Ethical Considerations for Responsible Data Curation

Jerone T. A. Andrews[*]
Sony AI, Tokyo

Dora Zhao[†]
Sony AI, New York

William Thong[†]
Sony AI, Zurich

Apostolos Modas[†]
Sony AI, Zurich

Orestis Papakyriakopoulos[†]
Sony AI, Zurich

Alice Xiang
Sony AI, Seattle

Abstract

Human-centric computer vision (HCCV) data curation practices often neglect privacy and bias concerns, leading to dataset retractions and unfair models. HCCV datasets constructed through nonconsensual web scraping lack crucial metadata for comprehensive fairness and robustness evaluations. Current remedies are post hoc, lack persuasive justification for adoption, or fail to provide proper contextualization for appropriate application. Our research focuses on proactive, domain-specific recommendations, covering purpose, privacy and consent, and diversity, for curating HCCV evaluation datasets, addressing privacy and bias concerns. We adopt an ante hoc reflective perspective, drawing from current practices, guidelines, dataset withdrawals, and audits, to inform our considerations and recommendations.

1 Introduction

Contemporary human-centric computer vision (HCCV) data curation practices, which prioritize dataset size and utility, have pushed issues related to privacy and bias to the periphery, resulting in dataset retractions and modifications [78, 126, 175, 216, 244, 320], as well as models that are unfair or rely on spurious correlations [22, 26, 112, 139, 146, 215, 272, 281]. HCCV datasets primarily rely on nonconsensual web scraping [99, 122, 124, 228, 260, 266, 310]. These datasets not only regard image subjects as free raw material [32], but also lack the ground-truth metadata required for fairness and robustness evaluations [91, 171, 196, 216]. This makes it challenging to obtain a comprehensive understanding of model blindspots and cascading harms [30, 85] across dimensions, such as data subjects, instruments, and environments, which are known to influence performance [222]. While, for example, image subject attributes can be inferred [7, 43, 170, 188, 198, 241, 267, 290, 343, 375], this is controversial for social constructs, notably race and gender [28, 132, 179, 180]. Inference introduces further biases [19, 107, 147, 263, 291] and can induce psychological harm when incorrect [47, 275].

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 [5, 27, 44, 94, 108, 149, 243, 258, 273, 307]. Similarly, initiatives by NeurIPS and ICML ask authors to consider the ethical and societal implications of their research after completion [257]. Further, dataset audits [247, 292] and bias detection tools [29, 340] expose dataset management issues and representational biases without offering guidance on responsible data collection. Although there are existing proposals for artificial intelligence (AI) and data design guidelines [35, 79, 116, 160, 202, 247], as well as calls to adopt methodologies from more established fields [154, 159, 166], general-purpose guidelines lack domain specificity and task-oriented guidance [307]. For example, remedies may prioritize privacy and governance [35] but overlook data composition and image content. Other recommended practices lack persuasive justification for adoption [116, 160] or fail to provide proper

---

[*]Correspondence to `jerone.andrews@sony.com`.
[†]Equal contribution; authors are listed in random order.

contextualization for appropriate application [252, 323, 367]. For instance, the *People + AI Guidebook* [116] suggests creating dataset specifications without explaining the rationale, and privacy methodologies are advocated without cognizant of privacy and data protection laws [252, 323, 367]. These efforts, which hold significance in promoting responsible practices, would benefit from being supplemented by proactive, domain-specific recommendations aimed at tackling privacy and bias concerns starting from the inception of a dataset.

Our research directly addresses these critical concerns 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. Our work, therefore, resonates with the call for domain-specific resources to operationalize fairness [68, 150, 305]. We draw insights from current practices [42, 170, 375], guidelines [31, 222, 231], dataset withdrawals [78, 126, 216], and audits [35, 36, 247], to motivate each recommendation, focusing on HCCV evaluation datasets that present unique challenges (e.g., visual leakage of personally identifiable information) and opportunities (e.g., leveraging image metadata for analysis). To guide curators towards more ethical yet resource-intensive curation, we provide a checklist in Appendix A.[3] This translates our considerations and recommendations into pre-curation questions, functioning as a catalyst for discussion and reflection.

While several of our recommendations can also be applied retroactively such measures cannot undo incurred harm, e.g., resulting from inappropriate uses, privacy violations, and unfair representation [129]. It is important to make clear that 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

HCCV should adhere to the most stringent ethical standards to address privacy and bias concerns. As stated in the NeurIPS Code of Ethics [1], it is essential to abide by established institutional research protocols, ensuring the safeguarding of human subjects. These protocols, initially designed for biomedical research, have, however, been met with confusion, resulting in inconsistencies when applied in the context of data-centric research [218]. For example, HCCV research often amasses millions of "public" images without obtaining informed consent or participation, disregarding serious privacy and ethical concerns [3, 35, 133, 245, 301]. This exemption from research ethics regulation is grounded in the limited definition of human-subjects research, which categorizes extant, publicly available data as minimal risk [218, 256]. Thus, numerous ethically-dubious HCCV datasets would not fall under Institutional Review Board (IRB) oversight [247]. What's more, the NeurIPS Code of Ethics only mandates following existing protocols when research involves "direct" interaction between human participants and researchers or technical systems. Even when research is subjected to supervision, IRBs are restricted from considering broader societal consequences beyond the immediate study context [217]. Compounding matters, CV-centric conferences are still to adopt ethics review practices [306].

These limitations are concerning, especially considering the potential for predictive privacy harms when seemingly non-identifiable data is combined [35, 70, 218] or when data is used for harmful downstream applications such as predicting sexual orientation [192, 344], crime propensity [358, 365], or emotion [10, 224]. Acknowledging this, our research study employed the same principles underpinning established guidelines [24, 331] for protecting human subjects in research to identify ethical issues in HCCV dataset design, namely autonomy, justice, beneficence, and non-maleficence. *Autonomy* respects individuals' self-determination—e.g., through informed consent and assent for HCCV datasets. *Justice* promotes the fair distribution of risks, costs, and benefits, guiding decisions on compensation, data accessibility, and diversity. *Beneficence* entails the proactive promotion of positive outcomes and well-being, e.g., by soliciting individuals' to self-identify, while *non-maleficence* centers on minimizing harm and risks during dataset design, e.g., by redacting privacy-leaking image regions and metadata.

To ensure comprehensive consideration, we harnessed diverse expertise, following contemporary, interdisciplinary practices [261, 270, 307]. Our team comprises researchers, practitioners, and

---

[3]The checklist can also be found at: `https://github.com/SonyResearch/responsible_data_curation`.

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 principles, we collectively discussed them, considering each author's background. After identifying key ethical issues in HCCV data curation practices, we iteratively refined them into an initial draft of ethical considerations. We extensively collected, analyzed, and discussed papers spanning a range of themes such as HCAI, HCCV datasets, data and model documentation, bias detection and mitigation, AI and data design, fairness, and critical AI. Our comprehensive literature review incorporated pertinent studies and datasets, resulting in refined considerations with detailed explanations and recommendations for responsible data curation. Additional details are provided in Appendix B.

3 Purpose

In ML, significant emphasis has been placed on the acquisition and utilization of "general-purpose" datasets [259]. Nevertheless, without a clearly defined task pre-data collection, it becomes challenging to effectively handle issues related to data composition, labeling, data collection methodologies, informed consent, and assessments related to data protection. 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 assessing fairness. For instance, the COCO dataset [196], focused on scene understanding, lacks subject information, making fairness assessments challenging. Researchers, consequently, resort to human annotators to infer, e.g., subject characteristics, limiting bias measurement to visually "inferable" attributes. This introduces annotation bias [56] and the potential for harmful inferences [47, 275].

Fairness-aware datasets are incompatible with common HCCV tasks. Industry practitioners stress the importance of carefully designed and collected "fairness-aware" datasets to detect bias issues [150]. Fabris et al. [93] found that out of 28 CV datasets used in fairness research between 2014 and 2021, only eight were specifically created with fairness in mind. Among these, seven were HCCV datasets (scraped from the web) [43, 170, 216, 308, 319, 342, 343], including five focused on facial analysis. Due to the limited availability and delimited task focus of fairness-aware datasets, researchers repurpose "fairness-unaware" datasets [120, 139, 196, 198, 208, 346, 373]. Fairness-aware datasets fall short in addressing the original tasks associated with well-known HCCV datasets, which encompass a range of tasks, such as segmentation [64, 209], pose estimation [13, 196], localization and detection [73, 91, 110], identity verification [153], action recognition [173], as well as reconstruction, synthesis and manipulation [114, 171]. The absence of fairness-aware datasets with task-specific labels hampers the practical evaluation of HCCV systems, despite their importance in domains such as healthcare [155, 220], autonomous vehicles [163], and sports [317]. Additionally, fairness-aware datasets lack self-identified annotations from image subjects, relying on inferred attributes, e.g., from online resources [43, 308, 319].

3.2 Practical Recommendations

Refrain from repurposing datasets. Existing datasets, repurposable but optimized for specific functions, can inadvertently perpetuate biases and undermine fairness [183]. 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 [181, 265]. Dirty data, including inferred data, can have significant downstream consequences, compromising the validity of research, policy, and decision-making [14, 63, 265, 341]. ML practitioners widely agree that a proactive approach to fairness is preferable, involving the direct collection of demographic information from the outset [150]. 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 [222].

Create purpose statements. Pre-data collection, dataset creators should establish *purpose statements*, focusing on motivation rather than cause [129]. Purpose statements address, e.g., data collection motivation, desired composition, permissible uses, and intended consumers. While dataset documentation [108, 258] covers similar questions, it is a *reflective* process and can be manipulated to fit the narrative of the collected data, as opposed to directing the narrative of the data to be collected. Purpose statements can play a crucial role in preventing both *hindsight bias* [51, 97, 176] and *purpose creep*, ensuring alignment with stakeholders' consent and intentions [186]. To enhance transparency and accountability, as recommended by Peng et al. [247], purpose statements can undergo peer review, similar to *registered reports* [238]. Registered reports, recognized by the UK 2021 Research Excellence Framework, incentivize rigorous research practices and can lead to increased institutional funding [51].

4   Consent and Privacy

Informed consent is crucial in research ethics involving humans [230, 235], ensuring participant safety, protection, and research integrity [59, 253]. Shaping data collection practices in various fields [35, 235], informed consent consists of three elements: *information* (i.e., the participant should have sufficient knowledge about the study to make their decision), *comprehension* (i.e., the information about the study should be conveyed in an understandable manner), and *voluntariness* (i.e., consent must be given free of coercion or undue influence). While consent is not the only legal basis for data processing, it is globally preferred for its legitimacy and ability to foster trust [82, 253]. We address concerns related to consent and privacy, and provide recommendations.

4.1   Ethical Considerations

Human-subjects research. As aforementioned in Section 2, HCCV datasets are frequently collected without informed consent or participation, primarily due to the classification of publicly available data as "minimal risk" within human-subjects research. However, beyond possible predictive privacy harms and unethical downstream uses, collecting data without informed consent hinders researchers and practitioners from fully understanding and addressing potential harms to data subjects [218, 333]. 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 [77, 223, 245, 253].

Creative Commons loophole. Some datasets have been created based on the misconception that the "unlocking [of] restrictive copyright" [35] through Creative Commons licenses implies data subject consent. However, the Illinois Biometric Information Privacy Act (BIPA) [161] mandates data subject consent, even for publicly available images [370]. In the UK and EU General Data Protection Regulation (GDPR) [88] 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) [233] 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 [370], it does not apply to image regions containing biometric data such as faces, which are protected by privacy and data protection laws [300].

Vulnerable persons. Nonconsensual data collection methods can result in the inclusion of vulnerable individuals unable to consent or oppose data processing due to power imbalances, limited capacity, or increased risks of harm [89, 207]. While scraping vulnerable individuals' biometric data may be incidental, some researchers actively target them, jeopardizing their sensitive information without guardian consent [128, 260].

Paradoxically, attempts to address racial bias in data have involved soliciting homeless persons of color, further compromising their vulnerability [103]. When participation is due to economic or situational vulnerability, as opposed to one's best interests, monetary offerings may be perceived as inducement [117]. Further ethical concerns manifest when it is unclear whether participants were adequately *informed* about a research study. For instance, in ethnicity recognition research [72], despite obtaining informed consent, criticism arose due to training a model that discriminates between Chinese Uyghur, Korean, and Tibetan faces. Although the study's focus is on the technology

itself [315], its potential use in enhancing surveillance on Chinese Uyghurs raises ethical questions due to the human rights violations against them [333].

Consent revocation. Dataset creators sometimes view autonomy as a challenge to collecting biometric data for HCCV, especially when data subjects prioritize privacy [214, 287, 297]. Nonetheless, informed consent emphasizes *voluntariness*, encompassing both the ability to give consent and the right to withdraw it at any time [74]. GDPR grants explicit revocation rights (Article 7) and the right to request erasure of personal data (Article 17) [350]. However, image subjects whose data is collected without consent are denied these rights. The nonconsensual FFHQ face dataset [171] 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 [81].

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 [46, 101, 194, 195, 213, 252, 311, 323, 362, 367]. Nonetheless, face detection algorithms used in obfuscation may raise legal concerns, particularly if they involve predicting facial landmarks, potentially violating BIPA [61, 370]. 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. Moreover, 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 [352]. This predictive inequity leads to allocative harm, denying certain groups opportunities and resources, including the rights to safety [322] and privacy [80].

It is important to note that face obfuscation may not guarantee privacy [145, 367]. The Visual Redactions dataset [242] 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 [239], indicating that solely obfuscating face regions might be insufficient under GDPR. Furthermore, image metadata can also disclose sensitive details, e.g., date, time, and location, as well as copyright information that may include names [11, 239]. This is worrisome for users of commonly targeted platforms like Flickr, which retain metadata by default.

4.2 Practical Recommendations

Obtain voluntary informed consent. Similar to recent consent-driven HCCV datasets [136, 254, 268], 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 [65, 379]. Following GDPR (Article 7), plain language consent and notice forms are recommended to address the lack of public understanding of AI technologies [199].

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 [182]. However, relying solely on guardian consent overlooks the views and dignity of the vulnerable person [141]. 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 [31, 327]. When employing appropriate language and tools, assent establishes the vulnerable person understands the use of their data and willingly participates [31]. 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* [60], 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 [117].

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 [174, 348]. 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 [174, 211, 314]. Alternatively, another recommended approach is to establish a steering board or charitable trust composed of representative dataset participants to make decisions regarding data use [255]. The feasibility of these proposals may vary based on a dataset's scale. Nonetheless, at a minimum, data subjects should be provided a simple and easily accessible method to revoke consent [136, 254, 268]. This aligns with guidance provided by the UK Information Commission's Office (ICO), emphasizing the need to provide alternatives to online-based revocation processes to accommodate varying levels of technology competency and internet access among data subjects [325].

Collect country of residence information. Anonymizing nonconsensual human subjects through face obfuscation, as done in datasets such as ImageNet [367], 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 [249, 268]. 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 [253].

Redact privacy leaking image regions and metadata. The European Data Protection Board emphasizes that anonymization of personal data must guard against re-identification risks such as singling out, linkability, and inference [76]. Re-identification remains possible even when nonconsensual subjects' faces are obfuscated, through other body parts or contextual information [242]. One solution is to redact all privacy-leaking regions related to nonconsensual subjects (including their entire bodies, clothing, and accessories) and text (excluding copyright owner information). However, 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 image regions with identifying or private information [367]. For nonconsensual individuals residing in certain jurisdictions (e.g., Illinois, California, Washington, Texas), automated region proposals requiring biometric identifiers should be avoided. Instead, human annotators should take the responsibility of generating these proposals.

Notwithstanding, 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 may involve replacing exact geolocation data with a more general representation, such as city and country, and excluding user-contributed details from metatags containing personally identifiable information, except when this action would 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 dataset diversity [13, 64, 78, 170, 171, 173, 196, 283, 361, 368], realism [110, 153, 164, 173, 196, 368], and difficulty [13, 16, 64, 73, 78, 91, 110, 173, 196, 198, 361, 368] to enhance fairness and robustness in real-world applications. Previous research has emphasized diversity across image subjects, environments, and instruments [43, 139, 222, 287], but there are many ethical complexities involved in specifying diversity criteria [14, 15]. This section examines taxonomy challenges 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 [67]. Many dataset creators often prioritize protected attributes, i.e., gender, race, and age, as key factors of dataset diversity [287]. Nevertheless, most HCCV datasets exhibit historical and representational biases [35, 166, 172, 312, 366]. These biases can be pernicious, particularly when models learn and *amplify* them. For instance, image captioning models may rely on contextual cues related to activities like shopping [377]

and laundry [376] to generate gendered captions. Spurious correlations are detrimental, as they are not causally related and perpetuate harmful associations [112, 281]. In addition, prominent examples in HCCV research demonstrate disparate algorithmic performance based on race and skin color [42, 43, 54, 123, 142–144, 148, 250, 271, 299, 318, 334, 375]. Most recently, autonomous robots have displayed racist, sexist, and physiognomic stereotypes [158]. Furthermore, face detection models have shown lower accuracy when processing images of older individuals compared to younger individuals [369]. While not endorsing these applications, discrepancies have also been observed in facial emotion recognition services for children in both commercial and research systems [152, 363], as well as age estimation [58, 115, 200].

Despite concerns regarding privacy, liability, and public relations, the collection of special and sensitive category data is crucial for bias assessments [15]. GDPR guidance from the UK ICO confirms that sensitive attributes can be collected for fairness purposes [324]. However, obtaining this information presents challenges, such as historical mistrust in clinical research among African-Americans [92, 191] or the social stigma of being photographed that some women face [166]. Nonetheless, marginalized communities may require explicit explanations and assurances about data usage to address concerns related to service provision, security, allocation, and representation [359]. This is particularly important as remaining *unseen* does not protect against being *mis-seen* [359].

The digital divide and accessibility. Healthcare datasets often lack representation of minority populations, compromising the reliability of automated decisions [356]. The World Health Organization (WHO) emphasizes the need for data accuracy, completeness, and diversity, particularly regarding age, in order to address ageism in AI [355]. ML systems may prioritize younger populations for resource allocation, assuming they would benefit the most in terms of life expectancy [355]. 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 [326, 354]. Insufficient access to digital technology hampers the representation of vulnerable persons in datasets [290], leading to *outcome homogenization*—i.e., the systematic failing of the same individuals or groups [39].

Confused taxonomies. Sex and gender are often used interchangeably, treating gender as a consequence of one's assigned sex at birth [95]. However, this approach erases intersex individuals who possess non-binary physiological sex characteristics [95]. Treating sex and gender as interchangeable perpetuates normative views by casting gender as binary, immutable, and solely based on biological sex [179]. 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* [45].

Similar to sex and gender, race and ethnicity are often used synonymously [332]. Nations employ diverse census questions to ascertain ethnic group composition, encompassing factors such as nationality, race, color, language, religion, customs, and tribe [328]. However, these categories and their definitions lack consistency over time and geography, often influenced by political agendas and socio-cultural shifts [286]. This variability makes it challenging to collect globally representative and meaningful data on ethnic groups. Consequently, several HCCV datasets have incorporated inconsistent and arbitrary racial categorization systems [7, 267, 343, 374]. For instance, the FairFace dataset [170] creators reference the US Census Bureau's racial categories without considering the social definition of race they represent [240]. The US 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 [53, 102, 125, 226, 236, 264, 277, 278, 374] or group labels [84, 105, 193, 302]. Age groupings are often preferred when collecting unconstrained images from the web, as human annotators must infer subjects' ages, which is challenging [48]. This is evident in crowdsourced annotations, where 40.2% of individuals in the OpenImages MIAP dataset [290] could not be categorized into an age group. Factors unrelated to age, such as facial expression [106, 237, 345] and makeup [83, 237, 313], influence age perception. Furthermore, annotators have exhibited lower accuracy when labeling people outside of their own demographic group [8, 9, 113, 279, 303, 335, 339].

Post hoc rationalization of the use of physiological markers. Gender information about data subjects is obtained through inference [53, 170, 187, 198, 236, 264, 277, 278, 290, 343, 374, 374]

or self-identification [136, 190, 203, 204, 375]. 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 [87, 127, 179]. 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 [277, 278]. Moreover, physical traits are just one of many dimensions, including posture, clothing, and vocal cues, used to infer not only gender but also race [100, 177, 179].

Erasure of nonstereotypical individuals. HCCV datasets frequently adopt a US-based racial schema [170, 190, 203, 204, 264, 343], which may oversimplify and essentialize groups [316]. This approach may not align with other more nuanced models, e.g., the continuum-based color system used in Brazil, which considers a wide range of physical characteristics. Nonconsensual image datasets rely on annotators to assign semantic categories, perpetuating stereotypes and disseminating them beyond their cultural context [180]. Notably, images without label consensus are often discarded [170, 267, 343], potentially excluding individuals who defy stereotypes, such as multi-ethnic individuals [276].

Phenotypic attributes. Protected attributes may not be the most appropriate criteria for evaluating HCCV models [43]. 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 [25, 48, 96, 138, 180, 347].

Environment and instrument. The image capture device and environmental conditions significantly influence model performance, and their impact should be considered [222]. Factors such as camera software, hardware, and environmental conditions affect HCCV model robustness in various settings [4, 140, 197, 221, 229, 353, 360, 364, 371]. 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 [62, 184, 185, 378], while available instruments or environmental co-occurrences may correlate with demographic attributes [139, 295, 376].

Annotator positionality. Psychological research highlights the influence of annotators' sociocultural background on their visual perception [19, 107, 147, 263, 275, 291]. However, recent empirical studies have evidenced a lack of regard for the impact an annotator's social identity has on data [79, 111]. Only a handful of HCCV datasets provide annotator demographic details [12, 56, 287, 375].

Recruitment and compensation. Data collected without consent patently lacks compensation. Balancing between excessive and deficient payment is crucial to avoid coercion and exploitation [231, 268]. An additional concern is the employment of remote workers from disadvantaged regions [248], often with low wages and fast-paced work conditions [71, 135, 162, 206]. This can lead to arbitrary denial of payment based on opaque quality criteria [98] and prevents union formation [206], creating a sense of invisibility and uncertainty for workers [321].

5.2 Practical Recommendations

Obtain self-reported annotations. Practitioners are cautious about inferring labels about people to avoid biases [15]. Moreover, data access request rights, e.g., as offered by GDPR, California Consumer Privacy Act, and PIPL, may require data holders to disclose inferred information. To avoid stereotypical annotations and minimize harm from misclassification [275], labels should be collected directly from image subjects, who inherently possess contextual knowledge of their environment 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 [156, 179, 274]. Open-ended questions provide more accurate answers but can be taxing, require extensive coding, and are harder to analyze [40, 109, 178, 298]. To balance this, closed-ended questions should be augmented with an open-ended response option, avoiding the term "other", which implies *othering* norms [285]. This gives subjects a *voice* [234, 296] and allows for future question design improvement.

Acknowledge the mutability and multiplicity of identity. *Identity shift*—the intentional self-transformation in mediated contexts [49]—is often overlooked. To address this, we propose collecting self-identified information on a per-image basis, acknowledging that identity is temporal and nonstatic. Specifically, for sensitive attributes, allowing the selection of multiple identity categories without limitations is preferable [304, 309]. This prevents oversimplification and marginalization. While we

acknowledge the potential burden of self-identification on fluid and dynamic identities, an image captures a single moment. Thus, evolving identity may not require metadata updates; however, we recommend providing subjects with mechanisms for updates when needed.

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 aggregation of the collected data. This is particularly 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 [157]. Self-identified pronouns are particularly important for sexual and gender minority communities as they "convey and affirm gender identity" [232]. Significantly, pronoun use is increasingly prevalent in social media platforms [86, 165, 167], workplaces [55], and education settings [20, 212], fostering gender inclusivity [21]. However, subjects should always have the option of not disclosing this information.

Finally, to address issues with ethnic and racial classification systems [180, 286], 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 [17]. The United Nations' M49 geoscheme can be used to operationalize ancestry [329], 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?". This avoids reliance on proxies, e.g., skin tone, that risk normalizing their inadequacies without reflecting their limitations [15].

Collect aggregate data for commonly ignored groups. Additional sensitive attributes should also be collected, such as disability and pregnancy status, when voluntarily disclosed by subjects. These attributes should be reported in aggregate data to reduce the safety concerns of subjects [309, 351]. Given that definitions of these attributes may be inconsistent and tied to culture, identity, and histories of oppression [37, 41], navigating tensions between benefits and risks is necessary. Despite potential reluctance, sourcing data from underrepresented communities contributes to dataset inclusivity [37, 168]. Regarding disability, the American Community Survey [330] 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 [57], e.g., skin color, eye color, hair type, hair color, height, and weight [19]. These attributes have enabled finer-grained analysis of model performance and biases [43, 75, 294, 318, 349, 372]. 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 [6, 18, 166, 284, 340].

Record environment and instrument information. Data should capture variations in environmental conditions and imaging devices, including factors such as 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 exchangeable image file format (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 [12, 79], rather than treating them as anonymous entities responsible for data labeling alone [52, 206]. While many datasets [78, 137, 196] neglect to report annotator demographics, assuming objectivity in annotation for visual categories is flawed [23, 169, 219]. Furthermore, using majority voting to reach the assumed ground truth, disregards minority opinions, treating them as noise [169]. Annotator characteristics, including pronouns, age, and ancestry, should be recorded and reported to quantify and address annotator perspectives and bias in datasets [12, 118]. Additionally, allowing annotators freedom in labeling helps to avoid replicating socially dominant viewpoints [219].

Fair treatment and compensation for contributors. In accordance with Australia's National Health and Medical Research Council [231] and the WHO [66], dataset contributors should not only be

guaranteed compensation above the minimum hourly wage of their country of residence [280], but also according to the complexity of tasks to be performed. However, alternative payment models, for example, based on the average hourly wage, may offer benefits in terms of promoting diversity by increasing the likelihood of higher socio-economic status contributors [251].

Besides payment, the implementation of direct communication channels and feedback mechanisms, such as anonymized feedback forms [246], can help to address issues faced by annotators while providing a level of protection from retribution. Furthermore, 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 the number of "glaring ethical lapses" [293].

6 Discussion and Conclusion

Supplementary to established ethical review protocols, we have provided proactive, domain-specific recommendations for curating HCCV evaluation datasets for fairness and robustness evaluations. However, encouraging change in ethical practice could encounter resistance or slow adoption due to established norms [218], inertia [33], diffusion of responsibility [151], and liability concerns [15]. To garner greater acceptance, platforms such as NeurIPS could adopt a model similar to the registered reports format, embraced by over 300 journals [51, 247]. This entails pre-acceptance of dataset proposals before curation, alleviating financial uncertainties associated with more ethical practices.

Nevertheless, seeking consent from all depicted individuals might give rise to logistical challenges. Resource requirements tied to the implementation and maintenance of consent management systems could emerge, potentially necessitating significant investment in technical infrastructure and dedicated personnel. Particularly for smaller organizations and academic research groups, these limitations could present considerable hurdles. A potential solution is forming *data consortia* [121, 166], which helps address operational challenges by pooling resources and knowledge.

Extending our recommendations to the curation of "democratizing" foundation model-sized training datasets [104, 119, 288, 289] poses an economic challenge. To put this into perspective, the GeoDE dataset of 62K crowdsourced object-centric images [262], without personally identifiable information, incurred a cost of $1.08 per image. While our recommendations may not seamlessly *scale* to the curation of fairness-aware, billion-sized image datasets, it is worth considering that "solutions which resist scale thinking are necessary to undo the social structures which lie at the heart of social inequality" [130]. Large-scale, nonconsensual datasets driven by scale thinking have included harmful and distressing content, including rape [35, 36], racist stereotypes [131], vulnerable persons [128], and derogatory taxonomies [34, 69, 129, 183]. Such content may further generate legal concerns [2]. We contend that these issues can be mitigated through the implementation of our recommendations.

Nonetheless, balancing resources between model development and data curation is value-laden, shaped by "social, political, and ethical values" [38]. While organizations readily invest significantly in model training [227, 336], compensation for data contributors often appears neglected [337, 338], disregarding that "most data represent or impact people" [380]. Remedial actions could be envisioned to bridge the gap between models developed with ethically curated data and those benefiting from expansive, nonconsensually crawled data. Reallocating research funds away from dominant data-hungry methods [38] would help to strike a balance between technological advancement and ethical imperatives.

However, the granularity and comprehensiveness of our diversity recommendations could be adapted beyond evaluation contexts, particularly when employing "fairness without demographics" [50, 134, 189, 210] training approaches, reducing financial costs. Nevertheless, the applicability of any proposed recommendation is intrinsically linked to the specific context [243]. Decisions should be guided by the social framework of a given application to ensure ethical and equitable data curation.

Just as the concepts of identity evolve, our recommendations must also evolve to ensure their ongoing relevance and sensitivity. Thus, we encourage dataset creators to tailor our recommendations to their *context*, fostering further discussions on responsible data curation.

Acknowledgments and Disclosure of Funding

References

[1] Announcing the NeurIPS Code of Ethics x2013; NeurIPS Blog — blog.neurips.cc. https://blog.neurips.cc/2023/04/20/announcing-the-neurips-code-of-ethics/. [Accessed August 14, 2023].

[2] OpenAI and Microsoft Sued for $3 Billion Over Alleged ChatGPT 'Privacy Violations' — vice.com. https://www.vice.com/en/article/wxjxgx/openai-and-microsoft-sued-for-dollar3-billion-over-alleged-chatgpt-privacy-violations. [Accessed August 14, 2023].

[3] Shared, but not up for grabs. *Nature Machine Intelligence*, 1(4):163–163, April 2019. doi: 10.1038/s42256-019-0047-y.

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

[5] Shazia Afzal, C Rajmohan, Manish Kesarwani, Sameep Mehta, and Hima Patel. Data readiness report. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pages 42–51. IEEE, 2021.

[6] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021.

[7] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 0–0, 2018.

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

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

[10] Nazanin Andalibi and Justin Buss. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2020.

[11] Jerone T. A. Andrews. The hidden fingerprint inside your photos. https://www.bbc.com/future/article/20210324-the-hidden-fingerprint-inside-your-photos, 2021. [Accessed June 30, 2022].

[12] Jerone T A Andrews, Przemyslaw Joniak, and Alice Xiang. A view from somewhere: Human-centric face representations. In *International Conference on Learning Representations (ICLR)*, 2023.

[13] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[14] McKane Andrus, Elena Spitzer, and Alice Xiang. Working to address algorithmic bias? don't overlook the role of demographic data. *Partnership on AI*, 2020.

[15] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 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)*, pages 249–260, 2021.

[16] Anelia Angelova, Yaser Abu-Mostafam, and Pietro Perona. Pruning training sets for learning of object categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 494–501, 2005.

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

[18] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer, 2021.

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

[20] Rich Barlow and Cydney Scott. Students can adjust their pronouns and gender identity in bu's updated data system. https://www.bu.edu/articles/2022/pronouns-and-gender-identities-in-updated-data-system/, November 2022.

[21] Dennis Baron. *What's Your Pronoun?: Beyond He and She*. Liveright Publishing, 2020.

[22] Alistair Barr. Google mistakenly tags Black people as 'gorillas,' showing limits of algorithms. *The Wall Street Journal*, 2015.

[23] Teanna Barrett, Quan Ze Chen, and Amy X Zhang. Skin deep: Investigating subjectivity in skin tone annotations for computer vision benchmark datasets. *arXiv preprint arXiv:2305.09072*, 2023.

[24] Tom Beauchamp and James Childress. Principles of biomedical ethics: marking its fortieth anniversary, 2019.

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

[26] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

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

[28] Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 289–298, 2019.

[29] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. Detecting discriminatory risk through data annotation based on bayesian inferences. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 794–804, 2021.

[30] A Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac. Representation in ai evaluations. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 519–533, 2023.

[31] Gabrielle Berman and Kerry Albright. Children and the data cycle: Rights and ethics in a big data world. *arXiv preprint arXiv:1710.06881*, 2017.

[32] Abeba Birhane. Algorithmic colonization of africa. *SCRIPTed*, 17:389, 2020.

[33] Abeba Birhane. Automating ambiguity: Challenges and pitfalls of artificial intelligence. *arXiv preprint arXiv:2206.04179*, 2022.

[34] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, 2021.

[35] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, 2021.

[36] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[37] Brianna Blaser and Richard E Ladner. 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, pages 1–8. IEEE, 2020.

[38] Borhane Blili-Hamelin and Leif Hancox-Li. Making intelligence: Ethical values in iq and ml benchmarks. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 271–284, 2023.

[39] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems (NeurIPS)*, 35:3663–3678, 2022.

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

[41] Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)*, 14(2):1–45, 2021.

[42] Joy Buolamwini. Gender shades. https://www.media.mit.edu/projects/gender-shades/overview/, n.d. [Accessed June 1, 2022].

[43] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 77–91. PMLR, 2018.

[44] Bradley Butcher, Vincent S Huang, Christopher Robinson, Jeremy Reffin, Sema K Sgaier, Grace Charles, and Novi Quadrianto. 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.

[45] Judith Butler. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal*, 40(4):519–531, 1988. ISSN 01922882, 1086332X.

[46] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[47] Mary E Campbell and Lisa Troyer. The implications of racial misclassification by observers. *American Sociological Review*, 72(5):750–765, 2007.

[48] Pierluigi Carcagnì, Marco Del Coco, Dario Cazzato, Marco Leo, and Cosimo Distante. 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.

[49] Caleb T Carr, Yeweon Kim, Jacob J Valov, Judith E Rosenbaum, Benjamin K Johnson, Jeffrey T Hancock, and Amy L Gonzales. An explication of identity shift theory. *Journal of Media Psychology*, 2021.

[50] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:19152–19164, 2022.

[51] Christopher D Chambers and Loukia Tzavella. The past, present and future of registered reports. *Nature human behaviour*, 6(1):29–42, 2022.

[52] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 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.

[53] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision (ECCV)*, pages 768–783. Springer, 2014.

[54] Brian X Chen. Hp investigates claims of 'racist' computers. https://www.wired.com/2009/12/hp-notebooks-racist/, December 2009.

[55] Te-Ping Chen. Why gender pronouns are becoming a big deal at work. *The Wall Street Journal. Retrieved October*, 15:2022, 2021.

[56] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14980–14991, 2021.

[57] Nicholas A Christakis and James H Fowler. Friendship and natural selection. *Proceedings of the National Academy of Sciences*, 111(supplement_3):10796–10801, 2014.

[58] Albert Clapés, Ozan Bilici, Dariia Temirova, Egils Avots, Gholamreza Anbarjafari, and Sergio Escalera. 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)*, pages 2373–2382, 2018.

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

[60] National Bioethics Advisory Commission et al. Ethical and policy issues in research involving human participants. 2001.

[61] 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.

[62] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. 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, January 2019. doi: 10.1109/tbiom.2019.2897801.

[63] A Feder Cooper, Ellen Abrams, and Na Na. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 46–54, 2021.

[64] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

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

[66] 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.

[67] 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].

[68] Kate Crawford. Artificial intelligence with very real biases. https://www.wsj.com/articles/artificial-intelligencewith-very-real-biases-1508252717, 2017. [Accessed May 15, 2022].

[69] Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021.

[70] Kate Crawford and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55:93, 2014.

[71] Nicolas Croce and Moh Musa. The new assembly lines: Why ai needs low-skilled workers too. https://www.weforum.org/agenda/2019/08/ai-low-skilled-workers/, August 2019.

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

[73] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. Ieee, 2005.

[74] Fida K Dankar, Marton Gergely, and Samar K Dankar. Informed consent in biomedical research. *Computational and structural biotechnology journal*, 17:463–474, 2019.

[75] Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 915–924, 2022.

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

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

[78] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[79] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

[80] Oliver Diggelmann and Maria Nicole Cleis. How the right to privacy became a human right. *Human Rights Law Review*, 14(3):441–458, 2014.

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

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

[83] Vincent Egan and Giray Cordan. 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.

[84] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

[85] Madeleine Clare Elish. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)*, 2019.

[86] Sonia Elks. Why twitter and instagram are inviting people to share their pronouns. https://www.context.news/big-tech/why-twitter-and-instagram-are-inviting-people-to-share-pronouns, October 2021.

[87] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. What people think ai should infer from faces. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 128–141, 2022.

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

[89] 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].

[90] European Union High-level Expert Group. Ethics guidelines for trustworthy ai: Building trust in human-centric ai. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.1.html, 2019. [Accessed August 7, 2023].

[91] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

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

[93] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.

[94] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Tackling documentation debt: a survey on algorithmic fairness datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13. 2022.

[95] Anne Fausto-Sterling. *Sexing the body: Gender politics and the construction of sexuality*. Basic books, 2000.

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

[97] Klaus Fiedler and Norbert Schwarz. Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1):45–52, 2016.

[98] Christian Fieseler, Eliane Bucher, and Christian Pieter Hoffmann. Unfairness by design? the perceived fairness of digital labor on crowdworking platforms. *Journal of Business Ethics*, 156:987–1005, 2019.

[99] Andrew P Founds, Nick Orlans, Whiddon Genevieve, and Craig I Watson. Nist special database 32-multiple encounter dataset ii (meds-ii). 2011.

[100] Jonathan B Freeman, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the part: Social status cues shape race perception. *PloS one*, 6(9):e25107, 2011.

[101] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2373–2380, 2009.

[102] Yun Fu, Ye Xu, and Thomas S Huang. Estimating human age by manifold analysis of face pictures and regression on aging features. In *2007 IEEE International Conference on Multimedia and Expo*, pages 1383–1386. IEEE, 2007.

[103] Sidney Fussell. 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].

[104] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

[105] Andrew C Gallagher and Tsuhan Chen. Understanding images of groups of people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 256–263, 2009.

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

[107] Denia Garcia and Maria Abascal. Colored perceptions: Racially distinctive names and assessments of skin color. *American Behavioral Scientist*, 60(4):420–441, 2016.

[108] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. 2018.

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

[110] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[111] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 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)*, pages 325–336, 2020.

[112] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[113] Patricia A George and Graham J Hole. Factors influencing the accuracy of age estimates of unfamiliar faces. *Perception*, 24(9):1059–1073, 1995.

[114] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. 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.

[115] Markos Georgopoulos, Yannis Panagakis, and Maja Pantic. Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and vision computing*, 102:103954, 2020.

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

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

[118] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Conference on Human Factors in Computing Systems (CHI)*, pages 1–19, 2022.

[119] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[120] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

[121] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[122] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011.

[123] Patrick J Grother, Mei L Ngan, Kayee K Hanaoka, et al. Face recognition vendor test part 3: demographic effects. 2019.

[124] Manuel Günther, Peiyun Hu, Christian Herrmann, Chi-Ho Chan, Min Jiang, Shufan Yang, Akshay Raj Dhamija, Deva Ramanan, Jürgen Beyerer, Josef Kittler, et al. Unconstrained face detection and open-set face recognition challenge. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 697–706. IEEE, 2017.

[125] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

[126] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102. Springer, 2016.

[127] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Conference on Human Factors in Computing Systems (CHI)*, pages 1–13, 2018.

[128] Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609, 2017.

[129] Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. An ethical highlighter for people-centric dataset creation. 2020.

[130] Alex Hanna and Tina M Park. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850*, 2020.

[131] Alex Hanna, Emily Denton, Razvan Amironesei, Andrew Smart, and Hilary Nicole. Lines of sight. https://logicmag.io/commons/lines-of-sight/, 2020. [Accessed February 7, 2023].

[132] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 501–512, 2020.

[133] Adam Harvey and Jules LaPlace. Exposing. ai, 2021.

[134] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[135] Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S Bernstein. 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*, pages 889–901, 2017.

[136] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[137] Caner Hazirbas, Yejin Bang, Tiezheng Yu, Parisa Assar, Bilal Porgali, Vítor Albiero, Stefan Hermanek, Jacqueline Pan, Emily McReynolds, Miranda Bogen, Pascale Fung, and Cristian Canton Ferrer. Casual conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness, 2022.

[138] Steven Y He, Charles E McCulloch, W John Boscardin, Mary-Margaret Chren, Eleni Linos, and Sarah T Arron. Self-reported pigmentary phenotypes and race are significant but incomplete predictors of fitzpatrick skin phototype in an ethnically diverse population. *Journal of the American Academy of Dermatology*, 71(4):731–737, 2014.

[139] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, pages 771–787, 2018.

[140] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

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

[142] Alex Hern. Flickr faces complaints over 'offensive' auto-tagging for photos. https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos, May 2015.

[143] Alex Hern. Google's solution to accidental algorithmic racism: Ban gorillas. https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people, January 2018.

[144] Alex Hern. Twitter apologises for 'racist' image-cropping algorithm. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm, September 2020.

[145] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021.

[146] Kashmir Hill. Wrongfully accused by an algorithm. *The New York Times*, 2020.

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

[148] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[149] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. 2018.

[150] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Conference on Human Factors in Computing Systems (CHI)*, pages 1–16, 2019.

[151] Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021.

[152] Ayanna Howard, Cha Zhang, and Eric Horvitz. 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)*, pages 1–7. IEEE, 2017.

[153] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 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.

[154] Han-Yin Huang and Cynthia CS Liem. Social inclusion in curated contexts: Insights from museum practices. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 300–309, 2022.

[155] Zhanyuan Huang, Yang Liu, Yajun Fang, and Berthold KP Horn. Video-based fall detection for seniors with human pose estimation. In *2018 4th international conference on Universal Village (UV)*, pages 1–4. IEEE, 2018.

[156] Jennifer L Hughes, Abigail A Camden, Tenzin Yangchen, 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.

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

[158] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 743–756, 2022.

[159] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 560–575, 2021.

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

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

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

[163] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, 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.

[164] Oliver Jesorsky, Klaus J Kirchberg, and Robert W Frischholz. 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*, pages 90–95. Springer, 2001.

[165] Julie Jiang, Emily Chen, Luca Luceri, Goran Murić, Francesco Pierri, Ho-Chun Herbert Chang, and Emilio Ferrara. What are your pronouns? examining gender pronoun usage on twitter. *arXiv preprint arXiv:2207.10894*, 2022.

[166] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 2020.

[167] Sonam Joshi. Why indians are sharing their pronouns on social media. https://timesofindia.indiatimes.com/india/why-indians-are-sharing-their-pronouns-on-social-media/articleshow/71669703.cms, October 2019.

[168] Rie Kamikubo, Utkarsh Dwivedi, and Hernisa Kacorri. Sharing practices for datasets related to accessibility and aging. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–16, 2021.

[169] Shivani Kapania, Alex S Taylor, and Ding Wang. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[170] Kimmo Karkkainen and Jungseock Joo. 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)*, pages 1548–1558, 2021.

[171] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[172] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Conference on Human Factors in Computing Systems (CHI)*, pages 3819–3828, 2015.

[173] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijaya-narasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[174] Jane Kaye, Edgar A Whitley, David Lund, Michael Morrison, Harriet Teare, and Karen Melham. Dynamic consent: a patient interface for twenty-first century research networks. *European journal of human genetics*, 23(2):141–146, 2015.

[175] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016.

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

[177] Suzanne J Kessler and Wendy McKenna. *Gender: An ethnomethodological approach*. University of Chicago Press, 1985.

[178] Florian Keusch. 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.

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

[180] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 587–597, 2021.

[181] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7:81–99, 2003.

[182] Jennifer Klima, Sara M Fitzgerald-Butt, Kelly J Kelleher, Deena J Chisolm, R Dawn Comstock, Amy K Ferketich, and Kim L McBride. Understanding of informed consent by parents of children enrolled in a genetic biobank. *Genetics in Medicine*, 16(2):141–148, 2014.

[183] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 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.

[184] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2093–2102, 2018.

[185] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 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.

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

[187] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 365–372, 2009.

[188] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, 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.

[189] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:728–740, 2020.

[190] Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S Ma. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12:161, 2021.

[191] Min Kyung Lee and Katherine Rich. 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)*, pages 1–14, 2021.

[192] John Leuner. A replication study: Machine learning models are capable of predicting sexual orientation from facial images. *arXiv preprint arXiv:1902.10739*, 2019.

[193] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015.

[194] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACM International Conference on Multimedia*, pages 3501–3509, 2021.

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

[196] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[197] Zhenyi Liu, Trisha Lian, Joyce Farrell, and Brian A. Wandell. Neural network generalization: The impact of camera parameters. *IEEE Access*, 8:10443–10454, 2020.

[198] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[199] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

[200] Zhongyu Lou, Fares Alnajar, Jose M Alvarez, Ninghang Hu, and Theo Gevers. Expression-invariant age estimation using structured learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):365–375, 2017.

[201] Mike Loukides, Hilary Mason, and D Patil. Of oaths and checklists. https://www.oreilly.com/radar/of-oaths-and-checklists/, 2018. [Accessed August 7, 2023].

[202] Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 199–212, 2022.

[203] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

[204] Debbie S Ma, Justin Kantner, and Bernd Wittenbrink. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53(3):1289–1300, 2021.

[205] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[206] Nicolas Malevé. On the data set's ruins. *AI & SOCIETY*, pages 1–15, 2020.

[207] Gianclaudio Malgieri and Jędrzej Niklas. Vulnerable data subjects. *Computer Law & Security Review*, 37:105415, 2020.

[208] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9562–9571, 2019.

[209] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 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)*, pages 416–423, 2001.

[210] Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind pareto fairness and subgroup robustness. In *International Conference on Machine Learning*, pages 7492–7501. PMLR, 2021.

[211] Deborah Mascalzoni, Roberto Melotti, Cristian Pattaro, Peter Paul Pramstaller, Martin Gögele, Alessandro De Grandi, and Roberta Biasiotto. 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.

[212] Anna McKie. South african university drops gender titles in student correspondence. https://www.timeshighereducation.com/news/south-african-university-drops-gender-titles-student-correspondence, July 2018.

[213] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

[214] Helen Meng, PC Ching, Tan Lee, Man Wai Mak, Brian Mak, Y Moon, Man-Hung Siu, Xiaoou Tang, H Hui, Andrew Lee, et al. The multi-biometric, multi-device and multilingual (m3) corpus. In *Proc. Workshop Multimodal User Authentication*, 2006.

[215] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2437–2445, 2020.

[216] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[217] Jacob Metcalf. "the study has been approved by the irb"': Gayface ai, research hype and the pervasive data ethics gap. *Pervade Team, Nov*, 2017.

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

[219] Milagros Miceli, Martin Schuessler, and Tianling Yang. 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.

[220] Alex Mihailidis, Brent Carmichael, and Jennifer Boger. 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.

[221] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3571–3583, 2021.

[222] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 220–229, 2019.

[223] Brent Daniel Mittelstadt and Luciano Floridi. The ethics of big data: current and foreseeable issues in biomedical contexts. *The ethics of biomedical big data*, pages 445–480, 2016.

[224] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[225] Carl-Maria Mörch, Abhishek Gupta, and Brian L Mishara. Canada protocol: An ethical checklist for the use of artificial intelligence in suicide prevention and mental health. *Artificial intelligence in medicine*, 108:e101934–e101934, 2020.

[226] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 51–59, 2017.

[227] Emad Mostaque (Stability AI). We actually used 256 a100s for this per the model card, 150k hours in total so at market price $600k. https://twitter.com/emostaque/status/1563870674111832066, August 2022. [Accessed August 12, 2023].

[228] Jessica Mudditt. 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].

[229] Guilherme Nascimento, Camila Laranjeira, Vinicius Braz, Anisio Lacerda, and Erickson R. Nascimento. A robust indoor scene recognition method based on sparse representation. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 408–415. Springer, 2018.

[230] 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.

[231] 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].

[232] 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].

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

[234] Cornelia Neuert, Katharina Meitinger, Dorothée Behr, and Matthias Schonlau. 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.

[235] Lokesh P Nijhawan, Manthan D Janodia, BS Muddukrishna, Krishna Moorthi Bhat, K Laxminarayana Bairy, Nayanabhirama Udupa, Prashant B Musmade, et al. Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research*, 4(3):134, 2013.

[236] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016.

[237] Roosa Norja, Linda Karlsson, Jan Antfolk, Thomas Nyman, and Julia Korkman. How old was she? the accuracy of assessing the age of adolescents' based on photos. *Nordic Psychology*, 74(1):70–85, 2022.

[238] Brian A Nosek and Daniël Lakens. Registered reports, 2014.

[239] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision (ECCV)*, pages 19–35. Springer, 2016.

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

[241] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *European Conference on Computer Vision (ECCV)*, pages 739–755. Springer, 2020.

[242] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8466–8475, 2018.

[243] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. Augmented datasheets for speech datasets and ethical decision-making. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 881–904, 2023.

[244] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[245] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[246] Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 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.

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

[248] Billy Perrigo. Inside facebook's african sweatshop. https://time.com/6147458/facebook-africa-content-moderation-employee-treatment/, February 2022.

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

[250] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

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

[252] AJ Piergiovanni and Michael Ryoo. Avid dataset: Anonymized videos from diverse countries. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16711–16721, 2020.

[253] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions. *Journal of cybersecurity*, 4(1):tyy001, 2018.

[254] Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. *arXiv preprint arXiv:2303.04838*, 2023.

[255] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

[256] Mattia Prosperi and Jiang Bian. Is it time to rethink institutional review boards for the era of big data? *Nature Machine Intelligence*, 1(6):260–260, 2019.

[257] Carina EA Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. Institutionalizing ethics in ai through broader impact requirements. *Nature Machine Intelligence*, 3(2):104–110, 2021.

[258] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT*, pages 1776–1826, 2022.

[259] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 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)*, 2021.

[260] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. 2021.

[261] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. You can't sit with us: exclusionary pedagogy in ai ethics education. In *ACM conference on Fairness, Accountability, and Transparency (FAccT)*, pages 515–525, 2021.

[262] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition. *arXiv preprint arXiv:2301.02560*, 2023.

[263] Daniel A Reid and Mark S Nixon. Using comparative human descriptions for soft biometrics. In *International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2011.

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

[265] Rashida Richardson, Jason M Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, 94:15, 2019.

[266] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision (ECCV)*, pages 17–35. Springer, 2016.

[267] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–1, 2020.

[268] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 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.

[269] Rata Rokhshad, Maxime Ducret, Akhilanand Chaurasia, Teodora Karteva, Miroslav Radenkovic, Jelena Roganovic, Manal Hamdan, Hossein Mohammad-Rahimi, Joachim Krois, Pierre Lahoud, et al. Ethical considerations on artificial intelligence in dentistry: A framework and checklist. *Journal of Dentistry*, page 104593, 2023.

[270] Norma RA Romm. Interdisciplinary practice as reflexivity. *Systemic Practice and Action Research*, 11:63–77, 1998.

[271] Adam Rose. Are face-detection cameras racist? http://content.time.com/time/business/article/0,8599,1954643-1,00.html, January 2010.

[272] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[273] Negar Rostamzadeh, Diana Mincu, Subhrajit Roy, Andrew Smart, Lauren Wilcox, Mahima Pushkarna, Jessica Schrouff, Razvan Amironesei, Nyalleng Moorosi, and Katherine Heller. Healthsheet: development of a transparency artifact for health datasets. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 1943–1961, 2022.

[274] Wendy Roth. *Race migrations: Latinos and the cultural transformation of race*. Stanford University Press, 2012.

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

[276] Myron Rothbart and Marjorie Taylor. Category labels and social reality: Do we view social categories as natural kinds? 1992.

[277] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 10–15, 2015.

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

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

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

[281] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

[282] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, may 2021. doi: 10.1145/3411764.3445518.

[283] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[284] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 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.

[285] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. Hci guidelines for gender equity and inclusivity. https://www.morgan-klaus.com/gender-guidelines.html, May 2020. [Accessed August 1, 2022].

[286] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 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, 2020.

[287] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

[288] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[289] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022.

[290] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Rebecca Pantofaru. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

[291] Marshall H Segall, Donald Thomas Campbell, and Melville Jean Herskovits. *The influence of culture on visual perception*. Bobbs-Merrill Indianapolis, 1966.

[292] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[293] Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 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)*, pages 3758–3769, 2021.

[294] Hera Siddiqui, Ajita Rattani, Karl Ricanek, and Twyla Hill. An examination of bias of facial analysis based bmi prediction models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2926–2935, 2022.

[295] Laura Silver. Smartphone ownership is growing rapidly around the world, but not always equally. https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/, August 2020.

[296] Eleanor Singer and Mick P Couper. 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.

[297] Richa Singh, Mayank Vatsa, Himanshu S Bhatt, Samarth Bharadwaj, Afzel Noore, and Shahin S Nooreyezdan. Plastic surgery: A new dimension to face recognition. *IEEE Transactions on Information Forensics and Security*, 5(3):441–448, 2010.

[298] Jolene D Smyth, Don A Dillman, Leah Melani Christian, and Mallory McBride. 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.

[299] Jacob Snow. Amazon's face recognition falsely matched 28 members of congress with mugshots. https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28, July 2018.

[300] Benjamin Sobel. 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.

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

[302] Gowri Somanath, MV Rohith, and Chandra Kambhamettu. Vadana: A dense dataset for facial image analysis. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2175–2182, 2011.

[303] Patrik Sörqvist, Linda Langeborg, and Mårten Eriksson. 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.

[304] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65, 2019.

[305] Aaron Springer, Jean Garcia-Gathright, and Henriette Cramer. Assessing and addressing algorithmic bias-but before we get there... In *AAAI Spring Symposia*, 2018.

[306] Madhulika Srikumar, Rebecca Finlay, Grace Abuhamad, Carolyn Ashurst, Rosie Campbell, Emily Campbell-Ratcliffe, Hudson Hongo, Sara R Jordan, Joseph Lindley, Aviv Ovadya, et al. Advancing ethics review practices in ai research. *Nature Machine Intelligence*, 4(12): 1061–1064, 2022.

[307] Ramya Srinivasan, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. Artsheets for art datasets. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2021.

[308] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 701–713, 2021.

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

[310] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2325–2333, 2016.

[311] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5050–5059, 2018.

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

[313] Keiko Tagai, Hitomi Ohtaka, and Hiroshi Nittono. Faces with light makeup are better recognized than faces with heavy makeup. *Frontiers in psychology*, 7:226, 2016.

[314] Harriet JA Teare, Megan Prictor, and Jane Kaye. Reflections on dynamic consent in biomedical research: the story so far. *European journal of human genetics*, 29(4):649–656, 2021.

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

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

[317] Graham Thomas, Rikke Gade, Thomas B Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, 2017.

[318] William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4903–4913, 2023.

[319] Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations. *arXiv preprint arXiv:2012.05463*, 2020.

[320] Antonio Torralba, Rob Fergus, and William T Freeman. 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.

[321] Carlos Toxtli, Siddharth Suri, and Saiph Savage. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–26, 2021.

[322] John Twigg. *The Right to Safety: some conceptual and practical issues*. Benfield Hazard Research Centre, 2003.

[323] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M Gavrila, 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)*, pages 10581–10590, 2019.

[324] 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].

[325] 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].

[326] 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].

[327] 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.

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

[329] 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].

[330] 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].

[331] 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.

[332] Tim Valentine, Michael B Lewis, and Peter J Hills. Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, 69(10):1996–2019, 2016.

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

[334] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, 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.

[335] Jenny Vestlund, Linda Langeborg, Patrik Sörqvist, and Mårten Eriksson. Experts on age estimation. *Scandinavian Journal of Psychology*, 50(4):301–307, 2009.

[336] James Vincent. 'an engine for the imagination': the rise of ai image generators. an interview with midjourney founder david holz. https://www.theverge.com/2022/8/2/23287173/ai-image-generation-art-midjourney-multiverse-interview-david-holz, August 2022. [Accessed August 12, 2023].

[337] James Vincent. Ai art tools stable diffusion and midjourney targeted with copyright lawsuit. https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart, January 2023. [Accessed February 10, 2023].

[338] James Vincent. Getty images is suing the creators of ai art tool stable diffusion for scraping its content. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit, January 2023. [Accessed February 10, 2022].

[339] Manuel C Voelkle, Natalie C Ebner, Ulman Lindenberger, and Michaela Riediger. 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.

[340] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. volume 130, pages 1790–1810. Springer, 2022.

[341] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 526–536, 2021.

[342] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9322–9331, 2020.

[343] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[344] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

[345] Ze Wang, Xin He, and Fan Liu. Examining the effect of smile intensity on age perceptions. *Psychological reports*, 117(1):188–205, 2015.

[346] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8919–8928, 2020.

[347] Olivia R Ware, Jessica E Dawson, Michi M Shinohara, and Susan C Taylor. Racial limitations of fitzpatrick skin type. *Cutis*, 105(2):77–80, 2020.

[348] Griffin M Weber, Kenneth D Mandl, and Isaac S Kohane. Finding the missing link for big biomedical data. *Jama*, 311(24):2479–2480, 2014.

[349] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K Denniston, Xiaoxuan Liu, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, 2022.

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

[351] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. Disability, bias, and ai. *AI Now Institute*, page 8, 2019.

[352] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[353] Lloyd Windrim, Arman Melkumyan, Richard Murphy, Anna Chlingaryan, and Juan Nieto. Unsupervised feature learning for illumination robustness. In *IEEE International Conference on Image Processing (ICIP)*, pages 4453–4457, 2016.

[354] 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].

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

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

[357] David Wright. A framework for the ethical impact assessment of information technology. *Ethics and information technology*, 13:199–226, 2011.

[358] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, pages 4038–4052, 2016.

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

[360] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[361] Yuanjun Xiong, Kai Zhu, Dahua Lin, and Xiaoou Tang. Recognize complex events from static images by fusing deep channels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1600–1609, 2015.

[362] Runhua Xu, Nathalie Baracaldo, and James Joshi. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*, 2021.

[363] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 506–523. Springer, 2020.

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

[365] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy's new clothes. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a, 2017. [Accessed October 22, 2022].

[366] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 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)*, pages 547–558, 2020.

[367] Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. In *International Conference on Machine Learning (ICML)*, pages 25313–25330. PMLR, 2022.

[368] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[369] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 813–822, 2022.

[370] Rui-Jie Yew and Alice Xiang. 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)*, page 1017–1027, 2022.

[371] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[372] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P Breckon. Measuring hidden bias within face recognition via racial phenotypes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 995–1004, 2022.

[373] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 94–108. Springer, 2014.

[374] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5818, 2017.

[375] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[376] Dora Zhao, Jerone T. A. Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning (ICML)*, 2023.

[377] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[378] Mingyuan Zhou, Haiting Lin, S. Susan Young, and Jingyi Yu. Hybrid sensing face detection and registration for low-light and unconstrained conditions. *Applied Optics*, 57(1):69–78, January 2018.

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

[380] Matthew Zook, Solon Barocas, Danah Boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, and Frank Pasquale. Ten simple rules for responsible big data research, 2017.

A Responsible Data Curation Checklist for Fairness and Robustness Evaluations

This checklist translates our HCCV data curation considerations and recommendations into action items for researchers and practitioners. Presented as a series of questions, these items are designed to stimulate discussions among data collection teams. The questions are purposefully worded to avoid binary responses, encouraging open-ended dialogues. The primary focus of the checklist is to underscore the ethical dimensions and ensure that teams address concerns encompassing purpose, consent and privacy, as well as diversity.

It is important to engage with the checklist as a preliminary exercise before beginning data collection. This approach promotes informed decision-making and minimizes risks, leading to more responsible and reliable outcomes.

Contextual diversity is acknowledged to avoid a one-size-fits-all approach. Moreover, customization is encouraged, as not all items apply universally; teams should modify or expand the checklist to align with their context and use case. As with existing AI ethics checklists [90, 201, 205, 225, 269, 357], it is important to recognize that the checklist is not a guarantee for ethical compliance; rather, it functions as a catalyst for discussion and reflection.

We understand that answering these questions is time-consuming, increasing the burden on data collection teams whose work is already undervalued [247, 282]. Therefore, when navigating through these lists, priority should be put on items related to the specific domain and task of interest. The level of engagement needed for each question will invariably differ. Keep in mind that the questions aim to spur ethical thinking during dataset development: "Ethics is often about finding a good or better, but not perfect, answer" [380].

A.1 Purpose

The questions in this section focus on eliciting strategies for curating HCCV evaluation datasets specifically for fairness and robustness assessments. They seek alignment with objectives and inquire about factors known to influence these assessments to ensure comprehensive evaluations. Moreover, the questions aim to assist in formulating clear dataset purpose statements, preventing ambiguity and misuse of data, as well as exploring external validation to enhance transparency and accountability.

Dataset Development Strategy

- Can you provide details about your strategy for developing a new dataset tailored specifically for conducting fairness and robustness assessments in the context of HCCV? How do you plan to ensure that this dataset is aligned with the objectives of evaluating fairness and robustness?

- Can you elaborate on the factors your dataset will encompass to comprehensively enable fairness and robustness evaluations for HCCV models? How do you intend to capture the primary factors, including data subjects, instruments, and environments, that influence these evaluations?

Dataset Purpose Statement

- Can you provide details about your plan to formulate a comprehensive dataset purpose statement? How will this statement effectively communicate the core motivations driving, e.g., data collection, outline the intended dataset composition, specify permissible uses of the data, and identify the specific audience you aim to serve with the dataset?

- Can you elaborate on your strategy for ensuring the accuracy and ethical alignment of your dataset's purpose statement? How do you plan to externally validate the content and ethical considerations of the statement?

- Can you provide insights into the benefits and implications of submitting your dataset's purpose statement as part of a research study proposal in the format of a registered report for your project?

A.2 Consent and Privacy

The questions in this section explore informed consent, legal compliance, and privacy protection measures within anonymization strategies. The questions emphasize clarity and voluntariness in consent processes to prevent coercion or misuse of data. Moreover, they attempt to elicit strategies for explaining data collection purposes, consent revocation, and accommodating diverse participation circumstances. Furthermore, the questions seek insights into addressing anonymization challenges, aiming to prevent re-identification risks, unauthorized exposure, and legal noncompliance, while preserving data utility and protecting data subjects' rights.

Informed and Voluntary Consent

- Can you elaborate on your approach to ensuring that you secure explicit, voluntary, and informed consent from all individuals who either appear in the dataset or can be discerned from it? How do you plan to handle consent for data annotators who may have disclosed personal information for the purposes of quantifying and addressing annotator perspectives and bias?

- Can you provide a comprehensive explanation of your strategy for conveying the purpose of data collection to the subjects? How do you intend to emphasize the utilization of their data, which includes various types of information such as facial, body, biometric images, as well as information about themselves and their environment, all in the context of assessing the fairness and robustness of HCCV systems?

- In what ways will you incorporate consent forms that are composed in plain language to enhance the understanding of AI technologies? How do you plan to make sure these forms effectively convey the intricacies of data usage?

- How do you plan to inform data subjects about their ability to withdraw consent at any given point during, or after, the data collection process? Can you provide details about the mechanisms you will have in place for facilitating this?

- Please provide insight into your strategy for collecting data from individuals below the age of majority or vulnerable individuals. How will you seek both guardian consent and voluntary informed assent in such cases?

- How do you plan to evaluate vulnerability along a continuous spectrum, taking into account contextual factors and recognizing that vulnerability is not solely binary or based solely on group affiliations, but can also be influenced by specific situations or circumstances?

- Can you also provide details about how you will consider the circumstances of participation, which might include the potential need for participatory design, assurances of compensation, provision of educational materials, and safeguards against authoritative structures? How will you address these various aspects in your approach?

- How do you intend to ensure that vulnerable individuals have a comprehensive understanding of the data usage and willingly provide informed assent? Can you outline the specific measures you intend to implement for this purpose?

- Can you elaborate on how you will respect the decision of a vulnerable individual who expresses dissent, regardless of the preferences of their guardian?

Consent Revocation Mechanisms

- How do you plan to integrate mechanisms that allow data subjects to easily withdraw their consent? Can you provide specifics on how this process will be designed and executed?

- Can you provide insights into the benefits and implications of implementing dynamic consent mechanisms that utilize personalized communication interfaces? How do you intend to ensure that these mechanisms adapt to the preferences and needs of individual data subjects?

- How do you intend to enable data subjects to actively participate in research activities and manage their consent preferences? Can you provide more details about the tools or processes you plan to put in place to achieve this?

- In what ways will you explore the feasibility of online platforms for consent management that are user-friendly and minimize complexity for data subjects? What steps will you take to ensure easy accessibility?

- Can you provide insights into the options you will provide to data subjects for granting consent? How will you offer choices between blanket consent, case-by-case selection, or opt-in based on specific data usage?

- Can you elaborate on your considerations regarding the formation of a steering board or charitable trust composed of representative subjects from the dataset? How do you envision this entity contributing to decision-making processes?

- How do you plan to empower data subjects to actively participate in decisions concerning the usage of their data? What mechanisms or channels will you establish to facilitate this involvement?

- Can you provide information about the method you will offer data subjects to easily and promptly revoke their consent? How will you ensure that this process is straightforward and accessible?

- How do you intend to address varying levels of technological know-how and internet access among data subjects? Can you detail the measures you will take to accommodate these variations?

- What alternatives do you plan to offer for revoking consent that do not rely solely on online-based processes? How will you ensure that individuals with different needs and preferences can effectively revoke their consent?

- How do you plan to assess the practicality and suitability of the chosen mechanisms for consent revocation, taking into account the expected dataset size and the resources available to you? What criteria will you use to evaluate their effectiveness?

Country of Residence Information

- How do you plan to address the fact that anonymization measures might not universally meet legal requirements in specific regions, necessitating additional considerations? Can you provide insights into your strategy for ensuring legal compliance while implementing anonymization?

- Can you elaborate on your approach to collecting information about the country of residence for each individual in your dataset? How do you intend to use this information to ensure legal compliance and address potential privacy concerns?

- How do you plan to familiarize yourself with the data protection laws that are applicable in the countries of residence of your data subjects? Can you provide details about your process for gaining this understanding and how you will apply it to your data curation project?

- How do you intend to prioritize safeguarding data subjects' rights as stipulated by the data protection laws in their respective countries? What steps will you take to ensure that the creation and utilization of the dataset strictly adhere to the relevant data protection regulations? Can you provide specifics about the measures you will put in place to achieve this?

- What mechanisms do you intend to implement to ensure the adaptability of your dataset management strategy to changing legislative requirements? Can you provide details about how you will monitor and accommodate legislative changes in your dataset management approach? Can you provide insights into how you will strike a balance between maintaining compliance and effective dataset management in dynamic legal environments?

Privacy-Sensitive Image Regions and Metadata

- How do you plan to implement measures that effectively safeguard against re-identification risks, encompassing singling out, linkability, and inference, within your anonymization approach?

- Can you elaborate on your strategy for redacting all image regions that could inadvertently disclose privacy-related information? How do you intend to comprehensively identify and address these regions?

- Can you elaborate on your strategy for the removal of elements such as body parts, clothing, and accessories for nonconsenting subjects to enhance privacy protection? Can you provide more details about the considerations and methods involved in this process?

- Can you elaborate on your strategy for the removal of text (possibly excluding copyright owner information) from the dataset's images to enhance privacy protection? Can you provide more details about the considerations and methods involved in this process?

- Can you explain your plan for empirically validating the chosen anonymization methods? How will you assess the methods' effectiveness in mitigating re-identification risks while preserving the utility of the data?

- Can you provide details about how human annotators will be engaged in the creation and verification of privacy leaking image region proposals for anonymization purposes? How will you ensure accuracy and consistency in this process?

- Can you provide details about how you intend to align region proposals predicted by algorithms with human judgment, addressing any potential failures or biases? Can you describe your strategy for maintaining a sensitive approach to these factors?

- What steps will you take to address jurisdiction-specific requirements that might necessitate human-generated proposals for biometric identifiers in order to comply with legal and regulatory standards?

- Can you elaborate on the measures you will take to prevent image metadata from inadvertently revealing unauthorized identifying information? How will you ensure that metadata remains privacy-conscious?

- How will you identify specific metadata elements that you intend to retain to ensure a comprehensive understanding during the evaluation process? Can you provide examples of the types of metadata you plan to retain for this purpose?

- How do you plan to replace or remove sensitive information within metadata while retaining its usefulness for fairness and robustness analyses? Can you provide insights into your approach for striking a balance in this regard?

A.3 Diversity

The questions in this section revolve around obtaining accurate image annotations related to identity, phenotype, environmental factors, and instruments, while upholding inclusivity, sensitivity, and privacy. Additionally, the questions attempt to elicit strategies for documenting identity, ensuring fair compensation, and effective (anonymous) communication.

Self-Reported Annotations

- How do you plan to acquire annotations for images directly from the data subjects, leveraging their self-awareness and contextual knowledge to enhance the accuracy and quality of annotations? Can you elaborate on the methods and strategies you intend to use for this purpose?

- Can you elaborate on your strategy for addressing biases and ensuring careful handling when inferring labels about individuals? Can you provide reasoning as to why labels about individuals will be inferred as opposed to being self-identified? How will you actively mitigate potential biases that may arise during the labeling process?

- How do you intend to consider the implications of inferred labels, for example, in relation to data access request rights?

Versatile and Inclusive Response Options

- How do you plan to enhance the accuracy and nuance of identity information collection by providing respondents with both closed-ended and open-ended response choices? Can you elaborate on your strategy for using open-ended responses to gather more detailed and comprehensive data?

- How do you intend to ensure inclusivity and prevent any potential implications of exclusion in the response choices you offer?

- Can you elaborate on your preparedness to manage the coding and analysis effort required for processing open-ended responses? What effective strategies do you plan to implement for managing and analyzing the data collected from open-ended questions? How will you handle the potential complexities and variations that can arise from these responses, ensuring that the insights and information derived can be accurately captured and utilized?

Dynamic Nature of Identity

- How do you plan to collect self-identified information on a per-image basis, accounting for the fact that identity is intrinsically contextual and temporal? Can you elaborate on your strategy for capturing nonstatic aspects of identity?

- Can you elaborate on your strategy for enabling data subjects to freely choose multiple identity categories without imposing any limitations? How will you ensure that subjects have the flexibility to express their identity in a comprehensive and unrestrictive manner?

- How do you intend to address potential requests for per-image updates to self-identified information provided by subjects over time, respecting their autonomy? What factors have you considered in relation to the potential effects of permitting updates?

Demographic Information

- How do you plan to collect precise biological age in years from data subjects to ensure an accurate representation of their age?

- Can you elaborate on your approach to gathering pronoun information from data subjects to enhance gender inclusivity and mitigate the risk of misgendering? How will you ensure that respondents feel comfortable providing this information?

- Can you explain your strategy for gathering consistent ancestry information from data subjects? How will you approach the collection of this information in a sensitive and inclusive manner?

- How do you intend to offer the option for data subjects not to disclose their sensitive attributes if they choose not to? Can you provide more details about how you will handle the sensitivity and privacy of these attributes?

Sensitive Attributes in Aggregate

- How do you plan to collect voluntarily disclosed sensitive attributes such as disability and pregnancy status? Can you elaborate on your approach to respecting the willingness of data subjects to provide these details?

- Can you provide insight into your strategy for reporting sensitive attributes, such as disability and pregnancy status, in aggregate data while safeguarding subjects' safety and privacy? How do you intend to ensure that individual identities are protected?

- Can you elaborate on your approach to relying on credible and appropriate sources for the categorization and definitions of sensitive attributes like disability or difficulty? How will you account for the potential variations in these definitions based on cultural, identity, and historical contexts?

Phenotypic and Neutral Performative Features

- How do you plan to collect phenotypic attributes, encompassing characteristics such as skin color, eye color, hair type, hair color, height, and weight? Can you provide insights into your strategy for obtaining these attributes in a sensitive and comprehensive manner?

- Can you elaborate on your approach to collecting a diverse range of neutral performative features, including aspects such as facial hair, hairstyle, cosmetics, clothing, and accessories? How do you intend to ensure inclusivity and accuracy in capturing these features?

Environment and Instrument Details

- How do you plan to gather data on environment-related factors, which encompass details such as image capture time, season, weather, ambient lighting, scene, geography, camera position, and camera distance? Can you provide insights into your strategy for capturing these factors accurately and comprehensively?

- Can you elaborate on your approach to collecting instrument-related factors concerning the imaging devices used, including aspects such as lens, sensor, stabilization, flash usage, and post-processing software? How do you intend to ensure accuracy in capturing these details?

- How do you plan to obtain environment- and instrument-related information? Can you provide more details about the methods you will use, such as self-reporting, annotator estimation, and sourcing information from Exif metadata? How will you leverage contextual knowledge from image subjects to enhance data quality?

- Can you explain your approach to handling information such as precise geolocation and user-added details in Exif metadata that might contain personally identifying information? How will you ensure compliance with copyright regulations (if applicable) while maintaining privacy and adhering to ethical considerations?

Annotators as Contributors

- How do you plan to document the identities of data annotators, including capturing demographic details such as pronouns, age, and ancestry? Can you provide insights into your strategy for gathering and preserving this information while respecting privacy and ensuring transparency?

- Can you elaborate on your approach to highlighting the contributions of annotators beyond data labeling in the dataset documentation after the curation process? How do you intend to accurately represent the multifaceted roles and contributions of annotators?

- How do you plan to report the demographic information of annotators to analyze potential sources of bias in dataset annotations? Can you provide more details about your proposed approach for conducting this analysis while ensuring privacy and ethical considerations?

Fair Treatment and Compensation

- How do you plan to ensure that all contributors receive compensation that exceeds the minimum hourly wage of their respective country or jurisdiction of residence? Can you provide insights into your compensation strategy to ensure fair and ethical remuneration?

- Can you elaborate on your approach to exploring alternative payment models, such as compensation based on the average hourly wage? How do you intend to determine a compensation structure that is both fair and reflective of contributors' efforts?

- How will you establish direct communication channels between dataset creators and contributors? Can you provide more details about the methods you intend to implement for effective and transparent communication?

- What communication methods do you plan to explore that maintain the anonymity of contributors? Can you provide insights into your approach to balancing communication and privacy needs, such as using anonymous feedback forms?

- Can you provide information about your strategy for developing clear and accessible plain language guides to facilitate various tasks, such as image submission and data annotation? How do you plan to ensure that these guides effectively assist contributors?

- How do you intend to ensure that contributors from diverse backgrounds can easily understand and follow any instructions provided? Can you elaborate on your approach to promoting inclusivity and accessibility in your communication and guidelines?

- Can you provide details about how you plan to subject your recruitment and compensation procedures to ethics review? What steps will you take to ensure that your procedures align with ethical considerations and best practices?

B Literature Review

Through a thematic search strategy, we identified relevant research studies and datasets, revealing deficiencies in current image data curation practices or proposing potential solutions. By utilizing Semantic Scholar and Google Scholar, we curated relevant papers covering a wide spectrum of themes, including:

- HCAI
- Human-subjects research
- HCCV datasets
- Dataset curation
- Ethical frameworks and considerations
- Data and model documentation
- Legal and regulatory considerations
- Privacy and data protection
- Consent
- Fairness
- Auditing and verification

- Guidelines and best practices
- Values in design
- Diversity and inclusion
- Representation
- Robustness and reliability
- Benchmarking and evaluation
- Bias detection and mitigation
- Critical AI
- Social implications
- Responsible AI

The themes were chosen based on our expertise and experience in designing CV datasets, training models, and developing ethical guidelines. To ensure a focused approach, we manually selected papers aligned with the scope of our study based on the relevance of a paper's title and abstract. This informed our initial categorization scheme, shown in Table 1, detailing key ethical considerations related to HCCV.

Initially broad, we further refined the categories to address the most prominent ethical issues pertaining to HCCV dataset curation, particularly for fairness and robustness evaluations. Consent and privacy categories were combined due to their interrelated nature and the influence of shared legal frameworks.

Table 1: Categorization scheme for ethical considerations in HCCV research

Category	Explanation
Purpose	The study discusses the underlying objectives and motivations for HCCV datasets.
Acquisition	The study discusses ethical considerations related to the acquisition, collection, and labeling of image data, including recruitment and compensation for contributors.
Consent	The study discusses consent and the responsible use of personal information.
Privacy	The study discusses privacy issues related to HCCV datasets or public data.
Ownership	The study discusses legal and ethical aspects of intellectual property rights in the context of HCCV datasets or public data.
Diversity	The study discusses factors concerning diversity, inclusion, and fair representation within HCCV datasets. This encompasses matters such as identifying and addressing biases, ensuring fairness, and mitigating discrimination.
Maintenance	The study discusses maintenance strategies for ensuring the integrity of HCCV datasets, including security measures.

Additionally, we integrated acquisition-related considerations into the categories of diversity, consent and privacy, as well as purpose, recognizing their interconnectedness in ethical image data collection, labeling, and usage. Maintenance-related matters were intentionally excluded from our scope, as these primarily pertain to post-dataset creation activities, while technical and organizational security measures are typically covered through consent forms. Ownership concerns, often intertwined with privacy issues, were incorporated into the consent and privacy category.

To establish a comprehensive view, we expanded our corpus as necessary. This encompassed examining cited works within our initial corpus, studies referencing our primary sources, and additional contributions by authors from our initial corpus. Our review was supplemented by incorporating publicly available resources from reputable sources, such as government bodies, private institutions, and reliable news organizations. In total, our analysis covered 500 research studies and online resources.