INTRODUCTION

Human-centric computer vision (HCCV) technologies,[1] including facial recognition, are some of the most controversial AI technologies. HCCV systems are among the few types of AI that have been subject to bans or moratoriums. Many U.S. jurisdictions have restricted the use of facial recognition technologies (FRT) by government entities, particularly law enforcement.[2] The recent E.U. proposed AI regulation categorizes all remote biometric identification (RBI) systems as high-risk (and thus subject to extensive regulatory requirements),[3] and prohibits the use of RBI by law enforcement (with some narrow carve-outs).[4] From a privacy perspective, the specter of mass surveillance, particularly by state actors, has led to significant criticism of the growing pervasiveness of FRT[5] and growing pushes for strengthening information privacy laws.

In addition, in recent years, there has been a growing awareness of the issues of bias in HCCV. The highly influential Gender Shades paper showed that many of the major commercial gender classification algorithms performed less well on women than men and less well on individuals with deeper skin tones than lighter skin tones.[6] Since then, subsequent studies,

---

[1] As will be discussed further in the Definitions section below, HCCV in this Article refers to computer vision systems that rely on images of humans for training and/or testing. This is a more specific subset of the "human-centered machine learning" models that Model Cards focus on. Margaret Mitchell et al., *Model Cards for Model Reporting*, PROC. OF THE 2019 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 220, https://dl.acm.org/doi/10.1145/3287560.3287596. HCCV is a more expansive term than Facial Processing Technologies (FPT), which encompasses "any task involving the identification and characterization of the face image of a human subject." Inioluwa Deborah Raji & Genevieve Fried, *About Face: A Survey of Facial Recognition Evaluation*, AAAI 2020 WORKSHOP ON AI EVALUATION, https://arxiv.org/abs/2102.00813. HCCV includes tasks involving human bodies and objects. HCCV can also be seen as any computer vision system relying on "people-centric" datasets. Margot Hanley et al., *An Ethical Highlighter for People-Centric Dataset Creation*, NAVIGATING THE BROADER IMPACTS OF AI RESEARCH WORKSHOP AT NEURIPS 2020, https://arxiv.org/abs/2011.13583.

[2] *See, e.g.,* Electronic Privacy Information Center, *State Facial Recognition Policy*, EPIC.ORG (last visited Jan. 3, 2022), https://epic.org/state-policy/facialrecognition/ (listing moratoriums or bans in California and Massachusetts); Grace Woodruff, *Maine Now Has the Toughest Facial Recognition Restrictions in the U.S.*, SLATE (July 2, 2021), https://slate.com/technology/2021/07/maine-facial-recognition-government-use-law.html (describing Maine's ban), *Facial Recognition Technology Ban Passed by King County Council*, KINGCOUNTY.GOV (June 1, 2021), https://kingcounty.gov/council/mainnews/2021/June/6-01-facial-recognition.aspx (describing King County's ban in Washington state).

[3] EUROPEAN COMMISSION, PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS (hereinafter "E.U. Proposed AI Regulation"), Annex III, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[4] EUROPEAN COMMISSION, PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, Title II, Article 5, 1(d), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[5] *See, e.g.*, Antoaneta Roussi, Resisting the Rise of Facial Recognition, 587 NATURE 350 (2020), https://www.nature.com/articles/d41586-020-03188-2; EDRi, *Facial Recognition & Biometric Mass Surveillance: Document Pool*, EDRI.ORG (Mar. 25, 2020), https://edri.org/our-work/facial-recognition-document-pool/; *Ban Dangerous Facial Recognition Technology That Amplifies Racist Policing*, AMNESTY INTERNATIONAL (Jan. 26, 2021), https://www.amnesty.org/en/latest/news/2021/01/ban-dangerous-facial-recognition-technology-that-amplifies-racist-policing/; *Facial Recognition Technology*, ACLU, https://www.aclu.org/issues/privacy-technology/surveillance-technologies/face-recognition-technology.

[6] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in*

including one by the National Institute of Standards and Technology (NIST), part of the U.S. Department of Commerce, have shown differences in performance on the basis of skin tone and gender for different HCCV systems.[7] These studies have attributed these biases to a lack of diversity in the datasets used to train these commercial AI systems.[8]

Simultaneously addressing these concerns around privacy and fairness, however, is difficult in practice.

Commercial Gender Classification, PROCEEDINGS OF MACHINE LEARNING RESEARCH 81:1–15, CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2018), http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[7] Patrick Grather et al., *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, NATL. INST. STAND. TECHNOL. INTERAG. INTERN. REP. 8280 (Dec. 2019). https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

[8] *Id*. *See* supra note 6.

[9] Michele Merler et al., *Diversity in Faces* (2019), https://arxiv.org/pdf/1901.10436.pdf.

[10] Kyle Wiggers, *IBM Releases Diversity in Faces, a Dataset of Over 1 Million Annotations to Help Reduce Facial Recognition Bias*, VENTUREBEAT (Jan. 29, 2019), https://venturebeat.com/2019/01/29/ibm-releases-diversity-in-faces-a-dataset-of-over-1-million-annotations-to-help-reduce-facial-recognition-bias/.

[11] Taylor Shankland, *IBM Stirs Controversy by Using Flickr Photos for AI Facial Recognition*, CNET (Mar. 13, 2019), https://www.cnet.com/news/ibm-stirs-controversy-by-sharing-photos-for-ai-facial-recognition/.

[12] Taylor Hatmaker, Lawsuits Allege Microsoft, Amazon and Google Violated Illinois Facial Recognition Privacy Law, TECH CRUNCH (Jul. 15, 2020), https://techcrunch.com/2020/07/15/facial-recognition-lawsuit-vance-janecyk-bipa/.

[13] Nicolas Rivero, *The Influential Project that Sparked the End of IBM's Facial Recognition Program*, QUARTZ (June 10, 2020), https://qz.com/1866848/why-ibm-abandoned-its-facial-recognition-program/.

[14] *See supra* note 12.

[15] *See supra* note 2 at 3.

I: DEFINITIONS

Throughout this Article, I will use the term "human-centric computer vision" (HCCV) to refer specifically to AI systems that rely on images of humans for their training and test data.[29] These are the AI technologies whose *development* is directly affected by biometric information privacy regulations that protect information extracted from human faces or bodies. I stress the word "development" because the human images I address in this Article are the images in the training set used to teach the HCCV system how to detect, recognize, or classify people or objects or the images in the test set used to evaluate the model's performance. These images used for development are typically distinct from the images the HCCV system perceives in deployment.[30] The question of which images developers should be allowed to process when the system is deployed is inextricably tied the highly context-specific exercise of determining which use cases of HCCV should be permitted vs. banned—although this is a highly important policy question, it is beyond the scope of this Article.

[29] *See supra* note 1 for discussion of related terms in the existing literature.

[30] An exception is the narrow case of active learning algorithms, which are continuously retrained using data gathered in the deployment context. This Article does not encourage expanding the deployment of HCCV solely for gathering more diverse data for future training.

The primary computer vision tasks motivating this piece are facial recognition, detection, verification, and classification, but I use the more expansive term of HCCV since many of my points also apply to body detection, pose estimation, and body recognition. Object detection and classification are also relevant insofar as developers use images of people and objects to train their models.

Although colloquially HCCV technologies are often referred to as "facial recognition technologies" (FRT), FRT is only a small subset of HCCV. HCCV encompasses *all* computer vision technologies whose development requires biometric information—thus confronting current information privacy laws—but these laws are typically motivated by the desire to tackle FRT specifically. In addressing the tensions between existing privacy laws and HCCV bias mitigation efforts, it is thus important to note that HCCV includes technologies, as enumerated below, that largely do not figure in policy conversations about biometric information privacy laws. Note that the paragraphs below do not seek to classify these technologies into "acceptable" vs. "unacceptable" bins, but rather to illustrate the wide variety of HCCV technologies.

Face detection involves detecting whether a human face is in an image and, if so, drawing a bounding box or other boundary around the face. This is one of the most frequently used face-related computer vision tasks and serves as the basis for the other face-related tasks (you must first detect a face before you can identify or analyze it). Face or body detection is often used to count people or to trigger a subsequent task. For example, an AI-assisted AC system for an office might only turn on if a human is detected as being in the room. An AI-assisted elevator might count the number of people in the elevator and not stop for additional people if the elevator is at capacity.

Face verification and recognition are related tasks for identifying a person. Face verification refers to a one-to-one comparison between a reference face and a new face. When unlocking a phone, a face verification algorithm is used to compare the face perceived by the camera with the reference face for the owner of the phone. Facial recognition refers to one-to-many comparisons; the perceived face is compared against a reference set of faces to identify which (if any) of the reference faces is a match. If police have an image of a suspect, they can run that image through a FRT system that compares the image to a reference set of driver's

[31] Illinois's BIPA, for example, defines "biometric identifiers" as "retina or iris scan, fingerprint, voiceprint, or scan of hand or face geometry." Typically, images used for training melanoma detection models are close-ups of the skin, so biometric information privacy laws would not apply. *See, e.g.*, Veronica Rotemberg et al., *A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context*, 8 SCIENTIFIC DATA 34 (2021), https://www.nature.com/articles/s41597-021-00815-z.

license photos to see if there is a match. FRT can also be used in social media applications to generate tag suggestions.

Face classification, also known as "facial analysis," refers to the task of automatically generating labels for a face. For example, the model might label faces as "male" or "female." This type of task can be fraught from an ethical perspective given concerns around how much information can be accurately discerned from someone's face. Gender classification has especially been criticized since gender cannot be assessed purely based on a photo, especially if an individual is transgender or non-binary.[32] In addition, controversial technologies like emotion recognition and character/fitness assessments fall under this category. Research suggests that emotion recognition is largely unreliable because people's facial expressions do not directly reflect their emotions—e.g., you might smile through discomfort or sadness.[33] In addition, efforts to use face classification to identify who might be a better job candidate or who might have a propensity to criminal behavior have been highly criticized as pseudoscientific.[34] That said, facial analysis can also be used for more benign purposes, such as a "smile setting" on a camera that waits until everyone in the frame is smiling before taking a photo.[35] AI-assisted medical analyses of a person's body or face can also fall into the classification category.

Body detection/verification/recognition/analysis tasks are analogous to the face-related tasks above, except that the focus is on the entire body rather than the face. Body detection, for example, might be used by an autonomous vehicle to detect and avoid pedestrians. Pose estimation is also a common task in this category and is used to estimate the spatial key points of a person's joints to determine whether an individual is doing a certain activity. In a security context, the goal might be to detect whether someone is shoplifting or making rapid movements that might be dangerous. Such technologies are also commonly used for augmented reality or CGI. Pose estimation typically does not involve identifying the person, but it can be used for such purposes. For example, gait recognition—leveraging the patterns unique to each person's gait to identify an individual—is recognized as a form of biometric identification, which is subject to relevant biometric information privacy laws in the U.S. and E.U.[36]

---

[32] *See* Os Keyes, *The Misgendering Machiens: Trans/HCI Implications of Automatic Gender Recognition*, PROC. OF THE ACM CONF. ON HUMAN-COMPUTER INTERACTION, Vo. 2, Issue CSCW (Nov. 2018), https://dl.acm.org/doi/10.1145/3274357.

[33] *See* Douglas Heaven, *Why Faces Don't Always Tell the Truth About Feelings*, 578 NATURE 502 (2020), https://www.nature.com/articles/d41586-020-00507-5; *Emotion Recognition: Can AI Detect Human Feelings From a Face?*, FINANCIAL TIMES (May 11, 2021), https://www.ft.com/content/c0b03d1d-f72f-48a8-b342-b4a926109452; Kate Crawford, *Artificial Intelligence is Misreading Human Emotion*, THE ATLANTIC (Apr. 27, 2021), https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696/.

[34] Blaise Aguera y Arcas et al., *Physiognomy's New Clothes*, MEDIUM (May 6, 2017), https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a; *Facial Recognition to "Predict Criminals" Sparks Row Over AI Bias*, BBC (June 24, 2020), https://www.bbc.com/news/technology-53165286; Jeremy Kahn, *HireVue Drops Facial Monitoring Amid A.I. Algorithm Audit*, FORTUNE (Jan. 19, 2021), https://fortune.com/2021/01/19/hirevue-drops-facial-monitoring-amid-a-i-algorithm-audit/

[35] Katherine Boehret, *New Cameras Guarantee A Smile on Your Face*, WALL ST. J. (Apr. 23, 2008), https://www.wsj.com/articles/SB120889435178135615.

[36] EUROPEAN COMMISSION, PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020XX1117%2801%29&qid=1627962106278; California Consumer Privacy Act of 2018, California Civil Code Title 1.81.5, https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

Moving to the core terms for this Article, being "seen" refers specifically to having images of your face and/or body collected and processed for *developing* HCCV systems. This definition encompasses computer vision contexts where there are privacy considerations under existing biometric information privacy laws, which will be further discussed in Section IV. Being "unseen" thus means *not* having your images or images of people like you collected or processed for *developing* HCCV (i.e., included in training or test sets). This includes images used to train the base model, which performs a more general task, and images collected in the specific domain for the specific task. Note that being "seen"/"unseen" focuses specifically on the how the HCCV system is *developed* since the tension highlighted in this paper is between privacy and the desire to *develop* more accurate and fairer HCCV systems. The focus is *not* on the images collected during the deployment of the HCCV system since those images are generally not useful for improving the fairness of the model, so there is usually no fairness vs. privacy tension in deployment. Another way to think of this dichotomy is images used for learning/evaluation vs.

(CCPA), New York Assembly Bill A6787D, https://www.nysenate.gov/legislation/bills/2019/a6787 (NY); Dan Cooper & Gemma Nash, *UK ICO Publishes New Guidance on Special Category Data*, COVINGTON (Nov. 29, 2019), https://www.insideprivacy.com/eu-data-protection/uk-ico-publishes-new-guidance-on-special-category-data/.

[37] COMMON OBJECTS IN CONTEXT (COCO) DATASET, https://cocodataset.org/.

[38] Note that recent research has shown that using privacy-preserving techniques like face blurring can enable object recognition to be trained on such datasets while reducing the privacy risk. Kaiyu Yang et al., *A Study of Face Obfuscation in ImageNet*, https://arxiv.org/abs/2103.06191 (2021). Face blurring will be discussed further in Section VII.D.

[39] *Algorithm*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/algorithm#note-1.

inference. Note that this distinction can break down in the context of active learning, where the model continues to learn from images collected in deployment. In such contexts,[40] the status quo prioritization of privacy is reasonable.

Being "mis-seen" refers to experiencing poor performance from a deployed HCCV system: this includes your face/body not being detected, being mis-recognized for someone else, someone else being mis-recognized for you, or having images/videos of you mis-classified or mis-characterized. This last category includes tasks like suspicious behavior detection, where you might be erroneously labeled as cheating on an exam or shoplifting. As will be explored in greater depth in Section VI, the harms of being mis-seen are both absolute and relative. An HCCV system can be harmful because it performs poorly in certain scenarios for all people or because it performs more poorly for specific subgroups, potentially perpetuating stereotypes or creating discriminatory disparities.

eing included vs. excluded in the training and test sets used to develop the model affects the accuracy of the model on individuals like you, not *whether* the system will be deployed on individuals like you or whether you will be included in a reference set. The excessive deployment of such technologies to surveil marginalized communities is what leads to these problems of hypervisibility.

Lastly, it is important to define the term "bias." Because "bias" is a catch-all term for many different types of disparities, some in the algorithmic fairness community have criticized the use of its term, arguing instead for more precise descriptions of the specific harms.[43] In this

[40] This approach is uncommon in deployment, however, given that it requires someone to label the new images collected to continue training the model.
[41] Isis H. Settles et al., *Scrutinized but not Recognized: (In)visibility & Hypervisibility Experiences of Faculty of Color*, J. OF VOCATIONAL BEHAVIOR (2018), https://www.icos.umich.edu/sites/default/files/lecturereadinglists/Settles%2C%20Buchanan%2C%20%26%20Dotson%202018%20Scrutinized%20but%20not%20recognized.pdf.
[42] Rasul A. Mowatt et al., *Black/Female/Body Hypervisibility & Invisibility: A Black Feminist Augmentation of Femist Leisure Research*, 45.5 J. OF LEISURE RESEARCH 644 (2013), https://www.nrpa.org/globalassets/journals/jlr/2013/volume-45/jlr-volume-45-number-5-pp-644-660.pdf
[43] *See* Barocas et al., *Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs*, PROC. OF 2021 AAI/ACM CONF. ON AI, ETHICS, AND SOCIETY 368 (2021), https://dl.acm.org/doi/abs/10.1145/3461702.3462610. *See also* Su Lin Blodgett et al., *Language (Technology) is*

Article, I will use the term "bias" to refer to disparate performance of the HCCV system (e.g., different rates of mis-recognition, mis-detection, or mis-classification) across different groups that might lead to disproportionate harm for specific groups. In Section VI, I break down the specific types of bias harms associated with being "mis-seen." "Fairness" in this Article will refer to the pursuit of bias mitigation. It is impossible for an AI system to be completely unbiased or "fair," but the goal is to minimize bias as much as possible while preserving privacy.

II: WHY WORRY ABOUT BEING MIS-SEEN?

Given that this Article focuses on the current tensions and imbalances between privacy and fairness when developing HCCV, it is important to address the basic question of why being "mis-seen" is such a problem.

The growing proliferation of HCCV in everyday life suggests that even small or subtle biases might accumulate into substantial harms.

Imagine, for example, being an individual of a minority demographic living in a world of HCCV designed for individuals in the majority group. Upon waking up, you check your phone, but it does not recognize you, so you have to manually input your passcode. Taking public transit to work, you try to use the facial recognition system to pay your fare, but it does not recognize you, so you must go through a special line with a human verifier and arrive late to work. You join your colleagues for coffee at a cafe, but again the payment system fails to recognize you. You are embarrassed as the automated system says your face does not match the bank account you are trying to access, and you have to ask the cafe staff to give you another method of

---

Power: A Critical Survey of "Bias," PROC. OF 58TH ANNUAL MEETING OF THE ASSOC. FOR COMPUTATIONAL LINGUISTICS 5454 (2020), https://aclanthology.org/2020.acl-main.485.pdf (example delineating specific harms).

[44] *Farmington Hills Father Sues Detroit Police Department for Wrongful Arrest Based on Faulty Facial Recognition Technology*, ACLU (Apr. 13, 2021), https://www.aclumich.org/en/press-releases/farmington-hills-father-sues-detroit-police-department-wrongful-arrest-based-faulty.

[45] *See, e.g.,* Title VII of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000e to 2000e-17 (2000 & Supp. 2004).

[46] *See, e.g.*, Fair Housing Act (FHA), Title VIII of the Civil Rights Act of 1968, 42 U.S.C. §§ 3601-3619.

[47] *See, e.g.,* Equal Credit Opportunity Act (ECOA), 15 U.S.C. § 1691 (2012).

[48] *See, e.g.,* Title VI of the Civil Rights Act of 1964, 42 U.S.C. 2000d et seq. ("Title VI").

payment. They unfortunately do not have any other methods of payment, so you need to ask a colleague to cover your tab. When you and your colleagues return to the office, you are unable to enter the building because the security system does not recognize you as one of the employees. While your colleagues are waiting for you, you call for a security guard to help you enter the building. The security guard is suspicious of your claim that you work in the office—the picture in the employee database looks like it *could* be someone else, and the AI system works extremely well for everyone else. Fortunately, your colleagues vouch for you, and the security guard lets you in. At the end of the workday, you stay late, after your colleagues have left, to finish a project. The lights and AC turn off, as the AI-enabled AC and lighting systems do not detect any people in the office. Sitting in the darkness, you are confronted with your own invisibility.

In the above scenario, I have only discussed a few of the possibly many instances of inconvenience, indignation, or embarrassment that might occur over the course of the day due to being "mis-seen" by HCCV. While most of the harms described would not be legally cognizable, together they amount to being treated as a second-class citizen, living in a world that cannot detect or recognize you. This sensation is similar to being a foreign tourist, forced to use alternative systems since you do not have a phone number, address, bank account card, etc. in the country, except that you cannot prevent these harms by simply setting up relevant accounts—you would need to change your face/body.

[49] McKane Andrus et al., *What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness*, PROC. OF 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 249 (2021), https://dl.acm.org/doi/10.1145/3442188.3445888.
[50] Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. 649 (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3650635.

IV: CHALLENGES TO ALGORITHMIC BIAS MITIGATION IN COMPUTER VISION

Collecting larger, more diverse training datasets and test datasets serves two aims: (i) improving the overall accuracy and robustness of the model and (ii) mitigating potential biases. While this Article addresses both aims, the focus is primarily on issues of bias since there are arguably sufficient existing commercial incentives to improve the overall performance of HCCV systems. Indeed, the accuracy of major commercial facial recognition technologies has improved dramatically over the past few years, while issues of bias persist.[72]

While the desire to build larger and more diverse datasets for training and testing computer vision systems is admirable, doing so immediately runs into complex questions of privacy, consent, money, and possible exploitation. Indeed, the computer vision community is infamous for blurring or crossing ethical lines to collect the large corpuses of data needed to train their systems. In the U.S., NIST uses mugshots and images of exploited children,[73] individuals crossing the border, and visa applicants in its test dataset, which is used by major companies to benchmark the performance of their commercial FRT.[74] In China, start-ups have developed facial analysis systems for identifying ethnic minorities for surveillance purposes using "face-

predicted recidivism rates) across groups. Correcting for this bias metric would involve ensuring that the ML model predicts proportional recidivism rates across groups, even if the training data suggest highly disproportionate rates. Other approaches to bias mitigation take a more nuanced approach, but most are analogous to affirmative action in contemplating some degree of rebalancing across groups for fairness rather than accuracy purposes. In the computer vision context, however, ground truth is more readily accessible, so it is easier to align fairness and accuracy. For example, if the task is to verify whether two faces are of the same person, and the test set includes unique identifiers for each of the individuals, then correcting for problems of bias (e.g., the model being worse at distinguishing between individuals of darker skin tones) directly improves accuracy as well.

[72] FACIAL RECOGNITION TECHNOLOGY: PRIVACY & ACCURACY ISSUES RELATED TO COMMERCIAL USES, U.S. GOVERNMENT ACCOUNTABILITY OFFICE REPORT TO CONGRESSIONAL REQUESTERS (July 2020), https://www.gao.gov/assets/gao-20-522.pdf.

[73] These images are used specifically to test the performance of face detection and recognition systems on children. *Chexia Face Recognition*, NIST.GOV (last accessed Jan. 5, 2022), https://www.nist.gov/programs-projects/chexia-face-recognition. Images of children are hard to come by in most datasets due to additional privacy restrictions.

[74] Os Keyes et al., *The Government Is Using the Most Vulnerable People to Test Facial Recognition Software*, SLATE (Mar. 17, 2019), https://slate.com/technology/2019/03/facial-recognition-nist-verification-testing-data-sets-children-immigrants-consent.html; Peter Grother et al., *Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification*, NISTIR 41-42, https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf; Peter Grother et al., *Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification*, NISTIR 5 https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf ("The evaluation uses six datasets: frontal mugshots, profile view mugshots, desktop webcam photos, visa-like immigration application photos, immigration lane photos, and registered traveler kiosk photos.")

image databases for people with criminal records, mental illnesses, records of drug use, and those who petitioned the government over grievances."[75]

While those datasets were collected by government entities, there are also many large publicly available human image datasets collected by academic or industry researchers. These typically rely on web-scraped photos. Some datasets focus on celebrities or public figures (e.g., MS-Celeb-1M[76]); others focus on a broader array of subjects through online platforms like Flickr (e.g., YFCC100M[77]), which made large numbers of images public and easily downloadable with Creative Commons licenses permitting their use for commercial purposes.

[75] Paul Mozur, *One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority*, N.Y. TIMES (Apr. 14, 2019), https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

[76] Yandong Guo et al., *MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition*, PROC. OF 2016 EURO. CONF. ON COMPUTER VISION, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/08/MSCeleb-1M-a.pdf (featuring 10 million face images or nearly 100,000 individuals).

[77] *Yahoo Flickr Creative Commons 100 Million (YFCC100m) Dataset*, http://projects.dfki.uni-kl.de/yfcc100m/ (featuring around 100 million images and videos).

[78] One artifact of using datasets exclusively of celebrities is that if you train a model to synthesize more feminine faces, it will do so by applying makeup to the face (specifically, a smokey eye and lipstick). *See* Joo & Karkkainen, *supra* note 92; Vidya Muthukumar, *Understanding Unequal Gender Classification Accuracy from Face Images*, PROC. OF 2019 IEEE/CVF CONF. ON COMPUTER VISION & PATTERN RECOGNITION WORKSHOPS (CVPRW), https://ieeexplore.ieee.org/document/9025567. Looking more feminine is thus conflated with wearing makeup. In contrast, the models that synthesize more masculine features actually change the features of the face to be more angular. Datasets like CelebA that include an "attractiveness" feature are also problematic in that they can replicate human biases around what looks attractive. One study illustrated this by increasing the "attractiveness" latent attribute of Barack Obama, only to find that it made him look like a young, blonde white woman. Vinay Prabhu, *Covering Up Bias in CelebA-like Datasets With Markov Blankets: A Post-Hoc Cure for Attribute Prior Avoidance*, https://arxiv.org/pdf/1907.12917.pdf.

[79] Aaron Nech & Ira Kemelmacher, *Level Playing Field for Million Scale Face Recognition*, PROC. OF 2017 IEEE CONF. ON COMPUTER VISION & PATTERN RECOGNITION (CVPR), https://www.researchgate.net/publication/320971151_Level_Playing_Field_for_Million_Scale_Face_Recognition; Merler et al., *supra* note 9; Tsung-Yi Lin et al., *Microsoft COCO: Common Objects in Context*, PROC. OF 2014 EURO. CONF. ON COMPUTER VISION (ECCV), https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

[80] *See* Nech & Kemelmacher, *supra* note 79.

[81] *See* Tsung-Yi Lin et al., *supra* note 79.

[82] *See* Merler et al., *supra* note 9.

B. HARMS OF BEING MIS-SEEN

In this Section, I will focus on four specific harms of being "mis-seen": differences in service provision, security threats, allocative harms, and representational harms. All these harms are caused by differences in the performance of the algorithmic system for different groups (e.g., lower accuracy rates or higher false positives/negatives for women or minorities), but they are distinguished by how this difference in performance affects the individuals.

First, differences in service provision refer to contexts where an algorithmic system performs a function less well for certain groups versus others. This is the most common and wide-ranging type of harm, applying to virtually all computer vision tasks. For example, if a facial verification system is used at border control to determine whether an individual's face matches the photo in their passport, but that system is less accurate for Middle Eastern individuals, then Middle Eastern individuals are more likely to be flagged and sent to a separate line for a human to conduct the verification.[141] In the face/body detection context, if an AI-assisted AC system is less proficient at detecting individuals with darker skin tones, those individuals might find that the AC often turns off even when they are still in the room.

A second category of harm is security threats. This type of harm is specific to the verification context. For example, if the face verification algorithm on your phone is not very good at distinguishing between different Asian people, and you are Asian, then other Asian people might be able to unlock your phone. This is particularly a concern in households, where, bias aside, family members can sometimes unlock each other's phones.[142] Increasingly, face verification is also used for building security and for payments,[143] so significant discrepancies in the ability of such systems to work for different groups could lead to substantial security risks (e.g., someone breaking into your home or using your credit card).

The third category of harms is allocative harms. This is when an inaccuracy leads to a misallocation of a good or opportunity. In the computer vision context, this is most relevant to recognition and classification tasks. The example of wrongful arrest due to a faulty facial recognition match is a very high-stakes example of allocative harm, as individuals are unjustly deprived of their liberty. In terms of classification tasks, algorithmic systems that seek to identify

---

[141] Given harmful stereotypes about Middle Eastern individuals in the airport security context post-911, such service provision harms could lead to allocative harms if the individual is falsely accused of carrying a passport not belonging to them.

[142] Karen Levy & Bruce Schneier, *Privacy Threats in Intimate Relationships*, 6 J. OF CYBERSECURITY 1 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3620883.

[143] *See, e.g.*, Caroline Spivack, *NYC Seeks to Curb Facial Recognition Technology in Homes & Businesses*, CURBED NEW YORK (Oct. 8, 2019), https://ny.curbed.com/2019/10/8/20903468/nyc-facial-recognition-technology-homes-businesses; Sam Dean, *Forget Credit Cards—Now You Can Pay With Your Face. Creepy or Cool?*, L.A. TIMES (Aug. 14, 2020), https://www.latimes.com/business/technology/story/2020-08-14/facial-recognition-payment-technology.

suspicious behavior or categorize an individual's mental state or ability can also lead to significant allocative harm. For example, a study found that eye tracking devices did not work as well for Asian participants as for other groups.[144] As such technology is increasingly used by educational institutions to determine whether students are paying attention and to detect cheating behavior,[145] such disparities in performance could lead to a higher risk of Asian students being incorrectly flagged for bad behavior.

Finally, we have representational harms, when algorithmic systems represent certain groups in negative, offensive, or other problematic ways. This type of harm is most relevant for classification tasks since such tasks involve applying a label to an image. A famous computer vision example of a representational harm was when Google Photos labelled an image of two Black individuals as an image of gorillas.[146] This harm can also occur with algorithms that determine which parts of images are the most relevant to focus on. In 2021, Twitter scrapped its image cropping algorithm following revelations that their algorithm was more likely to crop out black faces in favor of white faces.[147] Representational harms can also stem from existing biased trends in society. In the popular COCO dataset, images of women playing sports are more likely to be indoors, whereas images of men playing sports are more likely to be outdoors.[148] This can lead to HCCV models trained on COCO learning stereotyped representations. AI-powered image caption generators might consistently incorrectly label images of women playing outdoor sports as "men playing sports" and vice versa for men playing indoor sports, further perpetuating existing stereotypes.

While this Article primarily focuses on non-generative models, it is worth noting that representational harms are an especially relevant type of harm to consider when evaluating generative models. For example, Generative Adversarial Networks (GANs) trained to generate a synthetic image of an individual with longer hair have been shown to also feminize the facial features of the individual.[149] By conflating long hair with feminine facial features, the GAN perpetuates the stereotype that men have short hair and women long hair. Similarly, an app designed to make faces look more attractive could be offensive if it does so by making skin look lighter, an artifact of learning cultural biases that consider lighter complexion faces to be more attractive.[150] Generative language models have also been shown to be vulnerable to generating

---

[144] Pieter Blignaut & Daniel Jacobus Wium, *Eye-Tracking Data Quality as Affected By Ethnicity & Experimental Design*, 46 BEHAVIORAL RESEARCH METHODS 1 (2013), https://www.researchgate.net/publication/236266469_Eye-tracking_data_quality_as_affected_by_ethnicity_and_experimental_design.

[145] Todd Feathers & Janus Rose, *Students Are Rebelling Against Eye-Tracking Exam Surveillance Tools*, VICE (Sept. 24, 2020), https://www.vice.com/en/article/n7wxvd/students-are-rebelling-against-eye-tracking-exam-surveillance-tools,

[146] Notably this highly offensive harm seems to still not have been directly solved for. Tom Simonite, *When It Comes to Gorillas, Google Photos Remains Blind*, WIRED (Jan. 11, 2018), https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

[147] Rumman Chowdhury, *Sharing Learnings About Our Image Cropping Algorithm*, TWITTER (May 19, 2021), https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

[148] *See* Wang et al., *supra* note 54.

[149] G. Balakrishnan et al., *Towards Causal Benchmarking of Bias in Face Analysis Algorithms*, PROC. OF 16TH EURO. CONF. ON COMPUTER VISION (ECCV) 547 (2020), https://dl.acm.org/doi/abs/10.1007/978-3-030-58523-5_32.

[150] One study illustrated this by increasing the "attractiveness" latent attribute of Barack Obama, only to find that it made him look like a young, blonde white woman. Vinay Prabhu, *Covering Up Bias in CelebA-like Datasets With Markov Blankets: A Post-Hoc Cure for Attribute Prior Avoidance*, https://arxiv.org/pdf/1907.12917.pdf.

highly racist and offensive language. For example, Microsoft famously scrapped its chatbot Tay after the bot started making highly inflammatory statements.[151]

Most concerns about bias in computer vision apply primarily to contexts where images of humans are used, but bias can also manifest itself in object detection or recognition. As discussed previously, researchers at Facebook found that their tool had a harder time identifying objects in photos taken in developing countries.[152] Because their training data was disproportionately collected from developed countries, the model could only recognize toothpaste on a sink in a more affluent-looking bathroom. This is why, depending on the task, it is important not only to consider the demographic diversity of the people in the images, but also to consider factors like the geographic diversity of where the images are taken.

[151] Elle Hunt, Tay, *Microsoft's AI Chatbot, Gets A Crash Course in Racism From Twitter*, GUARDIAN (Mar. 24, 2016), https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech.
[152] *See* DeVries et al., *supra* note 60.
[153] *See* Andrus et al., *supra* note 49.
[154] Linsey Barrett, *Ban Facial Recognition Technologies for Children—And for Everyone Else*, 26 B.U. J. SCI. & TECH. L. 223 (2020), http://www.bu.edu/jostl/files/2020/08/1-Barrett.pdf; Sharon Nakar & Dov Greenbaun, *Now You See Me. Now You Still Do: Facial Recognition Technology & the Growing Lack of Privacy*, 23 B.U. J. SCI. & TECH. L. 88 (2017), http://www.bu.edu/jostl/files/2017/04/Greenbaum-Online.pdf; Joel R. Reidenberg, *Privacy in Public*, 69 U. MIAMI. L. REV. 141 (2014), https://repository.law.miami.edu/cgi/viewcontent.cgi?article=1345&context=umlr.