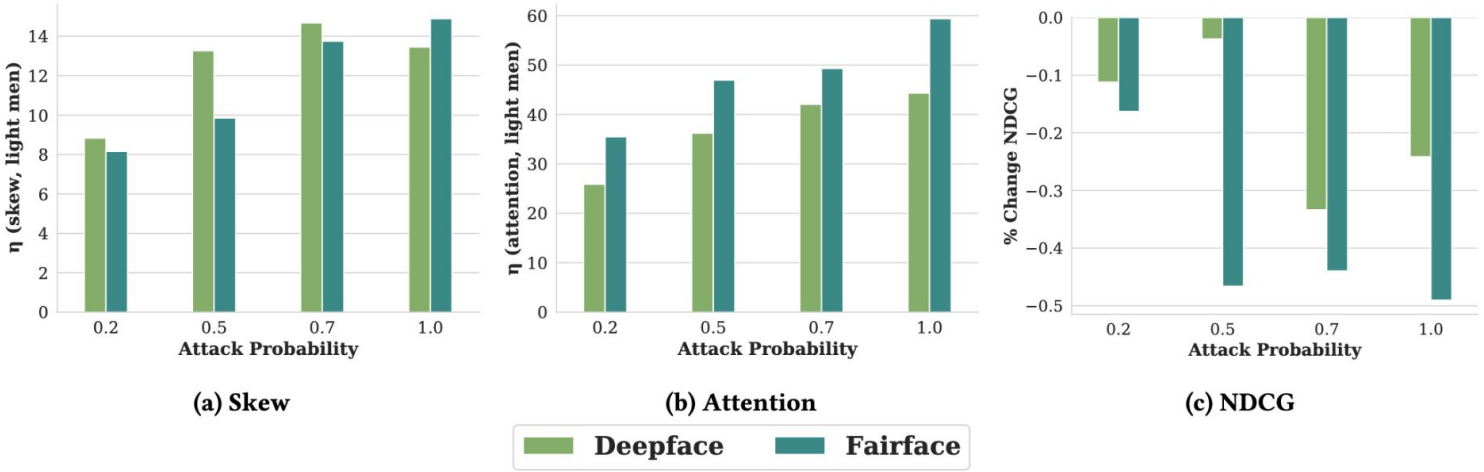
Effect of Top K and Attack Probability



Skew and Attention generally unfairly increases towards light men with increasing attack probability, however, NDCG is barely affected.

Results

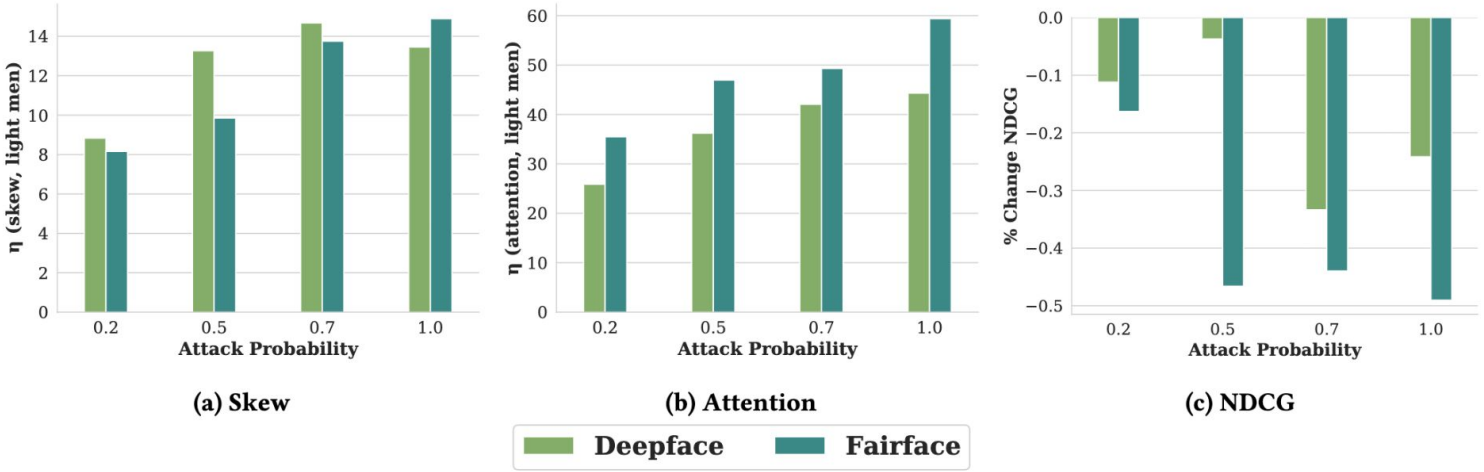
Effect of Training Model



We observe that the attack effectiveness is similar, no matter what the model used for training.

Results

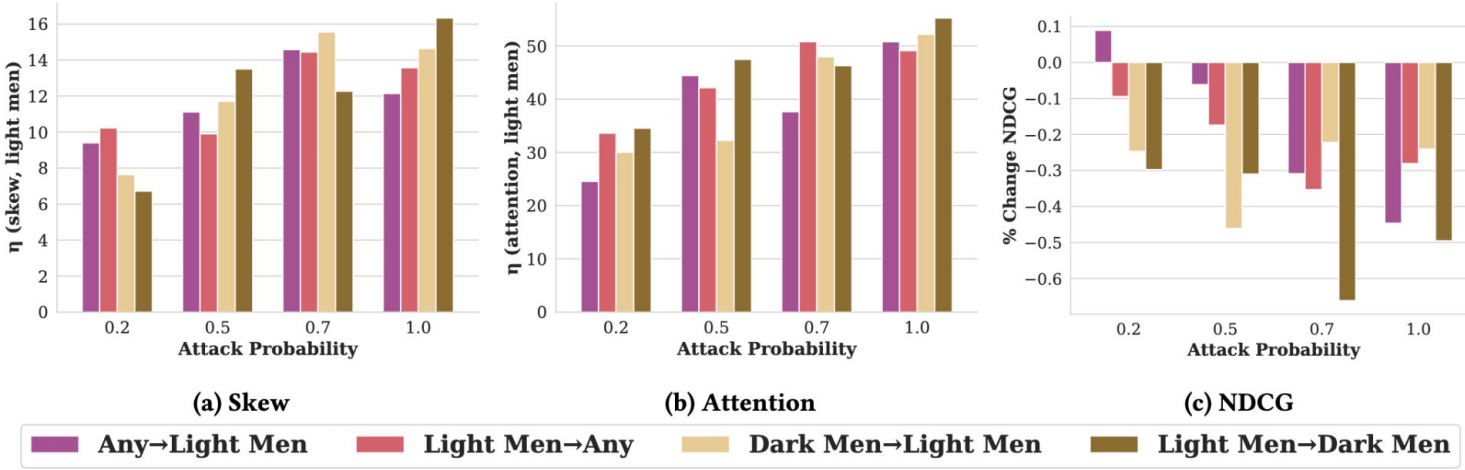
Effect of Training Model



We observe that the attack effectiveness is similar, no matter what the model used for training. The attack is also stronger as a higher percentage of images are perturbed.

Results

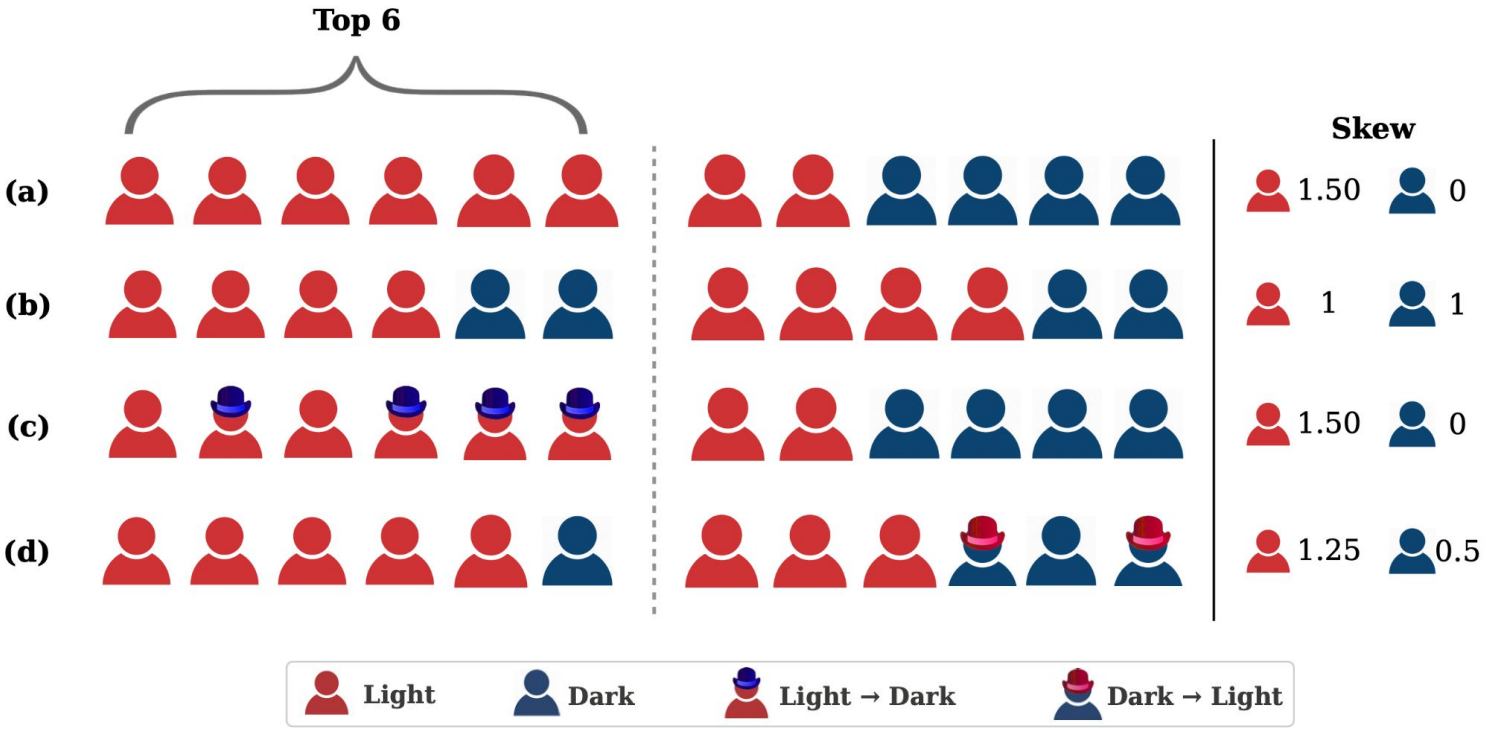
Effect of Training Objective



We observe that light men were getting unfair boosts no matter what the direction of the misprediction objective was. Why?

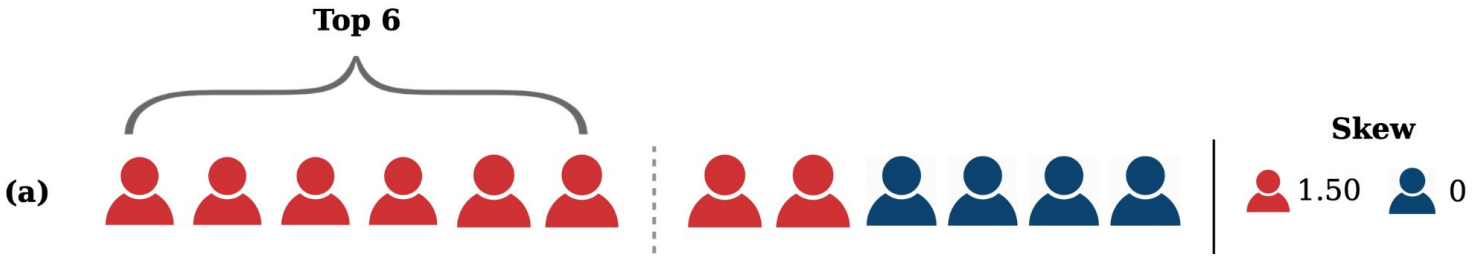
Minorities Always Harmed

Minorities always harmed



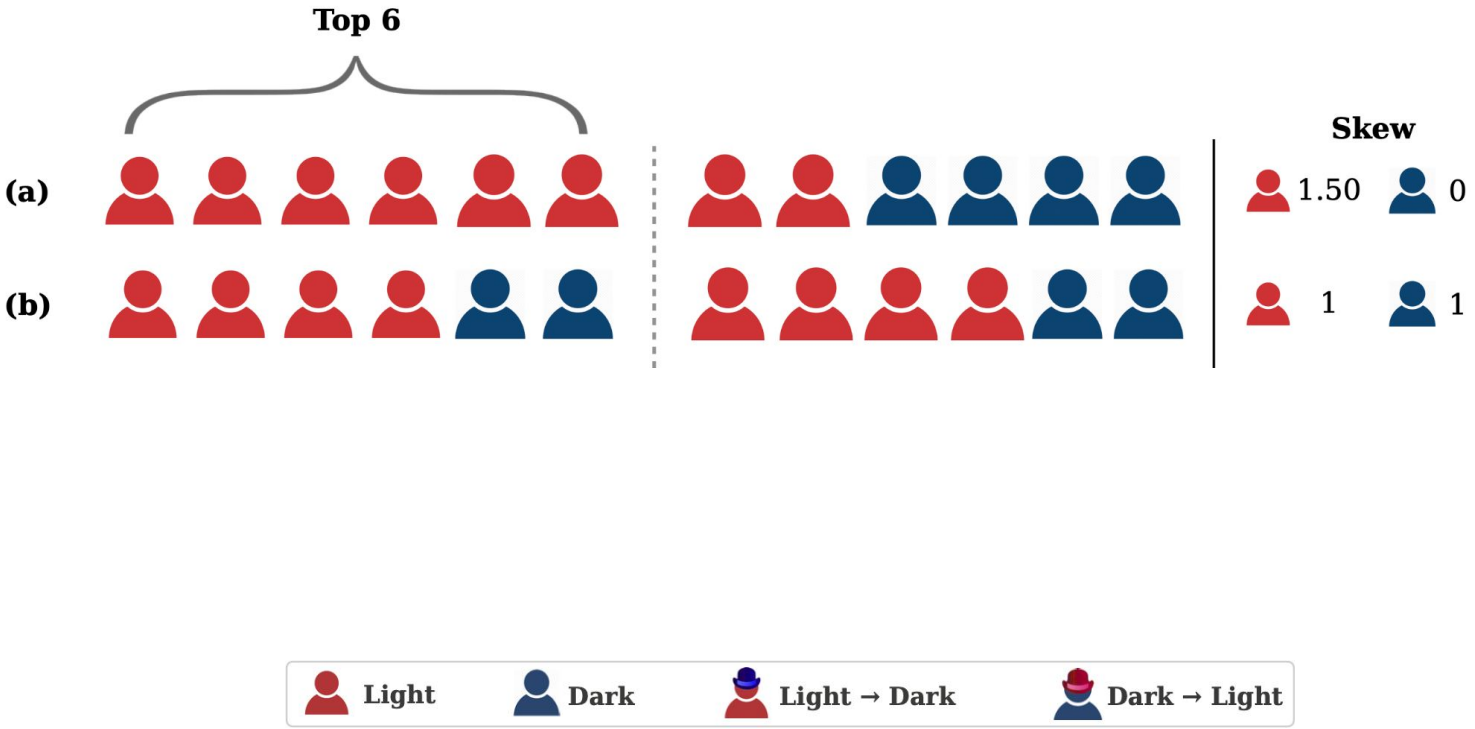
An example showing how incorrect group allocation in any direction always harms the minority group members in fair ranking.

Minorities always harmed



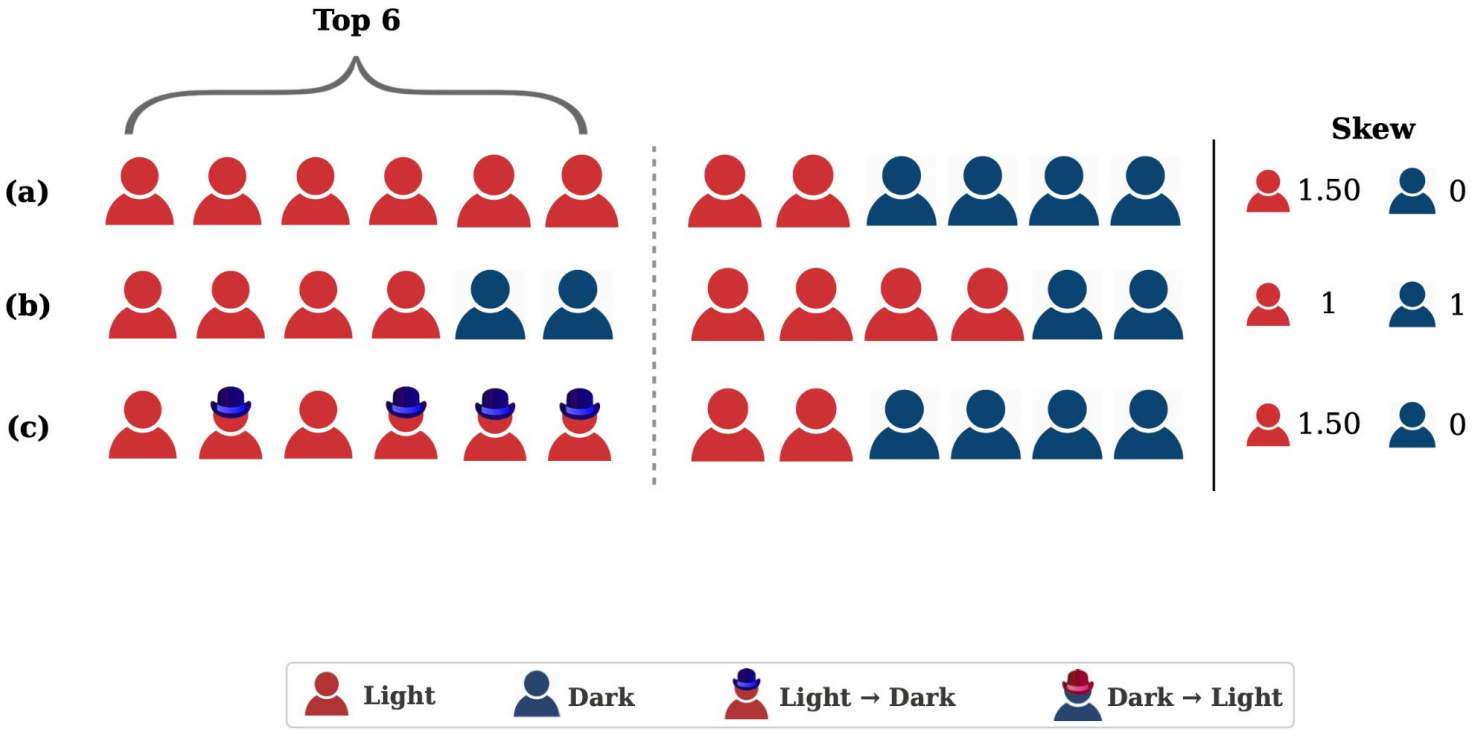
(a) shows a **baseline** unfair list, with all people sorted by relevance to the query and no dark people in the top 6.

Minorities always harmed



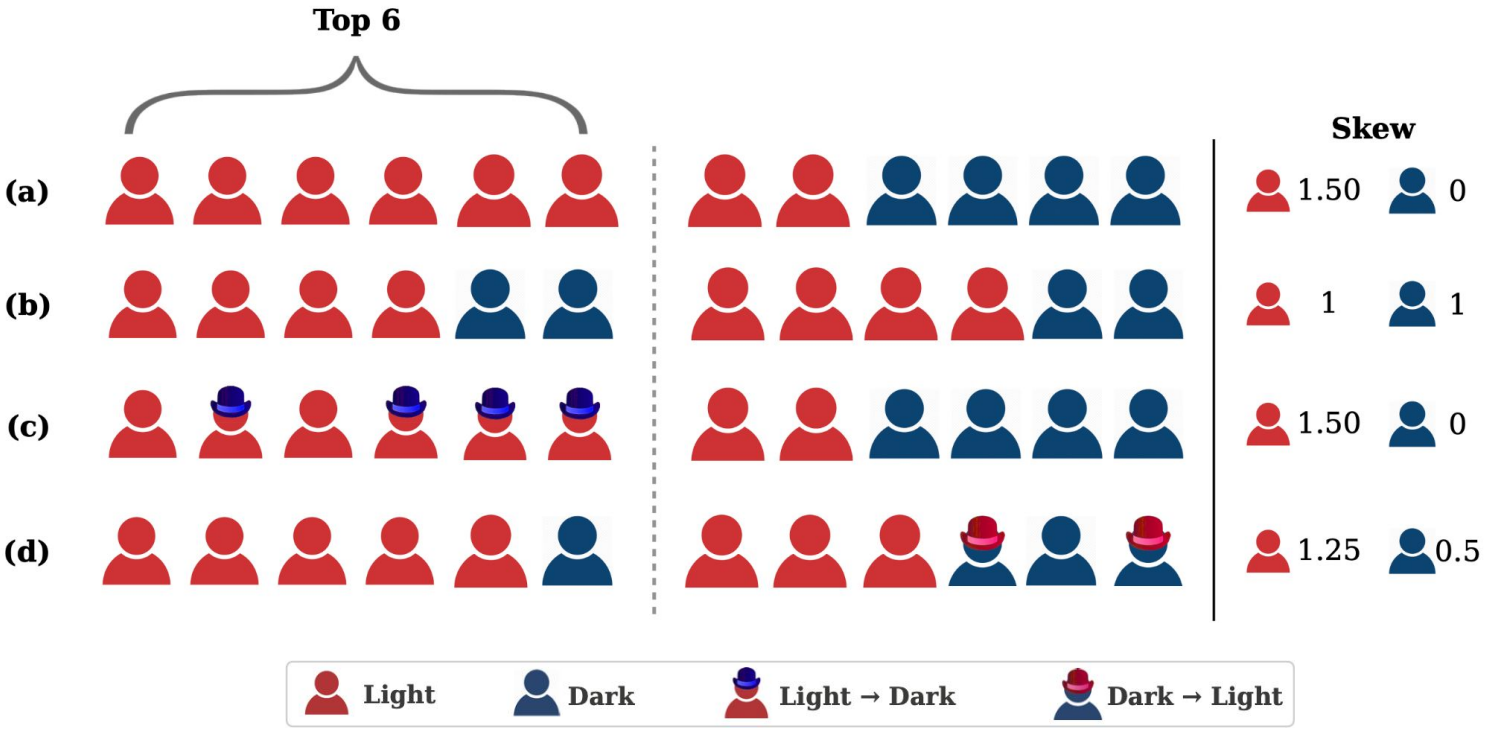
(b) shows a fair ranking produced by an algorithm, with the same proportion of light and dark people in the top 6 as the overall population.

Minorities always harmed



In **(c)**, Half of the light people are misgrouped with dark people. The fair ranker selects the "most relevant" dark skinned images, which are actually white people (see subfigure **a**)

Minorities always harmed



In **(d)**, half of the dark people are misgrouped as light people. To the fair algorithm, this appears to reduce the overall population of dark people, so it only needs to move one dark person into the top 6 to make the list proportionally fair. Note that if all light people were grouped as dark or all dark people were grouped as light, the ranking would remain the unfair baseline shown in **(a)**.

Conclusion

- The attacks can successfully confer **significant unfair advantage to people from the majority class** (light-skinned men, in the case study)—in terms of their overall representation and position in search results—relative to fairly-ranked baseline search results.

- The attack is **robust** across a number of variables, including the length of search result lists, the fraction of images that the adversary is able to perturb, the fairness algorithm used by the search engine, the demographic inference algorithm used to train the GAP models, and the training objective of the GAP models.

- The attacks are **stealthy**, i.e., they have close to zero impact on the relevance of search results, and the perturbations are invisible to the human eye.

References

[1] Julia Angwin et al. "Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. 2016". In: *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2019).

[2] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. 2018, pp. 77–91.

[3] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. "Ranking with Fairness Constraints". In: *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2018.

[4] Gregor Geigle et al. "Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval". In: *arXiv preprint* abs/2103.11920 (2021). arXiv: 2103.11920. URL: http://arxiv.org/abs/2103.11920.

[5] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2221–2231.

[6] Chen Karako and Putra Manggala. "Using image fairness representations in diversity-based re-ranking for recommendations". In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 2018, pp. 23–28.

[7] Vedant Nanda et al. "Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 466–477.

[8] Omid Poursaeed et al. "Generative adversarial perturbations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4422–4431.

[9] Dora Zhao, Angelina Wang, and Olga Russakovsky. "Understanding and Evaluating Racial Biases in Image Captioning". In: *International Conference on Computer Vision (ICCV)*. 2021.

Northeastern University

Thank you.