Subverting Fair Image Search with Generative Adversarial Perturbations

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

Matthew Jagielski
jagielski@google.com
Google Research
Mountain View, CA, USA

Christo Wilson
cbw@ccs.neu.edu
Northeastern University
Boston, MA, USA

ABSTRACT

In this work we explore the intersection fairness and robustness in the context of ranking: *when a ranking model has been calibrated to achieve some definition of fairness, is it possible for an external adversary to make the ranking model behave unfairly without having access to the model or training data?* To investigate this question, we present a case study in which we develop and then attack a state-of-the-art, fairness-aware image search engine using images that have been maliciously modified using a *Generative Adversarial Perturbation* (GAP) model [75]. These perturbations attempt to cause the fair re-ranking algorithm to unfairly boost the rank of images containing people from an adversary-selected subpopulation.

We present results from extensive experiments demonstrating that our attacks can successfully confer significant unfair advantage to people from the majority class relative to fairly-ranked baseline search results. We demonstrate that our attacks are robust across a number of variables, that they have close to zero impact on the relevance of search results, and that they succeed under a strict threat model. Our findings highlight the danger of deploying fair machine learning algorithms in-the-wild when (1) the data necessary to achieve fairness may be adversarially manipulated, and (2) the models themselves are not robust against attacks.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; • **Security and privacy**;

KEYWORDS

Information Retrieval, Fair Ranking, Adversarial Machine Learning, Demographic Inference

ACM Reference Format:
Avijit Ghosh, Matthew Jagielski, and Christo Wilson. 2022. Subverting Fair Image Search with Generative Adversarial Perturbations. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533128

1 INTRODUCTION

The machine learning (ML) community has awoken to concerns about the *fairness* of ML models, i.e., the elimination of unjustified bias against specific groups of people from models. There is now extensive literature documenting unfairness in deployed ML systems [7, 9, 21] as well as techniques for training fair classification [42, 47, 50, 65] and ranking [28, 86, 98] models. Companies are adopting and deploying fair ML systems in many real-world contexts [4, 10, 96].

Most ML systems that strive to achieve demographic fairness are dependent on high-quality demographic data to control for unjustified biases [13, 94]. Recent work has highlighted how critical this dependency is by showing how *unintentional errors* in demographic data can dramatically undermine the objectives of fair ranking algorithms [39].

Another serious concern in the ML community is model *robustness*, especially in the face of clever and dedicated adversaries. The field of adversarial ML has demonstrated that seemingly accurate models are brittle when presented with maliciously crafted inputs [24, 89], and that these attacks impact models across a variety of contexts [12, 31, 46, 89]. The existence of adversarial ML challenges the use of models in real-world deployments.

In this work we explore the intersection of these two concerns—fairness and robustness—in the context of ranking: *when a ranking model has been carefully calibrated to achieve some definition of fairness, is it possible for an external adversary to make the ranking model behave unfairly without having access to the model or training data?* In other words, can attackers *intentionally* weaponize demographic markers in data to subvert fairness guarantees?

To investigate this question, we present a case study in which we develop and then attack a fairness-aware image search engine using images that have been maliciously modified with *adversarial perturbations*. We chose this case study because image retrieval based on text queries is a popular, real-world use case for neural models (e.g., Google Image Search, iStock, Getty Images, etc.), and because prior work has shown that these models can potentially be fooled using adversarial perturbations [101] (although not in the context of fairness). To strengthen our case study, we adopt a strict threat model under which the adversary cannot *poison* training data [48] for the ranking model, and has no knowledge of the ranking model or fairness algorithm used by the victim search engine. Instead, the adversary can only add images into the victim's database *after* the image retrieval model is trained.

For our experiments, we develop an image search engine that uses a state-of-the-art MultiModal Transformer (MMT) [37] retrieval model and a fair re-ranking algorithm (FMMR [51]) that aims to achieve demographic group fairness on the ranked list of image

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533128

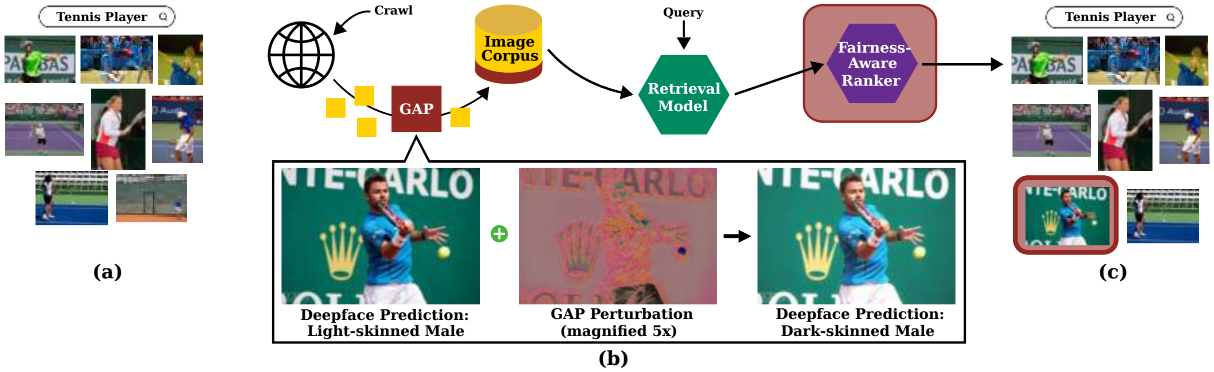

Tennis Player
(a)
Crawl
Image Corpus
GAP
Query
Retrieval Model
Fairness-Aware Ranker
Deepface Prediction: Light-skinned Male
GAP Perturbation (magnified 5x)
Deepface Prediction: Dark-skinned Male
(b)
Tennis Player
(c)

Figure 1: A diagram showing our attack approach. (a) shows example search results from an image search engine for the query "tennis player". This search engine attempts to provide demographically-fair results, and at this point no images in the corpus have been adversarially perturbed. (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. We show a real example of one image before and after applying the generated perturbation, which causes the Deepface model [90] to misclassify this person's skin tone. (c) in response to a future query for "tennis player", the retrieval model will identify relevant images, some of which are perturbed. 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.

query results without ever explicitly using demographic labels. Under normal circumstances, where the images are unperturbed, our search engine returns demographically balanced sets of images in response to free text queries. We then train a Generative Adversarial Perturbation (GAP) model [75] that learns from pretrained demographic classifiers to strategically insert human-imperceptible perturbations into images. These perturbations attempt to cause FMMR to unfairly boost the rank of images containing people from an adversary-selected subpopulation (e.g., light-skinned men). Figure 1 shows example image search results produced by our search engine in response to the query "Tennis Player", with and without our attack.

We present results from extensive experiments demonstrating that our attacks can successfully confer significant unfair advantage to people from the majority class (light-skinned men, in our case)—in terms of their overall representation and position in search results—relative to fairly-ranked baseline search results. We demonstrate that our 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 image embedding algorithm used by the search engine, the demographic inference algorithm used to train the GAP models, and the training objective of the GAP models. Additionally, our attacks are *stealthy*, i.e., they have close to zero impact on the relevance of search results.

In summary, we show that GAPs can be used to subvert fairness guarantees in the context of fair image retrieval. Further, our attack is successful under a highly restricted threat model, which suggests that more powerful adversaries will also be able to implement successful attacks. We hypothesize that similar attacks may be possible against other classes of ML-based systems that (1) rely on highly parameterized models and (2) make fairness decisions for inputs that are based on data controlled by adversaries.

The goal of our work is not to hinder or deter the adoption of fair ML techniques—we argue that fair ML techniques must be adopted in practice. Rather, our goal is to demonstrate that fairness guarantees can potentially be weaponized so that the research community will be energized to develop mitigations, e.g., by making models more robust, and by adopting high-quality sources of demographic data that are resistant to manipulation. To facilitate mitigation development without arming attackers, we plan to release our code and data to researchers on request.

2 BACKGROUND

2.1 Fairness in Ranking

Algorithmic decision making is permeating modern life, including high-stakes decisions like credit lending [11], bail granting [7], hiring [96], etc. While these systems are great at scaling up processes with human bottlenecks, they also have the unintended property of embedding and entrenching unfair social biases (e.g., sexism and homophobia). In response, there is a growing body of academic work on ways to detect algorithmic bias [9, 40] and develop classes of fair algorithms, for instance classification [42, 47, 50, 65], causal inference [62, 70], word embeddings [18, 20], regression [3, 15], and retrieval/ranking [28, 86, 98]. There is also a growing body of legal work and legislative action around the globe [1, 30, 33, 53, 73] to tackle algorithmic bias.

In this study we focus on fair Information Retrieval (IR) algorithms—a class of algorithms that have received comparatively less attention than classification algorithms in the literature. Initial studies that examined fair IR proposed to solve this in a binary context, i.e., make a ranked list fair between two groups [28, 98]. Subsequent work uses constrained learning to solve ranking problems using classic optimization methods [86]. There are also methods that use pairwise comparisons [16] and describe methods to achieve fairness in learning-to-rank contexts [68, 99].

In industrial settings, researchers at LinkedIn have proposed an algorithm that uses re-ranking in post-processing to achieve representational parity [38]. However, recent work by Ghosh et al. [39] shows how uncertainty due to incorrect inference of protected demographic attributes can undermine fairness guarantees in IR contexts. Fairness methods that do not require explicit demographic labels at runtime are, as of this writing, an emerging area of focus in classification [55] and ranking [51, 79]. One example that has been studied at large-scale is Shopify's Fair Maximal Marginal Relevance (FMMR) algorithm [51], which we describe in more detail in § 3.2.2. In this study, we examine the robustness of Shopify and LinkedIn's fair ranking algorithms.

2.2 Adversarial Machine Learning

Adversarial ML is a growing field of research that aims to develop methods and tools that can subvert the objectives of ML algorithms. For example, prior research has highlighted that deep learning models are often not robust when presented with inputs that have been intentionally, maliciously crafted [24, 43, 67, 74, 85, 95, 97].

Several proposed defenses against state-of-the-art adversarial ML attacks have been defeated [8, 91], and *adversarial examples* (i.e., maliciously crafted inputs) have been shown to transfer across models performing similar tasks [60, 92]. The most promising defense method, adversarial training, is computationally expensive and imperfect—it results in decreased standard accuracy while still having a somewhat low adversarial accuracy [35, 54, 84]. As such, adversarial ML presents a significant hurdle to deploying neural models in sensitive real-world systems.

Our work considers adversarial ML attacks on IR systems. Previous work has demonstrated successful attacks on image-to-image search systems [61, 101], allowing an adversary to control the results of targeted image queries using localized patches [19] or *universal adversarial perturbations*[1] [57]. Other work has demonstrated attacks on text-to-text retrieval systems [77] and personalized ranking systems [46]. Work by Zhou et al. [101] hypothesized that targeted attacks on selected items in a ranked list might be possible using universal adversarial perturbations. None of these works consider compromising text-to-image models or group fairness objectives, as we do in this study.

Prior work has demonstrated adversarial ML attacks against fairness objectives of ML systems at *training time*. In these attacks, the adversary supplies *poisoned* training data, which then results in models that either compromise the accuracy of targeted subgroups [48, 64] or exacerbate pre-existing unfairness between subgroups [29, 87]. Specific to classification, there exists theoretical work that shows how to learn fair models when sensitive attributes are noisy [25] or corrupted in a poisoning attack [27], but they do not consider ranking.

Adversarial ML attacks at *test time*—i.e., after training a model using non-malicious data—that we consider in this work are relatively unexplored in fairness settings. Nanda et al. [71] show that adversarial ML attacks can harm certain subpopulations more than others in classification tasks. However, while this is an important observation, the harms suggested by this work may be difficult to realize in practice, as they only involve disparity between examples that are adversarially corrupted. By contrast, our work shows that test-time attacks can harm fairness for benign data when launched in a ranking setting.

3 METHODOLOGY

We now present the plan for our study. First, we introduce our application context and the threat model under which an attacker will attempt to compromise the application. Second, we discuss the IR models and algorithms underlying our fairness-aware image search engine. Third, we discuss our strategy for attacking this search engine using GAPs.

3.1 Context and Threat Model

In this study, we consider the security of a fairness-aware image search engine. This search engine indexes images from around the web (either automatically via a crawler or from user-provided submissions) and provides a free-text interface to query the image database. Examples of image search engines include Google Image Search, iStock and Getty Images, and Giphy. In our case, the image search engine attempts to produce results that are both relevant to a given query and fair, according to some fairness objective. One example fairness objective is demographic representativeness, i.e., for search results that contain images of people.

We consider a malicious image curator (e.g., Imgur, 4chan, or similar) with a large database of *perturbed* images that are eventually scraped or uploaded into the victim image search engine's index.[2] Our adversarial image curator's goal is to perturb the images in their database to subvert the fairness guarantees of the downstream retrieval system. We assume that the adversary does not have any knowledge of the internals of the ranking system (e.g., what retrieval model is used, other images in the index, or which fairness algorithm is used).

This threat model constitutes a strict, but realistic, limitation on our adversary. Notice that this threat model would also apply if the image search engine was compromised, giving the adversary access to underlying models and the entire dataset of images. We consider both adversaries in our experiments. We also note that, if the adversary only seeks to target a small set of queries, they need only control a fraction of the images matching each query, rather than a fraction of the entire image database. This is useful for the adversary in the case that not all queries are equally sensitive.

3.2 Building an Image Search Engine

We now turn our attention to building a realistic image search engine that will serve as the victim for our attacks.

3.2.1 *Image Retrieval from Text Queries.*

The first choice we make for this study is to select an image retrieval model. There are several frameworks for image retrieval in the literature, starting from

[1] A Universal Adversarial Perturbation (UAP) is a an adversarial perturbation that generalizes to all classes of a target classifier, i.e., one patch to untargeted attack as many classes as possible.

[2] An adversarial image curator is also the threat model assumed for clean label poisoning attacks [83, 93]. This adversarial image curator may perturb copies of images taken from the web or original images that they author. This setup is also used by [85] as a defensive method against unauthorized models.

tag-based matching [56] to state-of-the-art vision-language transformers [58, 63]. For the purpose of this paper, we used a Multi-Modal Transformer (MMT) [37] 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. Concretely, the user provides a string q that queries into a database D of n images. For the retrieval step, the query string is encoded with an embedding function f_q to produce an embedding v_q, and all images in D have pre-computed embeddings from an embedding function f_I. The cosine distance between v_q and all embeddings of D are computed to collect some large set D_q of size $n' \leq n$ plausible image matches. These images are then ranked according to a joint model f_j that takes both the query and an image as input, returning scores $\{s_i\}_{i=1}^{n'}$ indicating how well each image $D_q[i]$ matches the query. These scores are used to produce the final ranking.

Note that the MMT model is not designed to be "fair" in any normative sense. To achieve fairness, results from the model must be re-ranked, which we describe in the next section. Thus, the MMT model is not the target of our attacks, since it is not responsible for implementing any fairness objectives.

3.2.2 Fairness-aware Re-ranking. The second choice we make for this study is selecting an algorithm that takes the output of the image retrieval model as input and produces a fair re-ranking of the items. In fairness-aware re-ranking, a ranking function $f_r(D, q)$ is post-processed to achieve fairness according to some subgroup labels on the dataset $D = \{s_i, x_i\}_{i=1}^{n}$, where s_i denotes the score of the i^{th} item (the heuristic score according to which the list is sorted) and x_i denotes the item to be ranked.

The re-ranking algorithm we adopt is Fair Maximal Marginal Relevance (FMMR) [51], which was developed and used at Shopify for representative ranking of images. FMMR builds on the Maximal Marginal Relevance [23] technique in IR that seeks to maximize the information in a ranked list by choosing the next retrieved item in the list to be as dissimilar to the current items present in the list as possible. MMR introduces a hyperparameter that allows the operator to choose the trade-off between similarity and relevance.

FMMR modifies the "similarity" heuristic from MMR to encode for similarity in terms of demographics, with the idea being that the next relevant item chosen to be placed in the re-ranked list will be as demographically different from the existing images as possible. Similarity is calculated using image embeddings, for which we examine three models: Faster R-CNN [78], InceptionV3 [88], and ResNet18 [45]. We fix the trade-off parameter λ at 0.14 as that is the value used by Karako and Manggala [51] in their FMMR paper.

It is notable that FMMR does not require demographic labels of people in images to perform fair re-ranking, since it uses a heuristic that only relies on embeddings. Indeed, FMMR comes from a class of fair ranking algorithms that all use the inherent latent representations of the objects for their re-ranking strategy [51, 79]. That said, since FMMR attempts to maximize the distance from the centroids of the embeddings of different demographic groups, it can be thought of as performing indirect demographic inference on individuals in images.

Additionally, we also evaluated our attacks against a second fairness-aware re-ranking algorithm, DetConstSort [38], developed by and deployed at LinkedIn in their talent search system. Unlike FMMR, DetConstSort explicitly requires demographic labels for the items it is trying to fairly re-rank. However, prior work [39] shows that DetConstSort has significant limitations when demographic inference is used rather than ground-truth demographic labels, making it unfair even without perturbed images. As a result, evaluating an attack against DetConstSort is not meaningful, and we defer our discussion of DetConstSort to § A.1.

3.3 Attack Construction

Having described our search engine, we are ready to turn our attention to our attack. First, we introduce the demographic inference models (Deepface [90] and FairFace [52]) that we use to train our attack. Next, we describe how we generate adversarial perturbations from a demographic inference model, modifying images in a way that is imperceptible to human eyes, yet significant enough to fool the fair re-ranking algorithm of our search engine.

3.3.1 Demographic Inference Algorithms. For large-scale datasets such as images scraped from the web, demographic meta-data for people in the images is (1) not readily available and (2) prohibitively expensive to collect through manual annotation [6, 17]. Pipelines using demographic inference are commonly used in practice when demographic labels are not available. For example, the Bayesian Improved Surname Geocoding (BISG) tool is used to measure fairness violations in lending decisions [2, 22], and it relies on inferred demographic information. This makes attacks on demographic inference models a natural candidate for adversely affecting ranking fairness.

We consider two image demographic inference models to train our attacks:

1. Deepface [90] is a face recognition model for gender and race inference developed by Facebook. We use its public wrapper [82], which includes models fine tuned on roughly 22,000 samples for race and gender classification.
2. FairFace [52] is a model designed for race and gender inference, trained on a diverse set of 108,000 images.

Since both of these models infer race/ethnicity, we used a mapping to infer skin tone, since we could not find commercially available algorithms to infer skin tones from human images.[3] We also use these models to infer demographics as input to the DetConstSort algorithm, matching the pipeline of [39], which we discuss in § A.1.

3.3.2 Subpopulation Generative Adversarial Perturbations. Recall our adversarial image curator's goal: to produce a database of malicious images that, when indexed by our image search engine, undermine its purported fairness guarantees. Concretely, this means fooling the fair re-ranker such that it believes a given set of search results is fair across two or more subgroups, when in fact the results are unfair because some subgroups are under- or over-represented. Additionally, these malicious images must (1) retain their relevance

[3] The mapping we used is: White, East Asian, Middle Eastern → Light, and Black, South Asian, Hispanic → Dark. We acknowledge that this is a crude mapping, but it enabled us to train a successful attack.

to a given query and (2) not be perceived as "manipulated" to human users of the search engine.

Prior work (see § 2.2) has demonstrated that neural image classification models can be fooled by adding *adversarial perturbations* to images. At a high-level, the adversary's goal is to train a model that can add noise to images such that specific latent characteristics of the images are altered. In our case, these altered characteristics should impact the image embeddings calculated by the image embedding model (e.g., InceptionV3) that FMMR relies upon to do fair re-ranking.

Running an adversarial perturbation algorithm on each of the images in the adversary's database would be prohibitive, as these algorithms involve computationally expensive optimization algorithms that are not practical at the scale of an entire database. We avoid this limitation by training a Generative Adversarial Perturbation (GAP) model [75]. A GAP model f_{GAP} takes a clean image as input and returns a perturbed image that is misclassified by some target model f_{targ}. This replaces the per-image optimization problem with a much less expensive forward pass of f_{GAP}. Training the GAP is a one time expense for the adversary, amortized over the large number of image perturbations done later. Universal Adversarial Perturbations (UAPs) [66] are another approach to amortizing runtime, but require all images to be the same dimensions—an unrealistic assumption for real-world image databases.[4]

Having motivated the choice of a GAP model for our attack, we now consider the problem of impacting fairness by attacking the fair re-ranking algorithm used by a victim search engine. We choose to design a GAP to target a demographic inference model f_{DI}.[5] This will produce perturbations that, to a deep image model, make an image of a person from one demographic group appear to be from a different demographic group. This attack would heavily impact a demographic-aware re-ranking algorithm such as DetConstSort [38] (see § A.1) if it used an accurate demographic inference algorithm to produce annotations.

Although FMMR does not use annotations, we show in § 5 that our attack is still successful at compromising FMMR's fairness guarantees. Our attack can be seen as an application of the *transferability property* of adversarial examples. Additionally, training our GAP against a demographic inference model causes our attack to be independent of the ranking algorithm and image corpus used by the victim search engine, both of which are strong adversarial assumptions.

In designing our GAP to compromise fairness, we first note that an attack that simply forces a f_{DI} to make arbitrarily many errors may not impact fairness. For example, suppose the image database contained two subpopulations, the advantaged class A and the disadvantaged class B. Suppose the attack causes f_{DI} to misclassify all members of B as A and all members of A as B. This is the best possible result of an attack on the demographic inference algorithm, but results in no changes to a fair ranking algorithm—it will simply consider A to be the disadvantaged class, and thus produce the same ranking! For this reason, our adversary must incorporate subpopulations into the attack. To do so, we propose the Class-Targeted Generative Adversarial Perturbation (CGAP):

Definition 1 (CGAP). *We consider a loss function ℓ, target model f_{targ}, distribution $\mathcal{D}$ over inputs x and outputs y. The adversary provides a source class y_s and target class y_t. Then the CGAP model f_{CGAP} is a model that takes as input an image x and returns an image x', minimizing the following loss functions:*

$$\ell^s_{CGAP}(\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_{CGAP}(x), y_t; f_{targ})|y = y_s],$$
$$\ell^r_{CGAP}(\mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_{CGAP}(x), y; f_{targ})|y \neq y_s].$$

That is, the CGAP should force 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.

We also consider two extensions of this definition. First, we permit the adversary to target multiple classes at once. In the extreme, an adversary may want all samples to be classified to the same class (this approach is proposed by [75]). For a demographic inference algorithm, all samples having the same demographic label will cause the fair re-ranking system to have similar performance to an unfair ranking system, as all points will appear to fall into the same subpopulation. The second extension is the untargeted attack, where the CGAP simply increases loss for points from class y_s, inducing arbitrary misclassifications. Simultaneously making both relaxations recovers the original untargeted GAP approach. We experiment with both relaxations independently, as well as multiple instantiations of CGAP as defined above.

4 EXPERIMENTS

In this section we introduce the dataset we used for our evaluation, describe the setup for our experiments, and define the metrics we use to evaluate our attacks.

4.1 Dataset, Annotation, and Preprocessing

We use Microsoft's Common Objects in Context (MS-COCO) [59] as our retrieval dataset, since it contains a variety of images with variable dimensions and depths. This closely mimics what a real-world image search dataset might contain.

To specifically measure for demographic bias, we filter the dataset, keeping only images that contain people. We also need the images to have demographic annotations for fair ranking, so we use an annotated subset of the COCO 2014 dataset, constructed by Zhao et al. [100]. Similar to prior work [26, 39], Zhao et al. crowdsource skin color (on the Fitzpatrick Skin Type Scale, which the authors simplified to Light and Dark) and binary perceived gender expression for 15,762 images. For the purposes of our experiments we only considered the 8,692 images that contain one person. After filtering, our final dataset consisted of 5,216 Light Men, 2,536 Light Women, 714 Dark Men, and 226 Dark Women.

4.2 Experimental Setup

As a starting point for our experiments, we need to collect ranked lists from our baseline, **unfair** retrieval system, as described in § 3.2.1. To do so, we run three different search queries on the retrieval system: "Tennis Player", "Person eating Pizza", and "Person at table". We chose these queries because they all reference a human

[4] A UAP can also be seen as a GAP, where $f_{\text{GAP}}(x) = x + \delta$ for a fixed δ. Therefore, we expect that a GAP will perform strictly better than a UAP.

[5] Recall that, per our threat model in § 3.1, the attacker does not know what fair re-ranking algorithm is used by the victim and thus cannot train against it directly.

Search Queries	Attack Training	Embedding	Training Objective	Attack Probability	Top k
"Tennis Player" "Person eating pizza" "Person at Table"	Deepface FairFace	F-RCNN InceptionV3 ResNet	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

Table 1: Variables and hyperparameters we used for evaluating our attack.

being, are ethnicity and gender neutral, and are well-supported in the COCO dataset (we picked popular object tags, see § A.2). We set the upper bound in the baseline retrieval system to be 200 images. The three queries return 131, 75, and 124 images, respectively, along with their relevance scores.

We show the distribution of the relevance scores and the skin color/gender distributions of the images within the top 40 search results for each query in Figure 2. As also shown by Zhao et al. [100], Light Men comprise the overwhelming majority in all three lists, and they also have high relevance scores across the board, meaning that the retrieval system places Light Men near the top of the search results. We call these lists the *baseline* lists.

We also need to produce fair versions of the baseline lists. To do so, we pass the baseline lists for each of our three queries through *FMMR* with the three embedding algorithms, without any adversarial perturbations. We refer to the nine lists obtained via the fair re-ranker (three queries times three image embedding models) as the *oracle* lists.

To train our adversarial attacks, we first remove the 330 images in our *oracle* lists from the original dataset, leaving 8,362 images. These 8,362 images were then split randomly into training and testing sets in an 8:2 ratio to train CGAP models for all possible combinations of training objectives and demographic inference algorithms f_{DI} (described in detail below). We ran our experiments on PyTorch with a CUDA backend on two NVIDIA RTX-A6000 GPUs, and trained CGAP models for 10 epochs each, with the L_∞ norm[7] bound set to 10.

We describe our different training and inference combinations below. Table 1 shows a summary of the different settings involved during the training and testing of our CGAP attacks.

4.2.1 *Embedding Algorithm.*

As we discuss in § 3.2.2, FMMR requires image embeddings. The authors of the original paper used a pretrained InceptionV3 model, which we also adopt. Additionally, we test the performance of FMMR using embeddings generated by pretrained Faster R-CNN and ResNet models. These models are trained for standard image classification tasks and have no inherent concept of demographic groups.

4.2.2 *Attack Training Algorithm.*

As detailed in § 3.3, we train CGAP models to induce adversary-selected misclassifications in two target demographic inference models, denoted as f_{DI}: Deepface [90] and FairFace [82]. These models are trained for demographic inference, and so do not overlap in training objective with the image embedding models for FMMR. The only similarity in architecture between the demographic inference and FMMR embedding models is that FairFace uses a ResNet architecture.

4.2.3 *Training Objectives.*

As discussed in § 3.3.2, we select certain subpopulations to be systematically misclassified by the two f_{DI} described above. The four CGAPs we train induce misclassifications with the following source-target pairs: Any→Light Men, where every subgroup was perturbed to be predicted as Light Men; Light Men→Any, where only Light Men are arbitrarily misclassified; Dark Men→Light Men, where only Dark Men are misclassified as Light Men; and Light Men→Dark Men, where only Light Men are misclassified as Dark Men.

4.2.4 *Attack Probability pr.*

It is a strong assumption that an adversary can perturb the entire image database of a victim search engine. This is only possible if the search engine itself is malicious or it is utterly compromised. Instead, we measure the effect of our attack when the attacker may perturb pr = 20%, 50%, 70% and 100% of the image database relevant to each query. If a small number of queries are targeted, only few images are required to run the attack.

4.2.5 *Top k.*

Ranking is very sensitive to position bias [39, 80], so we measure with different lengths k of the top list, ranging from top 10 to top 50, to gauge our attack's impact on the fair ranking algorithms as final list sizes vary.

4.3 Evaluation Metrics

To evaluate the impact of our attacks, we use three metrics that aim to measure (1) representation bias, (2) attention or exposure bias, and (3) loss in ranking utility due to re-ranking. Additionally, we introduce a summarizing meta-metric that enables us to clearly present the impact of our attacks with respect to each metric.

4.3.1 *Skew.*

The metric we use to measure the bias in representation is called Skew [38, 39]. For a ranked list τ, the Skew for attribute value a_i at position k is defined as:

$$\text{Skew}_{a_i}@k(\tau) = \frac{p_{\tau^k, a_i}}{p_{q, a_i}}. \tag{1}$$

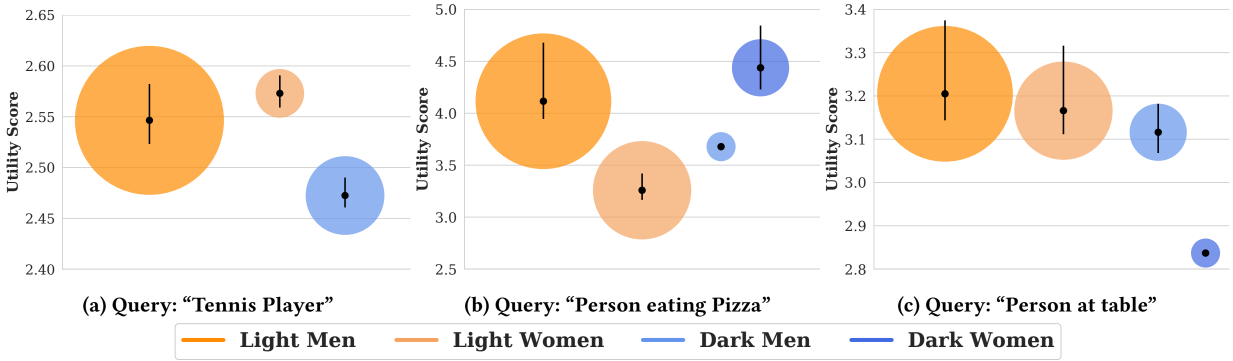
p_{τ^k, a_i} represents the fraction of members having the attribute a_i among the top k items in τ, and p_{q,a_i} represents the fraction of members from subgroup a_i in the overall population q. In an ideal, fair representation, the skew value for all subgroups is equal to 1, indicating that their representation among the top k items exactly matches their proportion in the overall population.

4.3.2 *Attention.*

Even if all subgroups were fairly represented in the top k ranked items of a list, the relative position of the ranked items adds another dimension of bias—unequal exposure. Previous

[6] Dark-skinned women do appear in the search results for the query "female tennis player". This seems to reflect stereotypical bias [36] within the learned-word representations in the MMT model.

[7] L_∞ is the absolute distance in pixel space any one pixel is changed, i.e. a pixel can at most change by a value of 10 in each color channel.



(a) Query: "Tennis Player"

(b) Query: "Person eating Pizza"

(c) Query: "Person at table"

Figure 2: Utility/Relevance score and group size distribution within the top 40 baseline search results for three queries. The black dots represent the average utility score for that group, while the circle size represents the group size. No dark-skinned women appear in the top 40 baseline results for the "tennis player" query.[6]

studies [69, 72] have shown that people's attention rapidly decreases as they scan down a list, with more attention given to the higher ranking items, ultimately dropping to zero attention.

In this study, we model attention decay using the geometric distribution as done in prior work by Sapiezynski et al. [80]. We compute attention at the k^{th} rank as:

$$\text{Attention}_p@k(\tau) = 100 \times (1-p)^{k-1} \times (p) \tag{2}$$

where p is the fraction of total attention given to the top search result. The choice of p is application specific—for this paper we fixed p to be 0.36, based on a study [44] that reported that the top result on Google Search receives 36.4% of the total clicks. We then calculate the average attention per subgroup:

$$\text{Average attention}_{a_i,\tau} = \frac{1}{|a_i|}\sum_{k=1}^{|\tau|} \text{Att(k) where } a_k^\tau = a_i. \tag{3}$$

Ideally, in a perfectly fair ranked list, all subgroups should receive equal average attention.

4.3.3 *Normalized Discounted Cumulative Gain.* NDCG is a widely used measure in IR to evaluate the quality of search rankings [49]. It is defined as

$$\text{NDCG}(\tau) = \frac{1}{Z}\sum_{i=1}^{|\tau|} \frac{s_i^\tau}{log_2(i+1)} \tag{4}$$

where s_i^τ is the utility score from the MMT retrieval model of the i^{th} element in the ranked list τ and $Z = \sum_{i=1}^{|\tau|} \frac{1}{log_2(i+1)}$. NDCG scores range from 0 to 1, with the latter capturing ideal search results.

4.3.4 *Summarizing Metric.* For the purpose of quantifying how much unfair advantage our attacks confer on members of the majority class relative to all other classes, we define a new meta-metric called Attack Effectiveness η. For a given metric $m \in \{$ Skew, Attention $\}$ and a subgroup g, it is defined as:

$$\begin{aligned}\eta(m,g) = \ &\text{\% change in } m \text{ for subgroup } g - \\ &\text{minimum \% change in } m \text{ over other subgroups.}\end{aligned} \tag{5}$$

We chose this formulation of η for two reasons. First, comparing percentage changes makes the metric scale invariant, which is useful since group sizes vary. Second, comparing to the group that gets the minimum boost ensures that the metric presents the widest fairness disparity, regardless of the total number of groups.

For the purposes of this paper, we set g as Light Men, because they are socially and historically the most advantaged group, and a large η for Light Men indicates that the attack causes their ranking to be unfairly boosted relative to the least privileged group. To make sure that the fairness impacts we observe are due to the effectiveness of our attack on the re-ranking algorithms only, the η values and the % change in NDCG are all measured against the *oracle* (i.e., fairly re-ranked) lists. Because we compare against the oracle list, all results with attack probability $pr = 0$ will have $\eta = 0$.

5 RESULTS

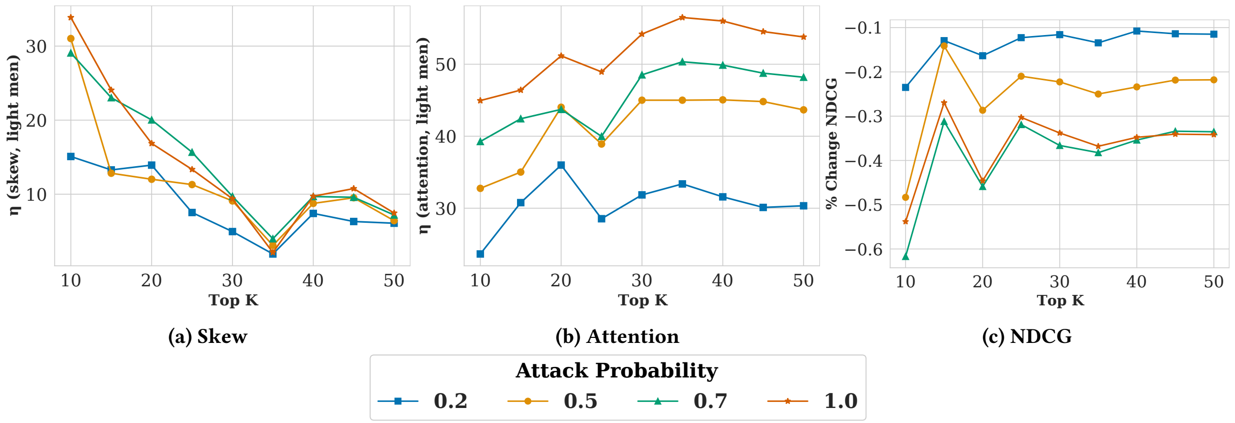
In this section, we evaluate the impact of our attacks on the fairness guarantees of FMMR. For each set of results we examine how attack effectiveness varies for one particular variable (e.g., top k, image embedding model, etc.) as the attack probability pr (i.e., the fraction of images under adversarial control) varies. When focusing on a particular variable, we present results that are averaged across all other variables and all three of our queries.

5.1 Top k and pr

We begin by evaluating the impact of our attacks as we vary the length of the top list k and the fraction of images in the query list under adversarial control pr, plotted in Figure 3.

Varying pr has the expected effect: as the adversary has more control over the image database, attacks become more effective, i.e., η for skew and attention increase. When the adversary is able to control 100% of images in the query list, attacks are especially strong—increasing attention unfairness by over 50% for some values of k. Even with only 20% control, the adversary can increase attention unfairness by ~30%. Recall that pr measures the fraction of each query list that is compromised, so as few as 35 images can be compromised at $pr = 0.5$ (for the "Person eating Pizza" query).

Varying k also impacts ranking fairness. As k increases, attention unfairness increases modestly and skew unfairness decreases. That skew unfairness decreases with k indicates that the composition of items in the search results becomes fairer as the length of the list

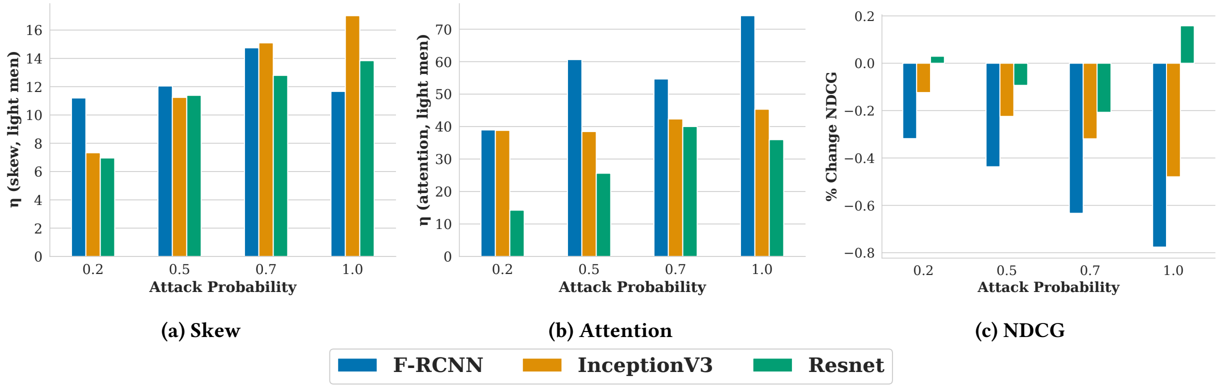


(a) Skew

(b) Attention

(c) NDCG

Figure 3: Attack effectiveness as a function of attack probability pr and list length k. Higher η is a more effective attack, i.e., the search results are more favorable to light-skinned men. Unfairness increases as pr increases, yet there is almost no impact on ranking quality (NDCG). As k increases skew is less impacted but attention is impacted somewhat more.



(a) Skew

(b) Attention

(c) NDCG

Figure 4: Attack effectiveness is stable when the model used for the FMMR embedding is changed. ResNet embeddings are slightly more robust to attack and F-RCNN are slightly less robust. Interestingly, the ResNet's robustness is in spite of it having the most similar model architecture to FairFace.

grows. However, our attack is able to cause FMMR to reorder the list such the top-most items remain unfair regardless of k, which is why attention unfairness exhibits less dependency on k.

Lastly, we observe that our attacks are stealthy. Regardless of k or pr, NDCG never changes more than 0.7%, meaning that our attack had effectively zero impact on search result relevance.

5.2 Choice of Training Objective

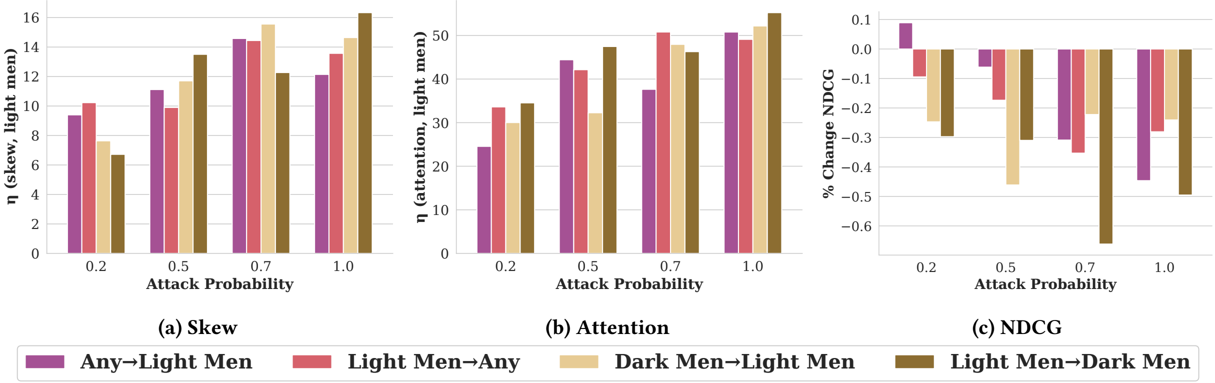
We evaluate our attack's impact on fairness with four CGAP models: one that misclassifies Dark Men as Light Men, one for misclassifying Light Men as Dark Men, and relaxed CGAP models that misclassify all people as Light Men and all Light Men as other groups. We show these attacks' effectiveness in Figure 5.

Each of these attacks performs similarly well at harming fairness in terms of skew and attention, and remaining stealthy in terms of NDCG. One surprising observation is that misclassifying Dark Men as Light Men performs similarly to the exact opposite attack: in both cases, Light Men end up with an significant, unfair advantage. We explain this seeming contradiction with an example in Figure 6. In essence, using a GAP to misclassify people from a minority group into the majority group reduces the minority group's overall share of the population. Since group fairness in this case is based on the overall population distribution, this causes FMMR to rerank fewer minority group members into the top of the search results.

Based on the results in Figure 5, it appears that there is no way to advantage a minority group with our attacks.

5.3 Choice of Attack Training Algorithm

We measure our attacks' effectiveness when the GAP models are trained on Deepface and FairFace demographic inference models. We observe that attack effectiveness is largely independent of the choice of inference model, and all attacks remain stealthy. We defer a plot of the results to the supplementary material, in Figure 8.



(a) Skew (b) Attention (c) NDCG

Figure 5: Attack effectiveness is relatively stable when the GAP training objective is changed.

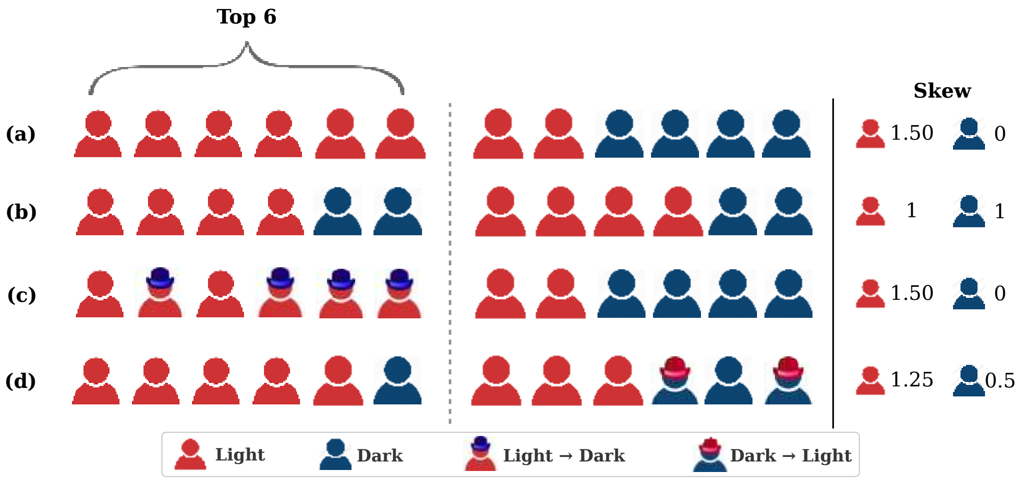


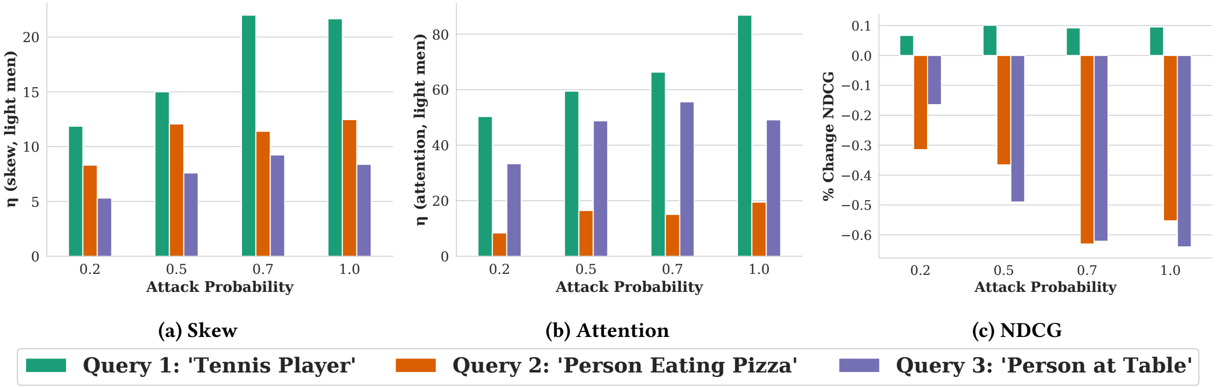
Figure 6: An example showing how incorrect group allocation in any direction always harms the minority group members in fair ranking. (a) shows a *baseline* unfair list, with all people sorted by relevance to the query and no dark people in the top 6. (b) shows the fair ranking produced by FMMR, with the same proportion of light and dark people in the top 6 as the overall population. In (c), light people's images are perturbed using a GAP so that half of them are grouped with dark people. FMMR moves the most relevant dark people into the top 6 to make the list fair, but in this case the most relevant "dark" people are really light skinned. In (d), half of the dark people are perturbed using a GAP to be grouped as light people. To FMMR, 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).

5.4 Choice of Query

Lastly, we examine the effectiveness of our attacks against three different queries and plot the results in Figure 7. We observe that all attacks were successful, but that effectiveness varies by query. The differences in attack effectiveness are explained by the underlying distributions of population and utility scores (see Figure 2). The "tennis" results exhibit the most unfairness post-attack because they were most unfair to begin with, i.e., the difference in utility scores between Light and Dark skinned people was greatest in the "tennis" results as compared to the other queries. In contrast, the "pizza" results exhibit the most robustness to attack in terms of attention because these were the only results among the queries where minority people had higher utility scores than majority people in the baseline results (Figure 2).

6 DISCUSSION

In this study, we develop a novel, adversarial ML attack against fair ranking algorithms, and use fairness-aware text-to-image retrieval as a case study to demonstrate our attack's effectiveness. Unfortunately, we find that our attack is very successful at subverting the fairness algorithm of the search engine—across an extensive set of attack variations—while having almost zero impact on search result relevance.



(a) Skew

(b) Attention

(c) NDCG

Figure 7: Attacks are effective against all three of our queries, but the effectiveness varies in relation to the underlying population and utility score distributions (see Figure 2).

Although we present a single case study, we argue that our attack is likely to generalize. We adopt a strong threat model and demonstrate that our attacks succeed even when the attacker cannot poison training data, access the victim's whole image corpus, or know what models are used by the victim. Thus, our attack is highly likely to succeed in cases where the threat model is more relaxed, e.g., when the fairness algorithm used by the victim is known.

Alarmingly, our work shows that an adversary can attack a fairness algorithm like FMMR *even when it does not explicitly rely on demographic inference*. Thus, it is highly likely that our attack will also succeed against any ranking algorithm that does rely on a demographic inference model, even if that model is highly accurate. We explore this possibility to the best of our ability in § A.1.

We hope that this research will raise awareness and spur further research into vulnerabilities in fair algorithms. Our results highlight how, in the absence of safeguards, fairness interventions can potentially be weaponized by malicious parties as a tool of oppression. In the absence of fair ML development methods and algorithms that are robust to adversarial attacks, it may not be possible for policymakers to safely mandate the use of fair ML algorithms in practice.

We believe that future work is needed to develop more robust fair ML interventions. We adopt a broad view of possible mitigations, spanning from value sensitive design [34] methods that help developers preemptively identify attack surface and plan defenses [32], to models that are hardened against adversarial perturbation techniques [5, 43], to auditing checklists [76] and tools that help developers notice and triage attacks.

Above all, this work highlights that achieving demographic fairness requires high-quality demographic data [6]. Allowing an adversary to influence demographic meta-data is the underlying flaw that enables our attack to succeed. Demographic data may be sourced from data subjects themselves, with full knowledge and consent, or from human labelers [10], with the caveat that these labels themselves will need to be de-biased [100].

6.1 Limitations

Our study has a number of limitations. First, our analysis is limited to two discrete racial and two discrete gender categories. Although our CGAP attack could be tailored to select any group, it is unclear how well our attack would perform in situations with > 4 discrete protected groups, groups with continuous attributes, people with multiple or partial group memberships, or with population distributions that varied significantly from our dataset. Second, while our dataset is sufficiently large to demonstrate our attack, it is smaller than the databases that real-world image search engines retrieve from. Third, our proof-of-concept was tuned to attack FairFace and Deepface. It is unclear how well our CGAP attack would generalize to other models or real-world deployed systems. Fourth, as we observe in Figure 5, our attack is only successful at generating unfairness in favor of already-advantaged groups. While this is a limitation, it in no way diminishes the potential real-world harm our attack could inflict on marginalized populations. Finally, as shown in Figure 7, our attack's effectiveness varies by query. In real world scenarios, an attacker could mitigate this to some extent by devoting more of their resources towards perturbing images that are relevant to high-value queries. It is unclear how much the attackers' effort would need to vary in practice given that we make no attempt to attack deployed search engines.

6.2 Ethics

In this work we present a concrete attack against a fair ranking system. Like all adversarial attack research, our methods can potentially be misused by bad actors. However, this also necessitates our research, since documenting vulnerabilities is the first step towards mitigating them. To the best of our knowledge, with the exception of Shopify and LinkedIn, few services are known to employ fair ranking systems in practice, meaning there currently exists a window of opportunity to preemptively identify attacks, raise awareness, and deploy mitigations.

Prior work on adversarial ML attacks against fairness made their source code publicly available [71]. However, because attack tools are dual-use, we have opted to take a more conservative approach: we will only share source code with researchers from (1) research universities (e.g., as identified by taxonomies like the Carnegie Classification) and (2) companies that develop potentially vulnerable products. Given that our attack can be used for legitimate, black-box algorithm auditing purposes, we opt to restrict who may access our source code rather than the uses it may be put towards. In our opinion, this process will facilitate follow-up research, mitigation development, and algorithm auditing without supplying bad actors with ready-made attack tools.

Like many works in the computer vision field, we rely on images with crowdsourced and inferred demographic labels. Both processes have been criticized for their lack of consent [41], the way they operationalize identity [81], and the harm they may cause through mis-identification [14]. These problems reinforce the need for high-quality, consensual demographic data as a means to improve ethical norms and defend against adversarial ML attacks.

ACKNOWLEDGMENTS

Funding/Support. This work was supported by a 2019 Sloan Fellowship and NSF grants IIS-1553088 and IIS-1910064. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the funders.

The legal department of Google participated in the review and approval of the manuscript. They asked that the authors add additional examples of image search engines besides Google Image Search, and we complied. Aside from the authors, Google had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; or preparation of the manuscript.

The authors thank Jane L. Adams, Michael Davinroy, Jeffrey Gleason, and the anonymous reviewers for their comments.

REFERENCES

[1] 116th Congress (2019-2020). [n.d.]. H.R.2231 - Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231.

[2] Dzifa Adjaye-Gbewonyo, Robert A Bednarczyk, Robert L Davis, and Saad B Omer. 2014. Using the Bayesian Improved Surname Geocoding Method (BISG) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. *Health services research* 49, 1 (2014), 268–283.

[3] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv:1905.12843* (2019).

[4] Facebook AI. 2021. How we’re using Fairness Flow to help build AI that works better for everyone. Facebook AI. https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/.

[5] Naveed Akhtar, Jian Liu, and Ajmal Mian. 2018. Defense against universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3389–3398.

[6] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can’t Measure, We Can’t Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT ’21)*. Association for Computing Machinery, New York, NY, USA, 249–260. https://doi.org/10.1145/3442188.3445888

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2019. Machine bias: There’s software used across the country to predict future criminals and it’s biased against blacks. 2016. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2019).

[8] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*. PMLR, 274–283.

[9] Solon Barocas and Andrew D Selbst. 2016. Big data’s disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[10] Sid Basu, Ruthie Berman, Adam Bloomston, John Campbell, Anne Diaz, Nanako Era, Benjamin Evans, Sukhada Palkar, and Skyler Wharton. 2020. Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data. AirBNB. https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf.

[11] Thorsten Beck, Patrick Behr, and Andreas Madestam. 2018. Sex and credit: Is there a gender bias in lending? *Journal of Banking and Finance* 87 (2018).

[12] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7345–7349.

[13] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[14] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. “It’s Complicated”: Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI ’21)*. Association for Computing Machinery, New York, NY, USA, Article 375, 19 pages. https://doi.org/10.1145/3411764.3445498

[15] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[16] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *KDD*. https://arxiv.org/pdf/1903.00780.pdf

[17] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 492–500.

[18] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[19] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[20] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*. 803–811.

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

[22] Consumer Financial Protection Bureau. 2014. Using publicly available information to proxy for unidentified race and ethnicity. *Report available at https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf* (2014).

[23] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[24] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE, 39–57.

[25] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*. PMLR, 1349–1361.

[26] L Elisa Celis and Vijay Keswani. 2020. Implicit Diversity in Image Summarization. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–28.

[27] L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. 2021. Fair Classification with Adversarial Perturbations. *arXiv preprint arXiv:2106.05964* (2021).

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

[29] Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. 2020. On Adversarial Bias and the Robustness of Fair Machine Learning. *arXiv preprint arXiv:2006.08669* (2020).

[30] European Commission. [n.d.]. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[31] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. 2018. Adversarial attack on graph structured data. In *International conference on machine learning*. PMLR, 1115–1124.

[32] Tamara Denning, Batya Friedman, and Tadayoshi Kohno. 2013. The Security Cards: A Security Threat Brainstorming Toolkit. University of Washington. https://securitycards.cs.washington.edu/.

[33] UK Office for Artificial Intelligence. [n.d.]. Ethics, Transparency and Accountability Framework for Automated Decision-Making. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making.

[34] Batya Friedman and David Hendry. 2019. *Value sensitive design: shaping technology with moral imagination*. MIT Press.

[35] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[36] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[37] Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. 2021. Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. *arXiv preprint* abs/2103.11920 (2021). arXiv:2103.11920 http://arxiv.org/abs/2103.11920

[38] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. 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*. 2221–2231.

[39] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. *When Fair Ranking Meets Uncertain Inference.* Association for Computing Machinery, New York, NY, USA, 1033–1043. https://doi.org/10.1145/3404835.3462850

[40] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing Intersectional Group Fairness with Worst-Case Comparisons. In *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI) (Proceedings of Machine Learning Research, Vol. 142)*, Deepti Lamba and William H. Hsu (Eds.). PMLR, 22–34. https://proceedings.mlr.press/v142/ghosh21a.html

[41] William Gies, James Overby, Nick Saraceno, Jordan Frome, Emily York, and Ahmad Salman. 2020. Restricting Data Sharing and Collection of Facial Recognition Data by the Consent of the User: A Systems Analysis. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*. 1–6. https://doi.org/10.1109/SIEDS49339.2020.9106661

[42] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[43] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[44] Danny Goodwin. 2011. Top Google Result Gets 36.4% of Clicks [Study]. Search Engine Watch. https://www.searchenginewatch.com/2011/04/21/top-google-result-gets-36-4-of-clicks-study/.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[46] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 355–364.

[47] Lingxiao Huang and Nisheeth K Vishnoi. 2019. Stable and fair classification. *arXiv preprint arXiv:1902.07823* (2019).

[48] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. 2020. Subpopulation data poisoning attacks. *arXiv preprint arXiv:2006.14026* (2020).

[49] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[50] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

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

[52] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1548–1558.

[53] Alistair Knott. [n.d.]. Moving Towards Responsible Government Use of AI in New Zealand). https://digitaltechitp.nz/2021/03/22/moving-towards-responsible-government-use-of-ai-in-new-zealand/.

[54] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).

[55] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. (2020).

[56] Kristina Lerman, Anon Plangprasopchok, and Chio Wong. 2007. Personalizing image search results on flickr. *Intelligent Information Personalization* (2007).

[57] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. 2019. Universal perturbation attack against image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4899–4908.

[58] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

[59] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[60] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770* (2016).

[61] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 2019. Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-Based Image Retrieval. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (Ottawa ON, Canada) *(ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 306–314. https://doi.org/10.1145/3323873.3325052

[62] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[63] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).

[64] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. 2020. Exacerbating Algorithmic Bias through Fairness Attacks. *arXiv preprint arXiv:2012.08723* (2020).

[65] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. 107–118.

[66] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. arXiv:1610.08401 [cs.CV]

[67] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.

[68] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. *arXiv preprint arXiv:2005.14713* (2020).

[69] Ankan Mullick, Sayan Ghosh, Ritam Dutt, Avijit Ghosh, and Abhijnan Chakraborty. 2019. Public Sphere 2.0: Targeted Commenting in Online News Media. In *European Conference on Information Retrieval*. Springer, 180–187.

[70] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

[71] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. 2021. Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 466–477.

[72] Jakob Nielsen. 2003. Usability 101: introduction to usability. Jakob Nielsen's Alertbox.

[73] Government of Canada. [n.d.]. Responsible use of artificial intelligence (AI). https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html.

[74] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, et al. 2016. Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768* (2016).

[75] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4422–4431.

[76] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proc. of FAT**.

[77] Nisarg Raval and Manisha Verma. 2020. One word at a time: adversarial attacks on retrieval models. *arXiv preprint arXiv:2008.02197* (2020).

[78] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[79] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. 2019. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *arXiv preprint arXiv:1904.05233* (2019).

[80] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference*. 553–562.

[81] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. 4, CSCW1, Article 058 (may 2020), 35 pages. https://doi.org/10.1145/3392866

[82] Sefik Ilkin Serengil and Alper Ozpinar. 2020. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 1–5.

[83] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. *arXiv preprint arXiv:1804.00792* (2018).

[84] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *arXiv preprint arXiv:1904.12843* (2019).

[85] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security 20)*. 1589–1604.

[86] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

[87] David Solans, Battista Biggio, and Carlos Castillo. 2020. Poisoning Attacks on Algorithmic Fairness. *arXiv preprint arXiv:2004.07401* (2020).

[88] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[89] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[90] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[91] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347* (2020).

[92] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. The Space of Transferable Adversarial Examples. *arXiv* (2017). https://arxiv.org/abs/1704.03453

[93] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2018. Clean-label backdoor attacks. (2018).

[94] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2773–2780.

[95] Yevgeniy Vorobeychik and Murat Kantarcioglu. 2018. Adversarial machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 12, 3 (2018), 1–169.

[96] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[97] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. 2020. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*. Springer, 1–17.

[98] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[99] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.

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

[101] Mo Zhou, Zhenxing Niu, Le Wang, Qilin Zhang, and Gang Hua. 2020. Adversarial Ranking Attack and Defense. *arXiv preprint arXiv:2002.11293* (2020).

A SUPPLEMENTARY MATERIAL

A.1 Comparison between DetConstSort and FMMR

In this section we compare the performance of two fair re-rankers in the presence of our GAP attack. We have already described the details of the first algorithm, FMMR, in § 3.2.2.

A.1.1 DetConstSort. The second algorithm, DetConstSort[38], was developed by and is currently deployed at LinkedIn in their talent search system. Unlike FMMR, DetConstSort requires access to the demographic labels of the items it is trying to fairly re-rank. DetConstSort rearranges a given list of items, such that for any particular rank k and for any attribute a_j, the attribute is present at least $\lfloor p_{a_j}.k \rfloor$ times in the ranked list, where p_{a_j} is the proportion of items in the list that have the attribute a_j. DetConstSort also re-sorts the items within the relevance criteria so that items with better utility scores are placed higher in the ranked list as much as possible, while maintaining the desired attribute ratio. It thus aims to solve a deterministic interval constrained sorting problem.

If ground-truth demographic labels are unavailable, DetConstSort may instead utilize labels sourced from a demographic inference model. Recent work, however, has shown that DetConstSort is sensitive to errors in demographic labels, with one example of such errors being inaccurate inferences [39].

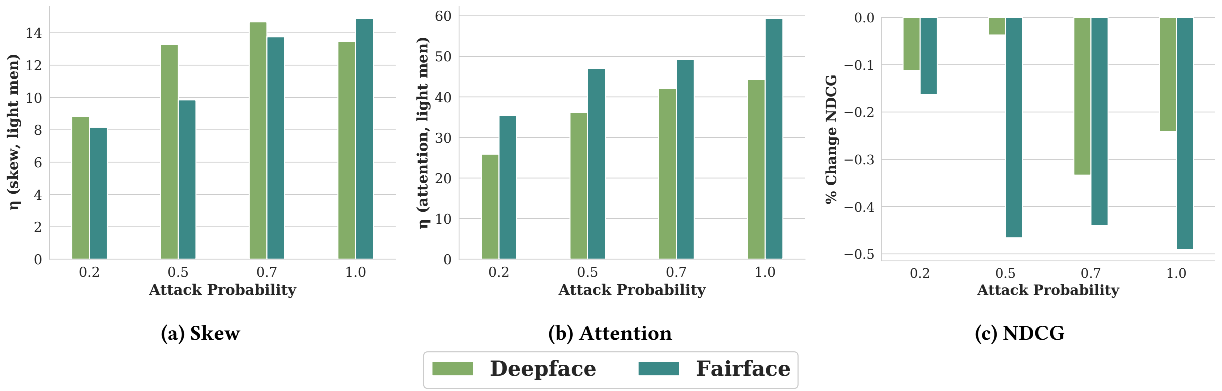
A.1.2 Evaluation Results. We present the results of our GAP attacks against our search engine when it uses DetConstSort and FMMR as the fair re-ranker, respectively, in Figure 9. As in § 5, these results are averaged across three queries, multiple values of k, etc.

For DetConstSort, the skew and attention metrics are not impacted by our attack. This can be clearly seen by comparing the η values when $pr = 0$ (i.e., there are no perturbed images) to other values of pr: for DetConstSort, η for skew and attention starts high (unfair) when $pr = 0$, and does not change as pr increases. The correct interpretation of these results is **not** that DetConstSort is resilient to our attack. Rather, the correct interpretation is that DetConstSort starts off unfair due to the use of inaccurate, inferred demographic data [39], and our attack is unable to make the unfairness worse.

Thus, we find that a prerequisite for evaluating the success of our attacks on DetConstSort is an accurate demographic inference model. Developing such models is still an active area of research, and is out-of-scope for our work. Should a more accurate demographic inference model be designed in the future, however, it must be designed with adversarial robustness in mind to prevent our attacks.

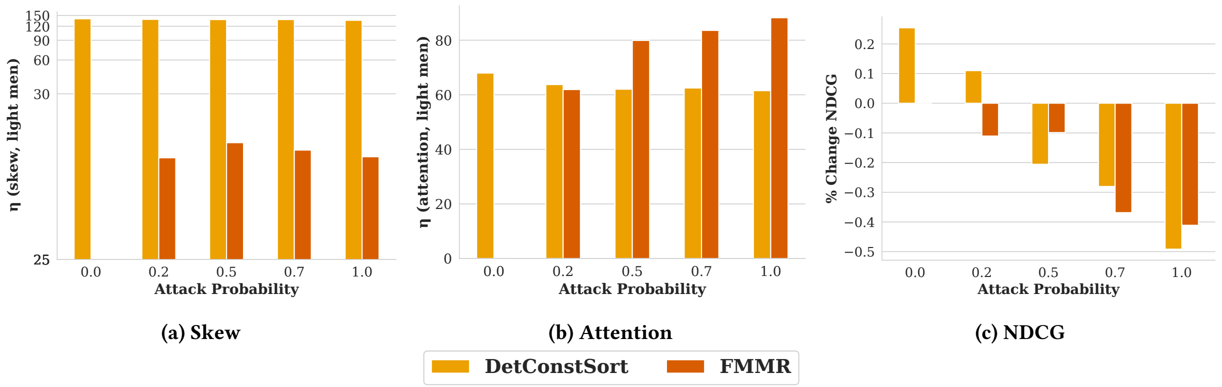
A.2 Choice of Queries

To facilitate our experiments, we chose to select search query terms that would provide a sizeable list of images. To do so, we looked at the list of terms in the COCO image captions (excluding English stop words and words related to ethnicity or gender). The following table shows some top terms. From this information, we composed our three queries given that "sitting", "tennis", "table", "person", "pizza", etc. were among the most popular terms.



(a) Skew (b) Attention (c) NDCG

Figure 8: GAP models trained on different demographic inference algorithms offer similar attack effectiveness.



(a) Skew (b) Attention (c) NDCG

Figure 9: DetConstSort has poor performance even without an attack, making our results uninteresting.

Term	Count
sitting	55084
standing	44121
people	42133
holding	29055
large	25305
person	25123
street	21609
table	20775
small	20661
tennis	19718
riding	18809
train	18287
young	17767
red	17522
baseball	15362
pizza	11163

Table 2: The most common (gender or race unrelated) caption terms in the evaluation dataset.