Principlism Guided Responsible Data Curation

Jerone T. A. Andrews [1] **Dora Zhao** [*2] **William Thong** [*3] **Apostolos Modas** [*3] **Orestis Papakyriakopoulos** [*3]
Alice Xiang [4]

Abstract

Human-centric computer vision (HCCV) data curation practices often neglect privacy and bias concerns, leading to dataset retractions and unfair models. Further, HCCV datasets constructed through nonconsensual web scraping lack the necessary metadata for comprehensive fairness and robustness evaluations. Current remedies address issues post hoc, lack persuasive justification for adoption, or fail to provide proper contextualization for appropriate application. Our research focuses on proactive, domain-specific recommendations for curating HCCV datasets, addressing privacy and bias. We adopt an ante hoc reflective perspective and draw from current practices and guidelines, guided by the ethical framework of principlism.

1. Introduction

Contemporary human-centric computer vision (HCCV) data curation practices which prioritize dataset features such as size and utility have pushed issues related to privacy and bias to the periphery, resulting in dataset retractions or modifications (Parkhi et al., 2015; Guo et al., 2016; Kemelmacher-Shlizerman et al., 2016; Merler et al., 2019; Torralba et al., 2008; Deng et al., 2009) as well as models that are unfair or rely on spurious correlations (Hendricks et al., 2018; Menon et al., 2020; Hill, 2020; Barr, 2015; Sagawa et al., 2020; Geirhos et al., 2020; Beery et al., 2018; Rosenfeld et al., 2018). HCCV datasets primarily rely on nonconsensual web scraping (Mudditt, 2022; Raji & Fried, 2021; Stewart et al., 2016; Ristani et al., 2016; Grgic et al., 2011; Günther et al., 2017; Founds et al., 2011). These datasets not only regard image subjects as free raw material (Birhane, 2020), but also lack the ground-truth metadata required for fairness and robustness evaluations (Karras et al., 2019; Lin et al., 2014; Merler et al., 2019; Everingham et al., 2010). This makes it challenging to obtain a comprehensive understanding of model performance across dimensions, such as data subjects, instruments, and environments, which are known to influence performance (Mitchell et al., 2019). While, for example, image subject attributes can be inferred (Karkkainen & Joo, 2021; Or-El et al., 2020; Liu et al., 2015; Kuznetsova et al., 2020; Schumann et al., 2021; Zhao et al., 2021; Buolamwini & Gebru, 2018; Alvi et al., 2018; Robinson et al., 2020; Wang et al., 2019), this is controversial for social constructs, notably race and gender (Hanna et al., 2020; Keyes, 2018; Benthall & Haynes, 2019; Khan & Fu, 2021). Inference introduces additional biases (Reid & Nixon, 2011; Segall et al., 1966; Balaresque & King, 2016; Hill, 2002; Garcia & Abascal, 2016) and can induce psychological harm when incorrect (Campbell & Troyer, 2007; Roth, 2016).

Recent efforts in machine learning (ML) to address these issues often rely on post hoc reflective processes. Dataset documentation focuses on interrogating and describing datasets after data collection (Holland et al., 2018; Bender & Friedman, 2018; Pushkarna et al., 2022; Srinivasan et al., 2021; Rostamzadeh et al., 2022; Butcher et al., 2021; Fabris et al., 2022b; Afzal et al., 2021; Papakyriakopoulos et al., 2023; Gebru et al., 2018). Similarly, initiatives by NeurIPS and ICML ask authors to consider the ethical and societal implications of their research after completion (Prunkl et al., 2021). Further, dataset audits (Shankar et al., 2017; Peng et al., 2021) and bias detection tools (Wang et al., 2022; Beretta et al., 2021) expose dataset management issues and representational biases without offering guidance on responsible data collection. In addition, proposals for artificial intelligence (AI) and data design guidelines (Peng et al., 2021; Denton et al., 2021; Luccioni et al., 2022; Google PAIR, 2019; IBM, 2019; Prabhu & Birhane, 2021) and adopting methodologies from more established fields exit (Hutchinson et al., 2021; Jo & Gebru, 2020; Huang & Liem, 2022). However, general-purpose guidelines lack specificity for specific domains and tasks (Srinivasan et al., 2021). For example, Prabhu & Birhane (2021)'s remedies focus on privacy and governance, disregarding data composition and image content. Other recommended practices lack persuasive justification for adoption (IBM, 2019) or fail to provide proper contextualization for appropriate application. For

[*]Equal contribution [1]Sony AI, Tokyo [2]Sony AI, New York [3]Sony AI, Zurich [4]Sony AI, Seattle. Correspondence to: Jerone T. A. Andrews <jerone.andrews@sony.com>.

2023 Data-centric Machine Learning Research (DMLR 2023) co-located with ICML 2023.

instance, the People + AI Guidebook (Google PAIR, 2019) suggests creating dataset specifications without explaining the rationale, and privacy methodologies are advocated without cognizant of privacy and data protection laws (Yang et al., 2022a; Piergiovanni & Ryoo, 2020; Uittenbogaard et al., 2019). While these efforts are important for responsible practices, they should be supplemented with proactive, domain-specific recommendations to address privacy and bias issues from the outset.

Our research directly addresses these critical issues by examining purpose (Section 3), consent and privacy (Section 4), and diversity (Section 5). Compared to recent scholarship, we adopt an ante hoc reflective perspective, offering considerations and recommendations for curating HCCV datasets for fairness and robustness evaluations. We draw insights from current practices, guidelines, dataset withdrawals, and audits to motivate each recommendation. Guided by the ethical framework of *principlism* (Beauchamp & Childress, 1994; Beever & Brightman, 2016), encompassing autonomy, beneficence, non-maleficence, and justice, our work aligns with the call for domain-specific resources to operationalize fairness (Holstein et al., 2019). We specifically focus on HCCV evaluation datasets with unique challenges (e.g., visual leakage of personally identifiable information) and opportunities (e.g., leveraging image metadata for performance analysis). Our proposals are not intended for the evaluation of HCCV systems that detect, predict, or label sensitive or objectionable attributes such as race, gender, sexual orientation, or disability.

2. Development Process

Principlism, derived from guidelines for protecting human subjects in research (Beauchamp & Childress, 2019; United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978), offers a practical approach to ethical dilemmas. It is based on four principles. *Autonomy* respects individuals' self-determination—e.g., through informed consent for HCCV datasets. *Beneficence* and *non-maleficence* require assessing harms and benefits during dataset design and considering broader implications for society. *Justice* promotes the fair distribution of risks, costs, and benefits, guiding decisions on compensation, data accessibility, and diversity. Thus, principlism provides a comprehensive framework for ethical dataset development and decision-making. To ensure comprehensive and consistent application of principlism, we harnessed diverse expertise and followed contemporary practices (Raji et al., 2021b; Romm, 1998; Srinivasan et al., 2021). Our team comprises researchers, practitioners, and lawyers with backgrounds in ML, CV, algorithmic fairness, philosophy, and social science. With a range of ethnic, cultural, and gender backgrounds, we bring extensive experience in designing CV datasets, training models, and developing ethical guidelines. To align our expertise with the ethical framework, we collaboratively discussed principlism's four pillars, considering each author's background. We identified key ethical issues in dataset design and refined them iteratively into an initial draft of ethical considerations. Through a comprehensive literature review, we incorporated relevant studies and datasets to revise the considerations, providing detailed explanations and recommendations for responsible data curation.

3. Purpose

In ML, significant emphasis has been placed on the collection, and utilization, of general-purpose datasets (Raji et al., 2021a). However, without a clearly defined task pre-data collection, it becomes difficult to address data composition, labels, data collection mechanisms, informed consent, and data protection assessments. This section addresses conflicting dataset motivations and provides recommendations.

3.1. Ethical considerations

Fairness-unaware datasets are inadequate for measuring fairness. Datasets lacking explicit fairness considerations are inadequate for mitigating or studying bias, as they often lack the necessary labels for fairness assessments. For instance, the COCO dataset (Lin et al., 2014), focused on scene understanding, lacks subject information, hindering fairness assessments. Researchers resort to human annotators to infer subject characteristics, limiting bias measurement to visually inferable attributes. However, this approach introduces annotation bias (Chen & Joo, 2021) and the potential for harmful inferences (Campbell & Troyer, 2007).

Fairness-aware datasets are incompatible with common computer vision tasks. Industry practitioners stress the importance of carefully designed and collected fairness-aware datasets to detect bias issues (Holstein et al., 2019). Fabris et al. (2022a) found that out of 28 computer vision datasets used in fairness research between 2014 and 2021, only eight were specifically created with fairness in mind. Among these, seven were (web scraped) HCCV datasets (Wang & Deng, 2020; Tong & Kagal, 2020; Buolamwini & Gebru, 2018; Steed & Caliskan, 2021; Karkkainen & Joo, 2021; Merler et al., 2019; Wang et al., 2019), including five focused on facial analysis. However, due to limited availability and narrow task focus, fairness-unaware datasets (Lin et al., 2014; Liu et al., 2015; Goyal et al., 2017; Zhang et al., 2014) are repurposed (Wang et al., 2020; Manjunatha et al., 2019; Hendricks et al., 2018). Fairness-aware datasets fall short in addressing the original tasks associated with well-known HCCV datasets, e.g., segmentation (Cordts et al., 2016; Martin et al., 2001), pose estimation (Lin et al., 2014; Andriluka et al., 2014), localization and detection (Everingham et al.,

2010; Dalal & Triggs, 2005; Geiger et al., 2012), identity verification (Huang et al., 2008), action recognition (Kay et al., 2017), as well as reconstruction, synthesis and manipulation (Karras et al., 2019; Georghiades et al., 2001). The absence of fairness-aware datasets with task-specific labels hampers the practical evaluation of these systems, despite their significance in applications such as healthcare (Mihailidis et al., 2004; Huang et al., 2018), autonomous vehicles (Janai et al., 2020), and sports (Thomas et al., 2017). Additionally, fairness-aware datasets lack self-identified annotations from image subjects, relying on inferred attributes, e.g., from online resources (Buolamwini & Gebru, 2018; Steed & Caliskan, 2021; Tong & Kagal, 2020).

3.2. Practical recommendations

Refrain from repurposing datasets. Existing datasets, repurposable but optimized for specific functions, can inadvertently perpetuate biases and undermine fairness (Koch et al., 2021). Repurposing fairness-unaware data for fairness evaluations can result in *dirty data*, characterized by missing or incorrect information and distorted by individual and societal biases (Kim et al., 2003; Richardson et al., 2019). Dirty data, including inferred data, can have significant downstream consequences for research, policy, and decision-making (Wang et al., 2021; Cooper et al., 2021; Richardson et al., 2019; Andrus et al., 2020). ML practitioners widely agree that a proactive approach to fairness is preferable, involving the collection of demographic information from the outset (Holstein et al., 2019). To mitigate epistemic risk, curated datasets should capture key dimensions influencing fairness and robustness evaluation of HCCV models, i.e., data subjects, instruments, and environments. Model Cards explicitly highlight the significance of these dimensions in fairness and robustness assessments (Mitchell et al., 2019).

Create purpose statements. Pre-data collection, dataset creators should establish purpose statements, focusing on motivation rather than cause (Hanley et al., 2020). Purpose statements address, e.g., data collection motivation, desired composition, permissible uses, and intended consumers. While dataset documentation (Gebru et al., 2018; Pushkarna et al., 2022) covers similar questions, it is a reflective process and can be manipulated to *fit* the narrative of the collected data, resulting in *hindsight bias* (Fiedler & Schwarz, 2016; Kerr, 1998; Chambers & Tzavella, 2022), as opposed to *directing* the narrative of the data to be collected. Purpose statements can prevent purpose creep, ensuring alignment with stakeholders' consent and intentions (Kugler, 2019). Enhancing transparency and accountability, as recommended by Peng et al. (2021), purpose statements can undergo peer review, similar to *registered reports* (Nosek & Lakens, 2014). Registered reports, recognized by the UK 2021 Research Excellence Framework, incentivize rigorous research practices and potential institutional funding (Chambers & Tzavella, 2022).

4. Consent and Privacy

Informed consent is crucial in research ethics involving humans (Nijhawan et al., 2013; National Commission for the Proptection of Human Subjects of Biomedicaland Behavioral Research, Bethesda, Md, 1978), ensuring participant safety, protection, and research integrity (Code, 1949; Politou et al., 2018). It consists of three elements: information, comprehension, and voluntariness, shaping data collection practices in various fields. While consent is not the only legal basis for data processing, it is globally preferred for its legitimacy and ability to foster trust (Politou et al., 2018; Edwards, 2016). We address concerns related to consent and privacy, and provide recommendations.

4.1. Ethical considerations

Human-subjects research. HCCV research often amasses millions of images without obtaining informed consent or participation, raising ethical concerns (Prabhu & Birhane, 2021; Harvey & LaPlace, 2021; Paullada et al., 2021; Solon, 2019). This exemption from research ethics principles stems from the perception that HCCV research does not fall under human-subjects research and publicly available data is considered low-risk for human subjects. Thus bypassing Institutional Review Board supervision. However, this approach is problematic due to the potential for predictive privacy harms when seemingly non-identifiable data is combined (Crawford & Schultz, 2014; Metcalf & Crawford, 2016; Prabhu & Birhane, 2021). Collecting data without informed consent hinders researchers and practitioners from fully understanding and addressing potential harms to image subjects (Van Noorden, 2020; Metcalf & Crawford, 2016). Some argue that consent is pivotal as it provides individuals with a last line of defense against the misuse of their personal information, particularly when it contradicts their interests or well-being (Mittelstadt & Floridi, 2016; De Hert & Papakonstantinou, 2016; Politou et al., 2018; Paullada et al., 2021).

Creative Commons loophole. Some datasets have been created based on the misconception that the "unlocking [of] restrictive copyright" (Prabhu & Birhane, 2021) through Creative Commons licenses implies data subject consent. However, the Illinois Biometric Information Privacy Act (BIPA) (Illinois Legislature, 2008) mandates data subject consent, even for publicly available images (Yew & Xiang, 2022). In the UK and EU General Data Protection Regulation (GDPR) (European Commission, 2016) Article 4(11), images containing faces are considered biometric data, requiring "freely given, specific, informed, and unambiguous" consent from data subjects for data processing. Similarly, in

China, the Personal Information Protection Law (PIPL) (National People's Congress, 2021) Article 29 mandates obtaining individual consent for processing sensitive personal information, including biometric data (Article 28). While a Creative Commons license may release copyright restrictions on specific artistic expressions within images (Yew & Xiang, 2022), it does not apply to image regions containing biometric data, e.g., faces, which are protected by privacy and data protection laws (Sobel, 2020).

Vulnerable persons. Nonconsensual data collection methods can result in the inclusion of vulnerable individuals who are unable to consent or oppose data processing due to power imbalances, limited capacity, or increased risks of harm (European Data Protection Board (Article 29 Working Party), 2017; Malgieri & Niklas, 2020). While scraping vulnerable individuals' biometric data may be incidental, some researchers actively target them, jeopardizing their sensitive information without guardian consent (Raji & Fried, 2021; Han et al., 2017). Paradoxically, attempts to address racial bias in data have involved soliciting homeless persons of color, further compromising their vulnerability (Fussell, 2019). When participation is due to economic, or situational, vulnerability, as opposed to one's best interests, monetary offerings may be perceived as inducement (Gordon, 2020). Further ethical concerns manifest when it is unclear whether participants were adequately *informed* about a research study. For instance, research on ethnicity recognition, despite obtaining informed consent, received criticism for training a model that discriminates between Chinese Uyghur, Korean, and Tibetan faces, considering the human rights violations against Chinese Uyghurs (Cunrui et al., 2019). Although the study's focus is on the technology itself (Tech Inquiry, 2019), its potential use in enhancing surveillance on Uyghur faces raises ethical questions (Van Noorden, 2020).

Consent revocation. Dataset creators sometimes view autonomy as a challenge to collecting biometric data for HCCV, especially when data subjects prioritize privacy (Scheuerman et al., 2021; Meng et al., 2006; Singh et al., 2010). Nonetheless, informed consent emphasizes *voluntariness*, encompassing both the ability to give consent and the right to withdraw it at any time (Dankar et al., 2019). GDPR grants explicit revocation rights (Article 7) and the right to request erasure of personal data (Article 17) (Whitley, 2009). However, image subjects whose data is collected without consent are denied these rights. The nonconsensual FFHQ face dataset (Karras et al., 2019) offers an opt-out mechanism, but since inclusion was involuntary, subjects may be unaware of their inclusion, rendering the revocation option hollow. Moreover, this burdens data subjects with tracking the usage of their data in datasets, primarily accessible by approved researchers (Dulhanty, 2020).

Image- and metadata-level privacy attributes. Researchers have focused on obfuscation techniques, e.g., blurring, inpainting, and overlaying, to reduce private information leakage of nonconsensual individuals (Xu et al., 2021; Frome et al., 2009; Uittenbogaard et al., 2019; Caesar et al., 2020; Piergiovanni & Ryoo, 2020; Yang et al., 2022a; Li et al., 2021; Sun et al., 2018; Li & Lin, 2019; McPherson et al., 2016). Face detection algorithms used in obfuscation may raise concerns, particularly if they involve predicting facial landmarks, potentially violating BIPA (Yew & Xiang, 2022; Complaint, Vance v. IBM, 2020). BIPA focuses on collecting and using face geometry scans regardless of identification capability, while GDPR protects any identifiable person, requiring data holders to safeguard the privacy of nonconsenting individuals. However, reliance on automated face detection methods raises ethical concerns, as demonstrated by the higher precision of pedestrian detection models on lighter skin types compared to darker skin types (Wilson et al., 2019). This predictive inequity leads to allocative harm, denying certain groups opportunities and resources, including the rights to safety (Twigg, 2003) and privacy (Diggelmann & Cleis, 2014). In addition, face obfuscation methods may not guarantee privacy (Yang et al., 2022a). The Visual Redactions dataset (Orekondy et al., 2018) includes 68 image-level privacy attributes, covering biometrics, sensitive attributes, tattoos, national identifiers, signatures, and contact information. Training faceless person recognition systems using full-body cues reveals higher than chance re-identification rates for face blurring and overlaying (Oh et al., 2016), indicating that simply perturbing or removing face information may be insufficient under GDPR. Furthermore, image metadata can disclose sensitive details, e.g., date, time, and location, as well as copyright information that may include names (Andrews, 2021; Oh et al., 2016). This is worrisome for users of commonly targeted platforms like Flickr, which retains metadata by default.

4.2. Practical recommendations

Obtain voluntary informed consent. Similar to recent consent-driven HCCV datasets (Hazirbas et al., 2021; Porgali et al., 2023; Rojas et al., 2022), explicit informed consent should be obtained from each person depicted in, or otherwise identifiable, in a dataset, allowing the sharing of their facial, body, and biometric information for evaluating the fairness and robustness of HCCV technologies. Datasets collected with consent *reduce* the risk of being fractured, however, data subjects may later revoke their consent over, e.g., privacy concerns they may not have been aware of at the time of providing consent or language nuances (Corrigan, 2003; Zimmer, 2010). Thus, following GDPR (Article 7), plain language consent and notice forms are recommended to address the lack of public understanding of AI technologies (Long & Magerko, 2020). When collecting images of individuals under the age of majority or those whose

ability to protect themselves is significantly impaired on account of disability, illness, or otherwise, guardian consent is necessary (Klima et al., 2014). However, relying solely on guardian consent overlooks the views and dignity of the vulnerable person (Henkelman & Everall, 2001). To address this, in addition to guardian consent, voluntary informed *assent* can be sought from a vulnerable person, in accordance with UNICEF's principlism-guided data collection procedures (UNICEF et al., 2015; Berman & Albright, 2017). When employing appropriate language and tools, assent establishes the vulnerable person understands the use of their data and willingly participates (Berman & Albright, 2017). If a vulnerable person expresses dissent or unwillingness to participate, their data should not be collected, regardless of guardian wishes. Informed by the U.S. National Bioethics Advisory Commission's *contextual vulnerability* framework (Commission et al., 2001), dataset creators should assess vulnerability on a continuous scale. That is, the circumstances of participation should be considered, which may require, e.g., a participatory design approach, assurances over compensation, supplementary educational materials, and insulation from hierarchical or authoritative systems (Gordon, 2020).

Adopt techniques for consent revocation. To permit consent revocation, dataset creators should implement an appropriate mechanism. One option is *dynamic consent*, where personalized communication interfaces enable participants to engage more actively in research activities (Kaye et al., 2015; Weber et al., 2014). This approach has been implemented successfully through online platforms, offering options for blanket consent, case-by-case selection, or opt-in depending on the data's use (Kaye et al., 2015; Mascalzoni et al., 2022; Teare et al., 2021). Another suggested option is to establish a steering board or charitable trust composed of representative dataset participants to make decisions regarding data use (Price & Cohen, 2019). The feasibility of these techniques may vary based on the dataset's scale. However, at a minimum, data subjects should be provided a simple and easily accessible one-step method to revoke consent (Hazirbas et al., 2021; Porgali et al., 2023; Rojas et al., 2022). As underscored by the UK Information Commission's Office, data subjects should be provided alternatives to online-based revocation processes to account for varying levels of technology competency and internet access (UK Information Commissioner's Office, n.d.).

Collect country of residence information. Anonymizing nonconsensual persons through face obfuscation, as done in datasets like ImageNet (Yang et al., 2022a), may not respect the privacy laws specific to the subjects' country of residence. To comply with relevant data protection laws, dataset curators should collect the country of residence from each data subject to determine their legal obligations, helping to ensure that data subjects' rights are protected and future legislative changes are addressed (Rojas et al., 2022; Phillips, 2018). For instance, GDPR Article 7(3) grants data subjects the right to withdraw consent at any time, which was not explicitly addressed in its predecessor (Politou et al., 2018).

Redact privacy leaking image regions and metadata. According to the European Data Protection Board (Article 29 Working Party), anonymization of personal data requires safeguards against re-identification risks, e.g., singling out, linkability, and inference (Data Protection Commission, 2019). As re-identification of nonconsensual human subjects whose faces have been obfuscated can still occur through other body parts or contextual information (Orekondy et al., 2018), one solution is therefore to completely redact all privacy-leaking regions. That is, the removal of regions related to nonconsensual image subjects (including their entire bodies, clothing, and accessories) and text (excluding copyright owner information). Anonymization approaches should be validated empirically, especially when using methods without formal privacy guarantees. Moreover, to mitigate algorithmic failures or biases, human annotators should be involved in creating region proposals, as well as verifying automatically generated proposals, for anonymizing image regions with identifying or private information (Yang et al., 2022a). It is important to note that for residents of certain jurisdictions (e.g., Illinois, California, Washington, and Texas), automated region proposals requiring biometric identifiers for nonconsensual subjects should be avoided, and human annotators should generate the proposals. To further protect privacy, dataset creators should take steps to ensure that image metadata does not reveal identifying information that data subjects did not consent to sharing. This includes removing or replacing exact geolocation with a more general representation (e.g., city and country) and removing user-added information from metatags if it includes personal identifying details, as long as it does not violate copyright. However, we do not advise blanket redaction of all metadata, as it contains valuable image capture information that can be useful for assessing model bias and robustness related to instrument factors.

5. Diversity

HCCV dataset creators widely acknowledge the significance of diversity, realism, and difficulty in datasets to enhance fairness and robustness in real-world applications (Karkkainen & Joo, 2021; Lin et al., 2014; Deng et al., 2009; Karras et al., 2019; Kay et al., 2017; Andriluka et al., 2014; Cordts et al., 2016; Sarkar et al., 2005; Xiong et al., 2015; Yang et al., 2016; Huang et al., 2008; Jesorsky et al., 2001; Dalal & Triggs, 2005; Angelova et al., 2005; Everingham et al., 2010; Liu et al., 2015; Geiger et al., 2012). Previous research (Buolamwini & Gebru, 2018; Liu et al.,

2020; Mitchell et al., 2019; Scheuerman et al., 2021) has emphasized diversity across image subjects, environments, and instruments, but there are many ethical complexities involved in specifying diversity criteria (Andrus et al., 2020; 2021). This section examines taxonomy challenges for these attributes and offers recommendations.

5.1. Ethical considerations

Representational and historical biases. The Council of Europe have expressed concerns about the threat posed by AI systems to equality and non-discrimination principles (Council of Europe, n.d.). Many dataset creators often prioritize protected attributes, i.e., gender, race, and age, as key factors of dataset diversity (Scheuerman et al., 2021). Nevertheless, most HCCV datasets exhibit historical and representational biases (Suresh & Guttag, 2021; Jo & Gebru, 2020; Yang et al., 2020; Kay et al., 2015; Prabhu & Birhane, 2021). These biases can be pernicious, particularly when models learn and reinforce them. For instance, image captioning models may rely on contextual cues related to activities like shopping (Zhao et al., 2017) and laundry (Zhao et al., 2023) to generate gendered captions. Spurious correlations are detrimental, as they are not causally related and perpetuate harmful associations (Sagawa et al., 2020; Geirhos et al., 2020). In addition, prominent examples in HCCV research demonstrate disparate algorithmic performance based on race and skin color (Grother et al., 2019; Vangara et al., 2019; Hirota et al., 2022; Zhao et al., 2021; Phillips et al., 2011; Buolamwini & Gebru, 2018; Buolamwini, n.d.; Snow, 2018; Rose, 2010; Chen, 2009; Hern, 2020). Most recently, autonomous robots have displayed racist, sexist, and physiognomic stereotypes (Hundt et al., 2022). Furthermore, face detection models have shown lower accuracy when processing images of older individuals compared to younger individuals (Yang et al., 2022b). While not endorsing these applications, discrepancies have been observed in facial emotion recognition services for children in both commercial and research systems (Howard et al., 2017; Xu et al., 2020), as well as age estimation (Lou et al., 2017; Clapés et al., 2018; Georgopoulos et al., 2020).

Despite concerns regarding privacy, liability, and public relations, the collection of special and sensitive category data is crucial for bias assessments (Andrus et al., 2021). GDPR guidance from the UK Information Commissioner's Office confirms that sensitive attributes can be collected for fairness purposes (UK Information Commissioner's Office, 2020). However, obtaining this information presents challenges, e.g., historical mistrust in clinical research among African-Americans (Eyal, 2014; Lee & Rich, 2021). In HCCV, there is a tension between privacy and fairness, as remaining unseen does not protect against being mis-seen (Xiang, 2022). Nonetheless, marginalized communities may require explicit explanations and assurances about data usage to address concerns related to service provision, security, allocation, and representation (Xiang, 2022).

The digital divide and accessibility. Healthcare datasets often lack representation of minority populations, compromising the reliability of automated decisions (World Health Organization and others, 2021). The World Health Organization (WHO) emphasizes the need for data accuracy, completeness, and diversity, particularly regarding age, in order to address ageism in AI (World Health Organization, 2022). ML systems may prioritize younger populations for resource allocation, assuming they would benefit the most in terms of life expectancy (World Health Organization, 2022). The digital divide further exacerbates the underrepresentation of vulnerable groups, including older generations, low-income school-aged children, and children in East Asia and the South Pacific who lack access to digital technology and learning opportunities (World Economic Forum, 2022; UNICEF, 2020). Insufficient access to digital technology hampers the representation of vulnerable individuals in datasets (Schumann et al., 2021).

Confused taxonomies. Sex and gender are often used interchangeably, equating gender as a consequence of one's assigned sex at birth (Fausto-Sterling, 2000). However, this approach erases intersex individuals who possess non-binary physiological sex characteristics (Fausto-Sterling, 2000). Treating sex and gender as interchangeable perpetuates normative views by casting gender as binary, immutable, and solely based on biological sex (Keyes, 2018). This perspective disregards transgender and gender nonconforming individuals. Moreover, sex, like gender, is a social construct, as sexed bodies do not exist outside of their social context (Butler, 1988). Similar to sex and gender, race and ethnicity are often used synonymously (Valentine et al., 2016). Nations employ diverse census questions to ascertain ethnic group composition, encompassing factors, e.g., nationality, race, color, language, religion, customs, and tribe (United Nations, 1998). However, these categories and their definitions lack consistency over time and geography, often influenced by political agendas and socio-cultural shifts (Scheuerman et al., 2020b). As a result, collecting globally representative and meaningful data on ethnic groups becomes challenging. Several HCCV datasets have incorporated inconsistent and arbitrary racial categorization systems (Wang et al., 2019; Zhang et al., 2017; Robinson et al., 2020; Alvi et al., 2018). For instance, the FairFace dataset (Karkkainen & Joo, 2021) creators reference the U.S. Census Bureau's racial categories without considering the social definition of race they represent (OpenReview, 2019). The U.S. Census Bureau explicitly states that their categories reflect a social definition rather than a biological, anthropological, or genetic one. Consequently, labeling the "physical race" of image subjects based on nonphysiological categories is contradictory. Furthermore, the FairFace creators do not disclose the

demographics or cultural compatibility of their annotators.

Own-anchor bias. HCCV approaches for encoding age in datasets vary, using either integer labels (Ricanek & Tesafaye, 2006; Guo et al., 2008; Fu et al., 2007; Rothe et al., 2018; 2015; Chen et al., 2014; Niu et al., 2016; Zhang et al., 2017; Moschoglou et al., 2017) or group labels (Gallagher & Chen, 2009; Somanath et al., 2011; Eidinger et al., 2014; Levi & Hassner, 2015). Age groupings are often preferred when collecting unconstrained images from the web, as human annotators must infer subjects' ages, which is challenging (Carcagnì et al., 2015). This is evident in crowdsourced annotations, where 40.2% of individuals in the OpenImages MIAP dataset (Schumann et al., 2021) could not be categorized into an age group. Factors unrelated to age, e.g., facial expression (Ganel, 2015; Wang et al., 2015; Norja et al., 2022) and makeup (Tagai et al., 2016; Egan & Cordan, 2009; Norja et al., 2022), influence age perception. Furthermore, annotators have exhibited lower accuracy when labeling people outside of their own demographic group (Anastasi & Rhodes, 2005; 2006; Sörqvist et al., 2011; Vestlund et al., 2009; Voelkle et al., 2012; George & Hole, 1995; Rowe, 2001).

Post hoc rationalization of the use of physiological markers. Gender information about data subjects is obtained through inference (Karkkainen & Joo, 2021; Wang et al., 2019; Zhang et al., 2017; Schumann et al., 2021; Rothe et al., 2018; 2015; Zhang et al., 2017; Niu et al., 2016; Chen et al., 2014; Kumar et al., 2009; Liu et al., 2015; Ricanek & Tesafaye, 2006) or self-identification (Ma et al., 2015; 2021; Hazirbas et al., 2021; Lakshmi et al., 2021; Zhao et al., 2021). Inference raises concerns as it assumes that gender can be determined solely from imagery without consent or consultation with the subject, which is noninclusive and harmful (Keyes, 2018; Hamidi et al., 2018; Engelmann et al., 2022). Even when combined with non-image-based information, inferred gender fails to account for the fluidity of identity, potentially mislabeling subjects at the time of image capture (Rothe et al., 2018; 2015). Moreover, physical traits are just one of many dimensions, including posture, clothing, and vocal cues, used to infer not only gender but also race (Kessler & McKenna, 1985; Freeman et al., 2011).

Erasure of nonstereotypical individuals. HCCV datasets often adopt a U.S.-based racial schema (Karkkainen & Joo, 2021; Wang et al., 2019; Ma et al., 2015; Lakshmi et al., 2021; Ma et al., 2021; Ricanek & Tesafaye, 2006), which can create disjoint and essentialized groups (Telles, 2002). However, this schema may not align with other models, e.g., the continuum-based color system used in Brazil, which considers a range of physical characteristics. Nonconsensual image datasets rely on annotators to assign semantic categories, perpetuating stereotypes and disseminating them beyond their cultural context (Khan & Fu, 2021). Notably, images without label consensus are often discarded (Karkkainen & Joo, 2021; Wang et al., 2019; Robinson et al., 2020), potentially excluding individuals who defy stereotypes, e.g., multi-ethnic individuals (Rothbart & Taylor, 1992).

Phenotypic attributes. Protected attributes may not be the most appropriate criteria for evaluating HCCV models (Buolamwini & Gebru, 2018). Social constructs like race and gender lack clear delineations for subgroup membership based on visible or invisible characteristics. These labels capture invisible aspects of identity that are not solely determined by visible appearance. Moreover, the phenotypic characteristics within and across subgroups exhibit significant variability (Becerra-Riera et al., 2019; Carcagnì et al., 2015; Feliciano, 2016; Khan & Fu, 2021; Ware et al., 2020).

Environment and instrument. The image capture device and environmental conditions significantly influence model performance, and their impact should be considered (Mitchell et al., 2019). Factors such as camera software, hardware, and environmental conditions affect HCCV model robustness in various settings (Windrim et al., 2016; Nascimento et al., 2018; Xie et al., 2019; Xu & Wang, 2019; Liu et al., 2020; Hendrycks & Dietterich, 2019; Afifi & Brown, 2019; Yin et al., 2019; Mintun et al., 2021). Understanding performance differences is crucial from ethical and scientific perspectives. For example, sensitivity to illumination or white balance may be linked to sensitive attributes, e.g., skin tone (Zhou et al., 2018; Cook et al., 2019; Kortylewski et al., 2018; 2019), while available instruments or environmental co-occurrences may correlate with demographic attributes (Silver, 2020; Hendricks et al., 2018).

Annotator positionality. Psychological research highlights the influence of annotators' sociocultural background on their visual perception (Reid & Nixon, 2011; Balaresque & King, 2016; Roth, 2016; Garcia & Abascal, 2016; Hill, 2002; Segall et al., 1966; Balaresque & King, 2016). However, recent empirical studies have evidenced a lack of regard for the impact an annotator's social identity has on data (Denton et al., 2021; Geiger et al., 2020) with only a handful of HCCV datasets providing annotator demographic details (Scheuerman et al., 2021; Chen & Joo, 2021; Zhao et al., 2021; Andrews et al., 2023).

Recruitment and compensation. Data collected without consent patently lacks compensation. Balancing between excessive and deficient payment is crucial to avoid coercion or exploitation (National Health and Medical Research Council, 2019; Rojas et al., 2022). An additional concern is the employment of remote workers from disadvantaged regions (Perrigo, 2022), often with low wages and fast-paced work conditions (Croce & Musa, 2019; Hata et al., 2017; Irani, 2015; Malevé, 2020). This can lead to arbitrary denial of payment based on opaque quality criteria (Fieseler et al., 2019) and prevents union formation (Malevé, 2020), creat-

ing a sense of invisibility and uncertainty for workers (Toxtli et al., 2021).

5.2. Practical recommendations

Obtain self-reported annotations. Practitioners are cautious about inferring labels about people to avoid biases (Andrus et al., 2021). Moreover, data access request rights, e.g., as offered by GDPR, CCPA, and PIPL, may require data holders to disclose inferred information. To avoid stereotypical annotations and minimize harm from misclassification (Roth, 2016), accurate labels should be collected directly from image subjects, who possess contextual knowledge and awareness of their own attributes.

Provide open-ended response options. Closed-ended questions, such as those on census forms, may lead to incongruous responses and inadequate options for self-identification (Roth, 2012; Hughes et al., 2016; Keyes, 2018). Open-ended questions provide more accurate answers but can be taxing, require extensive coding, and are harder to analyze (Bradburn, 1997; Keusch, 2014; Smyth et al., 2009; Geer, 1991). To strike a balance, closed-ended questions should be augmented with an open-ended response option, avoiding the term "other", which implies *othering* norms (Scheuerman et al., 2020a). This gives respondents a voice (Singer & Couper, 2017; Neuert et al., 2021) and allows for future question design improvement.

Acknowledge the mutability and multiplicity of identity. The concept of identity shift is often overlooked, i.e., the intentional self-transformation in mediated contexts (Carr et al., 2021). To address this, we propose collecting self-identified information on a per-image basis, recognizing that identity is temporal and non-static. In addition, particularly for sensitive attributes, the selection of multiple identity categories without limitations is preferable (Spiel et al., 2019; Stevens, n.d.).

Collect age, pronouns, and ancestry. First, to capture accurate age information, dataset curators should collect the exact biological age in years from image subjects, corresponding to their age at the time of image capture. This approach offers flexibility, insofar as permitting the appropriate disaggregation of the collected data. This is important given the lack of consistent age groupings in the literature. Second, dataset curators should consider opting to collect self-identified pronouns. This promotes mutual respect and common courtesy, reducing the likelihood of causing harm through misgendering (Human Rights Campaign Foundation, n.d.). Self-identified pronouns are particularly important for sexual and gender minority communities as they "convey and affirm gender identity" (National Institutes of Health – Division of Program Coordination, Planning and Strategic Initiatives, 2022). Significantly, pronoun use is increasingly prevalent in social media platforms (Jiang et al., 2022; Joshi, 2019; Elks, 2021), workplaces (Chen, 2021), and education settings (Barlow & Scott, 2022; McKie, 2018), fostering gender inclusivity (Baron, 2020). However, subjects should always have the option of not disclosing this information. Finally, to address issues with ethnic and racial classification systems (Scheuerman et al., 2020b; Khan & Fu, 2021), dataset creators should consider collecting ancestry information instead. Ancestry is defined by historically shaped borders and has been shown to offer a more stable and less confusing concept (Aspinall, 2001). The United Nations' M49 geoscheme can be used to operationalize ancestry (United Nations Statistics Division, n.d.), where subjects select regions that best describe their ancestry. To situate responses, subjects could be asked, e.g., "Where do your ancestors (e.g., great-grandparents) come from?". Proxies, e.g., skin tone, risk normalizing their inadequacies without reflecting their limitations (Andrus et al., 2021).

Collect aggregate data for commonly ignored groups. Additional sensitive attributes should also be collected, e.g., disability and pregnancy status, when voluntarily disclosed by subjects. These attributes should be reported in aggregate data to reduce the safety concerns of subjects (Stevens, n.d.; Whittaker et al., 2019). Given that definitions of these attributes may be inconsistent and tied to culture, identity, and histories of oppression (Blaser & Ladner, 2020; Bragg et al., 2021), navigating tensions between benefits and risks is necessary. Despite potential reluctance, sourcing data from underrepresented communities contributes to dataset inclusivity (Blaser & Ladner, 2020; Kamikubo et al., 2021). For disability, the American Community Survey (United States Census Bureau, 2021) covers categories related to hearing, vision, cognitive, ambulatory, self-care, and independent living difficulties.

Collect phenotypic and neutral performative features. Collecting phenotypic characteristics can serve as *objective* measures of diversity, i.e., attributes which, in evolutionary terms, contribute to individual-level recognition (Christakis & Fowler, 2014), e.g., skin color, eye color, hair type, hair color, height, and weight (Balaresque & King, 2016). These attributes have enabled finer-grained analysis of model performance and biases (Wen et al., 2022; Buolamwini & Gebru, 2018; Dash et al., 2022; Siddiqui et al., 2022; Yucer et al., 2022). Additionally, considering a multiplicity of neutral performative features, e.g., facial hair, hairstyle, cosmetics, clothing, and accessories, is important to surface the perpetuation of social stereotypes and spurious relationships in trained models (Scheuerman et al., 2019; Balakrishnan et al., 2021; Wang et al., 2022; Albiero et al., 2021).

Record environment and instrument information. Data should capture variations in environmental conditions and imaging devices, e.g., image capture time, season, weather,

ambient lighting, scene, geography, camera position, distance, lens, sensor, stabilization, use of flash, and post-processing software. Instrument-related factors may be easily captured, by restricting data collection to images with Exif metadata. The remaining factors, e.g., season and weather can be self-reported or coarsely estimated utilizing information such as image capture time and location.

Recontextualize annotators as contributors. Dataset creators should document the identities of annotators and their contributions to the dataset (Denton et al., 2021; Andrews et al., 2023), rather than treating them as anonymous entities responsible for data labeling alone (Malevé, 2020; Chancellor et al., 2019). While many datasets (Lin et al., 2014; Deng et al., 2009; Hazirbas et al., 2022) neglect to report annotator demographics, assuming objectivity in annotation for visual categories is flawed (Kapania et al., 2023; Miceli et al., 2020; Barrett et al., 2023). Furthermore, using majority voting to reach the assumed ground truth, disregards minority opinions, treating them as noise (Kapania et al., 2023). Annotator characteristics, including pronouns, age, and ancestry, should be recorded and reported to quantify and address annotator perspectives and bias in datasets (Gordon et al., 2022; Andrews et al., 2023). Additionally, allowing annotators freedom in labeling helps to avoid replicating socially dominant viewpoints (Miceli et al., 2020).

Fair treatment and compensation for contributors. Following guidance from Australia's National Health and Medical Research Council (National Health and Medical Research Council, 2019) and the WHO (Council for International Organizations of Medical Sciences and others, 2017), dataset contributors should not only be guaranteed compensation above the minimum hourly wage of their country of residence (Różyńska, 2022), but also according to the complexity of tasks to be performed. An alternative payment model based on the average hourly wage, however, may better promote justice and diversity by increasing the likelihood of higher socio-economic status contributors (Phillips, 2011). Besides payment, the implementation of direct communication channels and feedback mechanisms, e.g., anonymized feedback forms (Pavlichenko et al., 2021), can help to address issues faced by annotators while providing a level of protection from retribution. Complementarily, the creation of plain language guides can ease task completion and reduce quality control overheads. Ideally, recruitment and compensation processes should be well-documented and undergo ethics review, which can help to further reduce "glaring ethical lapses" (Shmueli et al., 2021).

6. Conclusion

Building upon recent scholarship addressing privacy and bias concerns, we have highlighted key ethical considerations and challenges in collecting HCCV data for fairness and robustness evaluations. Guided by principlism, we have concentrated on purpose, consent and privacy, as well as diversity, offering proactive recommendations that prioritize autonomy, beneficence, non-maleficence, and justice. While our recommendations hold broader relevance, we have placed specific emphasis on the distinctive attributes of HCCV datasets. We therefore encourage dataset creators to tailor these recommendations to suit their particular domain and task, fostering further discussions around responsible data curation.

Acknowledgements

This work was funded by Sony Research.

References

Afifi, M. and Brown, M. S. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

Afzal, S., Rajmohan, C., Kesarwani, M., Mehta, S., and Patel, H. Data readiness report. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pp. 42–51. IEEE, 2021.

Albiero, V., Zhang, K., King, M. C., and Bowyer, K. W. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021.

Alvi, M., Zisserman, A., and Nellåker, C. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision Workshops (ECCVW)*, pp. 0–0, 2018.

Anastasi, J. S. and Rhodes, M. G. An own-age bias in face recognition for children and older adults. *Psychonomic bulletin & review*, 12(6):1043–1047, 2005.

Anastasi, J. S. and Rhodes, M. G. Evidence for an own-age bias in face recognition. *North American Journal of Psychology*, 8(2), 2006.

Andrews, J. The hidden fingerprint inside your photos. https://www.bbc.com/future/article/20210324-the-hidden-fingerprint-inside-your-photos, 2021. Accessed June 30, 2022.

Andrews, J. T. A., Joniak, P., and Xiang, A. A view from somewhere: Human-centric face representations. In *International Conference on Learning Representations (ICLR)*, 2023.

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3686–3693, 2014.

Andrus, M., Spitzer, E., and Xiang, A. Working to address algorithmic bias? don't overlook the role of demographic data. *Partnership on AI*, 2020.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 249–260, 2021.

Angelova, A., Abu-Mostafam, Y., and Perona, P. Pruning training sets for learning of object categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 494–501, 2005.

Aspinall, P. J. Operationalising the collection of ethnicity data in studies of the sociology of health and illness. *Sociology of health & illness*, 23(6):829–862, 2001.

Balakrishnan, G., Xiong, Y., Xia, W., and Perona, P. Towards causal benchmarking of bias in face analysis algorithms. In *Deep Learning-Based Face Analytics*, pp. 327–359. Springer, 2021.

Balaresque, P. and King, T. Human phenotypic diversity. In *Genes and Evolution*, pp. 349–390. Elsevier, 2016. doi: 10.1016/bs.ctdb.2016.02.001.

Barlow, R. and Scott, C. Students can adjust their pronouns and gender identity in bu's updated data system, Nov 2022. URL https://www.bu.edu/articles/2022/pronouns-and-gender-identities-in-updated-data-system/.

Baron, D. *What's Your Pronoun?: Beyond He and She*. Liveright Publishing, 2020.

Barr, A. Google mistakenly tags Black people as 'gorillas,' showing limits of algorithms. *The Wall Street Journal*, 2015. URL https://www.wsj.com/articles/BL-DGB-42522.

Barrett, T., Chen, Q. Z., and Zhang, A. X. Skin deep: Investigating subjectivity in skin tone annotations for computer vision benchmark datasets. *arXiv preprint arXiv:2305.09072*, 2023.

Beauchamp, T. and Childress, J. Principles of biomedical ethics: marking its fortieth anniversary, 2019.

Beauchamp, T. L. and Childress, J. F. *Principles of biomedical ethics*. Edicoes Loyola, 1994.

Becerra-Riera, F., Morales-González, A., and Méndez-Vázquez, H. A survey on facial soft biometrics for video surveillance and forensic applications. *Artificial Intelligence Review*, 52(2):1155–1187, 2019.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Beever, J. and Brightman, A. O. Reflexive principlism as an effective approach for developing ethical reasoning in engineering. *Science and engineering ethics*, 22:275–291, 2016.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, dec 2018. doi: 10.1162/tacl_a_00041.

Benthall, S. and Haynes, B. D. Racial categories in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 289–298, 2019.

Beretta, E., Vetrò, A., Lepri, B., and Martin, J. C. D. Detecting discriminatory risk through data annotation based on bayesian inferences. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 794–804, 2021.

Berman, G. and Albright, K. Children and the data cycle: Rights and ethics in a big data world. *arXiv preprint arXiv:1710.06881*, 2017.

Birhane, A. Algorithmic colonization of africa. *SCRIPTed*, 17:389, 2020.

Blaser, B. and Ladner, R. E. Why is data on disability so hard to collect and understand? In *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, volume 1, pp. 1–8. IEEE, 2020.

Bradburn, N. Respondent burden: health survey research methods. In *Second Biennial Conference, Williamsburg, VA. Washington, DC: US Government Printing Office*, 1997.

Bragg, D., Caselli, N., Hochgesang, J. A., Huenerfauth, M., Katz-Hernandez, L., Koller, O., Kushalnagar, R., Vogler, C., and Ladner, R. E. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)*, 14(2):1–45, 2021.

Buolamwini, J., n.d. URL https://www.media.mit.edu/projects/gender-shades/overview/.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 77–91. PMLR, 2018.

Butcher, B., Huang, V. S., Robinson, C., Reffin, J., Sgaier, S. K., Charles, G., and Quadrianto, N. Causal datasheet for datasets: An evaluation guide for real-world data analysis and data collection design using bayesian networks. *Frontiers in Artificial Intelligence*, 4:612551, 2021.

Butler, J. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal*, 40(4):519–531, 1988. ISSN 01922882, 1086332X. URL http://www.jstor.org/stable/3207893.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, 2020.

Campbell, M. E. and Troyer, L. The implications of racial misclassification by observers. *American Sociological Review*, 72(5):750–765, 2007.

Carcagnì, P., Coco, M. D., Cazzato, D., Leo, M., and Distante, C. A study on different experimental configurations for age, race, and gender estimation problems. *EURASIP Journal on Image and Video Processing*, 2015(1):1–22, 2015.

Carr, C. T., Kim, Y., Valov, J. J., Rosenbaum, J. E., Johnson, B. K., Hancock, J. T., and Gonzales, A. L. An explication of identity shift theory. *Journal of Media Psychology*, 2021.

Chambers, C. D. and Tzavella, L. The past, present and future of registered reports. *Nature human behaviour*, 6 (1):29–42, 2022.

Chancellor, S., Baumer, E. P., and De Choudhury, M. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW):1–32, 2019.

Chen, B.-C., Chen, C.-S., and Hsu, W. H. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision (ECCV)*, pp. 768–783. Springer, 2014.

Chen, B. X. Hp investigates claims of 'racist' computers, Dec 2009. URL https://www.wired.com/2009/12/hp-notebooks-racist/.

Chen, T.-P. Why gender pronouns are becoming a big deal at work. *The Wall Street Journal. Retrieved October*, 15: 2022, 2021.

Chen, Y. and Joo, J. Understanding and mitigating annotation bias in facial expression recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14980–14991, 2021.

Christakis, N. A. and Fowler, J. H. Friendship and natural selection. *Proceedings of the National Academy of Sciences*, 111(supplement_3):10796–10801, 2014.

Clapés, A., Bilici, O., Temirova, D., Avots, E., Anbarjafari, G., and Escalera, S. From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2373–2382, 2018.

Code, N. The nuremberg code. *Trials of war criminals before the Nuremberg military tribunals under control council law*, 10(2):181–182, 1949.

Commission, N. B. A. et al. Ethical and policy issues in research involving human participants. 2001.

Complaint, Vance v. IBM. U.s. dist. lexis 168610, 2020 wl 5530134 (united states district court for the northern district of illinois, eastern division, january 14, 2020, filed), 2020.

Cook, C. M., Howard, J. J., Sirotin, Y. B., Tipton, J. L., and Vemury, A. R. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, jan 2019. doi: 10.1109/tbiom.2019.2897801.

Cooper, A. F., Abrams, E., and Na, N. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 46–54, 2021.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

Corrigan, O. Empty ethics: the problem with informed consent. *Sociology of Health & Illness*, 25(7):768–792, 2003.

Council for International Organizations of Medical Sciences and others. International ethical guidelines for health-related research involving humans. *International ethical guidelines for health-related research involving humans.*, 2017.

Council of Europe. Inclusion and anti-discrimination: Ai & discrimination. https://www.coe.int/en/web/inclusion-and-antidiscrimination/ai-and-discrimination, n.d. Accessed November 24, 2022.

Crawford, K. and Schultz, J. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55:93, 2014.

Croce, N. and Musa, M. The new assembly lines: Why ai needs low-skilled workers too, Aug 2019. URL https://www.weforum.org/agenda/2019/08/ai-low-skilled-workers/.

Cunrui, W., Zhang, Q., Liu, W., Liu, Y., and Miao, L. Facial feature discovery for ethnicity recognition. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(2):e1278, 2019.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 886–893. Ieee, 2005.

Dankar, F. K., Gergely, M., and Dankar, S. K. Informed consent in biomedical research. *Computational and structural biotechnology journal*, 17:463–474, 2019.

Dash, S., Balasubramanian, V. N., and Sharma, A. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 915–924, 2022.

Data Protection Commission, June 2019. URL https://www.dataprotection.ie/en/dpc-guidance/anonymisation-and-pseudonymisation. Accessed August 1, 2022.

De Hert, P. and Papakonstantinou, V. The new general data protection regulation: Still a sound system for the protection of individuals? *Computer law & security review*, 32(2):179–194, 2016.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

Diggelmann, O. and Cleis, M. N. How the right to privacy became a human right. *Human Rights Law Review*, 14 (3):441–458, 2014.

Dulhanty, C. Issues in computer vision data collection: Bias, consent, and label taxonomy. Master's thesis, University of Waterloo, 2020.

Edwards, L. Privacy, security and data protection in smart cities: A critical eu law perspective. *Eur. Data Prot. L. Rev.*, 2:28, 2016.

Egan, V. and Cordan, G. Barely legal: Is attraction and estimated age of young female faces disrupted by alcohol use, make up, and the sex of the observer? *British Journal of Psychology*, 100(2):415–427, 2009.

Eidinger, E., Enbar, R., and Hassner, T. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

Elks, S. Why twitter and instagram are inviting people to share their pronouns, Oct 2021. URL https://www.context.news/big-tech/why-twitter-and-instagram-are-inviting-people-to-share-pronouns.

Engelmann, S., Ullstein, C., Papakyriakopoulos, O., and Grossklags, J. What people think ai should infer from faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 128–141, 2022.

European Commission. General data protection regulation. https://gdpr-info.eu/, 2016. Accessed August 1, 2022.

European Data Protection Board (Article 29 Working Party). The workiung party on the protection of individuals with regard to the processing of personal data. https://ec.europa.eu/newsroom/document.cfm?doc_id=44137, 2017. Accessed August 1, 2022.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2):303–338, 2010.

Eyal, N. Using informed consent to save trust. *Journal of medical ethics*, 40(7):437–444, 2014.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022a.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. Tackling documentation debt: a survey on algorithmic fairness datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13. 2022b.

Fausto-Sterling, A. *Sexing the body: Gender politics and the construction of sexuality*. Basic books, 2000.

Feliciano, C. Shades of race: How phenotype and observer characteristics shape racial classification. *American Behavioral Scientist*, 60(4):390–419, 2016.

Fiedler, K. and Schwarz, N. Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1):45–52, 2016.

Fieseler, C., Bucher, E., and Hoffmann, C. P. Unfairness by design? the perceived fairness of digital labor on crowdworking platforms. *Journal of Business Ethics*, 156:987–1005, 2019.

Founds, A. P., Orlans, N., Genevieve, W., Watson, C. I., et al. Nist special databse 32-multiple encounter dataset ii (meds-ii). 2011.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., and Ambady, N. Looking the part: Social status cues shape race perception. *PloS one*, 6(9):e25107, 2011.

Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., and Vincent, L. Large-scale privacy protection in google street view. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2373–2380, 2009.

Fu, Y., Xu, Y., and Huang, T. S. Estimating human age by manifold analysis of face pictures and regression on aging features. In *2007 IEEE International Conference on Multimedia and Expo*, pp. 1383–1386. IEEE, 2007.

Fussell, S. How an attempt at correcting bias in tech goes wrong. https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/, 2019. Accessed June 30, 2022.

Gallagher, A. C. and Chen, T. Understanding images of groups of people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 256–263, 2009.

Ganel, T. Smiling makes you look older. *Psychonomic bulletin & review*, 22(6):1671–1677, 2015.

Garcia, D. and Abascal, M. Colored perceptions: Racially distinctive names and assessments of skin color. *American Behavioral Scientist*, 60(4):420–441, 2016.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for datasets. 2018.

Geer, J. G. Do open-ended questions measure "salient" issues? *Public Opinion Quarterly*, 55(3):360–370, 1991.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 325–336, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

George, P. A. and Hole, G. J. Factors influencing the accuracy of age estimates of unfamiliar faces. *Perception*, 24 (9):1059–1073, 1995.

Georghiades, A. S., Belhumeur, P. N., and Kriegman, D. J. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (6):643–660, 2001.

Georgopoulos, M., Panagakis, Y., and Pantic, M. Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and vision computing*, 102: 103954, 2020.

Google PAIR. Google pair. people + ai guidebook. https://pair.withgoogle.com/guidebook, 2019. Accessed February 1, 2023.

Gordon, B. G. Vulnerability in research: basic ethical concepts and general approach to review. *Ochsner Journal*, 20(1):34–38, 2020.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., and Bernstein, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *Conference on Human Factors in Computing Systems (CHI)*, pp. 1–19, 2022.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Grgic, M., Delac, K., and Grgic, S. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

Grother, P. J., Ngan, M. L., Hanaoka, K. K., et al. Face recognition vendor test part 3: demographic effects. 2019.

Günther, M., Hu, P., Herrmann, C., Chan, C.-H., Jiang, M., Yang, S., Dhamija, A. R., Ramanan, D., Beyerer, J., Kittler, J., et al. Unconstrained face detection and open-set face recognition challenge. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 697–706. IEEE, 2017.

Guo, G., Fu, Y., Dyer, C. R., and Huang, T. S. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, pp. 87–102. Springer, 2016.

Hamidi, F., Scheuerman, M. K., and Branham, S. M. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Conference on Human Factors in Computing Systems (CHI)*, pp. 1–13, 2018.

Han, H., Jain, A. K., Wang, F., Shan, S., and Chen, X. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609, 2017.

Hanley, M., Khandelwal, A., Averbuch-Elor, H., Snavely, N., and Nissenbaum, H. An ethical highlighter for people-centric dataset creation. 2020.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 501–512, 2020.

Harvey, A. and LaPlace, J. Exposing. ai, 2021.

Hata, K., Krishna, R., Fei-Fei, L., and Bernstein, M. S. A glimpse far into the future: Understanding long-term crowd worker quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 889–901, 2017.

Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., and Ferrer, C. C. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

Hazirbas, C., Bang, Y., Yu, T., Assar, P., Porgali, B., Albiero, V., Hermanek, S., Pan, J., McReynolds, E., Bogen, M., Fung, P., and Ferrer, C. C. Casual conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness, 2022.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, pp. 771–787, 2018.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Henkelman, J. J. and Everall, R. D. Informed consent with children: Ethical and practical implications. *Canadian Journal of Counselling and Psychotherapy*, 35(2), 2001.

Hern, A. Twitter apologises for 'racist' image-cropping algorithm, Sep 2020. URL https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm.

Hill, K. Wrongfully accused by an algorithm. *The New York Times*, 2020. URL https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hill, M. E. Race of the interviewer and perception of skin color: Evidence from the multi-city study of urban inequality. *American Sociological Review*, pp. 99–108, 2002.

Hirota, Y., Nakashima, Y., and Garcia, N. Quantifying societal bias amplification in image captioning. In *IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. 2018.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Conference on Human Factors in Computing Systems (CHI)*, pp. 1–16, 2019.

Howard, A., Zhang, C., and Horvitz, E. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pp. 1–7. IEEE, 2017.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Huang, H.-Y. and Liem, C. C. Social inclusion in curated contexts: Insights from museum practices. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 300–309, 2022.

Huang, Z., Liu, Y., Fang, Y., and Horn, B. K. Video-based fall detection for seniors with human pose estimation. In *2018 4th international conference on Universal Village (UV)*, pp. 1–4. IEEE, 2018.

Hughes, J. L., Camden, A. A., Yangchen, T., et al. Rethinking and updating demographic questions: Guidance to improve descriptions of research samples. *Psi Chi Journal of Psychological Research*, 21(3):138–151, 2016.

Human Rights Campaign Foundation. Talking about pronouns in the workplace, n.d. URL https://www.thehrcfoundation.org/professional-resources/talking-about-pronouns-in-the-workplace.

Hundt, A., Agnew, W., Zeng, V., Kacianka, S., and Gombolay, M. Robots enact malignant stereotypes. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 743–756, 2022.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 560–575, 2021.

IBM. Design for ai. https://www.ibm.com/design/ai, 2019. Accessed February 1, 2023.

Illinois Legislature. Biometric information privacy act. https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57, 2008. Accessed November 12, 2022.

Irani, L. The cultural work of microwork. *New media & society*, 17(5):720–739, 2015.

Janai, J., Güney, F., Behl, A., Geiger, A., et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020.

Jesorsky, O., Kirchberg, K. J., and Frischholz, R. W. Robust face detection using the hausdorff distance. In *Audio-and Video-Based Biometric Person Authentication: Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6–8, 2001 Proceedings 3*, pp. 90–95. Springer, 2001.

Jiang, J., Chen, E., Luceri, L., Murić, G., Pierri, F., Chang, H.-C. H., and Ferrara, E. What are your pronouns? examining gender pronoun usage on twitter. *arXiv preprint arXiv:2207.10894*, 2022.

Jo, E. S. and Gebru, T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2020.

Joshi, S. Why indians are sharing their pronouns on social media, Oct 2019. URL https://timesofindia.indiatimes.com/india/why-indians-are-sharing-their-pronouns-on-social-media/articleshow/71669703.cms.

Kamikubo, R., Dwivedi, U., and Kacorri, H. Sharing practices for datasets related to accessibility and aging. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–16, 2021.

Kapania, S., Taylor, A. S., and Wang, D. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2023.

Karkkainen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1548–1558, 2021.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Kay, M., Matuszek, C., and Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Conference on Human Factors in Computing Systems (CHI)*, pp. 3819–3828, 2015.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Kaye, J., Whitley, E. A., Lund, D., Morrison, M., Teare, H., and Melham, K. Dynamic consent: a patient interface for twenty-first century research networks. *European journal of human genetics*, 23(2):141–146, 2015.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., and Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4873–4882, 2016.

Kerr, N. L. Harking: Hypothesizing after the results are known. *Personality and social psychology review*, 2(3): 196–217, 1998.

Kessler, S. J. and McKenna, W. *Gender: An ethnomethodological approach*. University of Chicago Press, 1985.

Keusch, F. The influence of answer box format on response behavior on list-style open-ended questions. *Journal of Survey Statistics and Methodology*, 2(3):305–322, 2014.

Keyes, O. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

Khan, Z. and Fu, Y. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 587–597, 2021.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7:81–99, 2003.

Klima, J., Fitzgerald-Butt, S. M., Kelleher, K. J., Chisolm, D. J., Comstock, R. D., Ferketich, A. K., and McBride, K. L. Understanding of informed consent by parents of children enrolled in a genetic biobank. *Genetics in Medicine*, 16(2):141–148, 2014.

Koch, B., Denton, E., Hanna, A., and Foster, J. G. Reduced, reused and recycled: The life of a dataset in machine learning research. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., and Vetter, T. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2093–2102, 2018.

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., and Vetter, T. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Kugler, M. B. From identification to identity theft: Public perceptions of biometric privacy harms. *UC Irvine L. Rev.*, 10:107, 2019.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 365–372, 2009.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 128(7):1956–1981, 2020.

Lakshmi, A., Wittenbrink, B., Correll, J., and Ma, D. S. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12:161, 2021.

Lee, M. K. and Rich, K. Who is included in human perceptions of ai?: Trust and perceived fairness around healthcare ai and cultural mistrust. In *Conference on Human Factors in Computing Systems (CHI)*, pp. 1–14, 2021.

Levi, G. and Hassner, T. Age and gender classification using convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 34–42, 2015.

Li, J., Ma, S., Zhang, J., and Tao, D. Privacy-preserving portrait matting. In *ACM International Conference on Multimedia*, pp. 3501–3509, 2021.

Li, T. and Lin, L. Anonymousnet: Natural face de-identification with measurable privacy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Liu, Z., Lian, T., Farrell, J., and Wandell, B. A. Neural network generalization: The impact of camera parameters. *IEEE Access*, 8:10443–10454, 2020.

Long, D. and Magerko, B. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–16, 2020.

Lou, Z., Alnajar, F., Alvarez, J. M., Hu, N., and Gevers, T. Expression-invariant age estimation using structured learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):365–375, 2017.

Luccioni, A. S., Corry, F., Sridharan, H., Ananny, M., Schultz, J., and Crawford, K. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 199–212, 2022.

Ma, D. S., Correll, J., and Wittenbrink, B. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

Ma, D. S., Kantner, J., and Wittenbrink, B. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53(3):1289–1300, 2021.

Malevé, N. On the data set's ruins. *AI & SOCIETY*, pp. 1–15, 2020.

Malgieri, G. and Niklas, J. Vulnerable data subjects. *Computer Law & Security Review*, 37:105415, 2020.

Manjunatha, V., Saini, N., and Davis, L. S. Explicit bias discovery in visual question answering models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9562–9571, 2019.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 416–423, 2001.

Mascalzoni, D., Melotti, R., Pattaro, C., Pramstaller, P. P., Gögele, M., De Grandi, A., and Biasiotto, R. Ten years of dynamic consent in the chris study: informed consent as a dynamic process. *European Journal of Human Genetics*, 30(12):1391–1397, 2022.

McKie, A. South african university drops gender titles in student correspondence, Jul 2018. URL https://www.timeshighereducation.com/news/south-african-university-drops-gender-titles-student-correspondence.

McPherson, R., Shokri, R., and Shmatikov, V. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

Meng, H., Ching, P., Lee, T., Mak, M. W., Mak, B., Moon, Y., Siu, M.-H., Tang, X., Hui, H., Lee, A., et al. The multibiometric, multi-device and multilingual (m3) corpus. In *Proc. Workshop Multimodal User Authentication*, 2006.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2437–2445, 2020.

Merler, M., Ratha, N., Feris, R. S., and Smith, J. R. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

Metcalf, J. and Crawford, K. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211, 2016.

Miceli, M., Schuessler, M., and Yang, T. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020.

Mihailidis, A., Carmichael, B., and Boger, J. The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. *IEEE Transactions on information technology in biomedicine*, 8(3):238–247, 2004.

Mintun, E., Kirillov, A., and Xie, S. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3571–3583, 2021.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 220–229, 2019.

Mittelstadt, B. D. and Floridi, L. The ethics of big data: current and foreseeable issues in biomedical contexts. *The ethics of biomedical big data*, pp. 445–480, 2016.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. Agedb: the first manually collected, in-the-wild age database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 51–59, 2017.

Mudditt, J. The nation where your 'faceprint' is already being tracked. https://www.bbc.com/future/article/20220616-the-nation-where-your-faceprint-is-already-being-tracked, 2022. Accessed June 30, 2022.

Nascimento, G., Laranjeira, C., Braz, V., Lacerda, A., and Nascimento, E. R. A robust indoor scene recognition method based on sparse representation. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 408–415. Springer, 2018.

National Commission for the Proptection of Human Subjects of Biomedicaland Behavioral Research, Bethesda, Md. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Superintendent of Documents, 1978.

National Health and Medical Research Council. Payment of participation in research: information for researchers, hrecs and other ethics review bodies. https://www.nhmrc.gov.au/about-us/publications/payment-participants-research-information-researchers-hrecs-and-other-ethics-review-bodies, 2019. Accessed May 12, 2023.

National Institutes of Health – Division of Program Coordination, Planning and Strategic Initiatives. Gender pronouns & their use in workplace communications. https://dpcpsi.nih.gov/sgmro/gender-pronouns-resource, 2022. Accessed November 24, 2022.

National People's Congress. Personal information protection law. https://personalinformationprotectionlaw.com/, 2021. Accessed November 12, 2022.

Neuert, C., Meitinger, K., Behr, D., and Schonlau, M. The use of open-ended questions in surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, 15(1):3–6, 2021.

Nijhawan, L. P., Janodia, M. D., Muddukrishna, B., Bhat, K. M., Bairy, K. L., Udupa, N., Musmade, P. B., et al. Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research*, 4(3): 134, 2013.

Niu, Z., Zhou, M., Wang, L., Gao, X., and Hua, G. Ordinal regression with multiple output cnn for age estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4920–4928, 2016.

Norja, R., Karlsson, L., Antfolk, J., Nyman, T., and Korkman, J. How old was she? the accuracy of assessing the age of adolescents' based on photos. *Nordic Psychology*, 74(1):70–85, 2022.

Nosek, B. A. and Lakens, D. Registered reports, 2014.

Oh, S. J., Benenson, R., Fritz, M., and Schiele, B. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision (ECCV)*, pp. 19–35. Springer, 2016.

OpenReview. Fairface: A novel face attribute dataset for bias measurement and mitigation. https://openreview.net/forum?id=S1xSSTNKDB, 2019. Accessed August 1, 2022.

Or-El, R., Sengupta, S., Fried, O., Shechtman, E., and Kemelmacher-Shlizerman, I. Lifespan age transformation synthesis. In *European Conference on Computer Vision (ECCV)*, pp. 739–755. Springer, 2020.

Orekondy, T., Fritz, M., and Schiele, B. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8466–8475, 2018.

Papakyriakopoulos, O., Choi, A. S. G., Andrews, J., Bourke, R., Thong, W., Zhao, D., Xiang, A., and Koenecke, A. Augmented datasheets for speech datasets and ethical decision-making. *arXiv preprint arXiv:2305.04672*, 2023.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. Deep face recognition. 2015.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Pavlichenko, N., Stelmakh, I., and Ustalov, D. Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS D&B)*, 2021.

Peng, K., Mathur, A., and Narayanan, A. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Perrigo, B. Inside facebook's african sweatshop, Feb 2022. URL https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/.

Phillips, M. International data-sharing norms: from the oecd to the general data protection regulation (gdpr). *Human genetics*, 137:575–582, 2018.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., and O'Toole, A. J. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

Phillips, T. Exploitation in payments to research subjects. *Bioethics*, 25(4):209–219, 2011.

Piergiovanni, A. and Ryoo, M. Avid dataset: Anonymized videos from diverse countries. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 16711–16721, 2020.

Politou, E., Alepis, E., and Patsakis, C. Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of cybersecurity*, 4(1): tyy001, 2018.

Porgali, B., Albiero, V., Ryda, J., Ferrer, C. C., and Hazirbas, C. The casual conversations v2 dataset. *arXiv preprint arXiv:2303.04838*, 2023.

Prabhu, V. U. and Birhane, A. Large image datasets: A pyrrhic win for computer vision? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546, 2021.

Price, W. N. and Cohen, I. G. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

Prunkl, C. E., Ashurst, C., Anderljung, M., Webb, H., Leike, J., and Dafoe, A. Institutionalizing ethics in ai through broader impact requirements. *Nature Machine Intelligence*, 3(2):104–110, 2021.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT*, pp. 1776–1826, 2022.

Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. Ai and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS D&B)*, 2021a. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf.

Raji, I. D. and Fried, G. About face: A survey of facial recognition evaluation. 2021.

Raji, I. D., Scheuerman, M. K., and Amironesei, R. You can't sit with us: exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 515–525, 2021b.

Reid, D. A. and Nixon, M. S. Using comparative human descriptions for soft biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, pp. 1–6. IEEE, 2011.

Ricanek, K. and Tesafaye, T. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pp. 341–345. IEEE, 2006.

Richardson, R., Schultz, J. M., and Crawford, K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, 94:15, 2019.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision (ECCV)*, pp. 17–35. Springer, 2016.

Robinson, J. P., Livitz, G., Henon, Y., Qin, C., Fu, Y., and Timoner, S. Face recognition: too bias, or not too bias? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 0–1, 2020.

Rojas, W. A. G., Diamos, S., Kini, K. R., Kanter, D., Reddi, V. J., and Coleman, C. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2022. URL https://openreview.net/forum?id=qnfYsave0U4.

Romm, N. R. Interdisciplinary practice as reflexivity. *Systemic Practice and Action Research*, 11:63–77, 1998.

Rose, A. Are face-detection cameras racist?, Jan 2010. URL http://content.time.com/time/business/article/0,8599,1954643-1,00.html.

Rosenfeld, A., Zemel, R., and Tsotsos, J. K. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Rostamzadeh, N., Mincu, D., Roy, S., Smart, A., Wilcox, L., Pushkarna, M., Schrouff, J., Amironesei, R., Moorosi, N., and Heller, K. Healthsheet: development of a transparency artifact for health datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1943–1961, 2022.

Roth, W. *Race migrations: Latinos and the cultural transformation of race*. Stanford University Press, 2012.

Roth, W. D. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, 2016.

Rothbart, M. and Taylor, M. Category labels and social reality: Do we view social categories as natural kinds? 1992.

Rothe, R., Timofte, R., and Van Gool, L. Dex: Deep expectation of apparent age from a single image. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 10–15, 2015.

Rothe, R., Timofte, R., and Van Gool, L. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.

Rowe, Paul Willner, G. Alcohol servers' estimates of young people's ages. *Drugs: education, prevention and policy*, 8(4):375–383, 2001.

Różyńska, J. The ethical anatomy of payment for research participants. *Medicine, Health Care and Philosophy*, 25 (3):449–464, 2022.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Sarkar, S., Phillips, P. J., Liu, Z., Vega, I. R., Grother, P., and Bowyer, K. W. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

Scheuerman, M. K., Paul, J. M., and Brubaker, J. R. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33, 2019.

Scheuerman, M. K., Spiel, K., Haimson, O. L., Hamidi, F., and Branham, S. M. Hci guidelines for gender equity and inclusivity. https://www.morgan-klaus.com/gender-guidelines.html, May 2020a. Accessed August 1, 2022.

Scheuerman, M. K., Wade, K., Lustig, C., and Brubaker, J. R. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW1):1–35, 2020b.

Scheuerman, M. K., Hanna, A., and Denton, E. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Schumann, C., Ricco, S., Prabhu, U., Ferrari, V., and Pantofaru, C. R. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

Segall, M. H., Campbell, D. T., and Herskovits, M. J. *The influence of culture on visual perception*. Bobbs-Merrill Indianapolis, 1966.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Shmueli, B., Fell, J., Ray, S., and Ku, L.-W. Beyond fair pay: Ethical implications of nlp crowdsourcing. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3758–3769, 2021.

Siddiqui, H., Rattani, A., Ricanek, K., and Hill, T. An examination of bias of facial analysis based bmi prediction models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2926–2935, 2022.

Silver, L. Smartphone ownership is growing rapidly around the world, but not always equally, Aug 2020. URL https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/.

Singer, E. and Couper, M. P. Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, 11(2):115–134, 2017.

Singh, R., Vatsa, M., Bhatt, H. S., Bharadwaj, S., Noore, A., and Nooreyezdan, S. S. Plastic surgery: A new dimension to face recognition. *IEEE Transactions on Information Forensics and Security*, 5(3):441–448, 2010.

Smyth, J. D., Dillman, D. A., Christian, L. M., and McBride, M. Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, 73(2):325–337, 2009.

Snow, J. Amazon's face recognition falsely matched 28 members of congress with mugshots, Jul 2018. URL https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28.

Sobel, B. A taxonomy of training data: Disentangling the mismatched rights, remedies, and rationales for restricting machine learning. *Artificial Intelligence and Intellectual Property (Reto Hilty, Jyh-An Lee, Kung-Chung Liu, eds.), Oxford University Press, Forthcoming*, 2020.

Solon, O. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. *NBC News*, 2019.

Somanath, G., Rohith, M., and Kambhamettu, C. Vadana: A dense dataset for facial image analysis. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2175–2182, 2011.

Sörqvist, P., Langeborg, L., and Eriksson, M. Women assimilate across gender, men don't: The role of gender to the own-anchor effect in age, height, and weight estimates 1. *Journal of Applied Social Psychology*, 41(7):1733–1748, 2011.

Spiel, K., Haimson, O. L., and Lottridge, D. How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65, 2019.

Srinivasan, R., Denton, E., Famularo, J., Rostamzadeh, N., Diaz, F., and Coleman, B. Artsheets for art datasets. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

Steed, R. and Caliskan, A. Image representations learned with unsupervised pre-training contain human-like biases. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 701–713, 2021.

Stevens, N. Open demographics documentation. https://nikkistevens.com/open-demographics/index.htm, n.d. Accessed November 22, 2021.

Stewart, R., Andriluka, M., and Ng, A. Y. End-to-end people detection in crowded scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2325–2333, 2016.

Sun, Q., Ma, L., Oh, S. J., Van Gool, L., Schiele, B., and Fritz, M. Natural and effective obfuscation by head inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5050–5059, 2018.

Suresh, H. and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. 2021.

Tagai, K., Ohtaka, H., and Nittono, H. Faces with light makeup are better recognized than faces with heavy makeup. *Frontiers in psychology*, 7:226, 2016.

Teare, H. J., Prictor, M., and Kaye, J. Reflections on dynamic consent in biomedical research: the story so far. *European journal of human genetics*, 29(4):649–656, 2021.

Tech Inquiry. Official response from wiley. https://techinquiry.org/WileyResponse.html, 2019. Accessed June 30, 2022.

Telles, E. E. Racial ambiguity among the brazilian population. *Ethnic and racial studies*, 25(3):415–441, 2002.

Thomas, G., Gade, R., Moeslund, T. B., Carr, P., and Hilton, A. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, 2017.

Tong, S. and Kagal, L. Investigating bias in image classification using model explanations. *arXiv preprint arXiv:2012.05463*, 2020.

Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

Toxtli, C., Suri, S., and Savage, S. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–26, 2021.

Twigg, J. *The Right to Safety: some conceptual and practical issues*. Benfield Hazard Research Centre, 2003.

Uittenbogaard, R., Sebastian, C., Vijverberg, J., Boom, B., Gavrila, D. M., et al. Privacy protection in street-view panoramas using depth and multi-view imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10581–10590, 2019.

UK Information Commissioner's Office. What do we need to do to ensure lawfulness, fairness, and transparency in ai systems? https://ico.org.uk/for-organisations/guide-to-data-protection/key-dataprotection-themes/guidance-on-ai-and-data-protection/what-do-we-need-todo-to-ensure-lawfulness-fairness-and-transparency-in-ai-systems/, 2020. Accessed June 30, 2022.

UK Information Commissioner's Office. How should we obtain, record and manage consent? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/consent/how-should-we-obtain-record-and-manage-consent, n.d. Accessed May 1, 2023.

UNICEF. Bridging the digital divide for children and adolescents in east asia and pacific. https://www.unicef.org/eap/bridging-digital-divide-children-and-adolescents-east-asia-and-pacific, 2020. Accessed November 24, 2022.

UNICEF et al. Unicef procedure for ethical standards in research, evaluation, data collection and analysis. *Nueva York.(2012), Ethical Principles, Dilemmas, and Risks in Collecting Data on Violence against Children: A Review of Available Literature, Nueva York*, 2015.

United Nations. Principles and recommendations for population and housing censuses. *Statistical Papers*, No.67, Sales No E.98.XVII.8, 1998.

United Nations Statistics Division. Standard country or area codes for statistical use. https://unstats.un.org/unsd/methodology/m49/, n.d. Accessed May 1, 2021.

United States Census Bureau. How disability data are collected from the american community survey. https://www.census.gov/topics/health/disability/guidance/data-collection-acs.html, 2021. Accessed August 1, 2022.

United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

The Belmont report: ethical principles and guidelines for the protection of human subjects of research, volume 1. Department of Health, Education, and Welfare, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978.

Valentine, T., Lewis, M. B., and Hills, P. J. Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, 69(10): 1996–2019, 2016.

Van Noorden, R. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834):354–359, 2020.

Vangara, K., King, M. C., Albiero, V., Bowyer, K., et al. Characterizing the variability in face recognition accuracy relative to race. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Vestlund, J., Langeborg, L., Sörqvist, P., and Eriksson, M. Experts on age estimation. *Scandinavian Journal of Psychology*, 50(4):301–307, 2009.

Voelkle, M. C., Ebner, N. C., Lindenberger, U., and Riediger, M. Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age. *Psychology and aging*, 27(2):265, 2012.

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., and Russakovsky, O. Revise: A tool for measuring and mitigating bias in visual datasets. volume 130, pp. 1790–1810. Springer, 2022.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 526–536, 2021.

Wang, M. and Deng, W. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9322–9331, 2020.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Wang, Z., He, X., and Liu, F. Examining the effect of smile intensity on age perceptions. *Psychological reports*, 117 (1):188–205, 2015.

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., and Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8919–8928, 2020.

Ware, O. R., Dawson, J. E., Shinohara, M. M., and Taylor, S. C. Racial limitations of fitzpatrick skin type. *Cutis*, 105(2):77–80, 2020.

Weber, G. M., Mandl, K. D., and Kohane, I. S. Finding the missing link for big biomedical data. *Jama*, 311(24): 2479–2480, 2014.

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X., et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022.

Whitley, E. A. Informational privacy, consent and the "control" of personal data. *Information security technical report*, 14(3):154–159, 2009.

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., et al. Disability, bias, and ai. *AI Now Institute*, pp. 8, 2019.

Wilson, B., Hoffman, J., and Morgenstern, J. Predictive inequity in object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Windrim, L., Melkumyan, A., Murphy, R., Chlingaryan, A., and Nieto, J. Unsupervised feature learning for illumination robustness. In *IEEE International Conference on Image Processing (ICIP)*, pp. 4453–4457, 2016.

World Economic Forum. The digital revolution is leaving poorer kids behind. https://www.weforum.org/agenda/2022/04/the-digital-revolution-is-leaving-poorer-kids-behind/, 2022. Accessed November 24, 2022.

World Health Organization. Ageism in artificial intelligence for health. https://www.who.int/publications/i/item/9789240040793, 2022. Accessed November 24, 2022.

World Health Organization and others. Ethics and governance of artificial intelligence for health: Who guidance. 2021.

Xiang, A. Being'seen'vs.'mis-seen': Tensions between privacy and fairness in computer vision. *Harvard Journal of Law & Technology, Forthcoming*, 2022.

Xie, R., Yu, F., Wang, J., Wang, Y., and Zhang, L. Multi-level domain adaptive learning for cross-domain detection. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

Xiong, Y., Zhu, K., Lin, D., and Tang, X. Recognize complex events from static images by fusing deep channels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1600–1609, 2015.

Xu, R., Baracaldo, N., and Joshi, J. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*, 2021.

Xu, T., White, J., Kalkan, S., and Gunes, H. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 506–523. Springer, 2020.

Xu, X. and Wang, J. Extended non-local feature for visual saliency detection in low contrast images. In *European Conference on Computer Vision Workshops (ECCVW)*, 2019.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 547–558, 2020.

Yang, K., Yau, J. H., Fei-Fei, L., Deng, J., and Russakovsky, O. A study of face obfuscation in imagenet. In *International Conference on Machine Learning (ICML)*, pp. 25313–25330. PMLR, 2022a.

Yang, S., Luo, P., Loy, C.-C., and Tang, X. Wider face: A face detection benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, 2016.

Yang, Y., Gupta, A., Feng, J., Singhal, P., Yadav, V., Wu, Y., Natarajan, P., Hedau, V., and Joo, J. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 813–822, 2022b.

Yew, R.-J. and Xiang, A. Regulating facial processing technologies: Tensions between legal and technical considerations in the application of illinois bipa. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1017–1027, 2022.

Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yucer, S., Tektas, F., Al Moubayed, N., and Breckon, T. P. Measuring hidden bias within face recognition via racial phenotypes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 995–1004, 2022.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pp. 94–108. Springer, 2014.

Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5810–5818, 2017.

Zhao, D., Wang, A., and Russakovsky, O. Understanding and evaluating racial biases in image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhao, D., Andrews, J. T., and Xiang, A. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning (ICML)*, 2023.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Zhou, M., Lin, H., Young, S. S., and Yu, J. Hybrid sensing face detection and registration for low-light and unconstrained conditions. *Applied optics*, 57(1):69–78, 2018.

Zimmer, M. "but the data is already public": On the ethics of research in facebook. *Ethics and Information Technology*, 12(4):313–325, 2010.