Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Other titles in Foundations and Trends® in Computer Graphics and Vision

Computational Imaging Through Atmospheric Turbulence
Stanley H. Chan and Nicholas Chimitt
ISBN: 978-1-63828-999-9

Towards Better User Studies in Computer Graphics and Vision
Zoya Bylinskii, Laura Herman, Aaron Hertzmann, Stefanie Hutka and Yile Zhang
ISBN: 978-1-63828-172-6

An Introduction to Neural Data Compression
Yibo Yang, Stephan Mandt and Lucas Theis
ISBN: 978-1-63828-174-0

Learning-based Visual Compression
Ruolei Ji and Lina J. Karam
ISBN: 978-1-63828-112-2

Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends
Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu and Jianfeng Gao
ISBN: 978-1-63828-132-0

Semantic Image Segmentation: Two Decades of Research
Gabriela Csurka, Riccardo Volpi and Boris Chidlovskii
ISBN: 978-1-63828-076-7

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research

Timnit Gebru
The Distributed AI Research Institute (DAIR)
timnit@dair-institute.org

Remi Denton
Google Research
dentone@google.com

the essence of knowledge
Boston — Delft

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

T. Gebru and R. Denton. *Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research.* Foundations and Trends® in Computer Graphics and Vision, vol. 16, no. 3, pp. 215–321, 2024.

ISBN: 978-1-63828-355-3
© 2024 T. Gebru and R. Denton

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording or otherwise, without prior written permission of the publishers.

Photocopying. In the USA: This journal is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by now Publishers Inc for users registered with the Copyright Clearance Center (CCC). The 'services' for users can be found on the internet at: www.copyright.com

For those organizations that have been granted a photocopy license, a separate system of payment has been arranged. Authorization does not extend to other kinds of copying, such as that for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale. In the rest of the world: Permission to photocopy must be obtained from the copyright owner. Please apply to now Publishers Inc., PO Box 1024, Hanover, MA 02339, USA; Tel. +1 781 871 0245; www.nowpublishers.com; sales@nowpublishers.com

now Publishers Inc. has an exclusive license to publish this material worldwide. Permission to use this content must be obtained from the copyright license holder. Please apply to now Publishers, PO Box 179, 2600 AD Delft, The Netherlands, www.nowpublishers.com; e-mail: sales@nowpublishers.com

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Foundations and Trends® in Computer Graphics and Vision

Volume 16, Issue 3, 2024

Editorial Board

Editor-in-Chief

Aaron Hertzmann
Adobe Research

Editors

Marc Alexa
TU Berlin

Kavita Bala
Cornell

Ronen Basri
Weizmann Institute of Science

Peter Belhumeur
Columbia

Andrew Blake
Microsoft Research

Chris Bregler
Facebook-Oculus

Joachim Buhmann
ETH Zurich

Michael Cohen
Facebook

Brian Curless
University of Washington

Paul Debevec
USC Institute for Creative Technologies

Julie Dorsey
Yale

Fredo Durand
MIT

Olivier Faugeras
INRIA

Rob Fergus
NYU

William T. Freeman
MIT

Mike Gleicher
University of Wisconsin

Richard Hartley
Australian National University

Hugues Hoppe
Microsoft Research

C. Karen Liu
Stanford

David Lowe
University of British Columbia

Jitendra Malik
Berkeley

Steve Marschner
Cornell

Shree Nayar
Columbia

Tomas Pajdla
Czech Technical University

Pietro Perona
California Institute of Technology

Marc Pollefeys
ETH Zurich

Jean Ponce
Ecole Normale Superieure

Long Quan
HKUST

Cordelia Schmid
INRIA

Steve Seitz
University of Washington

Amnon Shashua
Hebrew University

Peter Shirley
University of Utah

Noah Snavely
Cornell

Stefano Soatto
UCLA

Richard Szeliski
Microsoft Research

Luc Van Gool
KU Leuven and ETH Zurich

Joachim Weickert
Saarland University

Song Chun Zhu
UCLA

Andrew Zisserman
Oxford

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Editorial Scope

Foundations and Trends® in Computer Graphics and Vision publishes survey and tutorial articles in the following topics:

- Rendering
- Shape
- Mesh simplification
- Animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and video retrieval
- Video analysis and event recognition
- Medical image analysis
- Robot localization and navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2024, Volume 16, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Contents

1 Introduction **3**

2 Positionality **7**

3 Overview of Risks and Harms Associated with Computer Vision Systems and Proposed Mitigation Strategies **9**

3.1 Representational Harms 9
3.2 Quality-of-Service and Allocative Harms 15
3.3 Interpersonal Harms 23
3.4 Societal Harms: System Destabilization and Exacerbating Inequalities 31

4 Frameworks and Principles for Computer Vision Researchers **52**

4.1 Guidelines for Responsible Data and Model Development . 53
4.2 Measurement Modeling 57
4.3 Reflexivity 59

5 Reorientations of Computer Vision Research **62**

5.1 Grounded in Historical Context and Considering Power Dynamics 62
5.2 Small, Task Specific 65
5.3 Community-Rooted 66

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

6 Systemic Change **71**
6.1 Collective Action and Whistleblowing 71
6.2 Refusal/The Right not to Build Something 73
6.3 Independent Funding Outside of Military and Multinational Corporations 75

7 Conclusion **77**

References **79**

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research

Timnit Gebru[1] and Remi Denton[2]

[1] *The Distributed AI Research Institute (DAIR), USA; timnit@dair-institute.org*
[2] *Google Research, USA; dentone@google.com*

ABSTRACT

The field of computer vision is now a multi-billion dollar enterprise, with its use in surveillance applications driving this large market share. In the last six years, computer vision researchers have started to discuss the risks and harms of some of these systems, mostly using the lens of fairness introduced in the machine learning literature to perform this analysis. While this lens is useful to uncover and mitigate a narrow segment of the harms that can be enacted through computer vision systems, it is only one of the toolkits that researchers have available to uncover and mitigate the harms of the systems they build.

In this monograph, we discuss a wide range of risks and harms that can be enacted through the development and deployment of computer vision systems. We also discuss some

Timnit Gebru and Remi Denton (2024), "Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research", Foundations and Trends® in Computer Graphics and Vision: Vol. 16, No. 3, pp 215–321. DOI: 10.1561/0600000102.
©2024 T. Gebru and R. Denton

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

existing technical approaches to mitigating these harms, as well as the shortcomings of these mitigation strategies. Then, we introduce computer vision researchers to harm mitigation strategies proposed by journalists, human rights activists, individuals harmed by computer vision systems, and researchers in disciplines ranging from sociology to physics. We conclude the monograph by listing principles that researchers can follow to build what we call community-rooted computer vision tools in the public interest, and give examples of such research directions. We hope that this monograph can serve as a starting point for researchers exploring the harms of current computer vision systems and attempting to steer the field into community-rooted work.

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

1

Introduction

Computer vision is no longer a purely academic endeavor, instead it is a field with significant industry market share. According to Fortune Business Insights, the global image recognition market size is projected to reach USD $178.13 billion by 2032,[1] with the "active adoption of facial recognition technologies to screen people amid the current COVID-19 pandemic ensuring "stable growth" of the market."[2] This market share estimate does not include the data annotation industry which is projected to reach $13.2 billion in market value by 2030.[3] Beyond industry applications, a number of computer vision researchers have also created tools in the public interest. For example, Sefala *et al.* (2021) analyzed satellite imagery to study the impacts of discriminatory

[1] https://www.fortunebusinessinsights.com/industry-reports/image-recognition-market-101855.

[2] https://www.globenewswire.com/news-release/2021/12/21/2355754/0/en/Image-Recognition-Market-to-Hit-86-32-Billion-by-2027-Rapid-Advancements-in-Face-Detection-Technologies-to-Drive-Market-Growth-Fortune-Business-Insights.html.

[3] https://www.businesswire.com/news/home/20220824005423/en/The-Worldwide-Data-Annotation-Tools-Industry-is-Expected-to-Reach-13.2-Billion-by-2030—ResearchAndMarkets.com.

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

policies, and Mutembesa *et al.* (2019) created a mobile app to help farmers monitor plant health on their farms.

In spite of some efforts to create computer vision tools in the public interest, the field's popularity and increase in industry and government use has led to negative consequences to marginalized groups, some of which have been discussed within the computer vision community, but many of which have not. This monograph gives an overview of the reported harms of computer vision systems and works that resist, refuse, mitigate or slow down these harms and provide an alternate path. We take a broad approach to surveying works, covering not only scholarship within computer vision and adjacent fields, but the important works of scholars in fields such as sociology, science and technology studies, communications, history, and surveillance studies who have done foundational work to contextualize the societal impacts of computer vision systems. Journalists, civil rights organizations and individuals who have been negatively impacted by computer vision systems have also played a major role in uncovering some of its harms, and workers have organized to curb a number of harmful use cases.

The rest of the monograph is structured as follows. Section 2 is a short positionality statement, following one of our recommendations to computer vision researchers (see Section 4.3). Section 3 gives an overview of the negative impacts that computer vision technologies can have on individuals, communities, and society. Our overview draws terminology and typologies from recent scholarship that has offered taxonomies of harm relating to machine learning technologies (e.g., Katzman *et al.*, 2023; Wang *et al.*, 2022a; Shelby *et al.*, 2023), but focuses specifically on harms associated with computer vision. We specifically use the taxonomy proposed by Shelby *et al.* (2023), which categorizes the harms posed by automated systems into 5 categories: representational harms, allocative harms, quality of service harms, interpersonal harms and social system harms. These taxonomies are not mutually exclusive however, and many examples can fall into multiple categories. Alongside our summary of reported harms, we discuss efforts to uncover and mitigate these harms, as well as the shortcomings of harm reduction strategies that solely depend on technical interventions.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

We then offer forward-looking guidance and recommendations for computer vision researchers and practitioners that we hope will guide the field towards more equitable and just outcomes. Section 4 summarizes actions individual computer vision researchers and practitioners can take, such as following recommended responsible AI frameworks, documentation standards, and other principles. Section 5 discusses the properties of computer vision systems that result in tools that help marginalized communities rather than surveil them, and research practices that enable the development of these types of tools. Section 6 takes a bigger-picture look at the institutions and power structures that scaffold computer vision research and development practices, and discusses avenues that researchers and practitioners can take to institute systemic change to encourage what we call community-rooted computer vision research.

This monograph is predominantly written for computer vision students, researchers, and practitioners. Within the computer vision community, and the machine learning community more generally, harms arising from computer vision and/or machine learning systems are often discussed using the narrow lenses of model fairness and bias. While this monograph does not specifically focus on these concepts, we give some examples of interventions that rely on them, and also elaborate on why these types of technical interventions can be limited. We encourage our readers to familiarize themselves with techniques used by the ML fairness community by reading the comprehensive book *Fairness and Machine Learning* (Barocas *et al.*, 2023), reviewing the many measures of fairness that have been introduced over the years (Narayanan, 2018), and reading surveys of debiasing methods used in neural-network based vision and language research (e.g., Parraga *et al.*, 2023).

In this monograph, we intentionally survey work that may be less familiar to computing researchers, including scholarship from a broad set of non-computing disciplines, and reports from journalists, human rights experts, and individuals directly impacted by computer vision systems. In doing so, we aim to introduce computer vision researchers to the broader discourse on the ethical implications of computer vision research, and deepen their awareness of the harms that can stem from the research, development, and deployment of a number of computer

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

vision technologies. We discuss interventions at the data and model level that can mitigate some of these harms (e.g., removing hate speech from an image captioning dataset to mitigate the risk of a model captioning an image with hateful language). However, many of the harms we discuss are ones that do not have purely "technical" interventions, because we aim to equip computer vision researchers and practitioners with a broader set of tools to understand and mitigate harms arising from the systems they build–including having the option of not building certain systems.

We recognize that our readers may feel that they have limited agency to prevent the negative impacts detailed in this monograph, particularly the societal-level harms discussed in Section 3.4 that may be understood as being "further" from model development and deployment practices. But we hope that the many examples of interventions provided in this monograph from critical interdisciplinary perspectives can serve as inspiration to students, researchers and practitioners in computer vision to move the field away from the surveillance of marginalized communities, and into serving their needs instead. We also recognize that due to the interdisciplinary nature of this monograph, many readers may be unfamiliar with some terminology and concepts that we introduce. As a result, we have striven to define terms external to the field of computer vision, and provide references that can be used as a foundation for our readers' future learning.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

2

Positionality

Although uncommon in computing disciplines, many academic authors write statements on positionality, acknowledging how their point of view impacts their work. We discuss this recommendation further in Section 4.3, and attempt to follow it in this section. We received our PhDs in computer vision from universities in the US and research groups that have been prominent within the field. One of us works at a BigTech company, and one of us was fired by a BigTech company and founded an institute seeking independence from BigTech influence. We are a duo of a cis Black woman who immigrated to the US as a refugee from Ethiopia, and a white transgender non-binary immigrant to the US from Canada. Both of us re-evaluated our work in computer vision after analyzing the harms that could be enabled through our own works. The lenses with which we analyze the harms perpetuated by computer vision systems, including the shortcomings of our analyses, are influenced by our life experiences, the communities we come from, our work experiences, and our academic training. Given that we both live in the US, some of the shortcomings of our analyses include having a higher focus on issues related to the US than other regions of the world. We remind computer vision researchers that computer vision,

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

and science, technology, engineering and math (STEM) in general, are political (Green, 2021). As such, while reading this monograph, we encourage our readers to think through which statements they find to be political and which statements they find to be "neutral," and how that relates to their positionality.

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

3

Overview of Risks and Harms Associated with Computer Vision Systems and Proposed Mitigation Strategies

This section discusses the many different types of reported harms that have been perpetuated with the help of computer vision technologies, and the strategies that computer vision researchers and practitioners, journalists, activists, and other scholars have pursued to uncover and mitigate these harms. To streamline our discussion, we follow the taxonomy of Shelby *et al.* (2023) to categorize these harms that range from stereotyping, to economic loss and increased warfare.

3.1 Representational Harms

Representational harms can occur when computer vision systems reinforce social hierarchies, often by influencing perceptions of and attitudes towards particular social groups (Katzman *et al.*, 2023; Shelby *et al.*, 2023). For example, image tagging and captioning systems can reinforce stereotypes about social groups by producing stereotype aligned predictions. When a model systematically tags images of female doctors with a label of "nurse" (Katzman *et al.*, 2023), or incorrectly assumes the gender of a person performing a gender-stereotyped activity (e.g., shopping, cleaning) (Zhao *et al.*, 2017; Hendricks *et al.*, 2018), it reinforces societal stereotypes about what activities are undertaken by

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

which genders. Image captioning systems are not the only types of computer vision technologies that have been documented to reinforce stereotypes, Safiya Noble's *Algorithms of Oppression* studied the issue of stereotyping in image recommendation systems, and was one of the first works to report sexualized results for the search query, "Black women" (Noble, 2018).

Given recent advances in generative models, a number of researchers have also studied stereotypical depictions and associations within text-to-image (T2I) models. Bianchi *et al.* (2023) and Fraser *et al.* (2023) examined the sociodemographic characteristics of people depicted in images generated from prompts referencing various terms that have stereotypical associations (e.g., terrorist, software engineer, untrustworthy person, etc.) and found evidence of harmful racial, ethnic, gendered, class, and intersectional stereotypes. Similarly, a 2023 investigation by Leonardo Nicoletti and Dina Bass of Bloomberg News found that Stable Diffusion amplifies stereotypes about race and gender (Nicoletti and Bass, 2023). Luccioni *et al.* (2024b) further expanded on the aforementioned analyses and developed multiple interactive visualization tools to help others perform similar analyses.

Motivated by calls to expand the evaluation of machine learning systems to non-Western and Global Majority settings (Sambasivan *et al.*, 2021a), Qadri *et al.* (2023) studied how text-to-image models represent South Asian cultures, and found that generated images contain harmful stereotypes and tropes. Bianchi *et al.* (2023) and Ghosh and Caliskan (2023) also explored T2I stereotypes associated with a broader range of national identity groups and found multiple harmful associations. Jha *et al.* (2024) extended these studies to explore geo-cultural stereotypes perpetuated by T2I models at a global scale, and introduced an evaluation dataset consisting of 135 regional identity groups and stereotypes associated with them.

Stereotype-aligned outputs of computer vision models can harm entire groups of people, beyond a single user interaction. For instance, several scholars have discussed how generative AI systems that erase or reductively represent marginalized cultures can perpetuate cultural hegemony by reinforcing a dominant (and often outsider) perspective on a culture (Jiang *et al.*, 2023; Qadri *et al.*, 2023). Senegalese artist

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Linda Dounia Rebeiz gives examples of outputs from OpenAI's text-to-image model, Dall-E, when prompting it to generate images of her hometown Dakar. The images depicted "arid desert landscapes and ruined buildings that Rebeiz said were nothing like the coastal homes in the Senegalese capital."[1]

There are a number of other documented representational harms that demean marginalized groups, including a pattern of oversexualization associated with certain identity groups. For example, Ungless *et al.* (2023) studied the representations of cisgender and non-cisgender identities in text-to-image models, finding that prompts for non-cisgender identities produced more stereotyped and sexualized imagery. Similarly, Wolfe *et al.* (2023) explored objectification biases in multimodal pre-trained models, and estimated that, for example, images generated with the prompt of "a [age] year old girl" produced sexualized images up to 73% of the time, while the same prompt with "boy" produced a maximum of 9% of sexualized images. Journalists have also commented on the sexualization biases of text-to-image models, noting the Lensa AI app's tendency to sexualize images of women (Demopoulos, 2022). Notably, Ghosh and Caliskan (2023) find that the T2I model Stable Diffusion, oversexualizes Latin American, Mexican, Indian and Egyptian women relative to women from other nationalities. We discuss additional risks of objectification enabled by computer vision technologies in Section 3.3 which covers interpersonal harms.

Computer vision systems can also produce outputs that dehumanize social groups, resulting in yet another form of representational harm. In image tagging or captioning systems, for instance, a model can associate hateful, dehumanizing, or offensive words with a particular identity group (Katzman *et al.*, 2023). This type of dehumanization can also be perpetuated through other types of systems, such as text-to-image-models. For instance, Mack *et al.* (2024) found that T2I models can produce offensive and dehumanizing images for prompts referencing disabled people, such as images with horror-like aesthetics.

The last form of representational harm that we discuss is erasure and omission, which has also been documented in different types of computer

[1] https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

vision technologies. Within image tagging systems, erasure may occur when a system fails to recognize, and appropriately label, culturally specific artifacts, activities, or attributes (Katzman *et al.*, 2023). Within generative AI systems, Qadri *et al.* (2023) explore erasure harms that stem from text-to-image systems failing to produce reasonable image responses to culturally specific prompts. In their qualitative study with South Asian participants, Qadri *et al.* (2023) found that state-of-the-art models frequently failed to depict important cultural subject matter that was referenced in prompts, ranging from art styles to important landmarks. Erasure harms can also stem from foundational design decisions embedded in a computer vision system, such as a categorical schema that inherently excludes or erases a social group. For example, gender classification systems defaulting to binary genders (Hamidi *et al.*, 2018) have been described as a form of erasure.

Researchers have developed several approaches to mitigate the aforementioned representational harms from computer vision technologies. Dataset filtering and rebalancing techniques offer one potential in-road for mitigating stereotypical associations by minimizing correlations between demographic characteristics and other labels in the training dataset. For example, the creators of ImageNet removed all categories that were deemed "non-imageable" in an attempt to minimize harmful stereotypes and associations (Yang *et al.*, 2020). Wang *et al.* (2020) developed the REVISE tool to support investigations into object-based, gender-based, and geography-based biases in visual datasets, since a number of representational harms stem from the inclusion of hateful or offensive labels within the training datasets. Indeed, audits of image datasets have uncovered millions of images labeled with offensive categories, including racial slurs and derogatory phrases (Crawford and Paglen, 2021; Birhane and Prabhu, 2020; Birhane *et al.*, 2021).

In some cases, the risk of representational harm can be partly mitigated by removing hateful and offensive labels from training data, or in extreme cases, deprecating the dataset entirely. For example, the creators of ImageNet removed all categories (and corresponding images) that were deemed offensive (Yang *et al.*, 2020). The widely used Tiny

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Images dataset was deprecated[2] after a dataset audit uncovered offensive and harmful labels in the dataset (Birhane and Prabhu, 2020). However, Peng *et al.* (2021) note that even after datasets have been taken down, there are many copies of previously downloaded versions that remain in circulation and continue to be used in research papers. For these reasons, the authors write that "dataset retraction has a limited effect on mitigating harms," and provide a set of recommendations distributed across dataset creators, and distributors to steward a dataset's use throughout its lifecycle (Peng *et al.*, 2021).

Generative AI systems like text-to-image-models pose new challenges for evaluating and mitigating demeaning or dehumanizing outputs, since the output vocabulary is generally unbounded (in contrast to image tagging systems which may have a fixed set of labels). Nevertheless, researchers have developed bias correction approaches for generative models (e.g., Grover *et al.*, 2019), and evaluation metrics that measure the diversity of generated samples (Friedman and Dieng, 2023). Wan *et al.* (2024) also survey approaches to defining, evaluating, and mitigating harmful biases in text-to-image models with respect to gender, skintone, and geo-culture. Some of the harm mitigation strategies taken by researchers include data filtering and model output filtering methods to address risks of hateful outputs (e.g., Hao *et al.*, 2023), and frameworks to support analyses of representations of people within multi-modal datasets used in generative AI systems (Diaz *et al.*, 2023).

While data interventions can be valuable, simple rebalancing techniques can sometimes fall short. For example, even when training datasets are constructed so that image labels co-occur at equal rates with each gender, image tagging models can still produce biased and stereotypical predictions (Wang *et al.*, 2019b). Hendricks *et al.* (2018) propose mitigating model amplification of gender biases present in training data, by adding losses that encourage the model to output gender neutral captions for a particular image, unless the model can confidently infer the gender of people present in the image. More broadly, one of the most popular approaches for mitigating model biases attempts to remove information that correlates with protected attributes such as

[2] https://groups.csail.mit.edu/vision/TinyImages/.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

gender or race, from the data or learned feature representation of the data (e.g., Alvi *et al.*, 2018; Zhang *et al.*, 2018; Wang *et al.*, 2019b; Zemel *et al.*, 2013; Creager *et al.*, 2019; Dwork *et al.*, 2018). See Wang *et al.* (2020) for benchmarking of a wide range of bias mitigation techniques for models performing visual recognition tasks.

Some representational harms cannot be mitigated through data or model interventions, but rather arise from the basic function of a technology, such as what sociologists Stark and Hutson refer to as "physiognomic AI" (Stark and Hutson, 2021). Before the advent of computer vision, one of the primary ways in which scientists profiled and stereotyped communities was through the pseudoscience of physiognomy, defined as "the study of the features of the face, or of the form of the body generally, as being supposedly indicative of character; the art of judging character from such study."[3] Despite being debunked many times, researchers have cautioned that modern computer vision is bringing about a resurgence of physiognomy (Stark and Hutson, 2021; Agüera y Arcas *et al.*, 2017), with a slew of companies currently offering profiling technologies reminiscent of physiognomy.[4,5,6] For instance, Faception claims to infer characteristics ranging from people's IQ, to their likelihood of "being a terrorist" to committing "white collar crime" or being a "pedophile" by simply looking at their faces.[7] Journalists have also reported that a number of governments are hoping to use such technology to infer the emotions of "allies and enemies."[8] and screen refugees.[9]

[3] https://www.oed.com/dictionary/physiognomy_n.

[4] https://www.wired.com/story/eye-scanning-lie-detector-polygraph-forging-a-dystopian-future.

[5] https://www.technologyreview.com/2020/03/13/905323/ai-lie-detectors-polygraph-silent-talker-iborderctrl-converus-neuroid.

[6] https://www.discernscience.com/avatar/.

[7] https://www.faception.com/.

[8] https://www.forbes.com/sites/thomasbrewster/2020/07/15/the-pentagons-1-million-question-can-ai-predict-an-enemys-emotions.

[9] https://www.newscientist.com/article/mg24032023-400-an-ai-lie-detector-will-interrogate-travellers-at-some-eu-borders/.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Similarly, a number of researchers have also cautioned against computer vision technologies that try to classify people into social categories such as race and gender. Attempting to infer or classify identity characteristics can cause harm by denying individuals an opportunity to self-identify (Shelby *et al.*, 2023). This harm is particularly salient in the context of gender classification systems that systematically misgender trans and non-binary individuals (Hamidi *et al.*, 2018). Moreover, such systems reinforce a flawed belief that gender, race, and other related socially constructed[10] categories are fixed attributes of an individual that are knowable from an image, which can in turn re-entrench a system of social stratification (West *et al.*, 2019; Stark and Hutson, 2021). While there is now some recognition of the risks of computer vision systems built with the explicit intent of classifying an individual's race or gender, there are some use cases for which there are ongoing debates about the risk/benefit tradeoff of automated labels or descriptions of race, gender, or other identity characteristics. For example, image descriptions are essential for screen reader users to access online images. Yet, there don't exist clear standards for how to write about people's appearances. This challenge is heightened as image captioning models are increasingly leveraged to write image descriptions. Based on interviews of screen reader users who are racial and gender minorities in the US, Bennett *et al.* (2021) offer best practices for providing descriptions (in English) pertaining to people's visual appearances in images.

3.2 Quality-of-Service and Allocative Harms

In this section we discuss two broad categories of harm, quality-of-service harm and allocative harm, which can occur when a computer vision system exhibits disparities in performance across different groups of users, typically defined along sociocultural lines such as race, gender, disability, age, ethnicity, etc. (Shelby *et al.*, 2023). These disparities in the performance of computer vision systems can lead to individuals being excluded from using a particular product or service, and/or result

[10] To say gender or race is socially constructed is to recognize that these categories do not exist inherently in the world. Rather, these categories have been created, over time, through social, cultural, and political forces (Hacking, 2000).

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

in an increased burden of labor on the user in order to effectively use a particular product or service (Shelby *et al.*, 2023). Moreover, when degraded quality-of-service is linked to one's identity, a user can experience feelings of alienation or inequity (Shelby *et al.*, 2023). Such harms are amplified for people who have repeated experiences of exclusion while interacting with computer vision and related technologies.

Quality-of-service harms can be serious and even deadly when they result in allocative harms, which occur when an individual is denied access to material resources or opportunities. The concept of allocative harms was introduced to the ML community in a talk by Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach (Barocas *et al.*, 2017), and is now part of the typology of harms discussed in computing literature (Suresh and Guttag, 2021; Shelby *et al.*, 2023).[11] In the context of computer vision systems, allocative harm is often connected to reduced quality of service. For example, there has been extensive media coverage of rideshare apps with automated face verification security protocols failing to recognize Black and transgender drivers in the US, leading to drivers being locked out and losing work.[12,13] Remote workers required to use company facial recognition systems have also reported being booted out of their work software "if they shifted in their seat, rested their eyes, adjusted their glasses, wore a headband or necklace, went to the bathroom or had a child walk through their room."[14] The use of a facial recognition based system to pay taxes and obtain government benefits such as unemployment in the US has also resulted in people misidentified by the systems unable to get these services.[15]

[11] See Virginia Eubanks' Automating Inequality for an in-depth look at how automated decision making systems can cause allocative harm. https://virginia-eubanks.com/automating-inequality/.

[12] https://www.fastcompany.com/90216258/uber-face-recognition-tool-has-locked-out-some-transgender-drivers

[13] https://time.com/6104844/uber-facial-recognition-racist/.

[14] https://www.washingtonpost.com/technology/2021/09/24/remote-work-from-home-surveillance/.

[15] https://www.theatlantic.com/ideas/archive/2022/01/irs-should-stop-using-facial-recognition/621386/.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

Given that an increasing number of organizations around the world are relying on automated recruiting tools to sift through job applicants, scholars warn that these types of tools can also be a single point of failure which perpetuate allocative harms at a mass scale (Ajunwa, 2021). Mathematician Kathy O'Neil notes that in the US, poor people are more likely to interface with automated tools while applying for jobs, which increases their chances of denied employment opportunities without ever interfacing with a human (O'Neil, 2017). The lack of a human that one can interface with in the interviewing process can also result in the discrimination of disabled people. For instance, an article in Wired magazine gives an example of a visually impaired interviewee who could not even read the interview questions.[16]

During recorded interviews of job applicants, allocative harms can also be perpetuated through the use of automated facial analysis tools which purport to give employers verbal and nonverbal cues about prospective employees (Ajunwa, 2021). That is, a job applicant could be denied employment at every organization using a specific recruiting tool, if that tool classifies a particular applicant as having an undesirable characteristic based on whatever the automated facial analysis tool infers about the person (Ajunwa, 2021).

More generally, the use of biometric IDs (e.g., those based on face recognition, iris scans, speech identification or fingerprint detection) has resulted in a number of reported allocative harms. For instance, since the launch of India's Aadhaar system in 2009, the world's largest biometric ID system, scholars and journalists have raised awareness about denied services to those who are elderly, disabled, or otherwise marginalized, who can be misidentified by the biometric systems. According to activists, denial of services, including food, based on the Aadhaar system was linked to half of the reported starvation deaths in India between 2015–2018.[17] Journalists have also reported on people with leprosy who do not have "fingers or eyes having their state benefits

[16] https://www.wired.com/story/ai-hiring-bias-disabled-people/.

[17] https://thewire.in/rights/of-42-hunger-related-deaths-since-2017-25-linked-to-aadhaar-issues.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

canceled because fingerprints and iris scans are mandatory."[18] Similar issues were also reported in Uganda (Cioffi *et al.*, 2021), where women and older persons were denied essential services like healthcare.

Many humanitarian organizations have instituted biometric IDs such that refugees or others need to give up their biometric information in order to receive assistance. In Jordan, the UNHCR requires refugees to "use eye scanners to receive cash assistance, established through a public-private partnership with the company Iris Guard" (Latonero *et al.*, 2019). Ethnographer Dragona Kaurin writes, "There is little independent auditing and oversight to ensure equitable outcomes in these systems (face recognition or other biometric systems)" (Kaurin, 2019). Such inequity can have dire consequences for refugees needing assistance, or those seeking asylum. For instance, in 2023, a US government application that asylum seekers are required to use to make asylum claims, used an automated facial analysis tool that could not identify Black asylum seekers, barring them from making their claims.[19]

Reports of allocative harms have also surfaced in the healthcare setting where computer vision systems have been used for diagnostic purposes. For example, a study from December 2020 found that "Black patients had nearly three times the frequency of occult hypoxemia that was not detected by pulse oximetry as White patients." This meant that the reliance of pulse oximeters to detect dangerously low oxygen levels, as was done during the COVID-19 pandemic, would put Black patients disproportionately at risk (Sjoding *et al.*, 2020).

To mitigate the risks of allocative and quality-of-service harms arising from computer vision systems with differing performances across various groups, researchers have started to disaggregate performance evaluations by different socio-demographic groups. For example, the Gender Shades study (Buolamwini and Gebru, 2018) examined commercially available facial analysis systems, disaggregating performance by intersectional groups defined by gender and skin-type. The study found higher error rates for darker skinned faces, with the highest error rates

[18] https://www.newindianexpress.com/states/karnataka/2017/dec/02/no-fingers-or-iris-for-aadhaar-bengaluru-woman-loses-pension-1716448.html.

[19] https://www.theguardian.com/us-news/2023/feb/08/us-immigration-cbp-one-app-facial-recognition-bias.

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

being on darker skinned women. The Gender Shades study popularized the method of intersectional disaggregated evaluations, with a number of works using this methodology to uncover bias in other computer vision tasks. Rhue (2018) uncovered racial bias within a range of commercially available emotion recognition APIs, and found that photos of Black faces were classified as displaying more negative emotions than white faces, even when controlling for the degree of smiling in the photos. Wilson *et al.* (2019) compared the performance of pedestrian detection systems across different subgroups defined by skin-tone, and found poorer performance for darker skinned pedestrians. And De Vries *et al.* (2019) uncovered geographic disparities in object classification systems by comparing performance on images depicting households from 50 countries, and finding uniformly poorer performance on images from lower-income countries.

While the idea itself is simple, developing an effective disaggregated analysis protocol can be challenging in practice (Barocas *et al.*, 2021; Madaio *et al.*, 2022). For example, deciding which groups to separate evaluations along, and subsequently defining and operationalizing these groups into computable forms, can be particularly challenging and rife with ethical tensions given the socially constructed nature of many social categories of interest (Hanna *et al.*, 2020; Khan and Fu, 2021). For instance, Buolamwini and Gebru (2018) grounded their analysis on legal scholar Kimberlé Crenshaw's seminal intersectionality theory, outlining how discrimination faced by people with multiple marginalized characteristics cannot be understood if studied in isolation (Crenshaw, 1997). Crenshaw notes how five Black American women lost a discrimination lawsuit against General Motors (GM) in 1976[20] (Crenshaw, 2016). In their decision, the judges noted that they did not find GM to discriminate against women or Black people. However, GM hired women, but only white people, for secretarial or front office positions–making Black women ineligible. Similarly, they hired Black people for industrial or maintenance related positions, but only men–once again discriminating against Black women. Any analysis that only looks at race and gender in isolation however, would miss the fact that Black women were

[20] https://law.justia.com/cases/federal/district-courts/FSupp/413/142/1660699/.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

discriminated against, hence them losing the lawsuit. Buolamwini and Gebru (2018) similarly found that accuracy metrics on subsets of the evaluation data partitioned solely by gender or skin-type did not expose the level of disparities in error rates between lighter skinned men and darker skinned women.

Following the Gender Shades study, researchers started to contribute new datasets to enable disaggregated evaluations of computer vision systems. For example, Wang *et al.* (2019a) developed Racial Faces in-the-Wild, a dataset of faces accompanied by race labels for the purpose of evaluating racial biases in face recognition systems. Schumann *et al.* (2021) developed the More Inclusive Annotations for People (MIAP) dataset, which adds person bounding boxes and perceived gender and age labels to the Open Images Extended collection. Schumann *et al.* (2024) developed the Monk Skin Tone Examples (MSTE) dataset to help train human annotators to effectively annotate skin tone, which in turn can enable more robust disaggregated analysis of computer vision systems and identify disparities in performance based on skin tone.

Although disaggregated testing is now seen as a crucial step in evaluating the performance of computer vision systems across different demographic groups, the data annotations required for this analysis often reference demographic categories that may be sensitive to collect or store in association with images, particularly when considering images of people (Villeneuve *et al.*, 2021; Andrus *et al.*, 2021). Ultimately, great care must be put into the framing and implementation of disaggregated analyses since design decisions will influence the precise model behaviors that are made visible (or invisible) within the disaggregated evaluation framework and the conclusions that can be drawn from the evaluations.

Due to ethical concerns that can arise in annotating demographic categories, and the fact that subgroups for which computer vision systems have high error rates may not be known in advance, some researchers have proposed unsupervised approaches to identify subgroups on which models underperform. This approach can be useful when subgroups are not labeled in the data or when novel subgroups on which the system underperforms need to be identified. For example, Krishnakumar *et al.* (2021) developed an unsupervised method based on a hierarchical clustering of dataset embeddings, that surfaces subgroups for which a model

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

underperforms alongside image features that may offer explanations for misclassification. Lee *et al.* (2022) extended this work by building a visualization tool to aid in the presentation and examination of the findings. And other researchers have proposed methods to evaluate the biases of automated facial analysis models, by measuring how model performance changes when images of faces are altered across attributes like skin-tone (Balakrishnan *et al.*, 2020; Denton *et al.*, 2019).

While evaluating the performance of computer vision systems, it is important to take into account that different groups can have varying perspectives on whether or not a particular system performs well. To study these varied perspectives, some works have employed qualitative evaluation methods to understand how users from different demographic groups perceive the quality of the results of a particular computer vision system. For example, Qadri *et al.* (2023) explored the perception of South Asian cultural representation of text-to-image models, with participants from Pakistan, India, and Bangladesh. Following a similar method, Mack *et al.* (2024) explored the perception of disability representation of T2I models with US participants holding a range of disability identities. In both studies, participants identified low quality results that closely related to their identities, and that may not have been easily identifiable by people without their lived experiences. These studies emphasize the complexity of determining what constitutes a high quality result for computer vision systems such as T2I models, and argue for participatory evaluation frameworks that directly engage impacted groups in the evaluation of a system.

The exclusion and underrepresentation of certain groups within datasets used for training and evaluating computer vision models, often results in models with worse performance on the underrepresented groups (Buolamwini and Gebru, 2018; Paullada *et al.*, 2021). In response, computer vision researchers have invested in developing new datasets that strive for greater diversity along a range of sociocultural dimensions. For example, a number of new facial analysis datasets have been produced that aim to depict a balanced number of people along dimensions such as gender, age, skin-tone, and racial or ethnic categories (Karkkainen and Joo, 2021; Merler *et al.*, 2019). Yang *et al.* (2020) used filtering and balancing techniques to update the ImageNet

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

dataset, in order to mitigate harms arising from unfair gender and racial bias. Researchers have also made some progress in mitigating disparities in model performance across geographic regions, by developing new geographically diverse training and evaluation datasets for a number of computer vision tasks ranging from object recognition to action recognition (e.g., Piergiovanni and Ryoo, 2020; Ramaswamy *et al.*, 2023). While diversifying datasets can help reduce performance disparities in some cases, certain data collection methods used to diversify datasets can introduce new harms such as privacy and consent violations,[21] which we discuss further in Section 3.3.2.

Beyond diversifying datasets and disaggregated evaluation, some researchers have proposed using available information about individuals during model training, to improve the fairness of the trained model (Dwork *et al.*, 2012). For example, Ryu *et al.* (2018) find that the accuracy of face attribute detection across race and gender subgroups can be improved by pre-training on demographic information. However, while this may be an effective approach to improving subgroup performance on that specific task, additional risks may be introduced through the demographic inference step as we discussed in Section 3.1. Some researchers have also formulated disparities in model performance across subgroups as a distribution shift problem, motivating the use of domain adaptation approaches to reduce differences in performance across contexts. For example, as discussed earlier in this section, image classification models trained predominantly on images from the west, may have lower accuracy rates for images from non-western contexts since the latter images were underrepresented at training time (De Vries *et al.*, 2019). Prabhu *et al.* (2022) explore the effectiveness of off-the-shelf domain adaptation approaches to mitigate performance gaps across geographical regions, and find these approaches to have limited efficacy, suggesting that more tailored approaches may be required.

While data- and model-oriented mitigations can reduce disparities in system performance in some contexts, scholars in disciplines outside of computer vision have written about the limitations of these approaches

[21] https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

Documents
Document Sharing Hub

Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

as an ethical intervention, particularly in cases where the benefits of the technology in question are highly contested, as we discuss further in subsequent sections. For example, a number of scholars have questioned the very aim of improving the accuracy of automated facial analysis tools, given the potential for abuse and misuse of these tools, and the documented harms perpetuated through these tools (Raji *et al.*, 2020; Gebru and Denton, 2020a). In such instances, diversifying training data may improve the performance of these systems, without addressing the underlying ethical concerns pertaining to how these technologies fit into broader social systems that can use these tools against marginalized groups.

3.3 Interpersonal Harms

Shelby *et al.* (2023) categorize harms that arise from technology-mediated interactions between people shaped by power dynamics, as interpersonal harms. They describe a number of consequences of interpersonal harms including loss of agency (e.g., due to algorithmic profiling), tech-facilitated violence, e.g., through systems that are used for harassment and violence, and diminished health and well-being due to a number of factors like emotional manipulation and privacy violations, including data collection without explicit and informed consent. As noted in Shelby *et al.* (2023), many of these macro and micro categories are not always neatly separated, and a particular type of harm can be categorized in multiple ways. There have been many documented interpersonal harms mediated through computer vision systems, which we discuss here.

3.3.1 Deepfakes and Harassment

The term deepfake was first used in 2017, by a Reddit user who shared pornographic videos where the faces of the actresses in the videos were swapped out for various female celebrities. The user performed this face-swapping by using generative adversarial networks (GANs),[22] and created a reddit community called r/deepfakes where others shared

[22] https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

similar videos and their methodologies for creating these videos, until the channel was shut down.[23] In the six years since, deepfakes have been used for harassment, revenge pornography, and mis/dis information campaigns (Franks and Waldman, 2019). For example, in 2018, Indian investigative journalist Rana Ayyub was harassed by Indian nationalists, after reporting on the rape of an eight-year-old Kashmiri girl during the same year. Internet trolls used a number of harassment tactics against the journalist, including widely sharing a deepfake porn video with her face on someone else's body. Ayyub writes: "It started with a misinformation campaign to discredit me as an investigative journalist. Then my face was edited into a porn video and they doxxed me . . . I started throwing up . . . Then, the fanpage of the BJP's leader shared the video and the whole thing snowballed. The video was shared 40,000 more times. It ended up on almost every phone in India. The next day, they doxxed me" (Ayyub, 2018). The harassment only subsided after the United Nations eventually intervened.

Deepfake enabled harassment campaigns are not isolated incidents, and the targets are disproportionately women. According to a TechCrunch article,[24] deep fake detection company Sensity estimated that there were a total of 49,000 deepfakes online in 2020, over 95% of which were porn with faces of women. There are even services to help people create such types of nonconsentual pornographic videos using deepfakes. For instance, in 2019, a service that creates pornographic videos with women's on demand faces was launched and then taken down after the publication of an MIT Tech Review article on the company providing that service.[25]

Harms perpetuated through deepfakes stand to be exacerbated with the advent and proliferation of diffusion model based systems that make it much easier for people to generate synthetic images with many different properties. As an example, artist Sarah Anderson (Anderson, 2020) describes how people edited her work "to reflect violently racist

[23] https://www.vice.com/en/article/neqb98/reddit-shuts-down-deepfakes.

[24] https://techcrunch.com/2022/08/24/deepfakes-for-all-uncensored-ai-art-model-prompts-ethics-questions/.

[25] https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

messages advocating genocide and Holocaust denial, complete with swastikas and the introduction of people getting pushed into ovens. The images proliferated online, with sites like Twitter and Reddit rarely taking them down . . . Through the bombardment of my social media with these images, the alt-right created a shadow version of me, a version that advocated neo-Nazi ideology . . . I received outraged messages and had to contact my publisher to make my stance against this ultraclear." Anderson discusses how she expects these issues to worsen with the advent of text-to-image models writing, "The notion that someone could type my name into a generator and produce an image in my style immediately disturbed me . . . I felt violated" (Anderson, 2020). Other artists have also described their loss of agency when people are able to mimic their styles without their consent, with Jiang *et al.* outlining many of the additional harms posed by image generators to artists, chief among them being economic loss (Jiang *et al.*, 2023).

3.3.2 Privacy Violations

Khan and Hanna define privacy violations as the "loss of control over one's information" (Khan and Hanna, 2023). This can occur in the context of computer vision systems, when people's data, like their biometric information, whereabouts, or other information which they did not consent to providing, is taken from them. These privacy violations can often result in many of the other types of harms that we discuss throughout Section 3.

There are a number of computer vision datasets consisting of photos of people that have resulted in privacy violations of the people depicted in the datasets. Given that one of the dominant modes of data collection for many computer vision datasets is scraping the web through image search, Common Crawl, or other means, there are image datasets that were created with these methodologies, that have images of people who did not consent to the inclusion of their data in computer vision datasets. For example, the Diversity in Faces dataset was developed with the aim of improving the fairness of automated facial analysis technologies (Merler *et al.*, 2019). However, despite this well-intentioned aim, the dataset was met with severe backlash due to non-consensual

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

practices of data collection. In 2020, two residents of the state of Illinois in the US sued IBM and others after finding their images in the Diversity in Faces Dataset without their consent.[26] The state's Biometric Information Privacy Act (BIPA)[27] prohibits the use of biometric information without consent. Moreover, when people uploaded their photos to the photosharing website Flickr, which is where the data was scraped from, many did not do it with the knowledge that their faces were going to be used in datasets to help train and evaluate automated facial analysis systems. The Diversity In Faces dataset has since been taken down. More generally, this example points to the importance of data subject and data creator agency over data collection. This is no less important when the aim of a data collection effort is to improve system performance for a previously under-represented subgroup in the system's training data.

There are many examples of privacy violations that were introduced during the dataset construction or product deployment processes. For example, MIT Tech Review reported that a house cleaning robot was sending images of people in bathrooms and children to data labelers to help train the system to navigate homes effectively.[28] And multiple research datasets that consisted of images from public cameras in cafes (Stewart *et al.*, 2016; Harvey and LaPlace, 2021) and university settings (Harvey and LaPlace, 2021) without the knowledge or consent of the people in the datasets, have since been removed in response to privacy concerns, although the papers introducing the datasets have not been retracted.

Given the proliferation of computer vision systems, a number of human rights leaders are also warning about increasing privacy violations that can be enabled by these systems. Brett Solomon, the executive director of digital rights NGO Access Now, warns about authoritarian regimes making use of near-perfect face recognition technologies,

[26] https://techcrunch.com/2020/07/15/facial-recognition-lawsuit-vance-janecyk-bipa/.

[27] https://techcrunch.com/2016/12/29/facebook-stalls-in-lawsuit-alleging-its-facial-recognition-tech-violates-illinois-law/.

[28] https://www.technologyreview.com/2022/12/19/1065306/roomba-irobot-robot-vacuums-artificial-intelligence-training-data-privacy/.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...

coupled with geo-location identifiers that can enable tracking in real time (Solomon, 2018). Echoing Brett Solomon's concerns, a number of journalists and researchers have cautioned that the quest to build "the internet of things," a world where most ordinary tools are equipped with sensors including cameras, radically increases the amount of personal data that can be collected from people. Similarly, egocentric vision, which envisions humans equipped with cameras and microphones that are constantly turned on, can record anyone's video and audio (e.g., Grauman *et al.*, 2022).

Although researchers have introduced some privacy safeguards into wearable camera data collection efforts as we discuss later on, tech journalist James Vincent notes that the development of augmented reality technology is outpacing the development of accompanying privacy safeguards (Vincent, 2021). In a Washington Post article titled "Surveillance will follow us into 'the metaverse,' and our bodies could be its new data source," Tatum Hunter writes about the myriads of privacy issues that are opened up by the metaverse, since Virtual Reality headsets can collect much more data than passive screens (Hunter, 2022). This data could in turn be used to profile individuals, sell them more ads, and even manipulate people's behavior. Author Shoshona Zuboff calls this practice of companies surveilling users to profile them and profit from attempting to infer their actions, surveillance capitalism (Zuboff, 2018).

With "the internet of things," surveillance that was somewhat confined to online interactions can extend to people's homes and other surroundings like physical stores, with biometric information gathered at every step. For instance, a number of multinational retail stores around the world have already deployed automated facial analysis tools or iris scanners used to collect payments.[29, 30] According to Axios, there are retail industries "experimenting with biometric technologies that" attempt to "interpret facial expressions, detect sweat on a person's skin or identify an elevated heart rate."[31] Brenda Leong from the Future of Privacy Foundation notes: "If someone walks into a drug store and

[29] https://www.wired.co.uk/article/supermarket-future-alibaba.

[30] https://www.medianama.com/2021/11/223-exclusive-telangana-facial-recognition-pds-scheme/.

[31] https://www.axios.com/2021/07/19/facial-recognition-retail-surge.

Documents
Document Sharing Hub
Full document details: IEEE Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and...