AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI

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

Severin Engelmann[*]
severin.engelmann@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

Michel Hohendanner[‡]
michel.hohendanner@hm.edu
University of Applied Sciences
Munich, Center for Digital Sciences
and AI & Faculty of Design
Munich, Germany

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

ABSTRACT

Recent advances in computer vision analysis have led to a debate about the kinds of conclusions artificial intelligence (AI) should make about people based on their faces. Some scholars have argued for supposedly "common sense" facial inferences that can be reliably drawn from faces using AI. Other scholars have raised concerns about an automated version of "physiognomic practices" that facial analysis AI could entail. We contribute to this multidisciplinary discussion by exploring how individuals with AI competence and laypeople evaluate facial analysis AI inference-making. Ethical considerations of both groups should inform the design of ethical computer vision AI. In a two-scenario vignette study, we explore how ethical evaluations of both groups differ across a low-stake advertisement and a high-stake hiring context. Next to a statistical analysis of AI inference ratings, we apply a mixed methods approach to evaluate the justification themes identified by a qualitative content analysis of participants' 2768 justifications. We find that people with AI competence (N=122) and laypeople (N=122; validation N=102) share many ethical perceptions about facial analysis AI. The application context has an effect on how AI inference-making from faces is perceived. While differences in AI competence did not have an effect on inference ratings, specific differences were observable for the ethical justifications. A validation laypeople dataset confirms these results. Our work offers a participatory AI ethics approach to the ongoing policy discussions on the normative dimensions and implications of computer vision AI. Our research seeks to inform, challenge, and complement conceptual and theoretical perspectives on computer vision AI ethics.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; • **Applied computing** → *Law, social and behavioral sciences*; • **Security and privacy** → **Human and societal aspects of security and privacy**; **Social aspects of security and privacy**.

KEYWORDS

artificial intelligence, computer vision, human faces, ethics, public participation

ACM Reference Format:
Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Michel Hohendanner, and Jens Grossklags. 2022. AI-Competent Individuals and Laypeople Tend to Oppose Facial Analysis AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3551624.3555294

1 INTRODUCTION

Companies and research institutes increasingly produce and release artificial intelligence (AI) applications that draw conclusions about individuals from human faces [22, 33, 34]. One task of such facial processing technologies is facial analysis (hereafter called *facial analysis AI*), which classifies facial characteristics as demographic or physical traits [82] and even personality traits from portrait images. Driven by scientific advances in the areas of face-based inferences on intelligence, trustworthiness, likability and other personality traits [1, 5, 98], as well as sexual orientation [63, 96], such AI products find application in various domains including human resources and advertising. In response, a community of critical data scientists has raised ethical concerns regarding the development of such facial analysis AI [e.g., 24, 69, 70, 81, 85].

In policy-making, researchers from various disciplines have argued that the veracity of inferences from faces is not significant enough to counterbalance negative consequences [10], and have pointed out the unreliability of human inferences from faces, such as trustworthiness or intelligence [94, 95]. Others have highlighted the variability and context-dependency of emotions depicted in pictures and videos showing faces [6]. Members of the European

[*] Denotes equal contribution.
[†] Currently affiliated with Sony AI, Switzerland.
[‡] Also affiliated with University of Wuppertal, Industrial Design, Germany.

EAAMO '22, October 6–9, 2022, Arlington, VA, USA

ACM ISBN 978-1-4503-9477-2/22/10.
https://doi.org/10.1145/3551624.3555294

Parliament recently called "for a ban on the use of private facial recognition databases" [32]. Moreover, serious misclassifications have been uncovered in commercial gender detection tools [12] and job candidate selection software [80, 90]. Nonetheless, many industry actors see an enormous market potential – the AI emotion recognition industry alone is predicted to become worth multiple billion dollars in the coming years [23].

Fundamental questions are how to draw a line between ethically permissible and impermissible AI facial inferences as well as who should be involved in making these decisions. These two questions are central to understand how AI systems and their regulatory frameworks can be developed in a socially-sustainable manner. We contribute to this research debate by exploring how laypeople and individuals with AI competence evaluate facial analysis AI inference-making. We believe that both groups, potential future designers of AI systems and subjects of facial analysis AI, should play a more critical role in the development of ethical computer vision AI.

Prior work has illustrated that the general population (i.e., laypeople) may be aware that facial analysis AI applications exist but that it has little knowledge of their technological characteristics [14]. Mainstream media and science fiction contribute to the propagation of AI narratives that create unrealistic expectations of AI capabilities [13–15, 17, 20, 35, 43], and pay little attention to their feasibility [66]. Hopes and fears are part of AI narratives [17] and although some argue that current perceptions are skewed or extreme [15] such perceptions can influence the acceptance and adoption of AI systems by the general public [13, 15, 17, 35, 43, 66]. How popular narratives on technology, including the role of AI, can influence the imagination of future societies has, for instance, been explored using research through design and narrative analysis [e.g., 16, 44].

It has become increasingly clear that challenges arising from AI systems do not have purely technical solutions. For example, the decision to use one fairness metric over another requires value judgments that cannot be solved by formalistic approaches. Normative decisions *always* attract support, skepticism or rejection by different groups in society. Achieving consensus on topics such as "algorithmic fairness will be difficult unless we understand why people disagree in the first place" [77, p.1]. In the context of facial analysis AI, we believe it is important to understand how individuals with AI competence perceive AI inference-making and how their perception differs from the perception of AI inference-making by laypeople. Overall, we ask the following research question:

How do ethical justifications of AI inference-making from faces differ between individuals with AI competence and laypeople?

We build this research on our prior work in which we explored a conceptualization of reasonable inference [30] and asked laypeople how they evaluate such inferences [31]. In this study, we extend this work and compare evaluations of AI inference-making of laypeople with those of individuals with AI competence. We first survey researchers and students studying AI or computer vision AI (N=122) for our sample of "individuals with AI competence". We then compare their ratings and open-text justifications to a laypeople dataset (N=122). Furthermore, we analyze whether a range of demographic factors correlates with differences in the ethical evaluation of AI inference-making from portrait pictures. We confirm the results using a validation laypeople dataset.

2 RELATED WORK

2.1 Research on AI inferences of social constructs and character traits from faces

Many companies have developed facial analysis products used for market research, customer targeting, health care or education. For instance, Face++ sells services that infer "face related attributes including age, gender, smile intensity, [...] emotion, beauty" [33]. EmoVu [29] and FaceReader by Noldus perform facial expression analysis and infer, amongst others, personal characteristics and the six basic emotions [28] "happy, sad, angry, surprised, scared, and disgusted" [73]. Betaface and SkyBiometry classify glasses, beard, mustache, mood, or ethnicity [8, 9]. Faception claims to be able to identify people with high IQ [34].

The foundation for these analyses stems from research on inferences from human faces by humans. Research in evolutionary anthropology and psychology presents findings that humans "cannot help" but form first facial impressions despite their proven inaccuracy [10, 27, 74, 92, 93]. In the past, organizational and institutional physiognomic practices relied on making inferences about character traits from visual appearance [25, 39, 82, 88, 89]. Well-known for their contributions to physiognomy, Francis Galton, Caspar Lavatar or Cesare Lombroso, amongst others, developed taxonomies of character interpretations and corresponding facial configurations (see [92] for physiognomy's history). Today, a line of research persists that advocates the accuracy of first facial impressions [47, 54, 71, 76]. Research in computer vision datasets, algorithms, and models is clearly aware of this line of research. Projects in computer vision AI have asserted to successfully infer sexual [63, 96] and political orientation [57, 97] or emotion intensity and emotion expression [7, 26] based on people's faces in images. Others claim to be able to infer a variety of latent traits in personality assessment, such as trustworthiness [98] or the big 5 personality traits [18, 36–38, 64, 78, 86, 87] from profile images. However, considerable evidence suggests that first facial impressions do not surpass a "kernel of truth" [10, 74, 75, 92, 93, 95].

Researchers in the field of critical data science highlight ethical concerns arising from classifying individuals with AI on the basis of their facial appearance. Image-based inferences about people can only represent visibly apparent factors of an inferred concept [42]. However, as such inferences are used today, they may be based on bold or questionable semiotic assumptions when predicting intentions, aims, and capabilities or characters of individuals based on their facial characteristics found in portrait images [25, 52]. Judgments of this kind are epistemologically unreliable [30, 90]. Some researchers have argued that such systems are morally objectionable because they treat individuals as categorized objects [42, 53], and others have proposed to abolish physiognomic AI [90].

2.2 Does knowledge of AI correlate with ethical perceptions of AI?

While prior research has investigated users' perceptions of AI-based systems, only a handful of research studies exist that investigate experts' ethical perceptions of AI systems [49, 77, 100]. Here, measuring AI knowledge has proven to be difficult. Approaches vary from attempts to identify actual AI knowledge over the recruitment of specific subject pools to measures involving programming and numeracy skills (see Appendix A.1 for an overview). Another difficulty in comparing the studies arises from the diversity of application contexts and the diversity of AI systems, e.g., "automated decision-making by AI" [3], "expert systems" [50], "algorithms" [65], "artificial intelligence" [99] , or "algorithmic decision-making" [49].

Some positive associations were observed: Araujo et al. [3] found that both higher levels of education and technical knowledge, including AI knowledge, have a positive association with perceived usefulness, but no significant association with perceived risk of AI decision-making. Higher technical knowledge levels show a positive association with AI fairness perceptions. Similarly, Kaufmann [50] reported that teachers with knowledge on expert systems perceive higher utility of advice from these systems compared to teachers lacking such knowledge; there was no relation between numeracy and acceptance of algorithmic advice. Logg et al. [65] found that less numerate people appreciate advice from algorithms less in the context of forecasting and estimation tasks.

In contrast, Zerfass et al. [99] found that AI expertise and perceptions on AI adoption were not related. Lee and Baykal [62] found that greater levels of computer programming knowledge decreased the perceived fairness of algorithmic decisions in the context of dividing household chores. The authors assumed that participants with higher levels of knowledge were either confronted with unexpected algorithmic decision-making results and/or had greater knowledge about the limitations of such systems. Generally, discussion-based decision outcomes were perceived as fairer than outcomes produced by algorithms. Audio-recorded interviews highlighted the importance of participation in decision-making – i.e., the ability to choose and to agree or disagree – as well as enhanced social transparency of decision outcomes via discussion of the perceptions of whether an outcome was fair or not. Logg et al. [65] observed that greater familiarity with algorithms led to less acceptance of advice from automated forecasting tasks.

Zhang et al. [100] found AI researchers to favor a prioritization of research on AI safety, to support pre-publication reviews to evaluate potential harms, to strongly disagree with AI research on lethal autonomous weapons, and, finally, to highly trust scientific and international organizations in shaping the development of AI applications for the public interest. Across three different scenarios (dynamically-priced premium of car insurance, re-routing of flight passengers, automatic loan allocation), Kasinidou et al. [49] did not find students' AI knowledge to influence ethical perceptions of AI. Instead, individual differences were observed between undergraduate and postgraduate participants. For the context of criminal justice, undergraduate computer science students changed their perceptions of algorithmic fairness after one discussion-intensive class [77]: After the intervention, students preferred adding the gender feature to the algorithms, which may be explained by weaknesses of the concept "fairness through blindness". They also preferred algorithms, as opposed to human judges, and favored algorithmic transparency as a general principle. However, consensus did not increase. Rather, opinions were more varied regarding some topics.

The literature reviewed above reveals mixed results regarding the influence of AI knowledge on AI perception. The present study contributes to this line of research by comparing how ethical perceptions of facial analysis in two different contexts vary between laypeople and individuals with AI competence.

3 STUDY PROCEDURE AND METHODS

3.1 Recruitment process and participants

We recruited 346 survey participants across three samples, one of which served validation purposes. We sampled AI-competent individuals at the end of 2021 and beginning of 2022 (N=122, female=27.05%, male=69.67%, other=3.28%). We targeted graduate and PhD students focusing on AI at two large European universities and one large European research institute via social media and news channels of computer science and data science study programs. We describe the exact filtering criteria to determine AI competence in Section 3.3 (and provide further data such as course experience in Appendix A.3.4). Each participant was compensated with a fixed payment of 5€. The mean duration was 16.31 minutes (min: 6.50, max: 32.25). The age distribution was: 46.72% with age 18-24, 49.18% with age 25-34, 2.46% with age 35-44, 0.82% with age 45-54, and 0.82% with age 55 or above (see Appendix A.4 for data cleaning).

We collected a laypeople sample at the end of 2019 and at the beginning of 2020 via Amazon Mechanical Turk (MT) in the course of another study [31]. Participation was limited to those registered in the United States. We produced a final sample of 3102 participants. For the present study, we randomly selected 122 laypeople (female=46.09%, male=48.36%, other=0%) from all participants who indicated to have either very little or novice AI knowledge (46.09% of the entire dataset). The mean duration was 9.98 minutes (min: 3.87, max: 25.08). The age distribution was: 8.20% with age 18-24, 36.07% with age 25-34, 23.77% with age 35-44, 13.93% with age 45-54, 9.02% with age 55-65, and 9.02% with age 65 or above.

We collected a validation laypeople sample in June of 2022 in a second semester undergraduate lecture at a large European university (N=102, female=18.63%, male=81.37%, other=0%). We excluded respondents with high AI competence from the sample. The mean duration was 21.88 minutes (min: 5.16, max: 37.4). We assume that the higher average duration was due to the perceived complexity of the AI knowledge quiz by participants who were not competent in AI. 99.02% were aged between 18-24, 0.98% were aged between 25-34. Survey completion was incentivized by being part of a number of voluntary tasks to become eligible for a grade bonus on the final exam. The validation dataset also allowed for a useful complementary comparison with the sample of AI-competent individuals due to their shared similarities in demographic features (gender balance, age and country of origin).

Our home institution does not require an ethics approval for questionnaire-based online studies. All participants in the dataset were informed about the procedure, the length and the basic premise of the study, and gave consent to the use of the data for research purposes. Participants could drop out at any point in the survey, or could exit the survey if they did not agree with the use of their data for research purposes. All analysis data was fully de-identified and the privacy of all subjects was preserved at all times during the study. The service used to collect the data guaranteed compliance with the European Union's General Data Protection Regulation

(GDPR). The compensation offered in the two paid studies was above minimum wage.

3.2 Vignette study

Experimental vignette studies are a common instrument to study people's perceptions and judgments in a variety of hypothetical scenarios [2, 4, 21, 40, 46, 55, 56, 68, 72]. The design of our factorial vignette study is based on our prior work [31]. It consists of two hypothetical decision scenarios: participants were either drawn into a low-stake advertisement (AD) or a high-stake hiring (HR) scenario. In both scenarios an AI system scans a portrait picture and makes a variety of inferences about an individual. Based on these and other inferences, in the AD context, a social media user will be shown a particular advertisement. In the HR context, an applicant will either be selected or rejected for a job position (see Figure 1 in Appendix A.2). Participants then rated on a 7-point Likert scale their level of agreement or disagreement (1 = "strongly agree", 7 = "strongly disagree") with eight distinct AI-made inferences from a portrait picture, drawn for the above described purpose of the application context: *gender*, *emotion expression*, *wearing glasses* and *skin color*, *intelligent*, *trustworthy*, *assertive*, and *likable.* These ratings are hereafter called *inference ratings.* After each inference rating and before proceeding to the next inference, participants were asked to justify their rating in one to two sentences.

3.3 Measuring AI competence

We developed an AI knowledge test with a total of nine questions. Four of them were directed at computer vision, out of which three were based on the computer vision textbook by Chollet [19]. The other six questions were based on an instrument designed to assess student's AI and machine learning knowledge by Rodríguez-García et al. [83]. Here, we adjusted questions for the purpose of this study and removed some items (see Appendix A.3). The AI knowledge test was first discussed with three researchers and the resulting feedback was implemented. The scale was evaluated via a pre-study with three participants, who had varying AI knowledge levels. The pre-study additionally included one question on the difficulty of each item. The pre-study illustrated that the AI knowledge test has easy, moderate and difficult questions, and was able to map out a variety of AI knowledge levels.

3.4 Mixed method analysis strategy

All analyses were performed in R and Python.

3.4.1 Content-structuring qualitative content analysis. The design of our research study followed an embedded design, which we analyzed using mixed methods by integrating qualitative and quantitative data [61, 79]. To analyze the application of justification themes, we applied content-structuring qualitative content analysis and developed a detailed category scheme to map justification patterns within the responses by participants [60, 61, 67, 79, 101]. First, one researcher labeled 15% of the two main datasets and formulated 57 detailed categories, which were discussed with a second researcher and grouped into 21 super-ordinate categories. Second, both researchers independently applied this category scheme to 10% [79] of both datasets using the instructions documented in the code book in Appendix C. The inter-coder reliability was above Krippendorff's $\alpha \geq 0.8$ for each of the inferences [59]. Differences were discussed with a third researcher. No further categories were included. Finally, one researcher labeled the entire dataset using the final category scheme. The coding occurred at the word level. This meant that as little as one word up to the entire answer could be assigned a code. Three researchers labeled the validation dataset applying the previously developed category scheme. They achieved Krippendorff's $\alpha \geq 0.7$ for each of the inferences. Differences were discussed and resolved among the three researchers.

3.4.2 Frequency and co-occurrence analysis of justification themes. We analyzed the justification themes using co-occurrence and frequency analysis. We compared the results for subgroups of the sample, e.g., AI-competent vs. laypeople, AD vs. HR context. First, the frequencies of the individual themes were analyzed independently of the co-occurrence with other themes. Second, the frequencies of all unique theme pairs, e.g., the likelihood of two themes being mentioned in combination with each other, were explored.

3.4.3 Factor analysis, Welch two-sample t-test and analysis of variances. To analyze subjects' ratings, we performed an exploratory factor analysis with orthogonal rotation (varimax), minres factor extraction and regression factor estimation for all three samples. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis [45, 48] and Barlett's test of sphericity indicated that the correlations between items were sufficiently large. For all samples, parallel analysis, BIC, the Velicer MAP and the Kaiser criterion, amongst other tests, suggested retaining two factors (see Appendix B.2 for details). Furthermore, Welch two-sample t-tests and analysis of variances (ANOVA) were computed to directly compare the inference ratings.

4 RESULTS

4.1 Inference ratings show no significant differences between AI-competent and laypeople.

4.1.1 Welch two-sample t-test results. Comparing the inference ratings of the two main samples, none of the Bonferroni-corrected Welch two-sample t-tests shows significant group differences (see Figure 1 and Appendix B.1). A robustness check of the results using Yuen's test for trimmed means confirms that there are no significant group differences. The validation laypeople dataset validates the absence of group differences for all inference ratings except for the inference *wearing glasses* ($p_{\text{Bonf.}}$=.04) in the AD context.

4.1.2 Exploratory factor analyses suggest all samples perceive the same two constructs underlying the eight inferences. Exploratory factor analyses produced the same structure of factor loadings, i.e. two factors, for all three samples. The first factor included the inferences *intelligent*, *trustworthy*, *assertive* and *likable*, which will be referred to as *character and personality traits* in the following. The second factor included the inferences *gender*, *emotion expression*, *wearing glasses* and *skin color*, which will be referred to as *social constructs and features.* Although prior tests (see Appendix B.2) proved the data to be appropriate, some factor loadings did not exceed 0.6 [41], and some of the items (e.g., *gender*) loaded on two factors [91]. We assume that this is due to our rather small sample sizes [41].

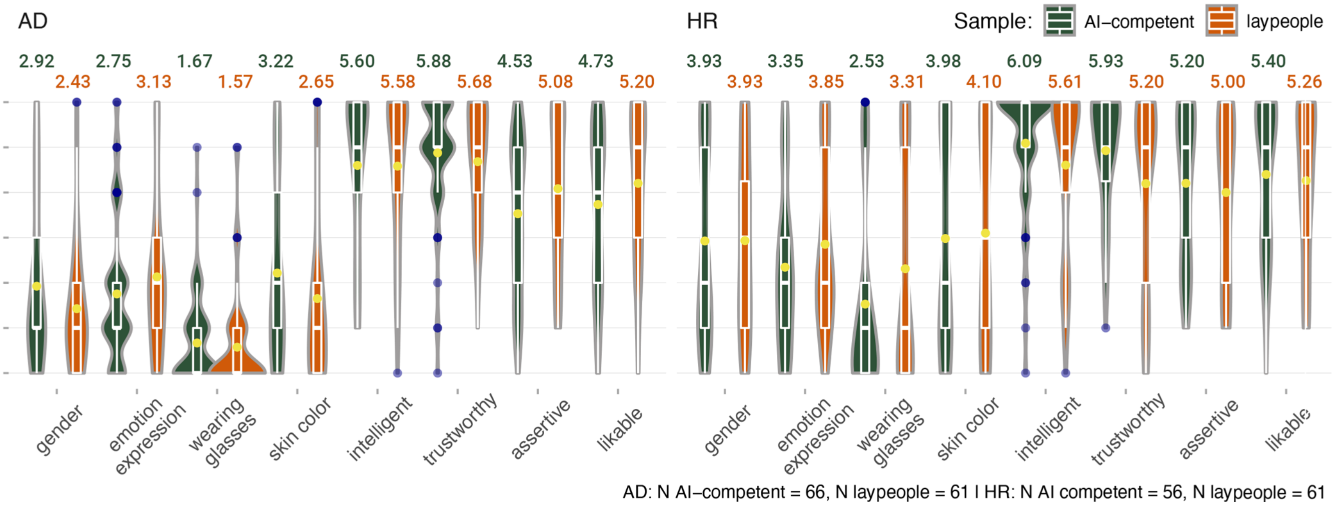

Figure 1: Mean inference ratings in AD vs. HR context by sample. Means of inference ratings for each inference by context and sample show that the AI-competent and laypeople (MT) largely agree in their ratings of facial AI inferences. Rating score 1: "strongly agree", rating score 7: "strongly disagree".

Next, we performed robustness checks by repeating the analysis on random sub-samples of 85% of the datasets. The robustness checks validated the findings. These results replicated findings with a large sample in [31]. The observations also confirmed the results from the Welch two-sample t-test: participants in both samples gave similar agreement-disagreement ratings to each of the inferences.

4.2 AI-competent and laypeople apply similar levels of complexity to their justifications.

To understand how AI-competent and laypeople justified their inference ratings, we first performed a complexity analysis of the open-text justifications. The analyzed justifications consisted of as little as one word up to a few sentences. Depending on the number of arguments embedded in the justification, we assigned a varying amount of themes during the labeling process. For instance, one participant gave the inference *likable* the rating "strongly disagree" and explained that one "absolutely can't tell if someone is likable because of the way they look. It's actually insulting and misleading and unfair to do that." This justification was labeled with the two themes "not sufficient/ good evidence (data) for task", and "bias/ stereotypes/ discrimination". We refer to justifications of this type as two-theme justifications. The use of fewer arguments could indicate that participants have a clear opinion regarding an inference. The use of more themes could indicate a more diverse and complex spectrum of viewpoints regarding an inference.

The analysis (Table 1) shows slight differences in the complexity of justifications by context and inference type. Subjects in the HR context and additionally laypeople in the AD context, provided somewhat more one-theme and less two-theme justifications when justifying their ratings on *character and personality trait* inferences than when justifying their ratings on *construct and feature* inferences. This suggests that evaluations were somewhat clearer for inferences on *character or personality traits.* In contrast, participants discussed inferences on *constructs and features* more diversely.

Table 1: Complexity of subject's justifications (in %)

	AI-competent		laypeople		validation$^+$	
Type	AD	HR	AD	HR	AD	HR
Inferences on constructs and features						
One theme	66.7	64.3	70.9	74.6	62.3	56.4
Two themes	29.2	31.2	27	23.8	30.9	29.4
Three themes	3.8	4.5	2	1.6	6.9	14.2
Four themes	0.4	-	-	-	-	-
# open text answers	*264	224	244	244	204	204
Inferences on character and personality traits						
One theme	66.3	76.8	79.5	80.7	58.8	64.7
Two themes	28.8	19.2	19.3	18.4	32.4	25
Three themes	4.9	2.7	0.8	0.8	8.8	10.3
Four themes	-	-	-	-	-	-
# open text answers	*264	224	244	244	204	204

* After cleaning of the data, more participants from the AI competent sample happened to be in the AD than HR context.
$^+$ More multi-theme justifications by the validation sample may be explained by the longer survey duration.

4.3 Context matters: People agree more with AI inferences in the AD than in the HR context.

We then turned our attention to the experimental variable *context* (AD context vs. HR context) to understand whether and how it influences ratings and justifications of participants.

4.3.1 *People agree more with AI inference-making in the low-stake AD context and less in the high-stake HR context.*

In all three samples, subjects in the HR context showed significantly less agreement with AI facial inferences than subjects in the AD context (AI-competent ($mean_{\text{AD}}$ =3.90, $mean_{\text{HR}}$ =4.54): $t_{\text{Welch}}(99.08)$ =-3.35, p<.01, $\hat{g}_{\text{Hedges}}$

=-0.62, $\mathrm{CI}_{95\%}$ [-0.99,-0.25]; laypeople ($mean_{\mathrm{AD}}$ =3.88, $mean_{\mathrm{HR}}$ =4.54): t_{Welch} (118.09) =-3.91, p<.01, $\hat{g}_{\mathrm{Hedges}}$ =-0.71, $\mathrm{CI}_{95\%}$ [-1.07,-0.34]; validation ($mean_{\mathrm{AD}}$ =4.06, $mean_{\mathrm{HR}}$ =4.71): t_{Welch} (98.86) =-3.35, p<.01, $\hat{g}_{\mathrm{Hedges}}$ =-0.66, $\mathrm{CI}_{95\%}$ [-1.06,-0.26]). These results indicate that the application context has an impact on participants' evaluations.

4.3.2 The decision context is the most influential factor in participants' ratings. We performed one six-way ANOVA for each of the eight inferences to analyze the effect of context on the inference rating while controlling for gender, age, education, country, and sample. The variable sample included the AI-competent and laypeople (MT) sample. Using Pillai's trace, ANOVAs with Bonferroni corrections for the eight tests showed that only the variable *context* had a statistically significant effect on inference ratings of *gender* (p<.001), *emotion expression* (p=.015), *wearing glasses* (p<.001) and *skin color* (p=.001). Bonferroni-corrected ANOVAs including the AI-competent and validation laypeople dataset confirmed these results, except for the inference *emotion expression*. We found no other significant effect for any other variable (see Appendix B.3).

4.3.3 Perceptions on the relevance of 'construct and feature' inferences are mixed; in the HR context, laypeople perceive inferences on 'character and personality traits' as relevant. The influence of the decision context was particularly evident when participants emphasized the "irrelevance" or "relevance" of *construct and feature* inferences (see Figure 2, light and dark orange). Participants evaluated these inferences as more "relevant" in the AD context and more "irrelevant" in the HR context. Similarly, participants used the theme "inference (only) sometimes relevant" more frequently in the HR context. This tendency was observed in all samples.

Both laypeople samples applied themes of "(ir)relevance" more frequently than participants with AI competence. Surprisingly, this was particularly the case for MTurk laypeople in the HR context for inferences on *character and personality traits* ("relevant": 15.7%, see Figure 2 light orange). For instance, participants from this sample justified that inferring *intelligence* "would give a hint as to how [...] [applicants] would perform on the job" or that inferring *trustworthiness* "in the workplace can be important and it's not wise to have a dishonest person around". For inferences on *constructs and features*, laypeople underlined the "irrelevance" of the inferences *wearing glasses* (26.2% of laypeople; 29.4% of validation laypeople) and *skin color* (27.9%; 39.2%) in the HR context and the "relevance" of the inferences *wearing glasses* (26.2%; 33.3%) and *gender* (26.2%; 29.4%) in the AD context. Some AI-competent subjects drawn into the AD context agreed that the inferences *wearing glasses* (21.2%) and *gender* (18.2%) are relevant to be inferred (see Appendix D.1).

4.4 Participants justify ratings on *construct and feature* inferences with a wide variety of themes; ratings on *character and personality* inferences with "insufficient data" themes.

Next, we analyzed whether specific themes were of special importance when justifying inference ratings on *constructs and features* or *character and personality traits*.

4.4.1 Ratings on 'construct and feature' inferences are explained by a variety of justification themes. As depicted in Figure 2, all subjects frequently applied themes highlighting "AI ability", "sufficiency" of the data, and – depending on the AD or HR context – the "relevance“ or "irrelevance" of an inference. AI-competent participants raised somewhat more "ethical and discriminatory concerns". Overall, justifications included a substantial variety of justification themes.

4.4.2 Ratings on 'character and personality trait' inferences are predominately explained by the "insufficiency" of a profile picture as evidence. The use of the "insufficiency" theme was particularly prevalent for laypeople in the HR context (AI-competent: 37.5%, laypeople: 56.7%; validation: 39.3%). Again, individuals with AI competence raised "ethical and discriminatory concerns" more often than participants in both laypeople samples. Furthermore, participants made references to the "subjectivity" of the inference task.

4.4.3 Participants believe "AI can infer" whether a person is wearing glasses on a portrait picture; they are skeptical about AI's ability to infer emotional expression. All three samples used the themes "technical ability of AI", "accurate and well working" models, and "easy to infer" most frequently to justify ratings on the inference *wearing glasses*. They applied the theme "can infer sometimes/ difficult in some situation" most often to justify ratings on *emotion expression* and *gender*. For instance, one participant explained that while "the majority of people can have a gender revealed through just a picture, not everyone fits that mold."

Some participants from both main samples believed that a "profile picture is good evidence" for the inferences *wearing glasses* and *emotion expression*. At the same time, there were critical voices stating that a profile picture is not sufficient evidence to infer *emotion expression*, e.g., "Emotion changes by the hour or minute. Can't make an inference based on that." The validation dataset supported these latter results.

4.5 Co-occurrence analysis: "AI (in)ability" and data-related themes co-occur most often with other themes.

We then analyzed the co-occurrence of themes with each other to identify patterns in the use of multiple justification themes (see Appendix D.2). We found that for inferences on *constructs and features*, the AI-competent raised concerns but acknowledged AI to be able to make certain inferences. Referring to inferences on *constructs and features*, people with AI competence raised "ethical and discriminatory concerns" in combination with almost all other justification themes, however, most frequently in combinations with themes on "AI ability" or the "sufficiency" of the profile picture as evidence (see Figure 5a and 5b-1 in the Appendix). This relationship reversed for justifications of ratings on *character and personality trait* inferences. Here, "ethical and discriminatory concerns" were most frequently brought forward in combination with themes on the "insufficiency" of a profile picture as evidence (see Figure 5a and 5b-3 in the Appendix).

For inferences on *character and personality traits*, laypeople often paired comments on the "(in)sufficiency" or "(in)adequacy" of the data with another theme. For *constructs and features*, a greater variety of theme combinations was observed.

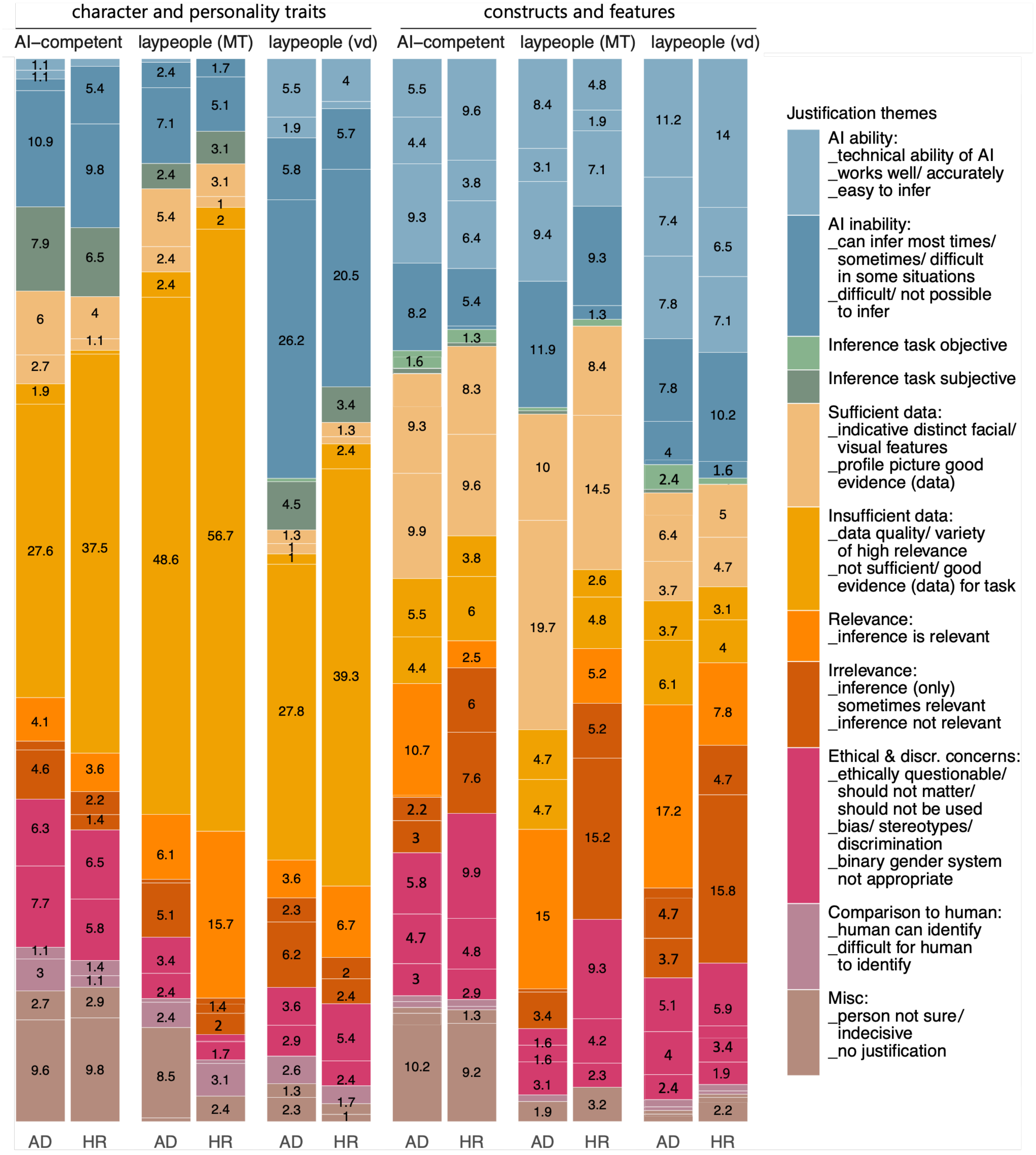

Figure 2: Percentages of individual themes grouped by super-ordinate topic, by context, and by sample. Stacked bars add up to 100% and represent the total of individual themes used by the specific sample. Only percentages > 1% are labeled on the graph.

4.6 Many inferences are based on questionable norms or resemble social constructs and societal stereotypes.

To understand participants' most critical concerns, we finally focused on themes related to "ethical and discriminatory concerns" and "AI inability" (see Figure 4 in the Appendix).

4.6.1 Individuals with AI competence perceive the inference likable as subjective. More than laypeople, individuals with AI competence and subjects from the validation sample described the inference *likable* as "relative", "based on sympathy", and "subjective", e.g., "Likability is a matter of perspective" and "depends on the observer." Comments also referred to other justification themes such as ethical concerns, e.g., "Likability is a highly subjective measure and inherently biased. In addition, it is highly unethical to have such type of decisions made by systems that are not capable of understanding the impact of this decisions" [sic] or "Likeability itself is an ill defined thing, predicting it from just portraits is wrong". Participants did not consider any other inference as equally subjective as *likable*.

4.6.2 Some subjects state that inferences on 'character and personality traits' cannot be inferred. However, approximately half of subjects highlight that the data is simply insufficient or inadequate. A considerable amount of subjects from all samples stated that a profile picture is "insufficient" data (26%-79% depending on inference, context, and sample) to infer *character and personality traits*. For instance, subjects commented that "[n]o facial features indicate trust", or that *intelligence* "is not quantifiable through visual data". At the same time, a minority (~15%) of the AI-competent, a small percentage of laypeople, and many participants from the validation dataset argued that AI cannot infer specific *character or personality traits*. An AI-competent participant explained that the "problem here is ill-posed", there "is no general understanding", and "no clear" or "objective definition of intelligence that everyone agrees with!" Given the lack of shared definitions, some asked "how is this measured? How is it implemented during training?", and "What are the parameters for identifying someone as intelligent?" These findings suggest that some participants evaluated inferences such as *intelligent* and *trustworthy* as social conceptualizations that require a common understanding before being used as inference in facial analysis AI.

4.6.3 Participants with AI competence believe that stereotypical judgments enable AI to draw 'character and personality traits'. Other people with AI competence worried about "stereotypes" embedded in the training data. They elaborated that, e.g., "a categorization of intelligence based on looks seems to correlate features that are not correlated" or that "the training data for trustworthiness depends on societal stereotypes and not actual trustworthyness" [sic]. Conversely, the existence of "stereotypes" was also used to argue in favor of AI being able to make an inference. For instance, a participant explained that the inference *likable* "makes sense because some people's appearance is appealing to more people. But, this inference can only be made on a statistical basis: Person is or is not likable on average." AI-competent participants stated justifications in relation to "bias, stereotypes and discrimination" most frequently when referring to the inferences *trustworthy*, *assertive*, and *likable*, e.g., one participant commented that "it's an unethical idea to give ai systems the ability to inference something so loosely defined and this will lead to biased choices made in the name of "science"." Laypeople did not show these levels of concern for any of these inferences.

4.6.4 A minority of participants raises concerns regarding the inference skin color. In the HR context, 23% of subjects from all samples raised "ethical concerns" regarding the inference *skin color*. One subject commented that *skin color* "should not be a criterion for job applications. Furthermore, being of a certain skin color should be a matter of self-description and not be determined by a computer program". Some participants also perceived the inference *skin color* to be based on biased data or to lead to discrimination: "Users will get predictions based on race and race-based stereotypes" or "if the model is biased towards skin color, it may not encourage a fair AI agent." Some subjects highlighted that *skin color* can be inferred but should not be done or used: "Color can be detected easily by computer vision frameworks (though this inference imposes certain ethical questions)" or "While it is possible to determine the skin color of a person from a portrait [...], it is ethically incorrect to base any decisions on skin color" or "Detecting skin colour should be trivial for the software, so it is reasonable to expect that inference. It is NOT reasonable that this information should be used to indicate whether someone is suitable for the job." These comments exemplify the diversity of normative evaluation of the inference *skin color*. Although suggesting that AI can infer *skin color*, this inference – which some specifically relate to "race" or "ethnicity" – was perceived as an impermissible inference by a considerable number of subjects.

4.6.5 A minority of participants highlights that binary gender norms are not appropriate and ethically questionable. Referring to the inference *gender*, some participants raised "ethical concerns" in the HR context (AI-competent: 16.1%; laypeople: 11.5%). In both contexts, 9% of participants with AI competence believed that inferences on gender are based on biased data: "The AI might learn to assign gender identity based on a heavily biased training data which are influenced by conventional gender identity norms hence making fateful inferences in the real world. Such inferences are unreasonable". Some subjects across all samples specifically highlighted that "gender norms are not appropriate" anymore: "This used to be a more 'objective' decision, however society has changed and persons can decide by themselves their gender, without being guided by their appearance. The most important part is, again, the inability of an AI system to understand the consequences of deciding something like this". Others commented that gender can be inferred but is not appropriate: "this is very apparent and thus somewhat alright, but then again, gender is a fluid concept". Some participants believed gender to be a social construct that is not binary as is often presupposed by facial analysis AI.

5 KEY OBSERVATIONS AND DISCUSSION

Overall, our study on the ethical perceptions of facial analysis AI suggests that there are no "common sense" facial analysis inferences. In all samples, there are participants who raise concerns, in particular, *ethical concerns* that inferences lack epistemic validity, should not matter or should not be used for the purpose of an application. In addition, we find that both AI-competent and laypeople express a variety of normative concerns regarding AI facial inferences. At the same time, only a minority of participants concluded that AI cannot, under any circumstance, make an inference from faces.

Regarding the facial inference *emotion expression*, participants note that a profile picture is only a snapshot and thus, "temporary and short-lived". Recently, emotion researchers have argued that emotion expression is more context-dependent and variable than commonly assumed. The *emotional state* of a person cannot be readily inferred from a person's facial expression [6]. Participants in both samples raised similar concerns. For example, one participant stated that there "are numerous people that tend to hide their emotions through pictures [...]".

Our analysis of justifications clearly shows that participants voice concerns regarding the classification of latent traits by facial analysis. Participants pointed out that the inference of attributes such as *intelligence* from facial information presupposed a highly simplified definition of a multidimensional concept. Similarly, participants mentioned potential problems related to the subjectivity associated with inferring attributes such as *likability* from faces.

We found that participants criticized the ethically problematic application of a binary conceptualization of *gender*. This finding aligns with recent critical data science research on computer vision. Here, authors, too, point to the fact that sensitive categories, such as gender and race, are often treated as "common sense categories" in computer vision datasets [25, 70, 80, 85].

On the other hand, a justification theme among both laypeople and people with AI competence pertains to the *possibility* of an AI inference provided that the "data is correct". This line of reasoning resembles narratives behind facial analysis AI research and commercial tools that try to solve issues with predictive power at the level of data *rather than question their epistemic foundations*. Some of the AI-competent and laypeople used entrenched stereotypical heuristics to evaluate AI facial inferences. While heuristics and stereotypes may initially help humans navigate through complex social interactions, research on the validity of human inferences from faces demonstrates that faces are no "strong and reliable indicator of people's underlying traits" [95, p.569].

Some specific differences between the two main samples could be observed. Both laypeople samples applied more pragmatic justifications referring to the "(ir)relevance" of the visual data for a decision-making procedure. For inferences on *character and personality traits*, more than half of laypeople (MT) described the data as "insufficient" for the inference task. People with AI competence mentioned themes related to "(ir)relevance" and "insufficiency" less frequently than laypeople, but raised "ethical concerns" more frequently than laypeople.

The complexities behind participants' justifications indicate a "struggle" for the power over the creation and attribution of meaning for visual data. Our study asks who can and should participate in this discourse. AI experts currently have free rein over the meaning that their datasets should be attributed with. However, politicians are aware of the complexities behind the meaning of visual data [e.g., 32] and we highlight again that more and more critics are voicing ethical concerns [e.g., 25, 42, 51, 80, 85, 89]. One of our main concerns is that the inference of perceived traits or features, e.g., "perceived trustworthiness" [e.g., 84] as opposed to "actual trustworthiness" by an AI system ultimately contributes to society remaining trapped in a cycle of stereotypes.

Taken together, we note that participants in all samples showed a tendency *to oppose* facial AI inference-making. Participants' evaluations underline many of the ethical complications of facial analysis AI that have recently been raised by critical data scientists and other scholars. Moreover, we see that people do not apply a consistent and universal justification profile for each of the facial inferences. Facial inferences are not simple constructs but overloaded with epistemic and pragmatic intuitions that are likely influenced by factors including cultural background.

We end by wondering how a justifiable ethical framework for facial AI inference-making could look like. What "standards" would a satisfactory justification fulfill? Given that we deal with *visual* inferences, we believe that they should first achieve reasonable epistemic validity and that this validity should be supported by scientific agreement over the quality of the evidence. The question then is what a reasonable level of scientific agreement should look like. We have pointed out that while a large majority of researchers underline the invalidity of first facial impressions, there is an ongoing stream of research publications that claim to present evidence on the validity of first impressions.

Participants in our samples disagreed with inferences common in human first impression-making (e.g., trustworthiness, likability etc.) by algorithmic systems. Indeed, one of the core findings of this work is that neither individuals with AI competence nor laypeople trust many of the inferences of facial analysis technology. With legislative attempts seeking to ban certain facial processing technologies, with a plethora of scholars pointing to the dangers of an automated version of physiognomy, and the different sample populations expressing their lack of trust toward such AI inference-making, we ask in what context and under what circumstances such facial analysis AI can be justified at all. It appears that, more often than not, there are better *reasons not to develop and deploy AI* that analyzes human faces to draw a variety of inferences that are then used for a particular decision-making context. Weaving together the argumentation threads from our previous results [31], critical remarks of data scientists and policy-makers, we take it that there is a strong case to be made that such AI inference-making is epistemically invalid, pragmatically of little use, and, overall, contributes and perpetuates stereotypes that stand in conflict with a society's welfare.

6 LIMITATIONS AND FUTURE DIRECTION

Our samples were composed of comparatively young people with AI competence that are not representative of all AI researchers. This may have introduced a bias in terms of the participants' understanding of and critiques on social constructs such as gender identities. In addition, this study does not include voices from industry. Future research should also survey corporate AI developers.

This research makes a methodological contribution by providing an AI knowledge instrument as an alternative to self-reported AI knowledge measures. We hope that the results from the application of the AI knowledge test will act as a starting point for the utilization of a more objective and reliable measure of knowledge on AI. It should be noted that given rapid advances in AI, the questions contained in the AI quiz should be regularly updated.

Our sample included participants from the United States (laypeople sample) and Europe (AI-competent and validation laypeople sample). We addressed the limitation of comparability of the two main samples by creating a validation dataset that shows substantial similarity in terms of demographics with the AI-competent sample. Given the international application of AI systems, diverse study participants are vital. Hence, future studies should explore whether cultural differences influence ethical concerns of facial processing technologies such as facial analysis AI. If there are no such cross-cultural differences then this could serve as evidence for the existence of culturally-universal ethical perceptions of facial inferences.

Whereas we evaluated the perception of AI inferences from profile pictures, future research should also evaluate perceptions of AI inferences from videos. Given that videos are used for a variety of inference tasks [11], the perception of somewhat more accurate results can be expected. However, it remains to be seen whether video data will influence whether such traits *should* be inferred.

7 CONCLUSION

As the use of AI grows in popularity and as the impact of AI inference-making on societies increases, so does the responsibility of those who develop such AI systems. A special focus must be placed on exploring the perspectives of a diverse group of people both who are potentially driving the implementation of computer vision and AI and those that are subjected to its inference-making.

This work provides insights into perceptions of AI inference-making by the general public compared to perceptions of individuals with high knowledge of AI. It suggests that, by and large, people with AI competence and the general public share many perceptions about AI inference-making and have distinct context- and task-dependent perceptual differences. Being aware of the perceptions and judgments of people with AI competence, on the one side, and users, on the other side, is essential to develop AI systems that are based on democratic discourse, accepted by society, and sustainable.

Concluding this research, we summarize that the application context does have an effect on how people perceive AI inference-making from faces. While differences in AI competence did not have an effect on the inference ratings, specific differences were observable for the ethical justifications. We found that both laypeople and people with AI knowledge showed more agreement with AI inference-making in the low-stake AD context than in the high-stake HR context. In both contexts, people with AI competence – although only a small minority – raised ethical and discriminatory concerns more frequently than laypeople. Laypeople made more references to themes related to the (ir)relevance of the inference for the context of application.

Having explored the question whether differences in AI knowledge account for changes in the perceptions of AI inference-making across two contexts, this work extends research in the field of perceptions of algorithmic systems and contributes to the nascent literature on AI experts' perceptions on AI inference-making. The results invite a deeper reflection on the similarities and differences in the perceptions of AI among different people within the general population. With this work, we aim to ultimately contribute to the development of sustainable AI systems that are supported, not only by their developers, but also by the general public.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments that improved the paper. We thank the study participants for taking part in this study. 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 (MM '14)*. Association for Computing Machinery, New York, NY, USA, 1153–1156. https://doi.org/10.1145/2647868.2655014

[2] Kwame Anthony Appiah. 2008. *Experiments in Ethics*. Harvard University Press, Cambridge, Massachusetts, USA.

[3] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[4] 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.

[5] 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

[6] 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

[7] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical and Electronics Engineers, New Jersey, USA, 5562–5570. https://doi.org/10.1109/CVPR.2016.600

[8] Betaface. 2021. Betaface API. https://www.betafaceapi.com/wpa/ Accessed: 2022-02-27.

[9] Sky Biometry. 2021. Face Recognition Demo. https://skybiometry.com/demo/face-recognition-demo/ Accessed: 2022-02-27.

[10] Jean-François Bonnefon, Astrid Hopfensitz, and Wim De Neys. 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

[11] Grzegorz Brodny, Agata Kołakowska, Agnieszka Landowska, Mariusz Szwoch, Wioleta Szwoch, and Michał R Wróbel. 2016. Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. In *2016 9th International Conference on Human System Interactions (HSI)*. IEEE, Portsmouth, UK, 397–404. https://doi.org/10.1109/HSI.2016.7529664

[12] 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 (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91.

[13] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2019. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2019), 220–239.

[14] Sarah Castell, Daniel Cameron, Steven Ginnis, Glenn Gottfried, and Kelly Maguire. 2017. Public views of machine learning – Findings from public research and engagement. 92 pages. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

[15] Stephen Cave, Kate Coughlan, and Kanta Dihal. 2019. "Scary robots": Examining public responses to AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 331–337. https://doi.org/10.1145/3306618.3314232

[16] Stephen Cave, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. Portrayals and perceptions of AI and why they matter. 28 pages. https://royalsociety.org/~/media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf

[17] Stephen Cave and Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1, 2 (2019), 74–78.

[18] 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 (MM '14)*. Association for Computing Machinery, New York, NY, USA, 1101–1104. https://doi.org/10.1145/2647868.2654977

[19] François Chollet. 2018. *Deep Learning with Python*. Manning Publications, Shelter Island, NY, USA. https://books.google.de/books?id=mjVKEAAAQBAJ

[20] Ching-Hua Chuan, Wan-Hsiu Sunny Tsai, and Su Yeon Cho. 2019. Framing artificial intelligence in American newspapers. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 339–344. https://doi.org/10.1145/3306618.3314285

[21] 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.

[22] Clearview.ai. 2022. Overview. https://www.clearview.ai/overview Accessed: 2022-02-24.

[23] Kate Crawford. 2021. Time to regulate AI that interprets human emotions. *Nature* 592, 7853 (2021), 167–167.

[24] 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. AI Now Institute.

[25] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY* 36 (2021), 1105–1116.

[26] Abhinav Dhall, O.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 (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 423–426. https://doi.org/10.1145/2818346.2829994

[27] Charles Efferson and Sonja Vogt. 2013. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports* 3, 1 (2013), 1–7. https://doi.org/10.1038/srep01047

[28] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129. https://doi.org/10.1037/h0030377

[29] EmoVu. n.d.. EmoVu Mobile. https://www.programmableweb.com/sdk/emovu-mobile Accessed: 2022-02-25.

[30] 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* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 301–307. https://doi.org/10.1145/3314183.3323846

[31] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What people think AI should infer from faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 128–141. https://doi.org/10.1145/3531146.3533080

[32] European Parliament. 2021. Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters. https://oeil.secure.europarl.europa.eu/oeil/popups/summary.do?id=1678184&t=e&l=en

[33] Face++. 2022. Face Attributes. https://www.faceplusplus.com/attributes/ Accessed: 2022-02-25.

[34] Faception. 2021. Our technology. https://www.faception.com/our-technology Accessed: 2022-02-24.

[35] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, Palo Alto, CA, USA, 963–969.

[36] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2015. Predicting personality traits with Instagram pictures. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015 (EMPIRE '15)*. Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/2809643.2809644

[37] Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. 2016. Using Instagram picture features to predict users' personality. In *MultiMedia Modeling*, Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu (Eds.). Springer International Publishing, Cham, 850–861. https://doi.org/10.1007/978-3-319-27671-7_71

[38] 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 (UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 157–161. https://doi.org/10.1145/3209219.3209248

[39] 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 (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 110–119. https://doi.org/10.1145/3287560.3287568

[40] Armin Granulo, Christoph Fuchs, and Stefano Puntoni. 2019. Psychological reactions to human versus robotic job replacement. *Nature Human Behaviour* 3, 10 (2019), 1062–1069.

[41] Edward Guadagnoli and Wayne F Velicer. 1988. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 103, 2 (1988), 265—-275. https://doi.org/10.1037/0033-2909.103.2.265

[42] 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* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 543–554. https://doi.org/10.1145/3461702.3462620

[43] Isabella Hermann. 2020. Beware of fictional AI narratives. *Nature Machine Intelligence* 2, 11 (2020), 654–654.

[44] Michel Hohendanner, Chiara Ullstein, and Mizuno Daijiro. 2021. Designing the exploration of common good within digital environments: A deliberative speculative design framework and the analysis of resulting narratives. In *Proceedings of the Swiss Design Network Symposium 2021 on Design as Common Good – Framing Design through Pluralism and Social Values*. SUPSI, HSLU, swissdesignnetwork, Lucerne, Switzerland, 566–580.

[45] Graeme D Hutcheson and Nick Sofroniou. 1999. *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. Sage Publications, New York City, NY, USA.

[46] Michael R Hyman and Susan D Steiner. 1996. The vignette method in business ethics research: Current uses, limitations, and recommendations. *Studies* 20, 100.0 (1996), 74–100.

[47] 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, 1 (2020), 1–11.

[48] Henry F Kaiser. 1974. An index of factorial simplicity. *Psychometrika* 39 (1974), 31–36. https://doi.org/10.1007/BF02291575

[49] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future developers' perception of fairness in algorithmic decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 690–700. https://doi.org/10.1145/3442188.3445931

[50] Esther Kaufmann. 2021. Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence* 2 (2021), 100028. https://doi.org/10.1016/j.caeai.2021.100028

[51] 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 (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 587–597. https://doi.org/10.1145/3442188.3445920

[52] 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*. Springer, Cham, 265–282.

[53] Owen C King. 2020. Presumptuous aim attribution, conformity, and the ethics of artificial social cognition. *Ethics and Information Technology* 22, 1 (2020), 25–37.

[54] 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), 1–7. https://doi.org/10.1371/journal.pone.0081237

[55] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194.

[56] 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, Stanford, CA, USA.

[57] Michal Kosinski. 2021. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports* 11, 1 (2021), 1–7.

[58] Steven R Kraaijeveld. 2021. Experimental philosophy of technology. *Philosophy & Technology* 34 (2021), 993–1012.

[59] Klaus Krippendorff. 2004. Reliability in content analysis. *Human Communication Research* 30, 3 (2004), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[60] Udo Kuckartz. 2009. *Evaluation online: internetgestützte Befragung in der Praxis.* Springer VS Wiesbaden, Wiesbaden, Germany.

[61] Udo Kuckartz. 2014. *Mixed methods: Methodologie, Forschungsdesigns und Analyseverfahren.* Springer VS Wiesbaden, Wiesbaden, Germany.

[62] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17).* Association for Computing Machinery, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230

[63] John Leuner. 2019. A replication study: Machine learning models are capable of predicting sexual orientation from facial images. arXiv:cs.CV/1902.10739

[64] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, Vol. 10. AAAI, Cologne, Germany, 211–220. https://ojs.aaai.org/index.php/ICWSM/article/view/14738

[65] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[66] [House Of Lords]. 2018. AI in the UK: Ready, willing and able? https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

[67] Philipp Mayring. 2015. *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12 ed.). Weinheim/Basel: Beltz Verlag.

[68] David E Melnikoff and Nina Strohminger. 2020. The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour* 4 (2020), 1–7.

[69] 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.

[70] 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.* Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/3442188.3445880

[71] 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%2F0146167209346309

[72] Shaun Nichols and Joshua Knobe. 2007. Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs* 41, 4 (2007), 663–685. https://doi.org/10.1111/j.1468-0068.2007.00666.x

[73] Noldus. [n.d.]. Facial Expression Analysis. https://www.noldus.com/facereader/facial-expression-analysis Accessed: 2022-02-27.

[74] 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

[75] 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), 20190435. https://doi.org/10.1098/rstb.2019.0435

[76] 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

[77] Emma Pierson. 2018. Demographics and Discussion Influence Views on Algorithmic Fairness. arXiv:cs.CY/1712.09124

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

[79] Stefan Rädiker and Udo Kuckartz. 2019. *Analyse qualitativer Daten mit MAXQDA.* Springer VS Wiesbaden, Wiesbaden, Germany.

[80] 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.* Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[81] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[82] 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.* Association for Computing Machinery, New York, NY, USA, 145–151. https://doi.org/10.1145/3375627.3375820

[83] Juan David Rodríguez-García, Jesús Moreno-León, Marcos Román-González, and Gregorio Robles. 2021. Evaluation of an online intervention to teach artificial intelligence with LearningML to 10-16-year-old students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education.* Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3408877.3432393

[84] Lou Safra, Coralie Chevallier, Julie Grèzes, and Nicolas Baumard. 2020. Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. *Nature Communications* 11, 1 (2020), 1–7. https://doi.org/10.1038/s41467-020-18566-7

[85] 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

[86] 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

[87] 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

[88] 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

[89] 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 (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 782–793. https://doi.org/10.1145/3442188.3445939

[90] Luke Stark and Jevan Hutson. 2022. Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal* 32, 4 (2022). https://ir.lawnet.fordham.edu/iplj/vol32/iss4/2

[91] Louis Leon Thurstone. 1947. *Multiple-factor Analysis; A Development and Expansion of The Vectors of Mind.* University of Chicago Press.

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

[93] 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

[94] Alexander Todorov, Friederike Funk, and Christopher Y Olivola. 2015. Response to Bonnefon et al.: Limited 'kernels of truth' in facial inferences. *Trends in Cognitive Sciences* 19, 8 (2015), 422–423. https://doi.org/10.1016/j.tics.2015.05.013

[95] 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/10.1146/annurev-psych-113011-143831

[96] 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://psycnet.apa.org/doi/10.1037/pspa0000098

[97] 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 International AAAI Conference on Web and Social Media*, Vol. 14. 726–737. https://ojs.aaai.org/index.php/ICWSM/article/view/7338

[98] 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 *MultiMedia Modeling*, Xiangjian He, Suhuai Luo, Dacheng Tao, Changsheng Xu, Jie Yang, and Muhammad Abul Hasan (Eds.). Springer International Publishing, Cham, 148–158. https://doi.org/10.1007/978-3-319-14442-9_13

[99] Ansgar Zerfass, Jens Hagelstein, and Ralph Tench. 2020. Artificial intelligence in communication management: A cross-national study on adoption and knowledge, impact, challenges and risks. *Journal of Communication Management* 24, 4 (2020), 377–389. https://doi.org/10.1108/JCOM-10-2019-0137

[100] Baobao Zhang, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C Horowitz, and Allan Dafoe. 2021. Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers. *Journal of Artificial Intelligence Research* 71 (2021), 591–666. https://doi.org/10.1613/jair.1.12895

[101] Cornelia Züll and Natalja Menold. 2019. Offene Fragen. In *Handbuch Methoden der empirischen Sozialforschung*, Nina Baur and Jörg Blasius (Eds.). Springer, 855–862.