BEING “SEEN” VS. “MIS-SEEN”: TENSIONS BETWEEN PRIVACY AND FAIRNESS IN COMPUTER VISION

Alice Xiang

ABSTRACT

The rise of facial recognition and related computer vision technologies has been met with growing anxiety over the potential for AI to create mass surveillance systems and further entrench societal biases. These concerns have led to calls for greater privacy protections and fairer, less biased algorithms. An under-appreciated tension, however, is that privacy protections and bias mitigation efforts can sometimes conflict in the context of AI. Reducing bias in human-centric computer vision systems (HCCV), including facial recognition, often involves collecting large, diverse, and candid image datasets, which can run counter to privacy protections.

It is intuitive to think that being “unseen” by AI is preferable—that being under-represented in the data used to develop facial recognition might somehow allow one to evade mass surveillance. As we have seen in the law enforcement context, however, just because facial recognition technologies are less reliable at identifying people of color has not meant that they have not been used to surveil these communities and deprive individuals of their liberty. Thus, being “unseen” by AI does not protect against being “mis-seen.” While in the law enforcement context this tension can simply be resolved by prohibiting the use of facial recognition technology, HCCV encompasses a much broader set of technologies, from face detection for a camera’s autofocus feature to pedestrian detection on a self-driving car.

The first contribution of this Article is to characterize this tension between privacy and fairness in the context of algorithmic bias mitigation for computer vision technologies. In particular, this Article argues that the irreducible paradox underlying current efforts to design less biased algorithms is the simultaneous desire to be “unseen” yet not “mis-seen” by AI. Second, the Article reviews the strategies that have been proposed for resolving this tension and evaluates their viability for adequately addressing the technical, operational, legal, and ethical challenges surfaced by this tension. These strategies include: using third-party trusted entities to collect data, using privacy-preserving techniques, generating synthetic data, obtaining informed consent, and expanding regulatory mandates or government audits. Finally, this Article argues that solving this paradox requires considering the importance of not being “mis-seen” by AI rather than simply being “unseen.” Detethering these notions (being seen vs. unseen vs. mis-seen) can help clarify what rights relevant laws and policies should seek to protect. For example, this Article will examine the implications of a right not to be disproportionately *mis-seen* by AI, in contrast to regulations around what data should remain *unseen* by AI. Given that privacy and fairness are both extremely important objectives for ethical AI, it is vital for lawmakers and technologists to address this tension head-on; approaches that rely purely on visibility or invisibility will likely fail to achieve either objective.

TABLE OF CONTENTS

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. In an effort to address bias in HCCV, researchers at IBM created the Diversity in Faces (DiF) dataset,[9] which was initially met with a positive reception for being far more diverse than previous face image datasets.[10] DiF, however, soon became embroiled in controversy once journalists highlighted the fact that the dataset consisted of images from Flickr.[11] The Flickr images had Creative Commons licenses, covering the copyright of the images, but the plaintiffs had not consented to having their images used in facial recognition training datasets.[12] In part due to this controversy, IBM announced it would be discontinuing its facial recognition program.[13] Microsoft, Amazon, and Google, which also used the DiF dataset, were also sued.[14] This example highlights the tension that HCCV technologies create between representation vs. surveillance, being "seen" vs. being "invisible." We want AI to recognize us, but we are uncomfortable with the idea of AI having access to data about us. While creating large, diverse human image datasets with informed consent is *not* impossible, as Section III.A will discuss, there are challenges that require further research and regulatory guidance.

This tension is further amplified when the need for sensitive attribute data is considered. For example, to even discern whether a training dataset is diverse, we need a taxonomy of demographic categories, some notion of an ideal distribution across that taxonomy, and labels of these demographic categories. The methods that have emerged to address these necessities are often discomfiting and raise further privacy concerns. In designing DiF, the researchers did not have a variable for race or ethnicity, so they used various computational methods to derive labels for different facial features to approximate differences across race, including metrics for skin color and craniofacial areas.[15] While these features were used in an effort to ensure racial diversity without access to direct data on race, these approaches do not capture the sociological nature of demographic labels and could be misused, as we have seen in the pseudoscience of

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.

physiognomy, which focuses on quantifying physical differences across races.[16] Other attempts at creating diverse face image datasets, like FairFace,[17] approach the challenge by having Mechanical Turkers (MTurkers) guess people's demographic attributes. If at least two MTurkers agree, then the label is considered ground truth; if there is no agreement, the image is discarded. This approach is concerning in that it relies on the ability of MTurkers to accurately assess people's demographic attributes, and it discards the images of people who might not fit neatly in the demographic taxonomy. This could, for example, lead to fewer multiracial, non-binary, or transgender individuals being represented in the data. Designing a taxonomy for demographic classification often relies on stereotypes and can impose and perpetuate existing power structures.

Existing privacy laws address this issue primarily by erring on the side of hiding people's personal data unless there is explicit informed consent. In fact, privacy law and anti-discrimination law are often viewed as symbiotic,[18] under the assumption that preventing companies from collecting protected attribute data helps to prevent discrimination. Evidence of bias in FRT, however, has contradicted this notion. There have been several cases of black men in the U.S. being wrongfully arrested due to faulty facial recognition matches.[19] In 2019, for example, Nijeer Parks, a Black man, was arrested due to a faulty facial recognition match.[20] He spent ten days in jail and paid around $5,000 to defend himself before the case was dismissed for lack of evidence.

To address such issues of bias in FRT, the policy response has centered around moratoriums on the usage of FRT by law enforcement and other public agencies.[21] While such moratoriums are reasonable given current problems with such technologies, they are limited to specific jurisdictions, do not apply to other domains for FRT, and do not address bias in other forms of HCCV. The lack of stronger regulatory incentives to address bias in HCCV is concerning given the growing use of such technology. While there are limited numbers about the broader HCCV market, the FRT market alone is projected to grow from $4.45 million in 2021 to $12.11 billion in 2028.[22] Even in North America, where FRT has been quite controversial, the

[16] *See* Blaise Agüera y Arcas et al., *Physiognomy's New Clothes*, MEDIUM (May 6, 2017), https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.
[17] Kimmo Kärkkäinen & Jungseock Joo, *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age*, PROC. OF THE IEEE/CVF WINTER CONF. ON APPLICATIONS OF COMPUTER VISION 1548 (2021), https://github.com/joojs/fairface.
[18] Jessica L. Roberts, *Protecting Privacy to Prevent Discrimination*, 56 WM. & MARY L. REV. 2097 (2015), https://scholarship.law.wm.edu/wmlr/vol56/iss6/4.
[19] *See, e.g.*, Kashmir Hill, Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match, N.Y. TIMES (Dec. 29, 2020), https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html; Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. TIMES (June 24., 2020), https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html; Elisha Anderson, *Controversial Detroit Facial Recognition Got Him Arrested for a Crime He Didn't Commit,* DETROIT FREE PRESS (July 10, 2020), https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002/.
[20] *See* "Another Arrest," *supra* note 19.
[21] *See supra* note 2.
[22] *Facial Recognition Market Size, Share & Trends Analysis by Technology (2D, 3D, Facial Analytics), by Application (Access Control, Security & Surveillance), by End-Use, by Region, and Segment Forecasts, 2021-2028*, GRAND VIEW RESEARCH (May 2021), https://www.grandviewresearch.com/industry-analysis/facial-recognition-market.

market for FRT is expected to double by 2027.[23] While recent moratoria on FRT for law enforcement suggest a strong discomfort with government use of the technology, the demand for private surveillance camera systems with FRT has continued to grow,[24] as has the use of this technology in everyday life. It is now common for people to open their phones with face verification or to automatically sort photos based on the people in the photos. Moreover, in regions where FRT has not faced as much controversy as in the U.S. or E.U.,[25] the technology is increasingly used by government authorities[26] and also for everyday verification purposes, such as payment[27] and entering establishments.[28] Outside of FRT, HCCV is increasingly common, with cameras using eye, face, or body detection for autofocus or for creating artificial bokeh effects (blurry background), robots using human/object detection to navigate real-world spaces, and CGI employing AI to create new fantastical scenes.

Thus, while privacy and bias concerns around FRT have manifested themselves in moratoriums on specific use cases in some jurisdictions, HCCV systems as a whole are unlikely to go away anytime soon. The focus of this Article is thus not on the line-drawing exercise of which HCCV systems should be banned vs. permitted but rather on the broader regulatory

[23] *In Charts: Facial Recognition Technology – and How Much Do We Trust It?*, FINANCIAL TIMES (May 16, 2021), https://www.ft.com/content/f6a9548a-a235-414e-b5e5-3e262e386722.

[24] *See, e.g.*, Lance Whitney, *Demand for Video Surveillance Cameras Expected to Skyrocket*, TECHREPUBLIC (July 24, 2020), https://www.techrepublic.com/article/demand-for-video-surveillance-cameras-expected-to-skyrocket/; Lauren Bridges, *Amazon's Ring is the Largest Civilian Surveillance Network the US Has Ever Seen*, THE GUARDIAN (May 18, 2021), https://www.theguardian.com/commentisfree/2021/may/18/amazon-ring-largest-civilian-surveillance-network-us; Edvardas Mikalauskas, *The Rise of the Private Surveillance Industry*, CYBERNEWS (Aug. 3, 2021), https://cybernews.com/privacy/the-rise-of-the-private-surveillance-industry/; *Global Video Surveillance Market Revenues to Exceed $24B by End of 2021*, CE PRO (Aug. 9, 2021), https://www.cepro.com/security/global-video-surveillance-market-revenues-exceed-24b-2021/.

[25] *See* Lea Steinacker et al., *Facial Recognition: A Cross-National Survey on Public Acceptance, Privacy, and Discrimination*, PROC. OF THE 37TH INTERNATIONAL CONF. ON MACHINE LEARNING, LAW & MACHINE LEARNING WORKSHOP (2020), https://arxiv.org/pdf/2008.07275.pdf. This is not to say there is no controversy around facial recognition in Asia. In fact, there is growing concern about biometric privacy in China. *See* Eva Dou, *China Built the World's Largest Facial Recognition System. Now, It's Getting Camera Shy*, WASH. POST (July 30, 2021), https://www.washingtonpost.com/world/facial-recognition-china-tech-data/2021/07/30/404c2e96-f049-11eb-81b2-9b7061a582d8_story.html; Sam Shead, *Chinese Residents Worry About Rise of Facial Recognition*, BBC (Dec. 5, 2019), https://www.bbc.com/news/technology-50674909; Stella Yifan Xie, *In China, Paying With Your Face Is Hard Sell*, WALL ST. J. (Sept. 20, 2020), https://www.wsj.com/articles/in-china-paying-with-your-face-is-hard-sell-11600597240. That said, the use of facial recognition is far more pervasive in China than in other countries. Paul Mozur, *Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras*, N.Y. TIMES (July 8, 2018), https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html

[26] *See, e.g.*, Aloysius Low, *In Singapore, Facial Recognition Is Getting Woven Into Everyday Life*, NBC NEWS (Oct. 12, 2020), https://www.nbcnews.com/tech/tech-news/singapore-facial-recognition-getting-woven-everyday-life-n1242945; Fresh Air, *Facial Recognition and Beyond: Journalist Ventures Inside China's "Surveillance State,"* NPR.ORG (Jan. 5, 2021), https://www.npr.org/2021/01/05/953515627/facial-recognition-and-beyond-journalist-ventures-inside-chinas-surveillance-sta.

[27] Richard Baimbridge, *Why Your Face Could Be Set to Replace Your Bank Card*, BBC (Jan. 24, 2021), https://www.bbc.com/news/business-55748964, *Four Japan Firms to Tie Up in Facial Recognition for Payment, Japan Times* (Aug. 2, 2021), https://www.japantimes.co.jp/news/2021/08/02/business/tech/facial-recognition-tie-up/.

[28] Kosuke Shimizu et al., *Japan in Race with China for Facial Recognition Supremacy*, NIKKEI ASIA (Dec. 20, 2019), https://asia.nikkei.com/Business/Business-trends/Japan-in-race-with-China-for-facial-recognition-supremacy; Chris Gallagher, *Masks No Obstacle for New NEC Facial Recognition System*, REUTERS (Jan. 6, 2021), https://www.reuters.com/article/us-health-coronavirus-japan-facial-recog/masks-no-obstacle-for-new-nec-facial-recognition-system-idUSKBN29C0JZ.

questions and tensions that need to be first addressed to judge what would constitute appropriate use cases. Policymakers and AI developers must assess these privacy, bias, and other ethical concerns in contexts where the technology is in use now or might be in use in the future. Unfortunately, as this Article will discuss, addressing both privacy and bias concerns in practice can be quite difficult—not only because each set of concerns entails addressing many sociotechnical challenges, but also because privacy and bias mitigation are often in tension in the algorithmic context. The goal of this Article is not to advocate for the increased or decreased usage of HCCV technologies, but rather to characterize this tension between privacy and bias mitigation efforts and to propose potential paths forward that respect both.

Section I lays out important definitions used throughout the Article to facilitate more nuanced discourse about HCCV. Section II discusses the importance of considering the harms of being "mis-seen" in a world where HCCV is increasingly pervasive. Section III explains what makes the HCCV context unique in terms of the privacy and fairness tensions it raises. Section IV discusses current challenges to mitigating algorithmic bias in HCCV, focusing particularly on the difficulties with collecting large, diverse datasets with informed consent. Section V discusses relevant privacy laws in the U.S. and E.U. Section VI elaborates on the harms associated with being "seen" vs. "mis-seen." Section VII evaluates potential solutions for better balancing protections against being "seen" vs. "mis-seen."

I: DEFINITIONS

Resolving the tensions between two very important ethical desiderata—privacy and fairness—requires a nuanced approach. The discourse around HCCV rarely distinguishes between the many different types of technologies implicated, which differ widely in terms of their potential societal harms. This Section will seek to provide the vocabulary needed for greater nuance by clarifying technologies and concepts that are often conflated.

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.

In characterizing the privacy concerns around HCCV, this Article will focus on biometric information privacy risks given that this is the primary area where there has been regulation and litigation around the images used to develop HCCV systems. Of course, even if images featuring biometric information are *not* involved, there might still be legal issues with copyright, and privacy risks might remain if photos taken inside people's homes are used. Note that certain medical use cases of computer vision (e.g., melanoma detection) might or might not count as HCCV for the purposes of this Article's discussion depending on whether the images used to develop the AI include faces or hands of the individuals.[31] If the AI is developed *without* images featuring biometric information subject to privacy protections, then there is no tension between existing privacy laws and algorithmic fairness.

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

Object detection and recognition is another major category of computer vision tasks. For example, a traffic camera might learn to detect and count the number of cars at an intersection. While such tasks might seem unrelated to HCCV, these technologies are often trained on images featuring humans. For example, in the research community, the COCO dataset is one of the most commonly used datasets for object-related tasks.[37] This dataset features around two hundred thousand images with humans and objects labelled. Using images with humans can be helpful given that, in the real-world, we are often interested in detecting objects that humans are interacting with. Training an object recognition model exclusively on images without humans might make it more difficult for the model to perform well in real-world contexts.[38] In addition, often the goal is not object recognition in isolation but rather to combine human detection with object recognition. A store might want a system to detect if someone is carrying a weapon, for example.

Thus, although FRT has animated much of the popular discourse around risks with emerging AI systems, HCCV more comprehensively encompasses a diverse array of computer vision models whose development is subject to biometric information privacy laws, raising difficult questions around privacy and fairness.

In terms of more general terms, throughout this Article, I will use the terms "model" and "algorithm" largely interchangeably to refer to the machine learning model being used. "Algorithm" is technically a more expansive term, referring to a "set of rules a machine (and especially a computer) follows to achieve a particular goal."[39] "Model" is a more precise term, but "algorithm" is more commonly used in colloquial discussions about AI. In general, I will use colloquial terms when referencing popular discourse that uses those terms. I will refer to "HCCV systems" to describe more expansively a particular product or service that includes an HCCV model.

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.

The tension between not wanting to be "seen" or "mis-seen" resembles the tension between "visibility" and "hypervisibility" in that it centers on the challenges of increasing representation in a way that is not harmful to the marginalized individuals represented.[41] These dichotomies are distinct, however, in that having one's images included in a dataset to train or test an HCCV model does not necessarily have the implications of hypervisibility in terms of heightened scrutiny or surveillance. Scrutiny or surveillance by HCCV comes at the point of deployment rather than development. Being 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. Moreover, whereas hypervisibility is often associated with tokenization and distorted visibility – the tendency to provide visibility disproportionately to negative representations of marginalized individuals[42]—algorithmic bias mitigation efforts in HCCV often revolve precisely around preventing models from learning stereotyped representations of people (e.g., that only men play outdoor sports), as will be discussed further in Sections III and VI.B. Thus, while the visibility vs. hypervisibility tension is highly relevant to discourse around whether, how, where, and for whom HCCV systems should be deployed, it is only indirectly related to the tension between being "seen" vs. "mis-seen."

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. While being “seen” by an HCCV system without informed consent is considered under privacy law to be a harm in and of itself, being “mis-seen” is only considered to be a harm if it leads to a separate legally cognizable harm. For example, when Robert Williams sued the Detroit police department after a faulty facial recognition match, he brought his action under the Fourth Amendment right to be free of unlawful seizures and the Elliot-Larsen Civil Rights Act, which protects against government entities “deny[ing] an individual full and equal enjoyment of public services on the basis of race.[44]

One could argue that this distinction is reasonable—that the harms of being mis-seen are already appropriately accounted for through existing anti-discrimination laws and other laws. Anti-discrimination law, however, primarily applies in specific, comparatively high-stakes contexts, like employment,[45] housing,[46] finance,[47] and the public sector.[48] While the limited domains of antidiscrimination law might be reasonable in the context of discrimination by human actors, algorithmic discrimination raises additional concerns. 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.

Of course, the scenario I described is *extreme* in that it is unlikely that most commercial AI systems would perform *so* consistently poorly for individuals in minority groups—occasional poor performance is much more likely. Nonetheless, currently the primary forces preventing such poor performance are the competitiveness of the market and the desire of companies to produce high-performing products. Such incentives might be insufficient if the system works very well for most people; those in the minority group might be seen as out-of-distribution edge cases that do not need to be specifically addressed. There is no legal protection for the individual in this scenario.

III: Why Computer Vision?

There is a general tension between privacy laws and algorithmic bias detection and mitigation efforts in that such efforts typically involve the use of protected class or sensitive attribute data (or proxies for such data). Prior works have discussed this empirically through interview methods[49] and in analyses of relevant antidiscrimination law prohibitions on the usage of such data.[50] This paper focuses on the context of bias mitigation in computer vision, given that here the concern is not simply with protected attributes or sensitive data, but rather with *all* the data used in developing such models. In the tabular or language data contexts, stripping the

[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.

dataset of personally identifiable information (PII) can significantly mitigate the privacy risks.[51] In contrast, for HCCV, even if the developer strips all the metadata, the face or body images themselves can constitute PII under certain laws.[52] Moreover, developing HCCV usually involves the processing of biometric information, which is subject to privacy protections, as will be discussed further in Section V. Section VII.D will discuss in greater depth the potential utility of face blurring and other image anonymization techniques, but, in short, depending on the type of HCCV being developed, such techniques cannot guarantee that the person cannot be identified while still preserving the ability to train an accurate model.

Not only are the privacy concerns stronger in the HCCV context, but also the need for wide-ranging data collection efforts is greater. While a simple logistic regression model with tabular data can be trained on thousands of instances, HCCV requires millions of images to train a base model that can do basic detection and recognition tasks.[53] Moreover, while dataset diversity is important in all contexts, bias in computer vision is particularly strongly connected with a lack of sufficient dataset diversity. The primary technical solution to addressing the harms discussed below in Section VI is to collect larger, more diverse, and more balanced datasets.[54] While there are other bias mitigation solutions that have been explored in the computer vision literature, these methods either rely on the generation of synthetic images to create a more diverse, balanced dataset[55] or address bias only indirectly.[56]

In the tabular data context, collecting data on more diverse individuals rarely solves issues of algorithmic bias. For example, in criminal justice data in the US, there is evidence that Black individuals have faced higher rates of arrest for drug-related crimes despite similar rates of

[51] For tabular data, removing unique identifiers, employing differential privacy techniques, and limiting the number of and types of features are all techniques that can significantly reduce privacy concerns. Similarly, for language data, stripping the dataset of identifiers and contextual information, and limiting the amount of data from individual conversations can significantly reduce the ability to tie specific language data to individuals.

[52] Erika McCallister et al., *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, NATIONAL INSTITUTE OF STANDARDS & TECHNOLOGY SPECIAL PUBLICATION 800-122, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf

[53] Note, however, that it is fine for some subset of these images to be synthesized. Iacopo Masi et al., *Do We Really Need to Collection Millions of Faces for Effective Facial Recognition?*, PROC. OF EURO. CONF. ON COMPUTER VISION (2016), https://arxiv.org/pdf/1603.07057.pdf.

[54] Angelina Wang et al., *REVISE: A Tool for Measuring & Mitigating Bias in Visual Datasets*, PROC. OF EURO. CONF. ON COMPUTER VISION (2020), https://arxiv.org/pdf/2004.07999.pdf; Buolamwini & Gebru, *supra* note 6; Patrick Grother et al., *Facial Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, NISTIR 8280, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf.

[55] *See, e.g.*, G. Balakrishnan et al., *Towards Causal Benchmarking of Bias in Face Analysis Algorithms*, PROC. OF 2020 EURO. CONF. ON COMPUTER VISION, https://www.semanticscholar.org/paper/Towards-causal-benchmarking-of-bias-in-face-Balakrishnan-Xiong/34f0db15fd2cd3aaa94f852a43b9e93956bb373a; Wood et al., *Fake It Till You Make It: Facial Analysis in the Wile Using Synthetic Data Alone*, Proc. of 2021 INTERNATIONAL CONF. ON COMPUTER VISION, https://microsoft.github.io/FaceSynthetics/.

[56] These types of methods work by focusing the model's attention on relevant features instead of irrelevant ones correlated with demographics. *See, e.g.*, Lisa Anne Hendricks et al., *Women Also Snowboard: Overcoming Bias in Captioning Models*, PROC. OF 2018 EUROPEAN CONFERENCE ON COMPUTER VISION, https://openaccess.thecvf.com/content_ECCV_2018/papers/Lisa_Anne_Hendricks_Women_also_Snowboard_ECCV_2018_paper.pdf; Wang et al., *Toward Fairness in Visual Recognition: Effective Strategies for Bias Mitigation*, PROC. OF THE CONFERENCE ON COMPUTER VISION & PATTERN RECOGNITION (2020), https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_Towards_Fairness_in_Visual_Recognition_Effective_Strategies_for_Bias_Mitigation_CVPR_2020_paper.pdf.

drug offenses.[57] Algorithmic risk assessment tools designed to predict recidivism thus can improperly learn to associate features correlated with being Black with higher rates of recidivism. The solution to such problems of biased historical data is *not* to gather more arrest data on Black individuals but rather to attempt to measure the impact biases have had on leading to higher arrest rates and counteract those biases in the data (e.g., through algorithmic rebalancing across groups[58] or trying to find less biased features for predicting criminal offense rather than arrest[59]).

Algorithmic bias in the HCCV context generally boils down to two problems: lack of representation[60] and spurious correlations.[61] The former refers to the lack of sufficient images of particular subgroups in a training dataset. This source of bias is also present in human facial recognition, where studies have shown that people have a harder time recognizing people of other races.[62] Psychological research has also shown that the ability of humans to recognize faces of people of other races improves with more contact with people of other races when they are growing up.[63] Similar to humans, facial recognition models also exhibit an "other-race effect," whereby algorithms developed in Western countries perform better for Caucasian faces and algorithms developed in East Asian countries perform better for East Asian faces.[64] If you think of the machine learning developer as the parent to the HCCV system, a parent who wants to ensure their "child" is able to equally recognize people of all different races, it is easy to understand the urgency for collecting a diverse set of faces for training the algorithm.

The other fundamental source of bias is spurious correlations, meaning that the training data contain misleading patterns, often due to societal biases.[65] For example, researchers have found that gender classification models are more likely to *incorrectly* predict that an individual in

[57] Sharad Goel et al., *Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy*, 10.1 ANNALS OF APPLIED STATISTICS 365 (2016), https://5harad.com/papers/stop-and-frisk.pdf; Kristian Lum & William Isaac, *To Predict and Serve?*, 13.5 SIGNIFICANCE 14 (2016), https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x; Emma Pierson et al, *A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States*, 4 NATURE HUMAN BEHAVIOUR 736 (2020), https://www.nature.com/articles/s41562-020-0858-1.

[58] SOLON BAROCAS ET AL., FAIRNESS & MACHINE LEARNING: LIMITATIONS & OPPORTUNITIES, https://fairmlbook.org/pdf/fairmlbook.pdf. Such methods can be pre-processing, in-processing, or post-processing methods.

[59] Riccardo Fogliato et al., *On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes*, PROC. OF 2021 AAAI/ACM CONF. ON AI, ETHICS, AND SOCIETY, https://arxiv.org/pdf/2105.04953.pdf.

[60] *See, e.g.,* Buolamwini & Gebru, supra note 6; DeVries et al., *Does Object Recognition Work for Everyone?*, PROC. OF THE IEEE/CVF CONF. ON COMPUTER VISION & PATTERN RECOGNITION WORKSHOPS 52 (2019), https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html.

[61] *See, e.g.,* Robert Geirhos et al., *Shortcut Learning in Deep Neural Networks*, 2 NATURE MACHINE INTELLIGENCE 665 (2020), https://www.nature.com/articles/s42256-020-00257-z; Hendricks, *supra* note 56.

[62] *See* Agata Blaszczak-Boxe, *Some People Suffer from Face Blindness for Other Races*, SCIENTIFIC AMERICAN (May 1, 2017), https://www.scientificamerican.com/article/some-people-suffer-from-face-blindness-for-other-races/

[63] Note, however, that this improvement only occurs up to the age of 12—greater social contact with people of other races in adulthood has little effect. Elinor McKone, *A Critical Period for Faces: Other-Race Face Recognition Is Improved by Childhood But Not Adult Social Contact*, SCIENTIFIC REPORTS (2019), https://www.nature.com/articles/s41598-019-49202-0.

[64] P. Jonathon Phillips et al., *An Other-Race Effect for Face Recognition Algorithms*, 8.2 ACM TRANSACTIONS ON APPLIED PERCEPTION 1 (2011), https://dl.acm.org/doi/10.1145/1870076.1870082.

[65] This is related to the problem of short-cut learning. *See* Geirhos, *supra* note 61.

a photo is female if the background is indoors and conversely for outdoor images,[66] perpetuating long-standing stereotypes of women inhabiting domestic spheres and men public spheres. Even though the background of an image should be irrelevant for discerning whether an individual is male or female, models learn to rely on such irrelevant factors when the training data disproportionately features images of females indoors and males outdoors. In another example, when researchers tried to develop a model to synthesize images where a person's hair was lengthened or shortened, the model learned to also feminize the person's features when lengthening the hair, suggesting a conflation between long hair and feminine facial features.[67] Thus, it is important to develop training datasets that are well-balanced and avoid spurious correlations. For example, the proportion of women indoors vs. outdoors should roughly match the proportion of men indoors vs. outdoors. Of course, it is impossible to account for all possible spurious correlations, so researchers typically focus on ones that are related to pernicious societal stereotypes. Collecting a balanced dataset in an unbalanced world, however, can be difficult in practice, as the next Section will discuss.

In addition to bias mitigation, the prosocial normative motivation for collecting large, diverse datasets in computer vision is particularly strong given that doing so can directly improve the accuracy of the model.[68] Outside of the HCCV context, bias mitigation itself can be a source of controversy.[69] For example, scholars have explored the ways in which many of the dominant approaches to bias mitigation in the tabular data ML context may actually violate antidiscrimination law because of their reliance on quotas, different thresholds, or other forms of rebalancing across the protected attribute.[70] In contrast, ensuring that your model recognizes people based on their facial features and not based on their clothing or the image background is important not only for reducing bias but also for increasing accuracy across a wider set of deployment contexts.[71]

[66] *See* Wang *supra* note 54.

[67] 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.

[68] Note that this is specifically true for verification and recognition tasks. For classification tasks, there can still be a trade-off between fairness and accuracy due to biases or stereotypes reflected in the classifications. Pinar Barlas et al., *To "See" Is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off*, PROC. OF ACM HUMAN-COMPUTER INTERACTION CSCW (2020), https://pure.mpg.de/rest/items/item_3286836/component/file_3286837/content.

[69] Nicol Turner Lee et al., *Algorithmic Bias Detection & Mitigation: Best Practices & Policies to Reduce Consumer Harms*, BROOKINGS INSTITUTE (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ (discussion about fairness-accuracy trade-off)

[70] Such approaches may be considered legally suspect affirmative action or reverse discrimination. Daniel E. Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, U.CHI. L. REV. ONLINE (Oct. 30, 2020), https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang/; Xiang, *supra* note 42; Jason R. Bent, *Is Algorithmic Affirmative Action Legal?,* 108 GEORGETOWN L. J. 803 (2020), https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2020/04/Is-Algorithmic-Affirmative-Action-Legal.pdf

[71] The reason for this distinction is that in the tabular data context, one of the fundamental problems is a lack of ground truth. For example, there are significant concerns around the use of ML recidivism risk assessment tools in the criminal justice context given that arrest is used as a proxy for re-offense. *See* Fogliato, *supra* note 59. Given the evidence that minority communities face disproportionately higher levels of policing and arrest, including wrongful arrest, arrest data itself is seen as suspect. The problem with having unreliable or biased arrest data, however, is that correcting for this bias requires some notion of what the data would look like in a fairer world. One extreme notion of this is the demographic parity fairness metric, which defines bias as disproportionate outcomes (in this case,

Thus, there are many aspects of computer vision that make the tensions between privacy and fairness particularly salient and difficult to untangle. That said, many of the insights from this Article are not unique to computer vision. If we can reconcile the tensions between privacy and fairness in HCCV, we might be able to apply analogous solutions to other forms of AI.

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.

Images of celebrities have especially assisted with the advancement of research into facial recognition and verification systems since such datasets include many images of the same person, at different times, angles, and settings. Such datasets, however, raise issues around consent and also biases introduced by only training algorithms to recognize celebrities, whose features are not representative of the general population.[78]

The use of Flickr images has been very pervasive in the computer vision community due to the uniquely diverse and candid nature of these images, which often include a wide variety of people and objects in each image. In fact, researchers who constructed large public datasets using Flickr images were often motivated to use Flickr to address the issues of bias that plague other datasets.[79] Flickr-based datasets feature photos of non-celebrities[80] from amateur photographers,[81] yielding a large amount of diversity.[82] Recently, however, there have been many lawsuits leveraging Illinois's Biometric Information Privacy Act (BIPA) against companies using such datasets since the individuals in the Flickr images did not consent to

[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.

having their photos used to train facial recognition algorithms.[83] Informed consent is thus a key consideration when collecting or using large image datasets for developing HCCV.

A. Why is Collecting Images with Informed Consent So Difficult?

The most obvious and reliable way to address the privacy concerns around collecting images for training HCCV systems is to obtain informed consent from the individuals in the photos. This is much easier said than done, however, given the need for millions of images with diverse subjects and conditions.

Social media or cloud service companies can collect large image datasets through products that incentivize individuals to upload photos. This is not to say they have always appropriately obtained informed consent, however. For example, Facebook recently reached a landmark settlement of $650 million in a BIPA case challenging their use of users' face images for training their face-tagging algorithm.[84] That said, for companies with a business model where individuals upload large numbers of diverse photos, the first step to solving the informed consent issue is comparatively straight-forward: Facebook now asks users whether they consent to having their images used for facial recognition.[85]

This is not to say that the problem is completely solved—users upload many photos of people other than themselves. Even if the user has consented to their photos being used for facial recognition, obtaining the consent of the individuals in the photos is still necessary. Even if the individuals have social media accounts where they have provided approval on their end, it is unclear how the social media platform can know whether the individuals in the photo has given consent without first attempting to recognize the individuals. Moreover, depending on the company's privacy policy, the images collected through the platform might or might not be eligible for use in developing HCCV.

For academic researchers, public sector entities, or companies without business models that incentivize organic data collection, the need to collect large, diverse datasets with informed consent poses additional difficulties. Companies can buy images from vendors that work with crowd workers who upload images of themselves to the platform in return for payment, but it is difficult to (i) obtain enough data and (ii) obtain sufficiently diverse and candid data. While social media companies do not have to pay users to upload thousands of pictures of themselves and their friends, a company using a vendor to collect images must pay for each image. To

[83] Olivia Solon, *Facial Recognition's "Dirty Little Secret": Millions of Online Photos Scraped Without Consent*, NBC (Mar. 17, 2019),https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921; Sara Merken, *IBM Can't Shake Facial Recognition Suit But Dodges Some Claims*, Reuters (Sept. 16, 2020), https://www.reuters.com/article/dataprivacy-ibm-biometrics/ibm-cant-shake-facial-recognition-suit-but-dodges-some-claims-idUSL1N2GD2JP; Taylor Hatmaker, *Lawsuits Allege Microsoft, Amazon & Google Violated Illinois Facial Recognition Privacy Law*, TechCrunch (July 15, 2020), https://techcrunch.com/2020/07/15/facial-recognition-lawsuit-vance-janecyk-bipa/

[84] *In re Facebook Biometric Information Privacy Litigation*, (N.D. Cal. 2021) (ORDER RE FINAL APPROVAL, ATTORNEYS' FEES AND COSTS, AND INCENTIVE AWARDS. Signed by Judge James Donato on 2/26/2021. (jdlc2S, COURT STAFF) (Filed on 2/26/2021)), https://www.govinfo.gov/app/details/USCOURTS-cand-3_15-cv-03747/USCOURTS-cand-3_15-cv-03747-16.

[85] *What is the Face Recognition Setting on Facebook and How Does It Work?*, Facebook Help Center (last accessed Jan. 21, 2022), https://www.facebook.com/help/122175507864081.

collect the millions of images needed to train from scratch a computer vision model with good performance, large funds are needed.

I emphasize this distinction between the challenges faced by companies with platforms where people upload images freely versus other companies because this creates competition concerns in addition to the privacy and bias concerns discussed elsewhere in this Article. There are very few companies that have the advantage of a large, global, diverse usership willing to upload billions of images for free. There are far more companies, academics, and public sector entities that either operate or seek to operate in the HCCV space.

In addition, when crowd workers are paid to upload images of themselves based on particular specifications (e.g., one front-facing photo, one side-facing photo, one photo indoors, one photo outdoors, one photo holding an object, one photo sitting/standing/running, one photo occluded by an object), the photos generally look staged.[86] In computer vision, there is a term "in the wild," which refers to "unconstrained" images that appear to be taken in a wide variety of everyday scenarios—similar to the contexts a deployed HCCV system would be working with.[87] When buying photos from crowd workers, however, it can be difficult to collect large numbers of unconstrained images. It is also difficult to ensure such photos meet detailed diversity specifications (e.g., gender, age, ethnicity, nationality, geography, scene, illumination, occlusion, camera type, camera angle), verify that all the individuals in the photos have consented to the use of the photo for training AI, and verify that the photographer has relinquished their copyright.

These challenges create a number of performance and bias concerns. First, an HCCV model trained on very staged selfies might struggle to perform in the real world, where there might be multiple people in an image, the lighting conditions might be more varied, the people might be smaller and blurrier, or the people might have a wider variety of poses, expressions, or occlusions (e.g., hats, masks, or sunglasses).[88] Moreover, if the dataset features images from only one country— often the case given the need for the crowd workers to sign a consent form based on the laws of their jurisdiction—that can exacerbate issues of bias in the dataset. Not only might there be insufficient demographic diversity, but also the backgrounds and objects in the photos might only reflect country-specific contexts. For example, research has shown that object recognition models trained predominantly on U.S. data struggle to accurately recognize common objects like soap and cooking equipment in developing country contexts.[89]

Moving beyond the necessity to collect large numbers of images, the need to collect a diverse, well-balanced dataset with minimal spurious correlations creates additional challenges. First, there is the challenge of defining what sufficient diversity would look like. Relevant dimensions of diversity from the computer vision literature include demographics (perceived gender, age, and ethnicity), hairstyles, clothing, lighting conditions, background, pose, and

[86] In the early days of developing computer vision datasets, researchers did stage the photos they collected, hiring actors and photographers, and manually designing the set-up. Inioluwa Deborah Raji & Genevieve Fried, *About Face: A Survey of Facial Recognition Evaluation*, ARXIV (2021) https://arxiv.org/pdf/2102.00813.pdf. This was a very labor-intensive and expensive process, so early datasets were quite small. The need for informed consent, however, raises the question of how we can adapt these more manual ways of collecting images to suit the needs of contemporary computer vision development.

[87] Gary B. Huang et al., *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, http://vis-www.cs.umass.edu/papers/lfw.pdf.

[88] This issue is known as "domain shift." *See, e.g.,* Daniel E. Ho et al., *Evaluating Facial Recognition Technology: A Protocol for Performance Assessment in New Domains*, 98 DENVER L. REV. 753(2021).

[89] *See* DeVries et al., *supra* note 60.

camera type.[90] Avoiding spurious correlations would mean ensuring that no unrelated attributes are inadvertently correlated. For example, if in the training dataset most images of computers feature men instead of women, then the model might learn to identify computers based on whether there is a man vs. woman in the image. In addition, determining the relevant subcategories within each group is a challenging task that AI developers are not necessarily the best equipped to determine. For example, how many gender or ethnicity categories should we ensure diversity along? Ten? Twenty? One hundred?

Even after these sociological questions are answered about the "ideal" taxonomy and distribution for the dataset, there is the challenge of fulfilling these specifications. When issues of bias are discovered in the human image dataset or in models trained on it, it can be difficult to augment the dataset to address these issues. For example, if a developer realizes that their model does not perform well for Native American individuals due to their training set not having any images of Native Americans, a natural solution would be to seek out images of Native Americans. Conducting that type of targeted recruitment can be very difficult. Especially when collecting data from historically marginalized communities, it is important to ensure that the data collection process is not exploitative and does not fall into the trap of predatory inclusion.[91]

Moreover, uncovering bias in the first place can be difficult since existing publicly available datasets typically do not include people's self-reported demographics, so researchers or developers who want to ensure dataset diversity take measures to guess or estimate the demographics. Datasets with celebrities sometimes have web-scraped data on nationality.[92] When that information is not available, common methods include having annotators look at the photos and guess people's demographics,[93] using skin tone or other features as a proxy for race,[94] or using automated race classifiers.[95] While it would be much more ideal to collect the self-reported demographics of the image subjects, collecting demographic data can present additional privacy concerns, as will be discussed further in Section V. Without such data, however, even doing a preliminary check to see if the dataset is diverse or if the model performs well for different demographic groups is difficult.[96]

[90] *See* Margaret Mitchell et al., *supra* note 1.

[91] "Predatory inclusion refers to a process whereby members of a marginalized group are provided with access to a good, service, or opportunity from which they have historically been excluded but under conditions that jeopardize the benefits of access." Louise Seamster et al., *Predatory Inclusion & Educational Debt: Rethinking the Racial Wealth Gap*, 4.3 SOCIAL CURRENTS 199 (2017), https://journals.sagepub.com/doi/pdf/10.1177/2329496516686620. The literature on this topic typically focuses on practices in the financial and educational domains.

[92] Mei Wang et al., *Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network*, PROC. OF INTERNATIONAL CONF. ON COMPUTER VISION (2019), https://openaccess.thecvf.com/content_ICCV_2019/papers/Wang_Racial_Faces_in_the_Wild_Reducing_Racial_Bias_by_Information_ICCV_2019_paper.pdf; Jungseock Joo & Kimmo Karkkainen, *Gender Slopes: Counterfactual Fairness for Computer Vision Models by Attribute Manipulation*, 2ND ACM INTERNATIONAL WORKSHOP ON FAIRNESS, ACCOUNTABILITY, TRANSPARENCY AND ETHICS IN MULTIMEDIA (2020). https://dl.acm.org/doi/pdf/10.1145/3422841.3423533.

[93] *See* Karkkainen & Joo, *supra* note 16.

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

[95] *See* Wang et al., *supra* note 92.

[96] Without such data, companies often rely on proxy variables. *See supra* note 49. For example, skin tone might be used as a proxy for race, or long hair as a proxy for gender. There are many downsides, however, to using such proxies. *See* Buolamwini & Gebru, *supra* note 6; (discussing shortcomings of using the Fitzpatrick skin tone scale as a proxy for race); Xiang, *supra* note 50 (discussing unintended consequences of using proxy variables for bias mitigation). https://arxiv.org/pdf/1812.00099.pdf (differences in performance are not due to skin tone)

Given all these data collection challenges, the computer vision research community is divided on how important informed consent should be for image datasets.[97] More than half of the respondents to a survey conducted by Nature did not think it was necessary to obtain informed consent from individuals before using their face images. Even researchers who believed in the importance of informed consent stated they would still use datasets that do not have appropriate informed consent. It was difficult for the researchers to see how they could conduct computer vision research and train accurate models otherwise.

Overall, the challenge of assembling large, diverse, and well-balanced human image datasets is a topic that requires more public awareness. When an AI system fails to work well for individuals from marginalized backgrounds, this often becomes a source of public outrage and used as evidence that developers do not care about such individuals. Even in situations where people do care deeply about making their products work well for everyone, however, collecting sufficiently large and diverse datasets is very difficult and runs directly into many privacy and other ethical challenges.

V: PRIVACY LAWS

There are two separate areas of privacy law that are relevant to the context of mitigating bias in computer vision systems: (i) the collection of biometric information and (ii) the collection of sensitive attributes. The former is generally relevant for the development of any HCCV system, but raises particular concerns in the context of attempting to collect more diverse datasets, focusing on marginalized groups. The latter is important for bias detection and mitigation; it is difficult to evaluate dataset diversity or performance across demographic groups without demographic information.

Some of the most salient privacy laws in the first category are U.S. state laws like BIPA[98] and the California Consumer Privacy Act (CCPA)[99] that regulate the processing of biometric information[100] and GDPR's restrictions around the processing of personally identifiable information (PII).[101] Biometric information in this context can be seen as a particularly sensitive subset of PII. BIPA, for example, regulates the collection, storage, and use of biometric identifiers and biometric information. Biometric identifiers include "scan[s] of hand or face geometry," which has been interpreted by courts to include both facial landmarks and facial templates, which are extracted for any computer vision task involving detecting, verifying, recognizing, or classifying faces. CCPA's protections of biometric information more expansively include face images themselves (not just biometric information extracted from them), images of hands or palms, and gait patterns.

[97] Richard Van Noorden, *The Ethical Questions That Haunt Facial Recognnition Research*, NATURE 587, 354-358 (2020), https://www.nature.com/articles/d41586-020-03187-3.

[98] 740 Ill. Comp. Stat. Ann. 14/15 [hereinafter BIPA].

[99] CAL. CIVIL CODE §1798.140(c) [hereinafter CCPA].

[100] Texas and Washington have also passed biometric information privacy laws. Tex. Bus. & Com. Code §503.001; Wash. Rev. Code Ann. §19.375.020.

[101] European Commission Regulation 2016/679, 2016 O.J. (L119) [hereinafter GDPR].

While each jurisdiction's biometric information privacy laws differ slightly in scope, they all seek to restrict the collection, storage, and use of images/videos of faces or bodies (or landmarks/templates extracted from these images/videos) that *could* in turn be used to identify a person (actual use for identification is not required). The laws vary in terms of the rights they provide; some provide a right to request and receive disclosures about information that has been collected, a right to request that the information be deleted, a prohibition on denying goods or services for exercising privacy rights, or a prohibition on sale of or profit from the information. The key protection this Article will focus on, however, is the requirement of informed consent in order to collect biometric information. While the type of notice and consent required varies under different laws,[102] some form of informed consent is the one constant across the various laws, and, as discussed above in Section IV, creates significant challenges in the development of HCCV.

It is worth briefly mentioning, however, that the right to revoke consent under GDPR and CCPA also creates significant challenges for HCCV development. Even if a company goes through the steps of ensuring that they obtain informed consent and compensate individuals for their images, the fact that the data subjects might later revoke their consent means that companies must design systems to deal with such a possibility. There is a lack of regulatory guidance, however, around the implications of such a revocation—does it only affect future models? What about future models derived from current models that used the data subject's image in development? Can the company require the data subject to refund the fee they were paid?[103] In addition, enabling data subjects to revoke their consent ironically requires more data retention—if the images are completely stripped of any identifying information, then if an image subject requests that their images be deleted, it will be difficult to determine which images feature them.

Most of the U.S. privacy cases about image collection for HCCV center on BIPA. BIPA was passed in 2008, making it much older than other comparable state laws. Over the past several years, BIPA's private right of action has made it a powerful tool for privacy advocates to challenge tech company data practices. Although state statutes like BIPA in theory are narrow in their jurisdictional scope, in practice the difficulty of determining whether images in a dataset are from Illinois residents has vastly expanded the influence of BIPA.[104] Section VI will feature a more in-depth discussion about the specific harms these laws seek to prevent and how courts have interpreted them.

In the second category—laws protecting sensitive attribute data—we again have GDPR, which regulates the processing of special categories of personal data like race.[105] We also have some U.S. privacy laws and antidiscrimination laws, like the Equal Credit Opportunity Act,

[102] Some laws like BIPA require the data subject to provide written consent. Others like CCPA only require notice with the right to access and delete any personal information collected.

[103] EU law requires consent to be freely given, so contractual provisions requiring the refund of the fee could be construed as undermining the extent to which the consent is completely voluntary.

[104] Some companies have tried to sidestep BIPA and other state information privacy laws by asking individuals what state they are residents of before giving them access to a product. *See, e.g.,* Jeffrey Neuburger, *Google App Disables Art-Selfie Biometric Comparison Tool in Illinois & Texas*, PROSKAUER (Jan. 18, 2018), https://newmedialaw.proskauer.com/2018/01/18/google-app-disables-art-selfie-biometric-comparison-tool-in-illinois-and-texas/. Note, however, that Google did ask for consent from users of the app before processing their selfies.

[105] Article 9, GDPR, https://gdpr-info.eu/art-9-gdpr/.

which place additional restrictions on the collection or consideration of sensitive demographic data in specific domains.[106] In practice, these restrictions have ironically erected significant barriers to both private and public sector entities attempting to audit their algorithmic systems for bias.[107] An interview study[108] of algorithmic fairness practitioners found that overwhelmingly companies across the AI industry, both small and large, struggle to check their AI systems for bias, let alone take remedial measures to address bias. Despite the growth in AI ethics, responsible AI, and algorithmic fairness teams in tech companies, these teams face practical challenges when attempting to convince their colleagues to collect sensitive attribute data to conduct bias assessments.[109] Often legal and compliance teams shut down efforts to collect, share, or use such data. In light of existing privacy laws, this knee-jerk reaction is understandable, but in practice, it makes progress toward less-biased AI more challenging.

There is evidence that policymakers are increasingly cognizant of this challenge. The E.U. proposed AI regulation creates a carve-out for processing sensitive data for bias monitoring, detection, and correction for high-risk AI systems.[110] In addition, the UK's Information Commissioner's Office (ICO) has released guidance suggesting that such data can and *should* be collected for the purposes of bias mitigation, and recommends pursuing the public good exception in GDPR.[111] There is less clarity on the U.S. side, however, on how to balance privacy and bias mitigation. While there have been growing calls for audits of tech company algorithms,[112] there has been less policy discussion around ways to enable and guide the data collection needed for audits. More generally, there seems to be less recognition of the existence of this tension between existing U.S. privacy and antidiscrimination laws and the pushes for less-biased facial recognition systems.[113] In short, while there is growing recognition in certain jurisdictions that sensitive attribute data might be needed for bias detection and mitigation,

[106] Regulation B, § 1002.5(b), https://www.consumerfinance.gov/rules-policy/regulations/1002/5/#a-4-vi.

[107] Scholars have shown how using anonymized smart-phone-based mobility data to inform COVID-19 response strategies can perpetuate demographic biases. These biases can be difficult to detect due to the fact that such data is aggregated up from the individual level for privacy reasons. Amanda Coston et al., *Leveraging Administrative Data for Bias Audits: Assessing*
Disparate Coverage with Mobility Data for COVID-19 Policy, PROC. OF ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY (2021), https://dho.stanford.edu/wp-content/uploads/Coston_etal.pdf.

[108] *See*, Andrus et al., *supra* note 49. Clavell et al. also discuss the challenge in practice of balancing data minimization under GDPR and bias audits. Clavell et al., *Auditing Algorithms: On Lessons Learned & the Risks of Data Minimization*, ARXIV (2020), https://dl.acm.org/doi/pdf/10.1145/3375627.3375852.

[109] Cloe Bakalar et al., *Fairness on the Ground: Applying Algorithmic Fairness Approaches to Production Systems*, FACEBOOK AI (2021), https://ai.facebook.com/research/publications/applying-algorithmic-fairness-approaches-to-production-systems.

[110] "To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data . . ." E.U. Proposed AI Regulation, Title III, Chapter 2, Article 10.5, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[111] *Guidance on AI & Data Protection*, UK INFORMATION COMMISSIONER'S OFFICE (last accessed Jan. 21, 2022), https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/.

[112] Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PENN. L. REV. ONLINE 189 (2017), https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1212&context=penn_law_review_online; James Guszcza et al., *Why We Need to Audit Algorithms*, HARVARD BUSINESS REV. (Nov. 28, 2018), https://hbr.org/2018/11/why-we-need-to-audit-algorithms; Alex Engler, *Auditing Employment Algorithms for Discrimination*, BROOKINGS (Mar. 12, 2021), https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/.

[113] *See* Andrus et al., *supra* note 49.

collecting large and diverse datasets that comply with privacy laws remains an area with little regulatory guidance.

VI: HARMS OF BEING SEEN VS. MIS-SEEN

One of the core contributions of this Article is to identify and characterize the tension between protecting against the harm of being "seen" by HCCV systems versus the harm of being "mis-seen" by such systems. The former is the primary concern of privacy law, whereas the latter is the primary concern of the algorithmic fairness community. Since both are important ethical considerations, this Section will focus on breaking down the specific harms of being "seen" and "mis-seen" to better delineate the potential trade-offs involved.

A. HARMS OF BEING SEEN

Privacy law is notorious for the ambiguity around the specific harms it envisions.[114] In the seminal article "The Right to Privacy," which is credited for essentially creating the U.S. common law privacy right,[115] Warren and Brandeis discussed privacy as "the right to be let alone."[116] The authors compared privacy to "the right not to be assaulted or beaten, the right not to be imprisoned, the right not to be maliciously prosecuted, the right not to be defamed."[117] In contrast to the laws governing those rights, the authors conceived of privacy as protecting against mental suffering, rather than simply reputational damage (as under defamation law) or infringements upon property (as under intellectual property law).[118] They justified privacy protections as an extension of common law's "secur[ing] to each individual the right of determining to what extent his thoughts, sentiments, and emotions shall be communicated to others."[119]

Modern U.S. consumer data privacy law is rooted in tort law, contract law (when companies employ privacy policies), property law, Section 5 of the FTC Act (prohibiting "unfair or deceptive acts or practices in or affecting commerce"), the Privacy Act of 1974 (applying to federal agencies), sectoral federal statutory regulation, and state statutory regulation.[120] As discussed in Section V, most relevant to our discussion are state biometric privacy laws like Illinois's BIPA and California's CCPA. These laws are notable for going beyond the sectoral nature of federal privacy laws, providing protections for biometric information or personal data regardless of the context of collection or use. While the right to privacy writ large might be conceived of as a right to be left alone, biometric privacy laws specifically protect an

[114] Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 BOSTON U. L. REV. __ (2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3782222.
[115] Daniel J. Solove & Paul M. Schwartz, Information Privacy Law 10-11 (7th ed.).
[116] Samel D. Warren & Louis D. Brandeis, The Right to Privacy, 4 Harv. L. Rev. 193 (1890).
[117] *Id.*
[118] *Id.*
[119] *Id.*
[120] Solove & Schwartz, *supra* note 115 at 812-813.

individual's control over their data, making informed consent the key requirement for collection, storage, or use.[121]

What, however, are the specific harms that laws like BIPA protect against? Constitutional standing requires a concrete, particularized harm.[122] In *Spokeo v. Robins*, the Court found that mere violation of BIPA alone did not constitute a concrete injury sufficient for standing; the requirements of standing had to be independently met.[123] In a subsequent case, *Patel v. Facebook*, the Ninth Circuit applied a test for whether a statutory violation caused a concrete injury: "(1) whether the statutory provisions at issue were established to protect concrete interests (as opposed to purely procedural rights) and, if so, (2) whether the specific procedural violations alleged in this case actually harm, or present a material risk of harm to, such interests."[124] In the context of Facebook using facial recognition in its "Tag Suggestions" technology, the Ninth Circuit determined that, "the development of a face template using facial-recognition technology without consent (as alleged here) invades an individual's private affairs and concrete interests." The court thus found that BIPA protects an individual's concrete privacy interests, such that violations of the procedures in BIPA "actually harm or pose a material risk of harm to those privacy interests."

Key to the Ninth Circuit's decision was the idea that common law protects an individual's "control of information concerning his or her person," such that lack of control over one's biometric information, as protected against by BIPA, constituted a concrete harm. Lack of control over data is still quite a broadly construed harm, however. In the HCCV cases litigated thus far in the U.S., there has been no allegation of a disclosure of information leading to mental harm similar to the gossiping press that was decried by Warren and Brandeis as the impetus for a right to privacy. Instead, in evaluating Article III standing, U.S. courts have found compelling the concrete harms of identity theft, use of electricity/processing power, and surveillance risk.

The concern around identity theft is that as face verification is increasingly used for security purposes (e.g., opening phones, accessing buildings, payment), face templates extracted from images could be used to gain unauthorized access. For example, in *Patel v. Facebook*, the court expressed concern that the face templates collected by Facebook could be used to unlock cell phones.[125] It is unclear, however, that extracting face templates or landmarks from face images in order to develop HCCV increases the security risks beyond simply storing the images themselves. Existing methods to hack face verification systems rely on generating 3D renderings using *publicly available* images of the individual being hacked.[126] This can be done regardless of whether the images are also used to develop HCCV. Especially if the developer is using publicly

[121] BIPA, for example, prohibits private entities from collecting, capturing, purchasing, receiving through trade, or otherwise obtaining a person's or a customer's biometric identifier or biometric information without first (i) informing the individual that the biometric identifier or information is being collected or stored, (ii) informing the individual of the length of time of the collection, storage, or use, and (iii) receiving written release from the individual.

[122] Lujan v. Defenders of Wildlife, 504 U.S. 555, 112 S. Ct. 2130 (1992).

[123] Spokeo v. Robins, 578 US _ (2016).

[124] Patel v. Facebook Inc., 290 F. Supp. 3d 948 (N.D. Cal. 2018) (quoting Robins v. Spokeo, Inc., 867 F.3d 1108, 1113 (9th Cir. 2017)).

[125] Patel v. Facebook, Inc., 932 F.3d 1264 (9th Cir. 2019).

[126] Lily Hay Newman, *Hackers Trick Facial-Recognition Logins With Photos From Facebook (What Else?)*, WIRED (Aug. 19, 2016), https://www.wired.com/2016/08/hackers-trick-facial-recognition-logins-photos-facebook-thanks-zuck/.

available images to develop the HCCV system, it is unclear that doing so would increase the risk of identity theft for the image subjects. This is not to say that having images publicized is not a harm in and of itself. Indeed, the identity theft harm described above is a result of having images of yourself shared publicly. Rather, it is important to distinguish the harm of having an image made public versus having that public image used to train or evaluate HCCV.

In recent lawsuits against TikTok[127] and Apple,[128] courts also considered the economic harm of using electricity and processing power. This is relevant in cases where companies train their HCCV systems in the background of the plaintiff's phones, computers, and other devices. This harm, however, is not characteristic of all HCCV systems—it depends on whether the company is using the edge device (e.g., the individual's phone, computer, or camera) for training vs. collecting the images and then training the model on their own servers. Training on the device requires more processing power. From a technical perspective, however, it is more privacy-preserving for companies to use techniques like federated learning that involve doing part of the training on the individual's edge device instead of on the cloud, so such a focus on processing power could ironically disincentivize the use of such privacy-preserving techniques.

While courts have appreciated the economic nature of identity theft and electricity/processing power harms, the most significant potential harm animating privacy fears around HCCV is the specter of mass surveillance. The harms in this context are related to safety concerns (e.g., a stalker finding your location) and chilling effects (e.g., self-censorship). In contexts where there is significant distrust of the government or disagreement about the appropriateness of the laws being enforced, being surveilled is also considered a societal harm. For example, there has been significant criticism of government efforts to surveil journalists[129] or opposition party members.[130] Moreover, one of the most controversial uses of mass surveillance is the Chinese government's tracking of Uyghur minorities.[131]

Indeed, the potential for mass surveillance was a concrete harm the Ninth Circuit found to be compelling in *Patel v. Facebook*.[132] The court expressed concern that, "Once a face template of an individual is created, Facebook can use it to identify that individual in any of the other hundreds of millions of photos uploaded to Facebook each day, as well as determine when

[127] In re: TikTok Inc. Consumer Privacy Litigation, Case No. 1:20-cv-04699 (U.S. Dist. Court for the N.D. of IL.).

[128] Hazlitt v. Apple Inc., 500 F. Supp. 3d 738, 2020 U.S. Dist. LEXIS 210963, 2020 WL 6681374 (United States District Court for the Southern District of Illinois, November 12, 2020, Filed).

[129] Basma Humadi, *Mass Surveillance Threatens Reporting That Relies on Confidential Sources*, Reporters Committee for Freedom of the Press (Sept. 30, 2019), https://www.rcfp.org/nsa-mass-surveillance-against-journalist/; Comment to Review Group on Intelligence & Communicatoins Technologies Regarding the Effects of Mass Surveillance on the Practice of Journalism (2013), https://www.dni.gov/files/documents/RG/Effect%20of%20mass%20surveillance%20on%20journalism.pdf.

[130] *See, e.g.,* Ozgun E. Topak, *The Making of a Totalitarian Surveillance Machine: Surveillance in Turkey Under AKP Rule*, 15 Surveillance & Society 535 (2017), https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/download/6614/6466/; Pinkaew Laungaramsri, *Mass Surveillance & the Militarization of Cyberspace in Post-Coup Thailand*, 9 Austrian J. of South-East Asian Studies 195 (2016), https://www.tde-journal.org/index.php/aseas/article/download/2648/2260.

[131] Drew Harwell & Eva Dou, *Huawei Tested AI Software That Could Recognize Uighur Minorities & Alert Police, Report Says*, Wash. Post (Dec. 8, 2020), https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/; 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.

[132] Patel v. Facebook, Inc., 932 F.3d 1264 (9th Cir. 2019).

the individual was present at a specific location. . . . It seems likely that a face-mapped individual could be identified from a surveillance photo taken on the streets or in an office building."[133]

It is important to note, however, that these harms of surveillance are specific to the data processed in the deployment of HCCV systems rather than in the development of such systems. There is an important distinction between the different types of datasets associated with HCCV. The harms discussed in *Patel* are specific to having one's image included in a reference set of images, against which new images are compared. As a result, this harm is only relevant for HCCV that involves recognition (rather than detection or classification) and does not apply to the data used to train or test the model. The *Patel* case can thus be contrasted with *Flores v. Motorola Solutions Inc.*, in which mugshot images were used as a reference set in a facial recognition tool sold to law enforcement. In that case, the use of people's images directly implicated potential surveillance harms.

Of course, data collected for one purpose can in theory be repurposed for another, so it is important to evaluate the extent to which the data can be used as a reference set. Images of random unidentified individuals would not be very useful for a reference set. There needs to be some identifying information or meaning to the reference set in order for the model's inference to be meaningful—e.g., this is an image of *Tom*, or this is an image of someone you previously took pictures of, or this is an image of the suspect. The key inquiry then is how difficult it is, given the data available to the developer, to match an anonymous image with relevant identifying or contextual information. If the developer only has access to a public dataset of anonymous images, the risk of surveillance is relatively low compared to if the developer has extensive access to sensitive information about individuals. Thus, while courts have frequently expressed concerns about the potential for HCCV to facilitate surveillance, it is important to consider these additional nuances to gauge the actual risk given a specific fact pattern.

Moreover, even though mass surveillance is often characterized as a core privacy harm—a result of being "seen"—in practice, the threat of mass surveillance encapsulates both concerns around being "seen" and "mis-seen." For example, much of the criticism of law enforcement use of FRT centers not on the harms of being "seen" but rather the harms of being "mis-seen." While FRT are constantly improving in accuracy,[134] the potential for higher rates of mis-recognition of women and minorities remains. Although humans also do not have perfect facial recognition accuracy—indeed, eyewitness testimony is notoriously flawed and manipulable[135]—critics of FRT have argued that there are currently insufficient safeguards in place to ensure that the technology is used appropriately and does not further compound existing trends of over-policing of minority communities.

When considering how to regulate the potential for HCCV to facilitate mass surveillance, a couple factors are important to consider. First, as emphasized previously, the harm of mass surveillance is tied specifically with the breadth of deployment of HCCV rather than the breadth of the data used to develop it. In this sense, it can be possible to mitigate the risk of people being "mis-seen" by biased HCCV without increasing the deployment of HCCV. In addition, while preventing mass surveillance has been a major motivation for strict privacy laws around

[133] Patel v. Facebook, Inc., 932 F.3d 1264, 1273 (9th Cir. 2019).

[134] Grother et al., *Face Recognition Vendor Test (FRVT) Part 2: Identification*, NISTIR 8271 DRAFT SUPPLEMENT (2021), https://pages.nist.gov/frvt/reports/1N/frvt_1N_report.pdf.

[135] Gary L. Wells & Elizabeth A. Olson, *Eyewitness Testimony*, 54 ANNUAL REV. PSYCH. 277 (2003), https://capitalpunishmentincontext.org/files/resources/innocence/annual_review_2003.pdf

processing biometric information, not all forms of HCCV facilitate mass surveillance. Face/body/object detection or classification do not directly enable mass surveillance since they do not involve identifying individuals. Moreover, whether recognition technologies enable mass surveillance depends on the degree to which the data on face/body matches is shared. If FRT is used only locally on your phone to sort your photos, and the matches are not shared with the company or anyone else, then such technology arguably does not enable surveillance. These and other nuances will be discussed further in Section VII, which proposes possible solutions for minimizing both the harms of being "seen" and "mis-seen."

This is not to dismiss the concerns that individuals might have with their images being used to develop HCCV. Indeed, zooming out from the specific harms considered by the courts, perhaps the most direct harm associated with being "seen" is having your biometric information used to create a technology that you do not support. This specific harm, however, and whether it is, by itself, legally cognizable, is underexplored. In their taxonomy of privacy harms, Citron & Solove discuss "lack of control" as one form of autonomy harm. As they discuss, however, courts have been inconsistent in recognizing loss of control as a privacy harm. Some courts have interpreted BIPA as only applying to the sharing of data with external parties—in *Rivera v. Google*, the court denied standing given that Google only stored (*not* shared) biometric data without consent.[136] While other courts have taken the opposite view, they did so by sidestepping the harm question by concluding that a violation of BIPA alone was sufficient for standing, even without actual injury or adverse effect.[137]

In *Vance v. Microsoft*, in which plaintiffs alleged that Microsoft had violated BIPA by using IBM's Diversity in Faces dataset "to improve the fairness and accuracy of its facial recognition products," the Western District of Washington dismissed the plaintiff's complaint under BIPA § 15(c) that Microsoft had "otherwise profit[ed] from" the biometric information of the plaintiffs. The court concluded that since the plaintiffs did not allege that Microsoft "disseminated or shared access to biometric data through its products," there was no alleged violation of BIPA's profit provision. This dismissal was notable since the fact that Microsoft used people's biometric information to improve their facial recognition technologies without permission is arguably at the crux of the "lack of control" harm. The same court, however, did not dismiss the BIPA profit complaint in a similar case against Amazon due to the possibility that Amazon used the images as part of the reference set for its facial recognition product.[138] Taken together, these cases suggest that the profit component of BIPA only applies in cases where the biometric information becomes part of the product itself.

Notably, this distinction is not very coherent in the context of HCCV given that the images used to train the model are arguably key to the product itself. Nonetheless, the way the court happened to apply this distinction in these cases suggests that the court had some intuition for the key distinction drawn throughout this Article between using individuals' data for development vs. deployment. Using the images simply for training or testing during

[136] Rivera v Google, 366 F. Supp. 3d 998 (N.D. Ill. 2018)

[137] Rosenbach v. Six Flags Entertainment Corporation et al., No. 123186, 2019 Ill. Lexis 7 (Ill. Jan. 25, 2019).

[138] It is highly unlikely, however, that Amazon incorporated the Diversity in Faces dataset into a reference dataset that it sold alongside its Rekognition product. The "known" faces used to make a match must be identified in some way in order to be useful—it is not helpful to say that a new face is the same as another face scraped from Flickr unless there is more information about the face from Flickr (e.g., name, ID, or evidence that the person committed a crime). A reference set of random anonymous faces is not useful for recognition purposes.

development, as in the Microsoft case, creates minimal privacy risks in comparison to including the images in a reference set used in deployment, as in the Amazon case. As discussed above, being included in the reference set directly implicates possible surveillance risks. As will be discussed in Section VII, making a distinction between development and deployment is one possible privacy law carve-out that could address the tensions discussed in this Article. Such a distinction would provide much greater clarity than the current guidance focusing on whether the data is part of the product.[139]

Given this lack of clear judicial guidance, there are two possible opposing stances regarding this potential harm. One would be to analogize the HCCV training process to the model "seeing" and learning from the world, as humans do. There is no legally cognizable harm associated with a human looking at non-explicit images that are readily available online. If the images have Creative Commons licenses that allow them to be used for commercial uses, the human might even be able to incorporate those images into a commercial product without harm. From this perspective, the fact that the way AI learns to "see" people involves the distillation of image pixels into biometric information does not inherently change the harm equation. Under this view, it should be fine for developers to use publicly available images (with appropriate licenses) to develop HCCV.

On the other hand, given how controversial some HCCV technologies are, there is a strong argument that people should have some control over whether their images are being used to develop such technology. There is a potential psychic harm associated with knowing that your personal data is contributing to technologies without your knowledge, particularly technologies you oppose. From a policy perspective, there are two ways to address this type of harm. One would be the current privacy regime, which emphasizes the importance of individual consent before one's images can be used for specific purposes. Another would be to increase the regulation of what is acceptable vs. unacceptable HCCV such that individuals could feel more assured that the technology developed with their data is considered (at least by legislative consensus) to be societally acceptable. The former has the advantage of punting the question of acceptable use cases to the individual to decide but has the disadvantage of conflating control with signing an informed consent form. Depending on the context, such agreements can be difficult to understand or impossible to negotiate.[140] The latter has the advantage of ensuring that data will not be used for certain purposes but the disadvantage of relying heavily on regulators to carefully draw the line between acceptable and unacceptable use cases.

Thus, despite the growth in regulations around biometric information, there are significant ambiguities that remain around the specific harms envisioned by such regulations and how they would manifest in the context of developing HCCV. The most concrete harms courts have found compelling (electricity, identity theft, and surveillance) depend on technical nuances that have not been addressed by courts and are distinct from the primary harm at hand—the use of data without consent to develop controversial technologies. This Section thus strived to clarify

[139] In both cases, however, the court did not dismiss the plaintiff's unjust enrichment claim, so it is still possible that this line of cases stemming from the Diversity in Faces dataset could provide more clarity in the future about the extent to which "lack of control" holds legal weight as a privacy harm. Thus far, however, the case law does not provide much guidance around such harms.

[140] Indeed, the fixation of modern privacy law on informed consent has been widely criticized in the literature. *See, e.g.*, Frederik Zuiderveen Borgesius, *Informed Consent: We Can Do Better to Defend Privacy*, 13.2 IEEE SECURITY & PRIVACY (Mar.-Apr. 2015), https://ieeexplore.ieee.org/abstract/document/7085952.

these harms and their relevance in this context. As courts consider such cases around the images used to develop of HCCV and as policymakers consider how to regulate HCCV, such distinctions will be highly salient.

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.

VII: Approaches to Balancing Privacy and Bias Mitigation

While privacy laws protect generally against the harms of being "seen" without consent, the harms of being "mis-seen" are not directly protected against. For practitioners charged with balancing the ethical desiderata of fairness and privacy, the threat of legal liability leans far more heavily in favor of protecting privacy than addressing algorithmic bias.[153] There are a few possible approaches for addressing this imbalance, as the subsections below will discuss.

One would be to create narrow carve-outs in the protections against being "seen" through privacy law. Another path would be to alleviate some of the concerns with being "seen" through participatory design, use of trusted third-parties to collect data, or privacy-preserving technological advances. A final approach would be to increase the protections against being "mis-seen." Note that this Section does not advocate for all of these options equally, but instead seeks to present a wide array of options and discuss the pros and cons of each.

A. Carve-Outs from Privacy Law

The idea of reducing privacy protections around FRT and other HCCV technologies might seem absurd at a time when there are calls for *stronger* privacy protections and the specter of mass surveillance seems increasingly threatening, with more and more deployment of HCCV technologies.[154] Indeed, some scholars have argued that we should instead be increasing privacy

[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.

protections in the U.S. in order to prevent the ethical and legal risks associated with FRT.[155] Countries like China have recently increased privacy protections in the FRT context. The Supreme People's Court issued a directive to lower courts to make "collection and analysis of facial data by companies an infringement of personal rights and interests if carried out without previous consent."[156] Nonetheless, given the tension presented in this Article between multiple ethical desiderata—privacy and fairness—and given efforts in the EU and UK to create privacy carve-outs for processing sensitive data in service of bias mitigation efforts (as discussed in Section VII.A), it is worth contemplating what possible surgical changes could be made to existing privacy regimes to balance these desiderata.

One possible carve-out is to make a distinction between images used to *develop* HCCV models versus images used during the *deployment* of an HCCV system. Training datasets are simply used to train the model to perform a specific task like detection, recognition, or classification in a particular context. While the HCCV system learns *how* to identify the individual, the goal is *not* to identify that individual. Similarly, test datasets are designed to assess the HCCV system's performance rather than to make use of an identification. In contrast, when the HCCV system is deployed, the goal is to detect/recognize/classify the individuals it encounters by comparing them to a reference list; being monitored by the HCCV system or being on the reference list thus presents the potential for more acute privacy harms. The collection and use of such images in deployment without informed consent is what directly enables mass surveillance.

Making this distinction between development and deployment has the benefit of enabling HCCV developers to use large corpuses of publicly available images and any other images they collect with appropriate licenses to train more accurate and less biased HCCV systems. This could promote the creation of larger, fairer publicly available datasets, leveling the playing field for smaller companies.

Of course, the drawback to this approach is that individuals would not be able to control what kinds of technologies their images are used to develop. Copyright would still apply, so the only images developers could use would be those that already have a license for commercial use, but many people would likely still feel uncomfortable if their images (even if publicly available with appropriate licensing) were used to develop HCCV. Indeed, given that the copyright belongs to the image taker rather than the image subject, copyright might not provide any protection for many individuals.

In addition, to the extent images processed in deployment are used for further training of the model, the lines between development and deployment might blur. In these cases, the images should retain privacy protections to prevent such a carve-out from enabling additional surveillant use cases without appropriate informed consent.

Another possible approach would be to make the privacy laws around biometric information more domain-specific or sectoral. Indeed, federal privacy laws in the US remain sectoral, protecting highly sensitive information in specific contexts, such as medical

[155] *See* Nakar & Greenbaum, *supra* note 154.
[156] *Ruling by Top China Court Respects Privacy*, SOUTH CHINA MORNING POST (Aug. 10, 2021), https://www.scmp.com/comment/opinion/article/3144579/ruling-top-china-court-respects-privacy.

information.[157] One of the primary sources of imbalance between privacy and fairness considerations in HCCV development is the fact that anti-discrimination protections are highly sectoral, whereas the state biometric privacy protections are not. The innovation of laws like BIPA was to protect specific types of information rather than information in a specific context. While this was motivated by the rationale that biometric information is uniquely immutable, this innovation significantly expanded the scope of such laws.

Indeed, even the recent E.U. proposed AI regulation, which some have criticized for being overly broad,[158] focuses specifically on prohibited use, high risk, or limited-risk cases of AI.[159] The regulation provides no requirements for other use cases. Similarly, privacy protections relevant to collecting and processing human images to develop HCCV could be limited to contexts like law enforcement, healthcare, finance, employment, education, and any other high-risk domain.

A likely critique of this approach, however, would be that it is difficult to control the contexts in which data are used. Companies often build general-purpose HCCV systems that can be tailored for a wide variety of different domains. Moreover, many concerns about surveillance specifically involve the use of these technologies in domains like retail that are not typically considered high-risk, like tracking people's movements in a mall for ad targeting. The fear is that inferences are being made about people without their knowledge, or that people might self-censor their behavior because of the possibility that they are being watched.[160]

A more promising approach would be to make identifiability a salient factor when evaluating the collection or processing of biometric information. This would have the advantage of incentivizing privacy-preserving techniques, such as blurring faces or manipulating them to be less recognizable, and efforts to silo data to prevent matching with identifying or sensitive information. This would not address the economic harms of electricity/processing power but would help address concerns around identity theft and surveillance. Such a carve-out, however, will require significant guidance. As will be discussed in Section VII.D, there are limitations to privacy-preserving techniques, such that the degree of identifiability that is relevant from a legal perspective will be a key question.

Finally, separate from the privacy protections of the images themselves are the protections around the sensitive attribute data of the image subjects. Making it easier for companies to collect demographic data for the exclusive purpose of conducting audits of their HCCV systems would only narrowly weaken privacy protections while enabling fairer HCCV development. The proposed EU AI regulation gestures in this direction with a carve-out for

[157] Daniel J. Solove & Paul M. Schwartz, Information Privacy Law 907 (7th ed.). The Health Information Portability and Accountability Act of 1996 (HIPAA) is a key example.
[158] *DigitalEurope's Initial Findings on the Proposed AI Act*, DIGITALEUROPE (Aug. 6, 2021), https://www.digitaleurope.org/wp/wp-content/uploads/2021/08/DIGITALEUROPEs-initial-findings-on-the-proposed-AI-Act.pdf; E*U Proposals to Regulate AI are Only Going to Hinder Innovation*, FINANCIAL TIMES (July 25, 2021), https://www.ft.com/content/a5970b6c-e731-45a7-b75b-721e90e32e1c; *Feedback on the Artificial Intelligence Act,* CENTER FOR DATA INNOVATION (2021), https://www2.datainnovation.org/2021-feedback-aia.pdf; Dan Whitehead, *Hogan Lovells Responds to the European Commission's Consultation*, HOGAN LOVELLS (Aug. 10, 2021), https://www.engage.hoganlovells.com/knowledgeservices/news/hogan-lovells-responds-to-the-european-commissions-consultation-on-the-ai-regulation.
[159] E.U. Proposed AI Regulation, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.
[160] *See* Nakar & Greenbaum, *supra* note 148 (discussing First Amendment challenges related to FRT); Reidenberg, supra note 154 (broader discussion of importance of privacy in public and anonymity for democratic values).

processing sensitive data for the purposes of complying with other provisions in the regulation.[161] A similar approach could be used in the U.S. to provide carve-outs for algorithmic bias detection and mitigation purposes.

Thus, while any proposals to limit the scope of current privacy protections around HCCV would likely be highly controversial, this Subsection illuminates some possible carve-outs that would make privacy and fairness more evenly incentivized from a regulatory perspective. Overall, however, privacy carve-outs are the category of potential solutions with the most significant trade-offs, so first pursuing the other possible solutions discussed below would be preferable.

B. Participatory Design

Another approach that scholars in the algorithmic fairness community have proposed is to look toward participatory design—methods that engage stakeholders who use or are affected by a technology in its design, in order to build greater trust between the data subjects and the data collectors. [162] In their piece discussing the parallels between data collection for AI and data collection for archives, Jo and Gebru emphasized the importance of establishing such community relationships and empowering communities to contribute to data collection efforts.[163]

This is an important approach to consider for bridging the gap between AI developers and communities affected by their development, but it faces many practical challenges to implementation. A key difference between archives and datasets for AI is the lack of incentive for most people to contribute to AI datasets. While contributing to an archive can be seen as an honor, a way to preserve the history of your family or community, contributing to an AI dataset is viewed with wariness. Many of the BIPA lawsuits against major US tech companies came after people realized that their Flickr photos were being used in training datasets. An artist even created a platform for people to check whether their images are included in the major publicly available datasets,[164] and journalists wrote of the creepiness of realizing their images were being used.[165]

The challenge for AI developers will thus be to establish trust with the communities from whom they are collecting images and create incentives for individuals to contribute to dataset

[161] Title III, Chapter 2, Article 10.5, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[162] Clay Spinuzzi, *The Methodology of Participatory Design*, 52.2 TECHNICAL COMMUNICATION (2005), https://repositories.lib.utexas.edu/bitstream/handle/2152/28277/SpinuzziTheMethodologyOfParticipatoryDesign.pdf?sequence=2; Michael Muller & Allison Druin, *Participatory Design: The Third Space in HCI* (2002), https://www.researchgate.net/profile/Michael-Muller-12/publication/279063895_Participatory_Design/links/5a82faa1aca272d6501c2deb/Participatory-Design.pdf

[163] Eun Seo Jo & Timnit Gebru, *Lessons From Archives: Strategies for Collecting Sociotechnical Data in Machine Learning*, PROC. OF CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY (FAT*) (2020), https://arxiv.org/pdf/1912.10389.pdf.

[164] EXPOSING.AI (last accessed Jan. 21, 2022), https://exposing.ai/; Cade Mez & Kashmir Hill, *Here's a Way to Learn if Facial Recognition Systems Used Your Photos*, N.Y. TIMES (Jan. 31, 2021), https://www.nytimes.com/2021/01/31/technology/facial-recognition-photo-tool.html.

[165] *See* Kashmir Hill & Aaron Krolik, *How Photos of Your Kids Are Powering Surveillance Technology*, N.Y. TIMES (Oct. 11, 2019), https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html.

collection initiatives. This is easier said than done. For one, the "community" in question might be the global human population if the goal is to ensure that the AI system works well for everyone. In addition, community trust will likely be predicated on the images only being used to develop HCCV systems that the individuals believe will benefit their communities. The vast majority of data used for training HCCV systems, however, is used to train base models that can perform general tasks—e.g., object, face, or body detection, recognition, and verification—not specific to particular use cases. Companies then adapt these base models to more specific contexts using transfer learning and smaller task- and deployment-specific datasets. Thus, while it might be possible for a company to partner with a specific community to develop an AI system that does a specific trusted task (e.g., a security system for the local school), the base model for such a system would be trained on many images from other communities. As a result, AI companies typically seek a more global consent for using individuals' photos to develop any computer vision system.

One way, however, to potentially reconcile the desire for both (i) close, carefully designed, and stakeholder-driven data-collection partnerships and (ii) a large breadth of such partnerships is through data consortia, which will be discussed in the next Section.

C. Trusted Third-Party Data Collection

One method for addressing these trust issues is to shift the responsibility for data collection and storage from private companies to third-party actors (governmental or non-governmental) that might be more trusted for data collection. Veale and Ruben, for example, have proposed this approach as a way to handle the privacy concerns around processing sensitive attribute data for bias mitigation.[166] This would have the advantage of creating large image datasets, using the pooled resources of companies, research institutions, and/or government entities, alleviating some of the challenges facing developers that do not have a pre-existing pipeline for images. In addition, provided that the third-party has strong transparency requirements and governance structures, the data collection process could be more easily evaluated and improved over time, building trust with data subjects. If this entity has sufficient funding and oversight, there should also be a greater incentive for it to uphold high standards and use the latest privacy and bias mitigation techniques. If the data consortium succeeds in being a trustworthy entity, then more people will likely be willing to contribute data to the entity in comparison to selling their data to companies with weaker ethical governance guarantees. The presence of such a trusted data consortium would also raise the ethical standards for data collection—even when companies are collecting their own data, their practices could be compared to those of the consortium.

Creating a third-party trusted data consortium to collect HCCV data would have to go beyond the proposal of Veale and Ruben, however, given the need not only to manage the sensitive attribute data used to audit an AI system, but also the fundamental building blocks of the HCCV system itself. The complications and challenges of ethical data collections discussed

[166] Michael Veale & Reuben Binns, *Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data*, 4.2 BIG DATA & SOCIETY (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3060763.

above persist. The third-party entity will have to struggle with questions of how to collect a globally representative dataset with adequate informed consent and sufficiently candid and diverse images. While this solution directly tackles the problem of trust, it does not necessarily solve for the other challenges.

Nonetheless, a third-party entity would arguably be better equipped to address some of these challenges. Economies of scale are quite useful for data collection generally, but especially for ethically and legally compliant sensitive data collection given the high overhead costs. For example, designing informed consent agreements that are compliant with all jurisdictions' privacy regulations is quite challenging, but the cost of designing such agreements is the same regardless of whether thousands or millions of images are collected. In addition, setting up systems for data subjects to be able to monitor how their data is being used and revoke their consent at will is quite difficult for small companies or academics, and could be more easily handled by a dedicated data consortium. More generally, there are high fixed costs to establishing a crowdsource platform where individuals around the world can upload their images and be compensated fairly. Even when large companies have tried to create such platforms, their initiatives have not been successful.[167]

It is difficult, however, to establish what actor would be sufficiently trustworthy to conduct such large-scale data collection, which could become the basis of many commercial HCCV systems. As mentioned previously, NIST uses mugshots and images of exploited children (among other marginalized populations) as the basis for their test dataset to evaluate commercial facial recognition systems,[168] so we cannot take for granted that government agencies will have easy access to more ethically collected data or that they will enforce the highest standards of informed consent in their data collection practices. A new entity might have to be created to take on this responsibility.[169]

As a result, such a solution will likely take years to develop, so it alone will not be a panacea. Moreover, even if such a trusted data consortium exists, companies will still need to collect some of their own data to tailor their models to specific tasks.[170] Thus, creating trusted third parties to collect data for HCCV development is a promising solution to pursue, alongside other approaches that tackle the core tensions addressed in this Article.

[167] Microsoft created a platform called Trove in 2020 to enable the responsible collection of images from crowd workers. Over 60,000 images were collected. Microsoft shut down the Trove project on November 16, 2021. *Microsoft Trove*, MICROSOFT (last accessed Jan. 21, 2022), https://www.microsoft.com/en-us/ai/trove-images-for-machine-learning.

[168] *See* Grother et al., *supra* note 134; *Chexia Face Recognition*, NIST (last accessed Jan. 21, 2022), https://www.nist.gov/programs-projects/chexia-face-recognition.

[169] It is also possible that existing non-profits could host an initiative like this. For example, Mozilla currently has an initiative for collecting voice data. *Common Voice*, MOZILLA (last accessed Feb. 1, 2022), https://commonvoice.mozilla.org/en.

[170] The large image dataset provided by a trusted third-party is useful for training a base model that is able to do common tasks like identification or verification, but AI developers still need to collect deployment-context data in order to tailor their models to the particular tasks at hand. For example, an HCCV model that is trying to identify shoplifting needs training data of images or videos of people shoplifting and not shoplifting. These datasets do not need to be as large, but the companies will still need to address the ethical dataset collection challenges discussed in this Article.

D. Technological Advances

Given the constant advances in HCCV technology, it is important to consider whether the problems addressed in this Article might be resolved over time purely through technological progress. In particular, could advances in privacy-preserving technologies and synthetic image generation address these issues?

Privacy-preserving technologies are generally helpful for mitigating privacy and security risks with HCCV datasets. Pixelization and blurring are the most well-known techniques, but do not provide any formal privacy guarantee that the original image cannot be reverse engineered.[171] While completely cutting out an individual's face can dramatically reduce the possibility of future identification, such images are only useful for developing non-face-related HCCV. Moreover, such techniques do not address the fundamental informed consent problem. For example, if you collect a large dataset of images from the internet, and then you use an algorithm to transform the faces or bodies to be less recognizable, you might still be processing biometric information without informed consent. The process of face blurring itself requires processing biometric information, creating a catch-22. More generally, there is always a trade-off between the level of privacy attained through such techniques and the utility of the data.[172] This is not to say that privacy-preserving techniques are not an important part of HCCV systems, but rather that they alone cannot solve the problems discussed in this Article.

Synthetic image generation is promising in that it can be used to generate images of people who are not real or of real people in new positions/settings, thus augmenting the training dataset. There are two general categories of synthetic image generation approaches relevant to this discussion: ones trained on images of real people and ones that are not. The former, which includes GANs, can modify specific features of an individual (e.g., skin tone, hair length, or perceived gender)[173] or "hallucinate" new people.[174] The models underlying these techniques, however, need to be trained on large numbers of human images, thus undermining the extent to which this approach can completely resolve the informed consent barrier. In addition, to the extent only a small number of people are scanned to form the basis of the synthetic images, the synthetic images might not accurately reflect the wide diversity of humanity. These approaches

[171] William L. Croft et al., *Obfuscation of Images via Differential Privacy: From Facial Images to General Images*, Proc. of International Cross-Domain Conf. for Machine Learning & Knowledge Extraction (CD-MAKE) 229 (2019), https://link.springer.com/chapter/10.1007%2F978-3-030-29726-8_15. Differential privacy, in contrast, is a mathematical criterion guaranteeing that the inclusion or exclusion of an individual cannot be distinguished, ensuring that individual-level information will not be leaked. *Differential Privacy*, Harvard University Privacy Tools Project (last accessed Jan. 21, 2022), https://privacytools.seas.harvard.edu/differential-privacy; Kobbi Nissim et al., *Differential Privacy: A Primer for a Non-Technical Audience* (2018), https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf. In practice, differential privacy is achieved by adding random noise from a carefully chosen distribution to the data. Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, Theory of Cryptography 265 (2006), https://link.springer.com/chapter/10.1007/11681878_14.

[172] *See, e.g.*, Croft et al. *supra* note 171.

[173] *See, e.g.,* Balakrishnan et al., *supra* note 149.

[174] *See, e.g.,* Blaz Meden et al., *k-Same-Net: Neural-Network-Based Face Deidentification*, Proc. of 2017 International Conf. & Workshop on Bioinspired Intelligence (IWOBI) (2017), https://ieeexplore.ieee.org/document/7985521.

can still be promising, however, as a means for augmenting existing datasets that have appropriate informed consent.

Almost always some images of real people are used somewhere in the pipeline of creating images of synthetic individuals, [175] but it is also possible to use techniques that do not use real human images. Such techniques are often used in animation, and use drawings, toy models, or 3D computer renderings. These approaches can directly circumvent privacy concerns. The primary downsides to this type of approach are i) the difficulties of creating large numbers of highly realistic and diverse images and ii) potential biases of the humans generating these images. The first concern will likely be mitigated over time with advances in this type of technology, propelled by the demand for ever-more realistic-looking images. The second issue is more complicated to address. It is inevitable that the people creating these images will have preconceptions of what are relevant types of people and contexts to feature, which has been a critique of start-ups that claim to use "zero data."[176] Creating a sufficiently diverse dataset to reflect the wide array of images an HCCV model is likely to encounter in the real world is a fundamentally challenging problem, even if you have the ability to create realistic images from scratch. Over time, these issues might be mitigated by engaging with diverse image creators and figuring out better ways to measure and audit image datasets for diversity, but for now this is still an open area for future research.

Aside from technologies that side-step or reduce the need for large numbers of human images, federated learning can also be beneficial for giving individuals more control over their data. Federated learning enables data across different parties to be used for training a model, without directly sharing that data between parties.[177] The data remains on the edge device or local server, where a local model is trained and then integrated into the larger model. This is beneficial in contexts where individuals consent to having their images used for training HCCV but are uncomfortable with directly sharing their images with the entity in question (e.g., someone who is comfortable with their photos being used for training a Apple's photo-sorting facial recognition model, but they do not want to directly share their photos with Apple out of concern over how else their photos might be used). Federated learning does not solve the fundamental issue that people might not want their images used for training HCCV, but for people who are supportive of the goal of supplying more diverse images to enable training better-performing, less biased HCCV, federated learning can ease some concerns around sharing their data. In recent lawsuits around HCCV, however, courts have considered the distinction between whether images are stored on the user's edge device versus the company's cloud to be irrelevant for reducing a company's liability (the company was still considered to have control over the data).[178] Moreover, as discussed above in Section VI.A, in analyses around the harms of using electricity/processing power, courts have arrived at conclusions that would disincentivize

[175] E.g., approaches to create realistic humans typically do involve at least some images of real people at the beginning to create the 3D models. Synthetic *Data Case Studies: It Just Works*, SYNTHESIS AI (June 17, 2021), https://synthesis.ai/2021/06/17/synthetic-data-case-studies-it-just-works/

[176] Sage Lazzaro, *AI Experts Refute Cvedia's Claim Its Synthetic Data Eliminates Bias*, VENTUREBEAT (July 6, 2021), https://venturebeat.com/2021/07/06/ai-experts-refute-cvedias-claim-its-synthetic-data-eliminates-bias/.

[177] Brendan McMahan & Daniel Ramage, *Federated Learning: Collaborative Machine Learning Without Centralized Training Data*, GOOGLE AI BLOG (April 6, 2017), https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

[178] Hazlitt v. Apple Inc., 500 F. Supp. 3d 738, 2020 U.S. Dist. LEXIS 210963, 2020 WL 6681374 (United States District Court for the Southern District of Illinois, November 12, 2020, Filed); In re: TikTok Inc. Consumer Privacy Litigation, Case No. 1:20-cv-04699 (U.S. Dist. Court for the N.D. of IL.).

the use of technologies like federated learning. While these cases are either are still in progress or were settled before a decision was made, they suggest that using techniques like federated learning might not resolve the privacy challenges to creating less biased HCCV.

Thus, the tension between privacy and fairness in HCCV data collection might be reduced in the medium- to long-term by technological advances. For now, however, current techniques still rely largely on images of real people and there remain fundamental unsolved questions around how to generate large numbers of diverse, realistic images without substantial bias.

E. Right Against Being Mis-Seen

The final approach to balancing the desire not to be "seen" or "mis-seen" would be to increase the protections against being "mis-seen." As discussed above in Section VI.B, currently there are only legal protections if being mis-seen triggers a separate legally cognizable harm. As a result, harms that manifest themselves as everyday inconveniences or indignities are unlikely to be protected against, even if the amalgamation of these harms leads to individuals living their lives like second-class citizens. This Section will explore possible instantiations of what a right *not* to be "mis-seen" might look like.

An initial inquiry is whether existing product liability law might be able to provide sufficient protection against being mis-seen by HCCV systems. After all, the harms of being mis-seen are caused by poor product performance, either for everyone or a specific subgroup. Unfortunately, there are several limitations to existing product liability doctrine that would render it unable to provide sufficient protections.

First, product liability law protects primarily against physical harm. If robots and autonomous vehicles become much more widely used in the future, there might be more risk of physical harm from HCCV systems, but for now such systems are primarily deployed in contexts where the potential for bodily harm is minimal (e.g., verifying someone's identity for security purposes, sorting photos, monitoring people, or providing entertainment on social media). While there is the potential to recover for emotional distress under product liability in cases where a bystander is distressed by witnessing a product physically harming another individual,[179] someone experiencing physical harm is still typically necessary.

A second limitation is that product liability law would not help plaintiffs who experienced algorithmic bias.[180] In cases where the product performs very well for the vast majority of people but poorly on particular subgroups, it would be difficult to establish that the product is unreasonably dangerous.[181] This is especially the case if the HCCV system still performed somewhat well for the subgroups despite a large gap in how it performed across groups. If the HCCV developer made false claims about the system being unbiased, the plaintiff

[179] Linda Trummer-Napolitano, *Emotional Distress in Products Liability: Distinguishing Users from Bystanders*, 50 FORDHAM L. REV. 291 (1981), https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=2508&context=flr.

[180] Selbst discusses the challenges of applying negligence law to contexts where there are unevenly distributed harms. Andrew D. Selbst, *Negligence & AI's Human Users*, 100 BOSTON U. L. REV. 1315, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3350508.

[181] *Id.*

might be able to succeed on a claim of false advertising, but there is no such thing as a completely unbiased AI model, so such litigation would likely simply lead companies to avoid making such outlandish claims.

A third limitation is the lack of robust standards for performance in the AI industry. Current HCCV systems are unlikely to be considered sufficiently inherently dangerous to trigger strict liability, so a negligence standard would likely apply. Industry standards are often relied upon in product liability law to evaluate whether a company has been negligent.[182] While NIST has created a Facial Recognition Vendor Test for companies to benchmark their facial recognition technologies,[183] there is no industry-wide consensus on a single benchmark for performance or what levels of performance are sufficient. Moreover, the need to tailor AI systems to specific deployment contexts suggests that any blanket benchmark or performance standard would be misleading.[184] For example, establishing that a facial recognition system performs well at matching mugshots does not imply it would work well at matching a driver's license photo with a surveillance camera image of a suspect. Surveillance camera images are typically much grainer and lower quality and rarely feature a clear frontal image of the suspect looking into the camera.

While consumer expectations are also often used as a benchmark for reasonableness, as a relatively new but rapidly evolving technology, consumer expectations for HCCV are particularly unstable.[185] This lack of clear consumer expectations has also made it easy for AI technologies to proliferate while providing minimal representations and warranties to consumers.[186] AI companies often avoid providing any details about how their technologies are developed or how well they perform on any standardized tests.[187]

Finally, there are many reasonable justifications for why companies do not do more to ensure their computer vision products are not highly biased, making it difficult to pursue a negligence case. As discussed above, both privacy and antidiscrimination laws discourage the collection of data that could be used to test the performance of the AI system across different demographic groups and to improve such performance. Current industry practices around preventing algorithmic bias are often minimal due to the lack of incentives (and the strong disincentives) to address this issue.[188]

Thus, new regulations on the state or federal level are likely needed to protect individuals against being "mis-seen." A right against being "mis-seen" would imply either a private right of action or government audits of HCCV systems. This right could be a general right for HCCV

[182] *Id.*

[183] *Face Recognition Vendor Test (FRVT)*, NIST (last accessed Jan. 21, 2022), https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt.

[184] Alicia Solow-Niederman et al., *The Institutional Life of Algorithmic Risk Assessment*, 34 BERKELEY TECH. L. J. 705 (2019), https://lawcat.berkeley.edu/record/1137216/files/01_Solow-Niederman_Web%5B1%5D.pdf. Notably, popular datasets like Labelled Faces in the Wild specifically warn that they should not be used for concluding whether an algorithm would be suitable for commercial purposes, citing lack of diversity along age, gender, ethnicity, lighting conditions, poses, occlusions, and photo resolution. LABELED FACES IN THE WILD, http://vis-www.cs.umass.edu/lfw/.

[185] Selbst, *supra* note 180.

[186] Stefano Puntoni et al., 85.1 *Consumers & Artificial Intelligence: An Experiential Perspective*, J. of Marketing (2020), https://journals.sagepub.com/doi/abs/10.1177/0022242920953847.

[187] Selbst discusses the importance of secrecy in AI development. Selbst, *supra* note 180.

[188] *See supra* note 49.

systems to have a minimum performance level, or an anti-discrimination right for the system to not have a significantly disproportionate performance for your subgroup. The former would be most related to negligence and product liability law. As discussed above, establishing standards for reasonableness in the HCCV context might be difficult in the short-term, so strengthening such a right might require the government developing specific HCCV regulations. Transparency obligations could further enable individuals to challenge the use of HCCV systems with poor performance.

The anti-discrimination right would be a new protection that acknowledges the fact that HCCV is increasingly pervasive and embedded into everyday life, creating the risk that those who are more likely to be mis-seen by such technology might find themselves living in a world not optimized for them. Given that HCCV systems lack intentionality,[189] the protection would be against disparate impact, a form of unintentional discrimination whereby facially neutral practices lead to disproportionate adverse effects on particular subgroups. While most anti-discrimination laws apply to specific domains, like employment, finance, or education, this protection would apply to a category of technology, HCCV. The justification for singling out HCCV for additional antidiscrimination protections would be that (i) bias mitigation for HCCV is particularly important but difficult (as discussed above in Section III) and (ii) the increasingly pervasiveness of HCCV in everyday contexts makes the lack of protections against bias in HCCV particularly pernicious, even in "low-stakes" contexts (as discussed in Section II). While the domain-specificity of many anti-discrimination laws is motivated by the high-stakes nature of those contexts, there are also anti-discrimination laws like Title II and Title III of the Civil Rights Act of 1964 that protect individuals in low-stakes but commonplace contexts like public accommodation.[190] In addition, the Americans with Disabilities Act of 1960 created accessibility and reasonable accommodation requirements to make it easier for individuals with disabilities to access public services and employment.

Having an anti-discrimination right against disparate impact in being "mis-seen" by HCCV technology would thus provide more incentive for companies to directly address issues of algorithmic bias. Of course, this would not directly solve the informed consent challenge posed by privacy laws, but creating such a right would better balance the ethical trade-offs around data collection. Policymakers would need to directly provide guidance more clearly defining the parameters for ethical data collection.

If this protection were enforced by an agency, then there should be resources allocated to conducting audits. This would be especially helpful since algorithmic bias can be very challenging for individuals to detect on their own. Without a concerted effort to gather information about other consumers' experiences and demographics, individuals cannot distinguish between a shoddy product and a biased one.

[189] This is not to say that intentional discrimination on the part of algorithmic developers does not exist, but the examples of algorithmic bias that have been publicly documented stem from unintentional discrimination, so it is important for protections against being "mis-seen" to prevent unintentional discrimination. If a developer does want to create a discriminatory algorithm, however, it is easy to mask their intentions. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L. J. 1043 (2019), https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3972&context=dlj.

[190] 42 U.S. Code § 2000a.

If the protection were instead enforced through a private right of action, then transparency requirements would be very helpful for enabling consumers to challenge potentially biased products. Of course, the most helpful information would be about the model's performance across different demographic groups.[191] In the absence of that information, however, the requirements should at least include information about the source and properties of the data, the annotation methods, and the testing procedure.[192]

CONCLUSION

Few technologies are currently as controversial as HCCV, prompting a flurry of privacy laws and moratoriums to be passed in the past several years. Arguments against HCCV generally center on its ability to facilitate mass surveillance and harm women and minority groups through faulty identifications or classifications. Not all HCCV enables mass surveillance, however, and the development of fair and accurate HCCV requires huge amounts of data, collected from diverse populations with balanced representations. As a result, efforts to improve the fairness and accuracy of HCCV often collide with efforts to enhance privacy protections.

This is not an insurmountable tension—indeed, this Article discusses many potential approaches to address it—but it is a difficult one that will require attention from policymakers and developers to address. Policymakers will need to consider the incentives that developers have under current laws and whether there are ways to both incentivize and enable more efforts to address algorithmic bias in HCCV. Researchers and developers in the HCCV community will need to direct efforts toward studying potential technical solutions to enable HCCV systems to be developed with maximal accuracy and minimal bias while being trained either on smaller, more carefully collected datasets, or on synthetic datasets. Researchers and developers will also need to focus on sociotechnical strategies for ethical data collection, including developing closer relationships of trust with the communities they seek to collect data from. There is no silver bullet for enabling more ethical HCCV systems that balances all of the concerns this Article surfaces. Breaking down these challenges and potential solutions, however, is an important first step.

More broadly, this Article provides a starting point for more nuanced debates about the appropriate development and use of HCCV. Implicit in the tensions addressed in this Article is the juxtaposition of the suspicion, anxiety, and fear people have toward HCCV and the strong demand for the services such technology can provide. The strategy of addressing the fears around HCCV exclusively through privacy laws and moratoriums is both over- and under-inclusive, increasing the barriers to developing more accurate and less biased HCCV technologies that bear no relation to mass surveillance while also disincentivizing companies

[191] A requirement for such disparate impact assessments was notably missing in the EU's proposed AI regulation. Mark MacCarthy & Kenneth Propp, *Machines Learn That Brussels Writes the Rules: The EU's New AI Regulation*, BROOKINGS (May 4, 2021), https://www.brookings.edu/blog/techtank/2021/05/04/machines-learn-that-brussels-writes-the-rules-the-eus-new-ai-regulation/.

[192] *See* Margaret Mitchell et al., *supra* note 1 for a more in-depth discussion about relevant model-related disclosures for transparency purposes. *See also* Clavell et al., supra note 108 (discussing the importance of early documentation for enabling audits).

from directly addressing issues of algorithmic bias. Instead, a multi-pronged policy strategy is needed, including support for trusted third-party data collection initiatives, greater legal protections against being “mis-seen,” and more clarity around acceptable uses for biometric and sensitive information for bias mitigation. Ultimately, we must balance the desire not to be “seen” with the desire not to be invisible.