What People Think AI Should Infer From Faces

Severin Engelmann[*]
severin.engelmann@tum.de
Technical University of Munich, Chair of Cyber Trust
Munich, Germany

Chiara Ullstein[*]
chiara.ullstein@tum.de
Technical University of Munich, Chair of Cyber Trust
Munich, Germany

Orestis Papakyriakopoulos
orestis@princeton.edu
Princeton University, Center for Information Technology Policy
Princeton, USA

Jens Grossklags
jens.grossklags@in.tum.de
Technical University of Munich, Chair of Cyber Trust
Munich, Germany

ABSTRACT

Faces play an indispensable role in human social life. At present, computer vision artificial intelligence (AI) captures and interprets human faces for a variety of digital applications and services. The ambiguity of facial information has recently led to a debate among scholars in different fields about the types of inferences AI should make about people based on their facial looks. AI research often justifies facial AI inference-making by referring to how people form impressions in first-encounter scenarios. Critics raise concerns about bias and discrimination and warn that facial analysis AI resembles an automated version of physiognomy. What has been missing from this debate, however, is an understanding of how "non-experts" in AI ethically evaluate facial AI inference-making. In a two-scenario vignette study with 24 treatment groups, we show that non-experts (N = 3745) reject facial AI inferences such as trustworthiness and likability from portrait images in a low-stake advertising and a high-stake hiring context. In contrast, non-experts agree with facial AI inferences such as skin color or gender in the advertising but not the hiring decision context. For each AI inference, we ask non-experts to justify their evaluation in a written response. Analyzing 29,760 written justifications, we find that non-experts are either "evidentialists" or "pragmatists": they assess the ethical status of a facial AI inference based on whether they think faces warrant sufficient or insufficient evidence for an inference (evidentialist justification) or whether making the inference results in beneficial or detrimental outcomes (pragmatist justification). Non-experts' justifications underscore the normative complexity behind facial AI inference-making. AI inferences with insufficient evidence can be rationalized by considerations of relevance while irrelevant inferences can be justified by reference to sufficient evidence. We argue that participatory approaches contribute valuable insights for the development of ethical AI in an increasingly *visual* data culture.

CCS CONCEPTS

• **Computing methodologies → Computer vision tasks**; • **Social and professional topics → User characteristics**; *Computing / technology policy*; • **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

artificial intelligence, computer vision, human faces, participatory AI ethics

ACM Reference Format:
Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What People Think AI Should Infer From Faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533080

1 INTRODUCTION

Human faces and the information they convey are essential in human interaction. When seeing a person for the first time, humans rapidly and automatically make a variety of judgments, such as whether a person looks trustworthy or likable [75, 76, 78, 99]. People's faces can play a significant role in some of society's most important decision-making scenarios: first facial impressions can determine hiring choices [76, 84], election outcomes [6, 59, 77], or jail sentences [26, 105, 109]. Yet, we are often told not to judge a book by its cover, an imperative that it is morally wrong to form beliefs about a person based on insufficient evidence. Indeed, inferring inner character traits based on looks had been foundational for once lauded physiognomic and phrenological practices in organizations and institutions [22, 35, 83, 92, 93].

Today, research in psychology and evolutionary anthropology shows that first facial impressions have an "irresistible" force, but are nonetheless largely inaccurate [13, 27, 78, 99, 100]. This line of research provides ample evidence that there is no relationship between how we look and how trustworthy or intelligent we actually are. Surprisingly, another body of research studies continues to suggest that first facial impressions are accurate or, at least, not completely invalid [45, 51, 56, 62, 72, 80]. Commonly recognizing this latter body of literature, computer vision artificial intelligence (AI) – the computerization of visual perception – has recently developed datasets, algorithms, and models to automate social perception tasks in fields such as affective computing (e.g., [19]) and social

[*] Denotes equal contribution.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

robotics [15, 97]. Using computer vision AI, studies have claimed to successfully infer emotion expression and intensity [10, 25], sexual [60, 103] and political orientation [54, 107], as well as a variety of latent traits in personality assessments based on people's faces in images [4, 16, 30–32, 81, 88, 89, 108]. AI research has established tools for feature extractions from faces (e.g., Face++ [1], EmoVu[2]) as well as for open training datasets (ImageNet[3], First Impression V2[4], PsychoFlickr dataset[5]) and models [1, 11] for facial analysis AI.

Computer vision AI drives software that helps "make sense" of user images on social media for advertising purposes, video interviews in hiring software, or mood detection in car systems. The AI emotion recognition industry alone is said to be worth US37$ billion by 2026 [20]. AI systems play an increasingly important role in the semantic interpretation of our world, and because faces have an indispensable social signaling function, they are taken to be particularly revealing of who we are. But how should AI interpret people's faces? All imagery is semantically ambiguous and computer vision AI inference-making necessarily follows from the semantic annotation of visual data by humans, in most cases, by crowd-sourced platform workers [67, 74, 95]. This complicated ethical question has led to debates between policymakers, researchers in computational and social sciences, and companies that develop or use such AI. A number of research papers, including from the FAccT research community, have pointed out ethical challenges with regard to computer vision AI inferences [21, 22, 29, 34, 35, 68, 69, 82, 83, 87, 92–94]. However, we believe that such an effort must at least be cognizant of how "ordinary" people, i.e., non-experts in AI, evaluate the normativity of computer vision inferences.

In this work, we follow calls for more empirically-informed AI ethics [55, 85] and investigate what non-experts (N = 3745) think AI should and should not infer from portrait images – images that only show a person's face. Using a two-scenario vignette study with 24 treatment groups, we show that non-experts find AI latent trait inferences (e.g., intelligence) morally impermissible regardless of the decision context for which the inference is used for (advertising & hiring). A majority of subjects evaluates inferences such as gender, skin color, and emotion expression as morally permissible in the low-stake decision context (advertising) but impermissible in the high-stake decision context (hiring). None of our framing effects influenced subjects' evaluations indicating a strong value disposition toward AI facial analysis. We use the transformer-based model RoBERTa [63] to analyze subjects' 29,760 written justifications for each AI inference. We find that subjects raise ethical concerns about all AI inferences in both contexts. When justifying the normativity of an AI inference, subjects use one of two meta-principles: an AI facial inference is permissible when facial information warrants sufficient evidence or when making the inference results in beneficial outcomes. Our analysis illustrates the normative complexity behind facial AI inferences, and provides guidance for forthcoming technology policy debates.

2 RELATED WORK: THE IMPOSITION OF MEANING IN A VISUAL DATA CULTURE

2.1 Power dynamics between requesters and data annotators

Recently, several authors have raised ethical questions regarding the creation, management, and application of computer vision datasets. Computer vision companies (also known as "requesters") hire data processing companies, most often located in "less developed" countries, to perform efficient and cost-effective dataset creation, including data annotation. The emergence of a visual data culture – across Facebook's services alone, 2 billion images are shared every day[6] – together with the need for manual, human semantic labeling has led to the establishment of a data annotation industry[7] [67, 95]. Critical data science (broadly speaking) highlights challenges related to accountability and transparency gaps resulting from the near-unbounded power of computer vision AI companies and AI research institutes (i.e, requesters) to determine the interpretative potential of visual content [33, 68–70, 85, 87].

Studies find that requesters face little pressure to justify data labeling projects when hiring data processing companies for dataset labeling [67–69, 85]. In a field study on two data processing companies, Miceli et al. concluded that the work of image annotators is largely guided by the interests of the requester organization [68]. The authors report that this power dynamic does not allow image annotators to voice ethical concerns during the data labeling process. The hierarchical managerial structure at data processing companies restricts the possibility for the deliberative input by annotators [69]. In [68], the authors assert that "the one who is paying has the right to the imposition of meaning". To increase transparency and accountability of dataset creation, researchers have developed proposals to standardize documentation. For example, Gebru et al. suggest that each dataset should have a corresponding datasheet, explaining, among others, the purpose for which the dataset was created, the description of the images (or other data types), procedural aspects such as data cleaning and labeling, as well as the tasks and their unique contexts that the dataset is intended to be used for [33]. Holland et al. propose a "Dataset Nutrition Label" that specifies different modules, including the data origin, dataset variables, and ground truth correlations [41]. These and other standardized documentation practices [e.g., 70] can help AI developers to select more suitable datasets for their model development. However, such documentation practices are currently voluntary and rely entirely on the initiative and implementation of dataset creators.

2.2 Faces as sources of meaning and means for classification?

Authors have raised critical questions regarding a second key ethical challenge that is the subject of this work: What kind of inferences should a computer vision AI make about people based on visual data? Moreover, how do we justify what differentiates

[1] https://www.faceplusplus.com/
[2] https://www.programmableweb.com/api/emovu
[3] https://www.image-net.org/
[4] http://chalearnlap.cvc.uab.es/dataset/24/description/
[5] https://figshare.com/articles/dataset/zahra_plos_data_zip/6469577
[6] *Using Artificial Intelligence to Help Blind People 'See' Facebook*: https://about.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/
[7] For a contribution by investigative journalists on the data annotation industry, see: *A.I. Is Learning From Humans. Many Humans.* https://www.nytimes.com/2019/08/16/technology/ai-humans.html

permissible from impermissible facial inferences when the context application changes? Given the inherently semantic ambiguity of visual data, fixing the large space of interpretive possibilities to a selection of target variables is an act of classification that inevitably demands an ethical justification [22, 40, 47, 82, 87, 93, 93]. This particularly applies to inferences about people based on their facial looks. Human faces are among the most frequently used "objects of interpretation" in computer vision AI. A recent review of nearly 500 prominent computer vision AI datasets found that 205 were "face-based": no other object was represented more often in computer vision datasets than human faces [85]. Social psychologists assert that humans are "obsessed" with faces and that they "cannot help but form impressions based on facial appearances" [99–101]. On first encounter, faces influence first impressions and shape whether we think someone *appears* trustworthy, intelligent, assertive, or attractive (among other traits) [76, 78, 101]. In many ancient cultures, and still today, there are persistent beliefs that faces are "a window to a person's true nature" [101], the idea that there is a reliable relationship between facial appearance and character[8]. The "irresistible influence" of faces can be consequential: first impressions can determine to whom we speak at a social gathering, whether we perceive a politician to be trustworthy, or whether we judge a job applicant as intelligent [100, 101, 110].

Recently, computer vision AI has purportedly inferred such first facial impressions for a variety of different contexts, for example in social media and for automatic hiring software [5, 11, 30–32, 39, 88, 89, 98, 108]. In the United States alone, millions of job applicants have participated in automatic hiring procedures that assess, among others, candidates' faces to produce an employability score [82, 94]. Sensitive categories such as gender and race are often treated as "commonsense categories" in computer vision datasets [22, 69, 82, 87]. However, a recent comparison between computer vision datasets presents findings that some racial categories show more variance than others across datasets despite nominally equivalent categorization [47]. Buolamwini and Gebru show that facial analysis AI produces the highest error rate for darker-skinned women and the lowest error rate for lighter-skinned males [14]. Critical perspectives warn that gender and skin color classification by facial analysis AI echoes colonial acts of "reading race onto the body" [86]. Facial analysis AI tends to rely on binary, cis-normative gender classifications [46, 86], thereby neglecting a trans-inclusive view of gender. Emotion recognition and sentiment analysis based on facial expressions have been the subject of multiple AI research projects and a plethora of digital companies – from large corporations to startups – use AI to infer facial emotion expression for social media, hiring, education, health, or security [93]. Other studies present facial analysis AI that is "better" at inferring sexual and political orientation from facial features than people [54, 103]. Others have organized yearly "first impression challenges" – competitions to create benchmark vision models for automatic first impression inferences in job candidate screening[9]. Computer vision AI studies often embrace research studies that underscore the apparent validity of first impressions or that, at least, assert that the invalidity of first impressions is inconclusive [45, 51, 56, 62, 72, 80, 102]. However, there is strong evidence that first facial impressions do not go beyond a "kernel of truth" [13, 78, 79, 99–101].

The conviction that facial configurations are indicative of a person's character inevitably rests on the pseudoscientific ideas of physiognomy and phrenology. Once celebrated scientific theories, prominent figures in the field of physiognomy such as Caspar Lavatar, Ceseare Lombroso, and Francis Galton developed entire taxonomies of facial configurations with what they believed to be corresponding character interpretations (for a historic account on physiognomy, see [99]). Critical data science research points to several ethical concerns resulting from the AI classification of people based on their facial appearance. Hanley et al. criticize that inferences about people based on visual data necessarily represent only those factors of an inference concept that are visibly discernible [40]. Similarly, Stark & Hoey underscore a "fixation on the visible" in their conceptual analysis on the ethics of emotion recognition AI [93]. Computer vision AI inference-making can be presumptuous when designed to predict aims or intentions of people in images [49]. Such systems are morally objectionable because they treat individuals as objects of categorization [40, 50]. Studying the influential ImageNet dataset, Crawford & Paglen find "highly questionable semiotic assumptions [that] echoe(s) of nineteenth-century phrenology" [22]. Other authors call for a ban on "Physiognomic AI" altogether [94].

Research in fairness, accountability, and transparency has successfully produced different formalizations of fairness metrics and approaches for de-biased datasets. However, when it comes to fair visual data inferences it is the selection of target variables that requires careful ethical consideration. If such ethical evaluations are "subjective" and "inescapably political", then how can we make progress in justifying a line between permissible and impermissible inferences? Contributing to this metaethical challenge, we analyze non-experts' ethical evaluations of specific computer vision AI inferences in a low-stake advertising and a high-stake hiring context. We argue that the input of non-experts (i.e., their moral intuitions) can help us critically advance the debate concerning fair computer vision inferences. We consider a participatory approach to be *at least* complementary to conceptual ethical analyses. For example, much of AI ethics in companies and research institutes is guided by "principlism": efforts of expert groups defining often vague ethical principles for algorithmic systems such as transparency, justice or responsibility [44]. Principlism has recently received criticism (e.g., [71]) arguing that abstract ethical principles too often leave room for interpretation and are therefore particularly susceptible to forms of "ethics washing" [12]. Relying on ethical principles alone critically fails to account for the influence of unique contextual factors on the ethical status of AI inference-making. Moreover, by democratic principle, whenever power hierarchies lead to an accountability vacuum, non-expert "users" should have – minimally – a voice in

[8] In evolutionary psychology, current research debates whether facial attributes (first impressions) are solely innate, evolutionary adaptive heuristics [99] or whether they also have a learned, cultural dimension [78, 79].

[9] ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results: https://hal.archives-ouvertes.fr/hal-01381149, 2017 Looking at People CVPR/IJCNN Competition: https://chalearnlap.cvc.uab.cat/challenge/23/description/

formulating values for the interpretative potential of visual data, including their own. We see this as one element of a holistic approach to advance computer vision AI ethics. For the purpose of the current study, we developed a factorial vignette study that we describe in more detail in the next section. Experimental vignette studies have been extensively used in different fields (including human computer interaction, psychology, experimental philosophy, business ethics) to elicit participants' explicit ethical judgments in a variety of hypothetical scenarios [2, 3, 18, 36, 42, 52, 53, 66, 73]. Our study follows calls for more survey-based AI computer vision ethics [85] and more experimentally-informed AI ethics in general [55]. For a review on the value of studying the "moral intuitions" of non-experts in ethics and philosophy more generally, see [53].

3 METHODS AND EXPERIMENTAL PROCEDURE

3.1 Data Collection

3745 subjects (male = 50.7%, female = 48.9%, other = 0.4%) participated in our study. Subjects were recruited via Amazon Mechanical Turk. Only "Turkers" with an approval rating above 95% were selected for the study. We deliberately chose to conduct our study via this platform because Turkers have been indispensable for the labeling of some of the most important datasets in computer vision [91, 106]. Besides the large subject pool required for our study, we were interested to understand how a community involved in the labeling of computer vision datasets would ethically evaluate AI facial inference-making.

Our home institution does not require an ethics approval for questionnaire-based online studies. When conducting the study and analyzing the data, we followed standard practices for ethical research: presenting detailed study procedures, obtaining consent, not collecting identifiable information or device data, and using a survey service[10] that guaranteed compliance with the European Union's General Data Protection Regulation. The study did not include any deceptive practices. Subjects could drop out of the study at any point. All data were fully anonymized, the privacy of all subjects was maintained at all times during the study. Following recommended principles of ethical crowdsourced research [104], we first ran a pre-study with 120 Turkers to determine the average time it would take to complete the survey and used this reference time to determine a payout above the US minimum wage (*mean*= 8.03 min). In our study (N = 3745), the *mean* was 10.4 min (min = 3.35 min, max = 31.55 min).

3.2 Vignette Study

The experiment was a between-subject design; each participant was randomly assigned to one of 24 groups. The 24 groups were composed of three experimentally altered variables: two decision contexts (advertising vs. hiring), six evaluative adjective terms (reasonable, fair, justifiable, acceptable, responsible, appropriate), and the presentation or absence of a dictionary definition of the evaluative adjective term. The use of different evaluative adjective terms with or without a dictionary definition accounted for framing effects and tested the robustness of subjects' conception of a normative AI inference [53, 55, 64].

First, subjects were randomly assigned to one of two hypothetical decision contexts: either a low-stake advertisement scenario (n = 1869; mean per group = 155) or a high-stake hiring scenario (n = 1876; mean per group = 156). In the hypothetical advertisement scenario, participants were told that an advertising company deployed computer vision AI to make a variety of judgments about social media users based on their portrait image. Participants were told that the inferences were used to show users more suitable *product advertisements*. We explicitly referred to product advertisements to avoid associations with political advertisements that could have raised the stakes of the decision context. In the hypothetical hiring scenario, a declared high-stake decision context by other studies on algorithmic perception [48, 90], participants were told that a company used computer vision AI to make a variety of judgments about applicants based on their application photo. Subjects were told that portrait inferences were used, together with other assessment metrics, to determine whether or not a candidate is suitable for a job. These scenarios presented curated, *hypothetical* decision contexts typical in vignette research on moral phenomena [3, 52, 53] and fulfilled one of our study's main purposes: to understand whether non-experts evaluate the same set of AI facial inferences differently across low-stake and high-stake contexts. The vignettes can be found in the Appendix in Figs. 1 and 2.

Second, past research has shown that vignettes can be prone to framing effects and that such effects can indicate weak value dispositions in morally-laden scenarios [17, 53]. In our vignettes, the *evaluative adjective term* that prompted subjects' normative deliberation prior to the primary rating task could have exerted a framing effect. To control for this potential framing effect, each participant was assigned *one* of six evaluative adjective terms – reasonable, fair, justifiable, acceptable, responsible, or appropriate – when performing the rating task: "Do you agree or disagree that this sort of inference made by a software using artificial intelligence is [evaluative adjective term]?". This increased the external validity of our vignette. Using only the evaluative term "fair" could have biased subjects' ratings and justifications. Some people (and in fact cultures) associate the term "reasonable" more descriptively with logical thinking and deliberation while other cultures associate it more prescriptively, such as being honest and responsible [37]. The same was found for people's intuitions about perceptions of normality (also part descriptive, part prescriptive) [9].

Third, studies in experimental philosophy have used "definition vs. no definition" conditions to understand whether subjects use their own intuitive concept when they evaluate essentially contested concepts (such as: what is a reasonable inference?) [52, 53, 55, 64]. Accordingly, half of subjects were presented with a generic dictionary definition of the evaluative adjective term assigned to them, the other half was not. For example: "What do we mean by fair? Something is fair if it's based on equality without favoritism or discrimination." All definitions were taken from the Cambridge Dictionary and were slightly adjusted for our context (see Appendix Table 1). The "definition vs. no definition" treatment allowed us to further test the robustness of subjects' normative evaluations for specific AI inferences: If non-experts' normative judgments were arbitrary to the extent that they could be manipulated by

[10] SoSci Survey: https://www.soscisurvey.de/

the presentation of a different evaluative adjective term (fair vs. reasonable, for example) or absence of a generic definition of that term, then this would indicate subjects' concept of a normative AI inference to lack robustness. Subjects would then have a low value disposition toward AI facial analysis inferences (studies in experimental philosophy typically use such and similar framing conditions see, for example, [17, 23, 53, 64]).

3.3 Facial inferences

To allow for comparison across contexts, inferences needed to have an acceptable degree of appropriateness for two very different decision contexts: advertising and hiring. To keep the cognitive load of our subjects at an acceptable level, we restricted the number of inferences rated and justified by each subject. We decided to present subjects with a total of eight inferences, first asking them to rate their agreement/disagreement and then to provide a short, written justification for each inference rating. We selected the inference "emotion expression" due to its prevalence in emotion detection AI [20, 93]. Similarly, the two inferences "skin color" and "gender" are common attributes in AI inference-making [14, 46]. Four inferences – "trustworthiness", "assertiveness", "intelligence", "likability" – were selected for their importance in studies on *human* first impression-making [76, 78, 79, 99–101]. Finally, we wanted to understand how subjects would evaluate a facial accessory. We chose "glasses" instead of piercings or tattoos, for example, because the latter two objects exist in more diverse forms. We constructed an 8-item scale to measure agreement with these eight facial inferences made by an AI on a 7-point Likert scale (1 = "strongly agree" to 7 = "strongly disagree", "can't answer"). We did not present subjects with sample portraits, since the impression they would have formed based on the face in the portrait would have likely influenced their normative judgments [99, 100]. The goal of the study was to explore non-experts' ethical evaluations of facial AI inferences *in principle*.

3.4 Classification of subjects' justifications

After rating each inference, subjects were asked to justify their evaluation in a written statement. This allowed us to understand the rationale behind subjects' inference ratings and increased data quality (e.g., understanding the plausibility and validity of evaluations, see, [57, 58]). While there is an entire research field dedicated to studying first impressions (e.g., [99–101]), we could not identify studies investigating people's *ethical evaluations* of such first impressions. This meant that we could not draw from an existing coding scheme for the classification of the 29,760 written justifications. Therefore, we derived the codes directly from the textual corpus. The manual coding process consisted of two iterative cycles. First, one researcher labelled 500 comments to discover major recurring types of reasoning. Another researcher labelled 250 of these comments with the same intent. The researchers then met to discuss and refine the set of identified "justification labels". In a second coding cycle, we randomly sampled 1,250 comments. Two researchers independently added a justification label to each comment. The intercoder reliability was high (Krippendorff's alpha = 0.953). In case of disagreement between the two coders, the comment was discussed with and reviewed by a third researcher. The final set of justification types consisted of the following: 1. "AI can tell", 2. "AI cannot tell", 3. "Inference relevant for decision", 4. "Inference not relevant for decision", 5. "Inference creates harm", 6. "AI has human biases", and 7. "Incomprehensible responses".

Based on this developed coding scheme, we used the language model RoBERTa [63] to analyze the remaining comments. RoBERTa is a more efficiently trained version of BERT [24], an NLP architecture designed for general-purpose language understanding. This required collecting 100 example comments for each justification type (i.e., code). One researcher collected 100 example comments for each justification type. A second researcher then verified classifications. Disagreement was resolved by a third researcher. We split our labeled dataset in 1,001 training and 250 test samples, and performed over-sampling of the smaller classes to create a balanced training dataset. The final optimized model had an overall accuracy on the test set of 95% and each label's F-1 score was higher than 0.94. For the optimization process, we used a learning rate of 3e-5, a maximum sequence length of 32 tokens, and warm-up initialization. We then predicted the labels of the remaining justifications based on the trained model. For the class overview with F-1 scores, see Appendix Table 7.

Our analysis strategy comprised statistical testing of subjects' inference ratings, an exploratory factor analysis, automated text classifications, and a multivariate analysis of variance with follow-up tests. Given the large number of subjects in our sample, we calculated the effect sizes for all significant (p<0.01) test results on subjects' ratings.

4 RESULTS

4.1 The consequentiality of the scenario influences non-experts' ethical evaluations of AI facial inferences

We first compared mean aggregate ratings of all inferences between the advertisement and the hiring scenario. A two-sided Welch two-sample t-test found subjects showed greater preference for the same set of inferences in the advertisement scenario (*mean*=3.85; *SE*=1.06) than in the hiring scenario (*mean*=4.41; *SE*=1.2). The difference was significant (t(3687.3)=-15.30; *P*<0.001; 95% *CI*: (-0.64, -0.49)) and represented a small to medium effect (*d*=0.50) (Fig. 1a).

We then compared mean ratings for each inference in the advertisement and hiring scenarios using a two-sided Welch two-sample t-test with Bonferroni corrections for eight tests (Fig. 1b). Subjects rated the inferences gender, emotion expression, wearing glasses, and skin color (e.g., skin color, *mean* AD=2.88, mean HR=4.19; *d*=0.60; *P*<0.001; 95% *CI*: (-1.44, -1.17)) significantly more positively in the low-stake advertisement than in the high-stake hiring scenario. In contrast, the inference ratings for intelligence, trustworthiness, and likability (e.g., likability, *mean* AD=5.04, *mean* HR=5.16; *d*=0.06; *P*=0.31; 95% *CI*: (-0.24, -0.006)) did not show a significant difference between the two scenarios. Ratings for the assertiveness inference were significantly different between the two scenarios, but the effect size was negligible (*mean* AD=4.69, *mean* HR=4.89; *d*=0.10; *P*=0.01; 95% *CI*: (-0.32, -0.078)).

To summarize, comparing the inference ratings solely based on the grouping variable *context*, the consequentiality of the decision context influenced subjects' ratings: in the hiring context, subjects showed significantly more disagreement with the AI inferences

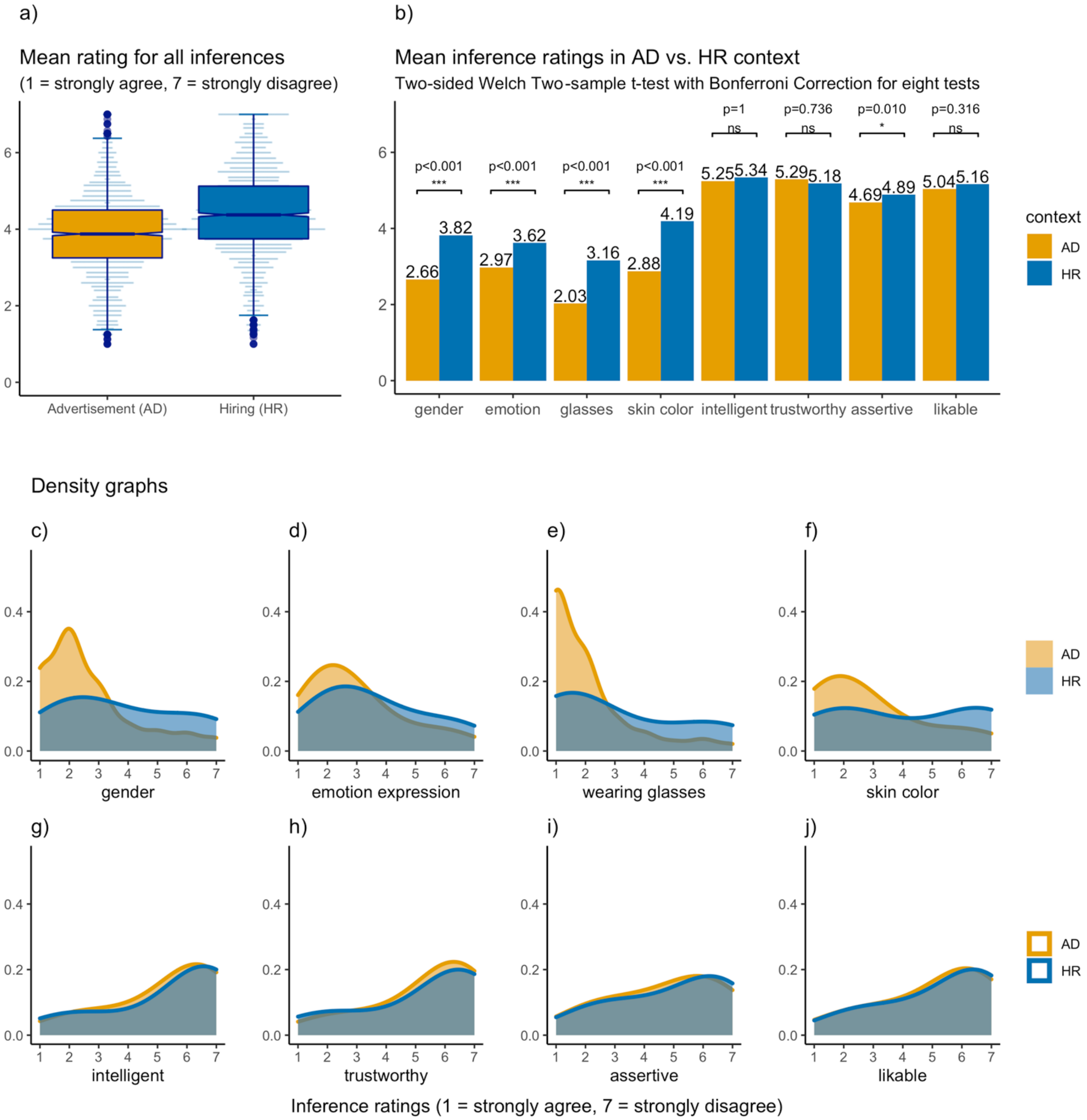

Figure 1: (a) Mean aggregate ratings for inferences were more positive in the advertising context than in the hiring context. (b) Participants rated the inferences gender, skin color, emotion expression, and wearing glasses significantly more positively in the low-stake advertisement than in the high-stake hiring scenario. Subjects rejected inferences intelligent, trustworthy, assertive, and likable regardless of the decision context: The inference ratings for intelligent, trustworthy, and likable did not show a significant difference between the two scenarios. Only ratings for the inference assertive were significantly different between the two scenarios, but the effect was negligible (see Appendix 5 for statistics). (c-j) Density plots of inference ratings. 1 = strongly agree; 7 = strongly disagree; 4 = neutral.

gender, skin color, emotion expression, and glasses than in the advertising context. Cohen's *d* was particularly large for ratings on gender, skin color, and wearing glasses between the two contexts. This difference did not replicate to ratings for the inferences trustworthiness, intelligence, assertiveness, and likability (Fig. 1).

4.2 Subjects differentiate between "first-order" and "second-order" inferences

To explore underlying constructs in our set of eight inferences, we conducted an exploratory factor analysis (EFA) (Appendix 6). Parallel analysis, scree plot, and the MAP criterion all suggested

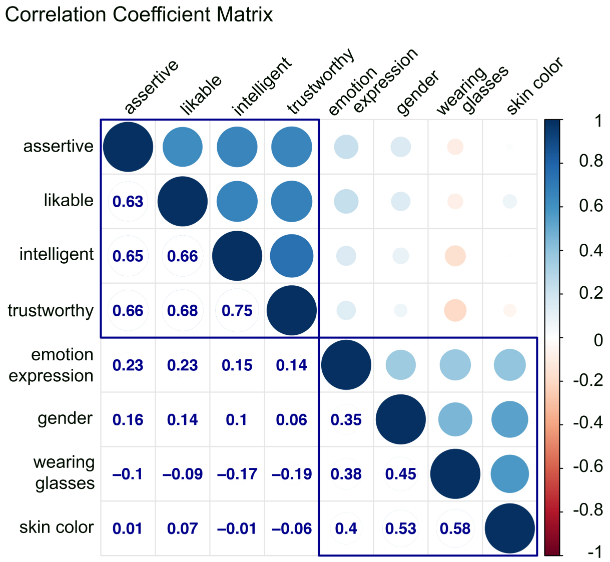

Figure 2: Exploratory factor analysis (EFA) resulted in two underlying constructs for subjects' ratings. One factor included the emotion expression, gender, wearing glasses, and skin color inferences. We termed this set of inferences *first-order inferences.* The other factor included the latent trait inferences assertive, likable, intelligent, and trustworthy. We termed this set of inferences *second-order inferences.*

two factors. One factor included the inferences gender, skin color, wearing glasses, and emotion expression. To use this group of inferences for further statistical comparison, we termed this construct *first-order inferences.* The other factor included the four latent trait inferences intelligence, trustworthiness, assertiveness, and likability. We termed this construct *second-order inferences.* We used these terms (first-order/second-order) as linguistic categories to reflect the statistical reality of subjects' ratings and less as an initial semantic interpretation of subjects' ethical evaluations. Both sub-scales had high reliability, the overall α was 0.89 for the factor labeled *second-order inferences* and 0.77 for the factor labeled *first-order inferences* (Fig. 2; see Appendix 6.6 for distribution of EFA factor scores).

4.3 Decision context only influences agreement with first-order inferences

We then extended our analysis to the entire set of treatment conditions. To test significant group differences among the 24 treatment groups on a combination of *first-order* and *second-order* factor scores from the EFA as a dependent variable, we computed a 2 (context: advertisement, hiring) x 6 (evaluative adjective terms) x 2 (definition, no definition) multivariate analysis of variance (MANOVA; Appendix 7). We controlled for main first-order justification theme, main second-order justification theme, AI knowledge, age, gender, occupation and education. Using Pillai's trace, there were significant main effects at an α-level of 0.01 for first-order justification (V=0.50, F(12, 6892)=190.76, P <.001, partial η^2 = 0.249), second-order justification (V=0.45, F(12, 6892)=164.60, P <.001, partial η^2 = 0.223), AI knowledge (V=0.03, F(8, 6892)=13.43, P <.001, partial η^2 = 0.015), and context (V=0.04, F(2, 3445)=73.68, P <.001, partial η^2 = 0.041) (Appendix Table 5).

Finally, univariate analysis with two separate ANOVAs on the *first-order* factor scores and on the *second-order* factor scores from the EFA revealed varying effect structures (Table 1; Appendix 7.2). With respect to the experimentally altered variables, *context* was the only significant treatment effect found, but only had an effect on ratings of first-order inferences (F(1, 3446) = 146.08, P <0.001, partial η^2 = 0.04). This finding supported the results from the two-sided Welch two-sample t-test. The experimental treatments *evaluative terms* and *definition vs. no definition* had no significant effect on subjects' ratings. This indicated that the subjects in our sample had a robust concept of a normative facial AI inference. AI knowledge had a small but significant effect on both inference ratings, whereas age had only a small effect on first-order ratings. Gender, occupation, and education did not have a statistically significant effect on subjects' ratings. Pairwise comparisons confirmed the results by identifying significant group differences between the advertisement and hiring context (Appendix 7.3).

4.4 Subjects find AI cannot tell second-order inferences in both contexts. Gender, skin color, and emotion expression produce more complex justifications.

4.4.1 Subjects evaluate the normativity of an AI inference according to two meta-principles. In their written evaluations, subjects considered whether or not an inference was proportional to the evidence (i.e., an epistemic justification) or whether making the inference resulted in positive or negative outcomes (i.e., a pragmatic justification). Representing epistemic principles, we introduced two codes: "AI can tell" and its opposite "AI cannot tell". For example, the comment *"I believe that someone's facial expressions can easily tell if they are assertive. I feel like facial expressions are easy to read and a computer could do that even better."* (assertiveness, HR) was classified as "AI can tell". The comment *"A person's intelligence is internal and based on learning, education, and other experiences. This can't be reflected in someone's looks."* was classified as "AI cannot tell" (intelligence, HR).

With the second meta-principle, subjects considered pragmatic reasons: we identified two contrary justification types "Inference relevant for decision" and "Inference not relevant for decision". The justification *"The reason I believe it is appropriate...is because this will help to select the potential candidate that possesses the assertiveness that could be useful for the job."* was classified as "Inference relevant for decision". The comment "*I don't think assertiveness makes or breaks a job applicant*" was classified as "Inference not relevant for decision" (both assertiveness, HR). A third justification type "Inference creates harm" classified comments stating AI inference-making could be harmful if used as part of the decision-making process (e.g., discrimination due to racism or sexism). For example, the justifications "*Seems like phrenology where intelligence and other traits were determined by the shape of someones head.*" (intelligence,

Table 1: Follow-up ANOVAs for factor scores from exploratory factor analysis (EFA)

	ANOVA for first-order					**ANOVA for second-order**				
	SS	df	F	Bonferroni	part. η^2	SS	df	F	Bonferroni	part. η^2
(Intercept)	7.32	1	22.22	**0.000**	0.006	5.135	1	15.399	**0.001**	0.004
Justifications										
first-order justifications	946.163	6	478.774	**0.000**	**0.455**	46.331	6	23.157	**0.000**	**0.039**
second-order justifications	18.785	6	9.506	**0.000**	**0.016**	844.717	6	422.212	**0.000**	**0.424**
Control Variables										
AI knowledge	14.069	4	10.679	**0.000**	**0.012**	26.058	4	19.537	**0.000**	**0.022**
age	9.939	5	6.035	**0.000**	0.009	5.648	5	3.387	0.052	0.005
gender	0.272	2	0.414	1.000	0.000	2.463	2	3.693	0.275	0.002
occupation	7.834	8	2.973	0.028	0.007	5.720	8	2.144	0.317	0.005
education	1.553	7	0.674	1.000	0.001	2.749	7	1.178	1.000	0.002
Experimental Variables										
context	48.115	1	146.081	**0.000**	**0.041**	2.325	1	6.972	0.092	0.002
terms	6.502	5	3.948	0.016	0.006	5.140	5	3.083	0.097	0.004
definition	0.161	1	0.487	1.000	0.000	0.293	1	0.880	1.000	0.000
Residuals	1135.010	3446				1149.065	3446			

Note:

All Bonferroni-corrected P-values are compared to a Bonferroni-corrected α = 0.005 for the computation of two ANOVAs.
Significant P-values and partial η^2 values of relevant size are marked in **bold**.
Partial η^2 = 0.01 small effect; partial η^2 = 0.06 medium effect; partial η^2 = 0.14 large effect.

AD) or *"Color should not matter in job hiring. This would be discrimination."* (skin color, HR) were classified as "Inference creates harm". Finally, a justification type that we called "AI has human biases" classified comments stating AI inference-making was flawed by biased human inference-making. Justifications in "AI has human biases" contained epistemic reasons (e.g., *"The software could be implanted with the bias of its creator"*; trustworthy, HR) or pragmatic reasons (e.g., *"The inference is unfair as the AI may be programmed to favor one sex over the other without context."*; gender, HR).

The classification results of subjects' written responses underline the semantic ambiguity of facial portraits: for each inference, we found a corpus of diverse explanations that fell back on epistemic and pragmatic accounts (the two meta-principles). We show the general line of subjects' justifications in Fig.3, where we map ratings (agreement/disagreement) to justification types. We complement subjects' general line of justifications with example comments. More example comments can be found in our "code book" in Appendix Table 8.

4.4.2 *Subjects believe AI second-order inferences are invalid inferences regardless of the decision-making context.*

The majority of subjects believed that faces do not provide sufficient evidence ("AI cannot tell") for inferences intelligence, trustworthiness, likability, and assertiveness (i.e., all second-order inferences) – regardless of the decision context. *"If you're just looking at a person and trying to determine if they're assertive, you're going to score no better than a random guess, I don't care how sophisticated this AI is."* (assertiveness, HR). Some subjects believed second-order inferences to be epistemically valid. *"Assertive people tend to have a set in their jaw, and eyes that is a bit more severe in the angles at the corners than those who are more passive...It might be possible to quantify those angles and measurements to have an AI program analyze the likelihood that they match those of assertive people...If you can come up with a mathematical formula to determine this, then the AI would be capable of measuring it."* (assertiveness, HR). The largest group of subjects agreeing with second-order inferences argued for their *relevance* in the hiring context (54.8%, "Inference relevant for decision"). Here, subjects did not express any epistemic reasoning, but asserted that such inferences were desirable qualities for employers. *"Almost always when you are working, you will work in teams and have to get along with others. You have to be likable to be successful on these teams - I would want the AI to try and assess this as best they could."* (likability, HR).

4.4.3 *Subjects believe first-order inferences are epistemically valid, but irrelevant and harmful in hiring.*

For the inferences emotion expression, wearing glasses, skin color, and gender, subjects' justification profile was more complex (Fig.3 e-j). The majority of subjects that agreed with these inferences believed in their epistemic validity in both contexts ("AI can tell"; AD: 62.6%, HR: 68.1%). However, in comparison to second-order inferences, the justification patterns differed between the advertising and hiring context: in the hiring

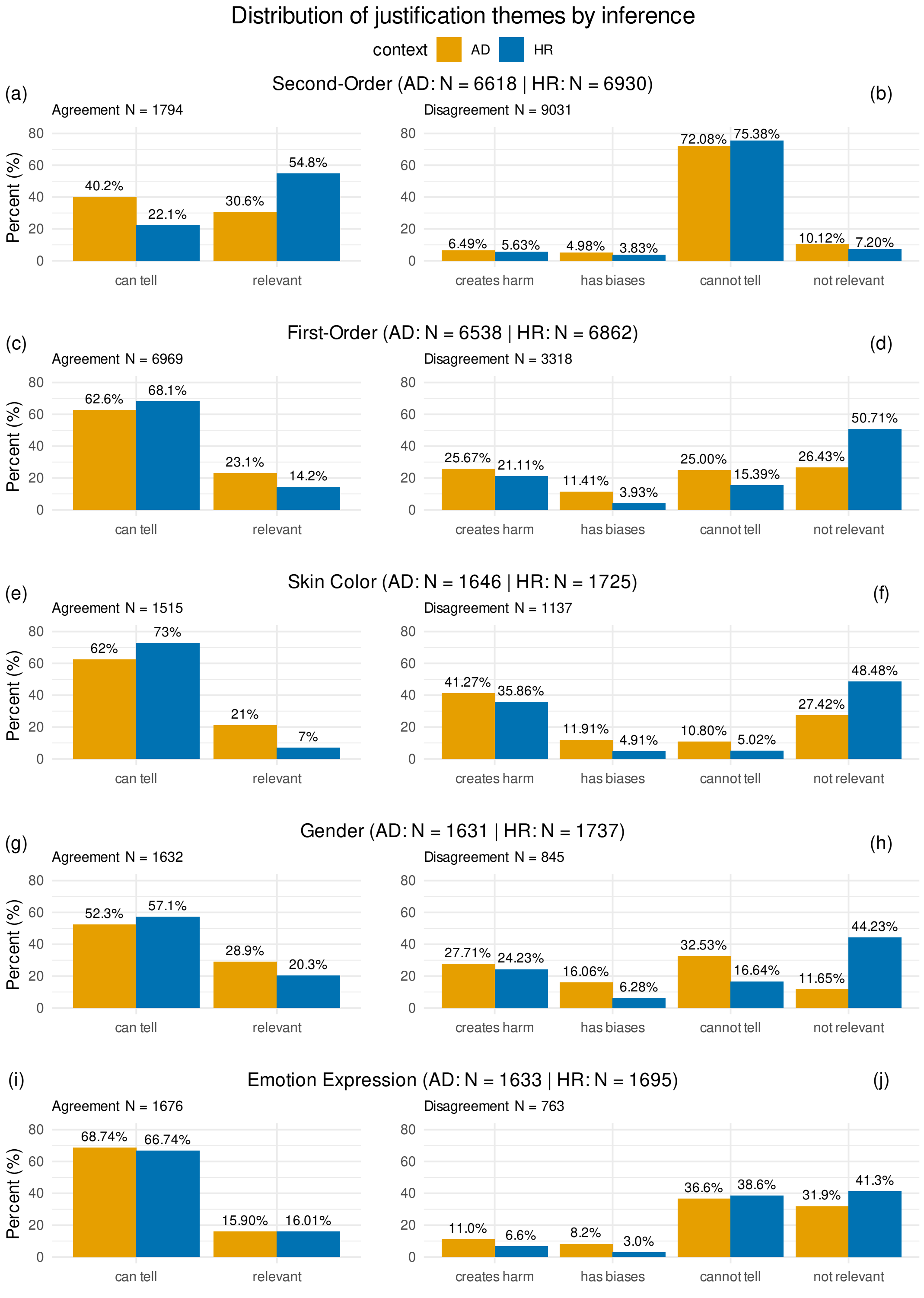

Figure 3: Distribution of justification types. Plots a) to o) present the proportions of the justification types used per context. E.g., for first-order ratings, 62.6% of participants in the AD context justified their agreement with an explanation allocated to the justification type "AI can tell" and 50.71% of respondents in the HR context justified their disagreement with an explanation related to the justification type "not relevant". The sum of N for AD and HR for an inference does not amount to the total N because the plot does not include individuals who neither agreed or disagreed. Percentages by context and agreement/disagreement do not sum up to 100%, since the visualization does not include a minority of individuals who provided a counter-intuitive justification based on their score.

context, considerations of relevance became more important reasons to reject an inference in comparison to the advertising context (Fig.3 c). The majority of subjects agreeing with skin color and gender in both contexts believed an "AI can tell" such inferences from facial information (Fig.3 e-h): *"Photos reveal this pretty easily assuming the photo is reasonably high rez. I would probably trust a computer to get this right more than some people."* (skin color, HR) or *"This is something that we, as humans can perceive with our sights, so an AI is definitely capable of inferring this."* (gender, AD). However, subjects that believed "AI can tell" skin color and gender still raised concerns in their written responses even when agreeing with these inferences. For example, subjects noted that accurately inferring skin color may be constrained by photo quality and lighting and may not be an indication of race or ethnicity as the following two comments illustrate: *"I believe a properly calibrated AI could estimate a person's skin color, but lighting, photo quality etc., would have to be accounted for. Also, skin color doesn't necessarily inform us about race."* (skin color, HR). *"Mixed feelings about this one – although skin color is something that can be visually seen in a photo, there is lots of room for error here depending on lighting in photo. Also, whether it's morally right is a whole different subject."* (skin color, AD). Likewise, for gender, subjects pointed to classification problems of non-binary gender identities: *"For the most part, male/female is an easy question, but there are many people that defy these binary categories that would be excluded."* (gender, HR).

Among the subjects rejecting skin color and gender in hiring, the most common justifications were "Inference not relevant for decision" (skin color: 48.48%; gender: 44.23%) and "Inference creates harm" (skin color: 35.86%; gender: 24.23%). With regard to skin color, most comments stated that skin color does not matter in hiring, while a few added that the inference was justifiable if it resulted in a more diverse workplace: *"This does not matter unless this information is being used to ensure a diverse workplace."* (skin color, HR). Subjects generally agreed that gender does not matter in hiring, however, some subjects asserted that some jobs may be more suitable for certain genders: *"Gender has nothing to do with how capable a person is to do a job unless the job itself requires a specific gender (which is very rare)."* (gender, HR). In contrast, subjects believed that both skin color (21%) and gender (28.9%) are a relevant AI inference in advertising: *"People with different skin colors need different products, and tend to shop for different styles, colors, and patterns."* (skin color, AD) or *"I think this is a 50/50 subject, but I believe personally that this is fair...Perhaps men wouldn't like to see advertisements for bras which would be avoided with this scan."* (gender, AD).

4.4.4 A majority of subjects believe emotion expression indicates emotion sensation. For emotion expression (Fig.3 i-j), subjects' agreement or disagreement mainly depended on whether or not they believed facial expressions to be a valid indicator for emotion sensation. Comments classified as "AI can tell" (agreement, AD: 68.74%, HR: 66.74%) claimed internal emotional states could be expressed via the face: *"It is reasonable to judge emotions by looking at a person's face, humans do it all the time. Though some faces can be more expressive than others."* (emotion expression, HR). Given that many Turkers have engaged in portrait image labelling tasks, we also found comments that highlighted the possibility of AI emotion expression inference based on previously conducted labelling tasks: *"A person's emotion can be seen pretty well by looking at a picture as I have done surveys in the past deciding emotion through facial expressions"* (emotion expression, AD). Comments classified as "AI cannot tell" (disagreement, HR: 38.6%, AD: 36.6%) stated the opposite. *"An emotion could be expressed, but the person may not actually be expressing it. In other words, the emotion viewed externally could be one of joy, but, inside the actual person, they may have a different emotion from what is outwardly being expressed."* (emotion expression, HR). The difficult relationship between emotion expression and emotion inference was also evident in comments with the justification types "Inference relevant for decision" (agreement, AD: 15.9%, HR: 16.01%) and "Inference not relevant for decision" (disagreement, AD: 31.9%, HR: 41.3%). To give one example, in comments classified as "Inference relevant for decision" in hiring, subjects claimed that employers may seek employees that need to be friendly, particularly in jobs involving customer interaction: *"Depending on the job emotional expressiveness may be a requirement, you don't want a person in a customer service position who's monotonous and robotic."* (emotion expression, HR).

5 KEY OBSERVATIONS & FINAL DISCUSSION

The vast abundance of digital imagery together with recent advances in computer vision analysis have raised concerns about the kinds of conclusions AI should make about people based on their face. How do we design computer vision AI in such a way that it will incorporate those preferences and values that are ethically desirable? We explored non-experts' normative preferences of AI portrait inferences in a two-scenario vignette study with 24 treatment groups. One MANOVA and two ANOVAs found that none of our framing effects influenced subjects' ratings, indicating that subjects have a robust, intuitive concept of a normative AI inference for both contexts. Future studies need to further explore how strong this normative concept is in light of other trade-offs such as cost-efficiency, narratives of bias-free technology, or success of the decision outcome, for example.

Conducting an exploratory factor analysis on subjects' evaluations of eight AI facial inferences, two inference categories emerge: we term one category of inferences first-order inferences and the other second-order inferences. Factor loadings of emotion expression as a first-order inference together with subjects' justifications suggest that a majority of the subjects in our sample subscribe to the so-called "Basic View" of emotions [28], which proposes that facial expressions (or "facial action units") are reliable indicators of emotion. Note that this perspective has recently been challenged by emotion researchers arguing that contextual and social factors lead to variability in facial emotion expression that make such inferences unreliable and unspecific [7, 93]. Nonetheless, subjects are aware of the volatility of AI emotion inference from facial expression. They assert that emotion expression as social signaling can be different from the internal phenomenological experience.

Finally, independent of the decision context, subjects believe AI should not draw inferences common in human first facial impression-making due to their epistemic invalidity, i.e., intelligence, likability, assertiveness, and trustworthiness [99–101]. Subjects raised concerns about all AI inferences in both contexts, even for the – perhaps intuitively – non-problematic "glasses" inference in the low-stake advertising context (Appendix Fig. 7). This leads us to assume that other facial AI inferences, such as beauty, sexual orientation, or political stance, that all have been inferred from faces using AI will likely draw their own justification profiles.

Our analysis highlights the normative complexity behind facial AI inferences. We find that some subjects use a *pragmatic* rationalization of AI facial inferences when they believe that an AI inference is relevant for (i.e., has a supposedly positive effect on) a decision's outcome. However, why should the normativity of a *vision*-based inference be evaluated by criteria other than evidence? The decision context does not have any bearing on the relationship between evidence and inference and therefore should not lead to a different normative evaluation. Thus, our results show that epistemically invalid AI vision inferences can be rationalized by considerations of relevance. The fact that AI research organizations, academic and commercial, commission data annotation companies to label visual data relevant for a specific application purpose necessarily creates a conflicting negotiation between epistemic and pragmatic considerations. Taken together, over-reliance on AI capabilities, narratives of bias-free technological decision-making, and beliefs in the relevance of an inference for the decision context may form a line of reasoning that supports justification of epistemically invalid AI inference-making. The ongoing publication of research studies that purportedly find a significant correlation between second-order inferences and facial information produces a quasi-epistemic legitimization of first-impression AI. Our study provides evidence that a vast majority of non-expert subjects do not form a justification of AI inference-making along these lines of reasoning.

Finally, how would experts differ in their justification of AI inference-making in comparison to non-experts? Indeed, critical data scientists argue that facial inferences are not reasonable because of their lack of scientific validity (evidentialists) [20, 92], while some AI experts deploying computer vision AI point to positive outcomes in terms of efficiency, cost-reduction, and flexibility that AI inference-making will facilitate [8, 43, 61, 65, 96]. Future studies will need to provide evidence for a unique ethical justification profile of AI vision inferences among AI expert groups. Other future studies should explore to what extent cultural factors play a role in evaluating the normativity of AI inferences based on visual data. We also believe it would be valuable to understand whether subjects evaluate AI video analysis inferences differently than AI image inferences. In fact, AI video analysis interprets visual content at the level of individual frames (i.e., decomposed as a collection of single images) [38].

We hope that the present study underlines the importance of including non-experts in the process of arguing for and against ethically permissible and non-permissible computer vision inferences. We expect norms regarding AI inference-making to shift over time. Allowing non-experts to engage in the formulation of goals and values for AI helps identify such shifts in sociocultural norms. Our study lays an important foundation for determining what types of inferences machines should and should not make about one of the most significant characteristics of us and our place in the social world: our faces.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments that improved the paper. For their valuable feedback we thank the participants of the 2021 CEPE/International Association of Computing and Philosophy Conference, the participants of the 2021 Ethics and Technology Lecture Series of the Munich Center for Technology in Society, and the participants of the Venice 2019 Metaethics of AI & Self-learning Robots Workshop.

FUNDING & SUPPORT

This research was conducted with the help of a Volkswagen Foundation Planning Grant. The Volkswagen Foundation had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no competing or financial interests.

REFERENCES

[1] Noura Al Moubayed, Yolanda Vazquez-Alvarez, Alex McKay, and Alessandro Vinciarelli. 2014. Face-based automatic personality perception. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 1153–1156. https://doi.org/10.1145/2647868.2655014

[2] Kwame Anthony Appiah. 2008. *Experiments in Ethics*. Harvard University Press.

[3] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6, 3 (2010), 128–138. https://doi.org/10.1027/1614-2241/a000014

[4] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences* 124 (2018), 150–159. https://doi.org/10.1016/j.paid.2017.12.018

[5] Mitja Back, Juliane Stopfer, Simine Vazire, Sam Gaddis, Stefan Schmukle, Boris Egloff, and Samuel Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science* 21, 3 (2010), 372–374. https://doi.org/10.1177/0956797609360756

[6] Charles C. Ballew and Alexander Todorov. 2007. Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17948–17953. https://doi.org/10.1073/pnas.0705435104

[7] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. https://doi.org/10.1177/1529100619832930

[8] Johannes M. Basch and Klaus G. Melchers. 2019. Fair and flexible?! Explanations can improve applicant reactions toward asynchronous video interviews. *Personnel Assessment and Decisions* 5, 3 (2019), Article 2. https://doi.org/10.25035/pad.2019.03.002

[9] Adam Bear and Joshua Knobe. 2017. Normality: Part descriptive, part prescriptive. *Cognition* 167 (2017), 25–37. https://doi.org/10.1016/j.cognition.2016.10.024

[10] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. 2016. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5562–5570. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Benitez-Quiroz_EmotioNet_An_Accurate_CVPR_2016_paper.html

[11] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. Facetube: Predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. 53–56. https://doi.org/10.1145/2388676.2388689

[12] Elettra Bietti. 2020. From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 210–219. https://dl.acm.org/doi/abs/10.1145/3351095.3372860

[13] Jean-François Bonnefon, Astrid Hopfensitz, Wim De Neys, et al. 2015. Face-ism and kernels of truth in facial inferences. *Trends in Cognitive Sciences* 19, 8 (2015), 421–422. https://doi.org/10.1016/j.tics.2015.05.002

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[15] Filippo Cavallo, Francesco Semeraro, Laura Fiorini, Gergely Magyar, Peter Sinčák, and Paolo Dario. 2018. Emotion modelling for social robotics applications: A review. *Journal of Bionic Engineering* 15, 2 (2018), 185–203. https://doi.org/10.1007/s42235-018-0015-y

[16] Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic personality and interaction style recognition from Facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 1101–1104. https://doi.org/10.1145/2647868.2654977

[17] Dennis Chong and James N. Druckman. 2007. Framing Theory. *Annual Review of Political Science* 10, 1 (2007), 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

[18] Cory J. Clark, Jamie B. Luguri, Peter H. Ditto, Joshua Knobe, Azim F. Shariff, and Roy F. Baumeister. 2014. Free to punish: A motivated account of free will belief. *Journal of Personality and Social Psychology* 106, 4 (2014), 501–513. https://doi.org/10.1037/a0035880

[19] Jeff F. Cohn and Fernando De la Torre. 2015. Automated face analysis for affective computing. In *The Oxford Handbook of Affective Computing*, Rafael A Calvo, Sidney D'Mello, Jonathan Matthew Gratch, and Arvid Kappas (Eds.). Oxford University Press, 131–150. https://doi.org/10.1093/oxfordhb/9780199942237.013.020

[20] Kate Crawford. 2021. Time to regulate AI that interprets human emotions. *Nature* 592, 7853 (2021), 167–167. https://doi.org/10.1038/d41586-021-00868-5

[21] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, A. Sánchez, et al. 2019. *AI Now 2019 report*. Research Report. The AI Now Institute, NYU. https://ainowinstitute.org/AI_Now_2019_Report.pdf

[22] Kate Crawford and Trevor Paglen. 2019. *Excavating AI: The politics of images in machine learning training sets*. Research Report. The AI Now Institute, NYU. https://excavating.ai/

[23] Joanna Demaree-Cotton. 2016. Do framing effects make moral intuitions unreliable? *Philosophical Psychology* 29, 1 (2016), 1–22. https://doi.org/10.1080/09515089.2014.989967

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* (2018). arXiv:1810.04805

[25] Abhinav Dhall, Oruganti V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. 2015. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 423–426. https://doi.org/10.1145/2818346.2829994

[26] Rafaële Dumas and Benoît Testé. 2006. The influence of criminal facial stereotypes on juridic judgments. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie* 65, 4 (2006), 237–244. https://doi.org/10.1024/1421-0185.65.4.237

[27] Charles Efferson and Sonja Vogt. 2013. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports* 3 (2013), Article 1047. https://doi.org/10.1038/srep01047

[28] Paul Ekman and Wallace V. Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues*. Malor Books.

[29] Severin Engelmann and Jens Grossklags. 2019. Setting the stage: Towards principles for reasonable image inferences. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 301–307. https://doi.org/10.1145/3314183.3323846

[30] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with Instagram pictures. In *Proceedings of the Workshop on Emotions and Personality in Personalized Systems*. 7–10. https://doi.org/10.1145/2809643.2809644

[31] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2016. Using Instagram picture features to predict users' personality. In *Proceedings of the 22nd International Conference on Multimedia Modeling*. 850–861. https://doi.org/10.1007/978-3-319-27671-7_71

[32] Bruce Ferwerda and Marko Tkalcic. 2018. Predicting users' personality from Instagram pictures: Using visual and/or content features?. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 157–161. https://doi.org/10.1145/3209219.3209248

[33] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[34] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336. https://doi.org/10.1145/3351095.3372862

[35] Jake Goldenfein. 2019. The profiling potential of computer vision and the challenge of computational empiricism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 110–119. https://doi.org/10.1145/3287560.3287568

[36] Armin Granulo, Christoph Fuchs, and Stefano Puntoni. 2019. Psychological reactions to human versus robotic job replacement. *Nature Human Behaviour* 3, 10 (2019), 1062–1069. https://doi.org/10.1038/s41562-019-0670-y

[37] Igor Grossmann, Richard P. Eibach, Jacklyn Koyama, and Qaisar B. Sahi. 2020. Folk standards of sound judgment: Rationality versus reasonableness. *Science Advances* 6, 2 (2020), Article eaaz0289. https://doi.org/10.1126/sciadv.aaz0289

[38] Yağmur Güçlütürk, Umut Güçlü, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel A.J. Van Gerven, and Rob Van Lier. 2017. Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing* 9, 3 (2017), 316–329. https://doi.org/10.1109/TAFFC.2017.2751469

[39] Sharath Chandra Guntuku, Weisi Lin, Jordan Carpenter, Wee Keong Ng, Lyle H. Ungar, and Daniel Preoţiuc-Pietro. 2017. Studying personality through the content of posted and liked images on Twitter. In *Proceedings of the 2017 ACM on Web Science Conference*. 223–227. https://doi.org/10.1145/3091478.3091522

[40] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 543–554. https://doi.org/10.48550/arXiv.2105.12754

[41] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint* (2018). https://doi.org/10.48550/arXiv.1805.03677

[42] Michael R. Hyman and Susan D. Steiner. 1996. The vignette method in business ethics research: Current uses, limitations, and recommendations. In *Proceedings of the Annual Meeting of the Southern Marketing Association*. 261–265.

[43] Srirang K. Jha, Shweta Jha, and Manoj Kumar Gupta. 2020. Leveraging artificial intelligence for effective recruitment and selection processes. In *Proceedings of the International Conference on Communication, Computing and Electronics Systems*. 287–293. https://doi.org/10.1007/978-981-15-2612-1_27

[44] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[45] Alexander Kachur, Evgeny Osin, Denis Davydov, Konstantin Shutilov, and Alexey Novokshonov. 2020. Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports* 10 (2020), Article 8487. https://doi.org/10.1038/s41598-020-65358-6

[46] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22. https://doi.org/10.1145/3274357

[47] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 587–597. https://doi.org/10.48550/arXiv.2102.02320

[48] Kimon Kieslich, Marco Lünich, and Frank Marcinkowski. 2021. The threats of artificial intelligence scale (TAI). *International Journal of Social Robotics* 13 (2021), 1563–1577. https://doi.org/10.1007/s12369-020-00734-w

[49] Owen C. King. 2019. Machine learning and irresponsible inference: Morally assessing the training data for image recognition systems. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Don Berkich and Matteo V. d'Alfonso (Eds.). Springer, 265–282. https://doi.org/10.1007/978-3-030-01800-9_14

[50] Owen C. King. 2020. Presumptuous aim attribution, conformity, and the ethics of artificial social cognition. *Ethics and Information Technology* 22, 1 (2020), 25–37. https://doi.org/10.1007/s10676-019-09512-3

[51] Karel Kleisner, Veronika Chvátalová, and Jaroslav Flegr. 2014. Perceived intelligence is associated with measured intelligence in men but not women. *PloS ONE* 9, 3 (2014), Article e81237. https://doi.org/10.1371/journal.pone.0081237

[52] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194. https://www.jstor.org/stable/3329308

[53] Joshua Knobe and Shaun Nichols. 2017. Experimental Philosophy. In *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[54] Michal Kosinski. 2021. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports* 11, 100 (2021). https://doi.org/10.1038/s41598-020-79310-1

[55] Steven R. Kraaijeveld. 2021. Experimental philosophy of technology. *Philosophy & Technology* 34 (2021), 993–1012. https://doi.org/10.1007/s13347-021-00447-6

[56] Robin S. Kramer and Robert Ward. 2010. Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology* 63, 11 (2010), 2273–2287. https://doi.org/10.1080/17470211003770912

[57] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement?. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 805–814. https://doi.org/10.1145/3209978.3210033

[58] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial*

Intelligence Research 69 (2020), 143–189. https://doi.org/10.1613/jair.1.12012

[59] Gabriel S. Lenz and Chappell Lawson. 2011. Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science* 55, 3 (2011), 574–589. https://doi.org/10.1111/j.1540-5907.2011.00511.x

[60] John Leuner. 2019. A replication study: Machine learning models are capable of predicting sexual orientation from facial images. *arXiv preprint arXiv:1902.10739* (2019). https://doi.org/10.48550/arXiv.1902.10739

[61] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 166–176. https://doi.org/10.1145/3461702.3462531

[62] Anthony C. Little and David I. Perrett. 2007. Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology* 98, 1 (2007), 111–126. https://doi.org/10.1348/000712606X109648

[63] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* (2019). arXiv:1907.11692

[64] Bertram F. Malle and Joshua Knobe. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology* 33, 2 (1997), 101–121. https://doi.org/10.1006/jesp.1996.1314

[65] Julie M. McCarthy, Talya N. Bauer, Donald M. Truxillo, Neil R. Anderson, Ana Cristina Costa, and Sara M. Ahmed. 2017. Applicant perspectives during selection: A review addressing "So what?,""What's new?," and "Where to next?". *Journal of Management* 43, 6 (2017), 1693–1725. https://doi.org/10.1177/0149206316681846

[66] David E. Melnikoff and Nina Strohminger. 2020. The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour* 4 (2020), 1258–1264. https://doi.org/10.1038/s41562-020-00943-3

[67] Milagros Miceli and Julian Posada. 2021. Wisdom for the crowd: Discoursive power in annotation instructions for computer vision. *arXiv preprint* (2021). https://doi.org/10.48550/arXiv.2105.10990

[68] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25. https://doi.org/10.1145/3415186

[69] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 161–172. https://doi.org/10.1145/3442188.3445880

[70] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 220–229. https://doi.org/10.1145/3287560.3287596

[71] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[72] Laura Naumann, Simine Vazire, Peter Rentfrow, and Samuel Gosling. 2009. Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin* 35, 12 (2009), 1661–1671. https://doi.org/10.1177/0146167209346309

[73] Shaun Nichols and Joshua Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41, 4 (2007), 663–685. https://www.jstor.org/stable/4494554

[74] Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval.* 557–566. https://doi.org/10.1145/1743384.1743478

[75] Christopher Y. Olivola, Dawn L. Eubanks, and Jeffrey B. Lovelace. 2014. The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly* 25, 5 (2014), 817–834. https://doi.org/10.1016/j.leaqua.2014.06.002

[76] Christopher Y. Olivola, Friederike Funk, and Alexander Todorov. 2014. Social attributions from faces bias human choices. *Trends in Cognitive Sciences* 18, 11 (2014), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

[77] Christopher Y. Olivola, Abigail B. Sussman, Konstantinos Tsetsos, Olivia E. Kang, and Alexander Todorov. 2012. Republicans prefer Republican-looking leaders: Political facial stereotypes predict candidate electoral success among right-leaning voters. *Social Psychological and Personality Science* 3, 5 (2012), 605–613. https://doi.org/10.1177/1948550611432770

[78] Harriet Over and Richard Cook. 2018. Where do spontaneous first impressions of faces come from? *Cognition* 170 (2018), 190–200. https://doi.org/10.1016/j.cognition.2017.10.002

[79] Harriet Over, Adam Eggleston, and Richard Cook. 2020. Ritual and the origins of first impressions. *Philosophical Transactions of the Royal Society B* 375, 1805 (2020), Article 20190435. https://doi.org/10.1098/rstb.2019.0435

[80] Ian S. Penton-Voak, Nicholas Pound, Anthony C. Little, and David I. Perrett. 2006. Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition* 24, 5 (2006), 607–640. https://doi.org/10.1521/soco.2006.24.5.607

[81] Lin Qiu, Jiahui Lu, Shanshan Yang, Weina Qu, and Tingshao Zhu. 2015. What does your selfie say about you? *Computers in Human Behavior* 52 (2015), 443–449. https://doi.org/10.1016/j.chb.2015.06.032

[82] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 469–481. https://doi.org/10.1145/3351095.3372828

[83] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* 145–151. https://doi.org/10.1145/3375627.3375820

[84] Nicholas O. Rule and Nalini Ambady. 2008. The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological Science* 19, 2 (2008), 109–111. https://doi.org/10.1111/j.1467-9280.2008.02054.x

[85] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37. https://doi.org/10.1145/3476058

[86] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8, 2 (2021), Article 20539517211053712. https://doi.org/10.1177/20539517211053712

[87] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. 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 (2020), 1–35. https://doi.org/10.1145/3392866

[88] Cristina Segalin, Dong Seon Cheng, and Marco Cristani. 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding* 156 (2017), 34–50. https://doi.org/10.1016/j.cviu.2016.10.013

[89] Crisitina Segalin, Alessandro Perina, Marco Cristani, and Alessandro Vinciarelli. 2016. The pictures we like are our image: Continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Transactions on Affective Computing* 8, 2 (2016), 268–285. https://doi.org/10.1109/TAFFC.2016.2516994

[90] Aaron Smith, Lee Rainie, Kenneth Olmstead, Jingjing Jiang, Andrew Perrin, Paul Hitlin, and Meg Hefferon. 2018. Public attitudes toward computer algorithms. *Pew Research Center* 16 (2018). https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2018/11/PI_2018.11.19_algorithms_FINAL.pdf

[91] Alexander Sorokin and David Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. In *Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* 1–8. https://doi.org/10.1109/CVPRW.2008.4562953

[92] Luke Stark. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 50–55. https://doi.org/10.1145/3313129

[93] Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 782–793. https://doi.org/10.1145/3442188.3445939

[94] Luke Stark and Jevan Hutson. 2021. Physiognomic artificial intelligence. *Available at SSRN* (2021). https://doi.org/10.2139/ssrn.3927300

[95] Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In *Proceedings of the Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.* 40–46. https://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/download/5350/5599

[96] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61, 4 (2019), 15–42. https://doi.org/10.1177/0008125619867910

[97] Adriana Tapus, Antonio Bandera, Ricardo Vazquez-Martin, and Luis V. Calderita. 2019. Perceiving the person and their interactions with the others for social robotics – A review. *Pattern Recognition Letters* 118 (2019), 3–13. https://doi.org/10.1016/j.patrec.2018.03.006

[98] Thales Teixeira, Michel Wedel, and Rik Pieters. 2012. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research* 49, 2 (2012), 144–159. https://doi.org/10.1509/jmr.10.0207

[99] Alexander Todorov. 2017. *Face Value: The Irresistible Influence of First Impressions.* Princeton University Press.

[100] Alexander Todorov, Sean G. Baron, and Nikolaas N. Oosterhof. 2008. Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience* 3, 2 (2008), 119–127. https://doi.org/10.1093/scan/nsn009

[101] Alexander Todorov, Christopher Y. Olivola, Ron Dotsch, and Peter Mende-Siedlecki. 2015. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology* 66 (2015), 519–545. https://doi.org/10.1016/j.tics.2014.09.007

[102] Richard J.W. Vernon, Clare A.M. Sutherland, Andrew W. Young, and Tom Hartley. 2014. Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences* 111, 32 (2014), E3353–E3361. https://doi.org/10.1073/pnas.1409860111

[103] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114, 2 (2018), 246–257. https://doi.org/10.1037/pspa0000098

[104] Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics* 49, 1 (2016), 77–81. https://doi.org/10.1017/S104909651500116X

[105] John Paul Wilson and Nicholas O. Rule. 2015. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychological Science* 26, 8 (2015), 1325–1331. https://doi.org/10.1177/0956797615590992

[106] Gerhard Wohlgenannt. 2016. A comparison of domain experts and crowdsourcing regarding concept relevance evaluation in ontology learning. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*. Springer, 243–254. https://doi.org/10.1007/978-3-319-49397-821

[107] Nan Xi, Di Ma, Marcus Liou, Zachary C. Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*. 726–737. https://ojs.aaai.org/index.php/ICWSM/article/view/7338

[108] Yan Yan, Jie Nie, Lei Huang, Zhen Li, Qinglei Cao, and Zhiqiang Wei. 2015. Is your first impression reliable? Trustworthy analysis using facial traits in portraits. In *Proceedings of the 21st International Conference on Multimedia Modeling*. 148–158. https://doi.org/10.1007/978-3-319-14442-9_13

[109] Leslie A. Zebrowitz and Susan M. McDonald. 1991. The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior* 15, 6 (1991), 603–623. https://doi.org/10.1007/BF01065855

[110] Leslie A. Zebrowitz and Joann M. Montepare. 2008. Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass* 2, 3 (2008), 1497–1517. https://doi.org/10.1111/j.1751-9004.2008.00109.x