Attitudes Toward Facial Analysis AI: A Cross-National Study Comparing Argentina, Kenya, Japan, and the USA

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

Severin Engelmann
severin.engelmann@cornell.edu
Cornell Tech,
Digital Life Initiative
New York City, USA

Orestis Papakyriakopoulos
orestis.p@tum.de
Technical University of Munich,
Professorship of Societal Computing
Munich, Germany

Yuko Ikkatai
y.ikkatai@staff.kanazawa-u.ac.jp
Kanazawa University
Kanazawa, Japan

Naira-Paola Arnez-Jordan
naira.arnez@tum.de
Technical University of Munich
Munich, Germany

Rose Caleno
calenorose@gmail.com
Nairobi, Kenya

Brian Mboya
brian.mboya@protonmail.com
Dedan Kimathi University of
Technology, Computer Science
Nairobi, Kenya

Shuichiro Higuma
shuichiro-higuma@g.ecc.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Tilman Hartwig
Tilman.Hartwig@uba.de
Umweltbundesamt,
AI Lab
Leipzig, Germany

Hiromi Yokoyama
hiromi.yokoyama@ipmu.jp
The University of Tokyo,
Kavli IPMU, CD3
Tokyo, Japan

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

ABSTRACT

Computer vision AI systems present one of the most radical technical transformations of our time. Such systems are given unparalleled epistemic power to impose meaning on visual data, despite their inherent semantic ambiguity. This epistemic power is particularly evident in computer vision AI that interprets the meaning of human faces. The goal of this work is to empirically document laypeople's perceptions of the epistemic and ethical complexity of computer vision AI through a large-scale qualitative study with participants in Argentina, Japan, Kenya, and the USA (N=4,468). We developed a vignette scenario about a fictitious company that analyzes people's portraits using computer vision AI to make a variety of inferences about people based on their faces. For each inference that the fictitious company draws (e.g., age, skin color, intelligence), we ask participants from all countries to reason about how they evaluate computer vision AI inference-making. In a series of workshops, we collaborated as a multinational research team to develop a codebook that captures people's different justifications of facial analysis AI inferences to create a comprehensive justification portfolio. Our study reveals similarities in justification patterns, but also significant intra-country and inter-country diversity in response to different facial inferences. For example, participants from Argentina, Japan, Kenya, and the USA vastly disagree over the reasonableness of AI classifications such as *beautiful* or *skin color*. They tend to agree in their opposition to AI-drawn inferences *intelligence* and *trustworthiness*. Adding much-needed non-Western perspectives to debates on computer vision ethics, our results suggest that, contrary to popular justifications for facial classification technologies, there is no such thing as a "common sense" facial classification that accords simply with a general, homogeneous "human intuition."

CCS CONCEPTS

• **Computing methodologies** → *Computer vision*; • **Social and professional topics** → *User characteristics*; • **Security and privacy** → Human and societal aspects of security and privacy.

KEYWORDS

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

ACM Reference Format:
Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Yuko Ikkatai, Naira-Paola Arnez-Jordan, Rose Caleno, Brian Mboya, Shuichiro Higuma, Tilman Hartwig, Hiromi Yokoyama, and Jens Grossklags. 2024. Attitudes Toward Facial Analysis AI: A Cross-National Study Comparing Argentina, Kenya, Japan, and the USA. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3630106.3659038

FAccT '24, June 3–6, 2024, Rio de Janeiro, Brazil

1 INTRODUCTION

Computer vision AI is a prominent AI subdiscipline that develops vision models for a diverse set of application contexts. These models are deployed for cancer prediction [56], mood detection [44], self-driving cars [49], or social robotics [15]. Among the best-known examples of computer vision AI are projects that classify, recognize, and analyze human faces from visual data [14, 20, 22]. A comprehensive review of close to 500 computer vision AI datasets found that 205 were "face-based": no other object was represented more often in computer vision datasets than human faces [82]. Collecting, processing, and analyzing facial data is controversial, and it raises complex ethical concerns regarding privacy, consent, and potential misuse. In the age of AI, questions of justifiable practices and standards around facial image analysis have become a fixture in public debates. Critical data scientists – including members of the FAccT community – have called facial recognition the "plutonium of AI" [90] and demonstrated the potential adverse impacts of such projects due to biased misrepresentations [e.g., 40, 41, 91].

In our work, we zoom in on *facial analysis AI*, a branch of computer vision AI that aims to infer semantic meaning about the depicted individuals based on their facial appearance. In facial analysis AI, a model learns the relationship between features in the facial image, usually represented as a vector space, and a pre-specified target variable such as age, gender, or skin color [3, 5, 29, 42, 76]. While facial recognition typically attempts to identify or verify a person's face by locating matching faces in a database, facial analysis operates under the assumption that faces bear meaning that goes beyond the detection, identification, or verification of a particular face. To develop a training dataset, a human determines a set of target variables and defines a notion of ground truth. For example, annotated facial datasets usually distinguish gender as binary female/male, following a historically rooted practice [83]. Labelers are then instructed to follow guidelines that accord to the human-defined ground truth when annotating a training dataset of facial images.

Such target variables in commercial facial analysis AI tools go beyond gender and include, e.g., age, emotions, beauty, or intelligence (further outlined in Section 2.2). Pointing to benchmark accuracy measures [11], research projects in facial analysis have claimed their models can reliably infer people's ethnicity [67], sexual orientation [58, 101], political ideology and orientation [53, 77, 104], emotion expression and intensity [8, 24], or personality traits [5, 85] based on facial images only. But what semantics should facial analysis AI justifiably infer from faces? How do we define meta-criteria that govern the selection process of reasonable and unreasonable facial image inferences? These questions are not only central to research and development efforts but also of significant societal concern, given that facial analysis AI directly affects the populace. With our study, we contribute to this debate by exploring how laypeople from different countries argue for or against the justifiability of facial analysis AI.

An illustrative example of a conceptual justification of facial analysis is Paul Ekman's theory of the basic six emotions [28] that has become a widely applied conceptual foundation for the justification of inferring emotion (expressions and states) using visual cues through so-called facial action units [59]. Even though there is substantive evidence that there is cultural variation in the mapping between facial action units and their meaning [e.g., 18], AI emotion recognition tools often operate based on a monocultural interpretation around Ekman's six basic emotions.

Facial analysis AI models and the products that they drive are often developed under the notion that *humans* regularly make inferences from facial information. While we are often told not to judge persons by their outer appearance, first impressions have a remarkable effect on election outcomes [6, 57, 69], employment decisions [68, 81], or jail sentences [25, 103, 105]. Applied to computer vision, AI systems are given the power to semantically interpret the visual data about our identities and characteristics. However, research from psychology and anthropology has shown that first impressions of others are often inaccurate [12, 27, 70, 95, 96]. For example, individuals are incapable of reliably assessing a stranger's trustworthiness solely from their facial features when no other information is present [48]. Against the view of six basic emotions, studies from psychology and cognitive sciences indicate that facial expressions and the way people interpret emotions are influenced by the context, can be similar or dissimilar across regions and, thus, go beyond Ekman's basic facial expressions of emotions [19, 55].

Also, voices from the computer science community and the political sphere have raised concerns regarding different practices of facial analysis AI [21, 62, 63, 75, 84, 90, 91, 100]. For example, Miceli et al. [62] find that target variables and their arbitrary levels are hardly ever questioned, and are rather imposed on data through the annotators. Scheuerman et al. [84] find that image databases rarely describe how variables, here race and gender, are defined or how they have been annotated. Vemou et al. [100] highlight data protection issues and risks associated with facial emotion recognition tool use, including inaccuracies of inferred emotional states, less precise classification results for people with darker skin tones as well as a lack of transparency and control of what facial images are analyzed, for what purpose and by whom.

Despite these findings and raised concerns, computer vision AI has claimed not only being able to infer personality traits [50], emotions [8] or political orientation [53] from still images but also to create artificial faces that humans perceive as more trustworthy than real faces [65]. This marks the closing of a cycle in which AI systems infer attributes from faces through predetermined mechanisms, only to subsequently generate faces that conform to those predefined attributes once more – as D'Ignazio and Klein [26, p.104] put it, a classification system "becomes naturalized as 'the way things are'." For similar arguments see [13, 30, 31, 87]. Advances in facial analysis AI and the use of results of facial analysis AI in other computer vision fields raise again the fundamental question of what should be inferred from a face and why.

Reacting to raised concerns, the EU Artificial Intelligence Act (AI Act) [34][1] prohibits biometric categorization based on biometric data for the specific purpose of inferring "race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation" [34, Article 5(1), point (g)]. However, the "labelling or filtering of lawfully acquired biometric datasets [...] in the area of law enforcement" [34, Article 5(1), point (g)] is

[1]Please note: The EU AI Act has not yet been published in the Official Journal. We refer to the Corrigendum of the European Parliament's First Reading of 19 April 2024.

exempted from this prohibition. This prohibition does not apply to AI systems for which the categorization of facial features is an "ancillary feature" of a product or service. Examples include filters used on online marketplaces to try on products or tools on online social network sites to add or modify pictures [34, Recital 16]. The AI Act also prohibits "AI systems to infer emotions [...] in the areas of workplace and education institutions" [34, Article 5(1), point (f)]. Where such systems serve a safety or medical purpose, they are exempted from the prohibition. All other emotion recognition systems are to be classified as high-risk [34, Recital 54]. These political agreements between the European Parliament and the Council of the European Union are compromises on diverging perspectives and, thus, testify to the difficulty of regulating AI for facial analysis [97].

Our work represents a comprehensive effort to document laypeople's perceptions of the epistemic and ethical complexity of facial analysis AI, and to counter hype-driven narratives about AI promoting an "anything goes" approach when imposing meaning on visual data. For that purpose, we conduct a large-scale survey study with demographically representative samples from four countries. We build on previous research by members of our research team [32, 98] showing that German AI-competent people and laypeople from the USA are critical of facial AI inferences such as trustworthiness or intelligence in low-risk and high-risk contexts. Perceptions differ between contexts for facial AI inferences such as skin color or gender. In our current work, we follow calls for adding cross-national and non-Western perspectives to the debate on profound challenges in AI ethics [86].

Our study makes the following contributions to the field:

We perform a large-scale cross-national survey study to document participants' ethical reasoning about facial analysis AI. We report the commonalities and differences of justification norms toward facial analysis AI by laypeople from Argentina, Japan, Kenya, and the USA (N=4,468). Participants from each country were asked to write a justification for a total of eight facial AI inferences (*age*, *gender*, *skin color*, *beautiful*, *wearing glasses*, *trustworthy*, *intelligent*, and *emotion expression*) drawn by a fictitious AI company called ImageInsight.

We develop a coding scheme for interpreting participants' justifications for or against AI facial inferences from four countries. With an interdisciplinary and multinational research team, we manually coded, automatically classified, and subsequently analyzed more than 35,500 written justifications. We developed a comprehensive coding scheme through four iterative phases: ideating, testing, refining, and validating, until achieving full consensus. The manually classified multilingual database forms the basis for a classifier following an inter-lingual ensemble approach. We analyzed both participants' ratings and justification codes. This mixed-methods approach enabled us to gain deeper insights into people's perceptions of AI inference-making from facial data.

AI inferences from human faces are neither epistemically nor normatively intuitive constructs. AI inferences have been portrayed as intuitive epistemic constructs in computer vision AI research and development. In contrast, we find that participants' justifications counter such a convenient and straightforward legitimization of AI inferences highlighting the justificatory complexity of each inference drawn.

Participants evaluate inferences epistemically, pragmatically, and technologically. While there are inter-country and intra-country differences, participants evaluate inferences along three main types of justification: First, whether an inference is epistemically sound. Second, whether AI, as a "technology of discovery", is able to produce an inference. And third, whether an inference is useful for supporting a decision, i.e., whether it is relevant for the purpose of a decision-making context.

Perceptual differences are specific to inference and context. Participants tend to agree that inferring *intelligent* and *trustworthy* is unjustifiable, while the inferences *wearing glasses*, *skin color*, and *emotion expression* can, in principle, be inferred by AI from a portrait. While these inferences and the inferences *gender*, *age*, and *beautiful* warrant some epistemic justifiability, participants also perceived them as subjective, irrelevant to the decision context, or associated with negative consequences. Japanese participants are most critical of AI-generated inferences, whereas Kenyans are most affirmative. Negative views on inferences are consistently more prevalent in the hiring than in the advertising context across countries.

We advance AI ethics research with a cross-national research study. So far, cross-national survey studies have been rare at FAccT. In providing a detailed description of our methodological process, we hope to support other research studies that document ethics perceptions of AI from non-Western perspectives. We make our data (participants' justifications) open and freely available.[2]

We proceed as follows. In Section 2, we present our methodological approach. We explain the process of setting up a multinational research cooperation and designing our vignette study. We outline our approach of developing a coding scheme and compiling a dataset for a classification model and discuss how we implement this model. In Section 3, we present our findings. We focus on documenting written justifications for or against facial analysis AI inferences offered by participants from all four countries. We highlight commonalities and differences in how participants reason about these inferences. In Section 4, we offer a discussion of our results and a reflection on our study's limitations and researcher positionality. In Section 5, we conclude with final remarks.

2 METHODS

2.1 Initiating and coordinating a cross-national AI ethics study

We surveyed participants from Argentina, Kenya, Japan, and the USA. The selection of these countries was guided by an exploratory rationale to document justifications from a diverse set of countries. Our unit of analysis was at the country level rather than the culture level. This allowed us to define our study as cross-national [2, 80, 99], sometimes also considered a sub-category of cross-cultural research [99]. We acknowledge that by using membership to a nation-state as a grouping variable, we did not focus on variations in cultures that exist across a group of countries or within individual countries [1]. Our study maintained a descriptive focus, highlighting potential differences and commonalities in ethical judgments across and within countries. In this work, we refrained from offering simplistic explanations based on a narrow understanding of culture relative to

[2]We make the data available here: https://osf.io/brq7h/

national boundaries. We anticipated heterogeneity in justification toward facial analysis AI within and between the four countries.

Best practices to conduct cross-national studies require a research team that reflects the linguistic and national backgrounds of participants from the studied countries [72]. It took several months to build a research team that could satisfy these requirements. Our final research team consisted of researchers that were either from the studied countries, from a country belonging to the same linguistic area, or had previously lived in a studied country for a long period (see also our positionality statement in Section 4.3). We lay out our methodological procedures in detail to support other AI ethics scholars in designing and conducting future studies with participants from multiple countries.

2.2 Vignette design and survey procedure

In this paper, we present a follow-up study on two prior works by members of the research team [32, 98]. Our primary focus was to carefully extend the scope to a cross-national study. We further altered the content of the vignette study regarding information on a hypothetical company. We ran a between-subject experimental vignette study [4] with a quasi-experimental component (see Appendix A.1 for the exact wording of the vignettes). Each participant was assigned to one of 24 groups, which resulted from the variation of three variables: four *countries* (Argentina, Japan, Kenya, USA), two *decision contexts* (low-stakes advertisement, high-stakes hiring), and three *informational settings*. The latter settings varied in the information about the fictitious AI company described in the vignette. For each of the four countries, participants were randomly assigned to the experimental groups based on decision contexts and informational settings. The variable country (from which the study participants originated) constituted the quasi-experimental component. In this paper, we direct our attention to the experimental variables *decision context* and *country* and analyze how participants justify their perception of AI inference-making from portraits.

The vignettes introduced a fictitious company called ImageInsight. Subjects were told that "*ImageInsight has developed software that uses artificial intelligence (AI) to analyze images.*" Participants randomly assigned to the advertisement context were informed that the "*software analyzes portraits of users on social media in order to show social media users suitable advertisements for products. How does that work? ImageInsight's deep learning AI is presented with a portrait of a user showing only the user's face but nothing else. The AI scans the user's face and makes a variety of inferences about the user. Based on these and other inferences a user will be shown personalized advertising material on the social media platform.*" Participants assigned to the hiring context read that ImageInsight's AI "*analyzes portraits of job applicants in order to select suitable candidates during hiring procedures. [...] Based on these and other inferences an applicant will be invited for a job interview.*" We used a written description of the vignette that did not show any exemplary portraits to avoid possible framing effects due to the visual stimulus of a particular portrait.

With the differentiation of the two *decision contexts*, roughly half of the subjects were assigned to a hypothetical advertisement situation that might not be perceived as particularly detrimental, and the other half of the subjects into a hypothetical hiring situation that might be perceived as highly consequential. Past studies on algorithmic perception suggest individuals view the employment context as more critical compared to other decision contexts [32, 51, 88, 98]. We, therefore, refer to the hiring scenario as the "high-stakes" context, and to the advertisement scenario as the "low-stakes" context. The description of *fictitious* scenarios is a common device for analyzing moral dispositions in vignette studies [4, 52].

After being presented with one of the two vignettes, participants were prompted to evaluate, in subsequent steps, a total of eight AI inferences: *age, gender, skin color, wearing glasses, emotion expression, beautiful, intelligent,* and *trustworthy*. These were presented in random order, each on a separate survey page. The inferences were identical in the product advertisement and the hiring context. The eight inferences were chosen based on their use in real-world computer vision applications. Most facial analysis AI tools claim to be capable of inferring gender [e.g., 10, 35, 60, 64, 94], age [e.g., 10, 35, 60, 64, 92], and accessories like glasses [e.g., 10, 64]. There is an entire industry focused on making inferences about emotion expression and interpreting (perceived) emotions [e.g., 35, 60, 64]. Other tools claim to infer how pretty someone is [e.g., 7, 36], while the beauty industry is making use of facial analysis to suggest specific makeup products [e.g., 93]. Social media platforms, such as TikTok, provide the possibility to apply filters that transform facial features and display an alteration of one's face based on predefined beauty metrics [39]. Other facial analysis tools claim to infer ethnicity [e.g., 10, 17, 73], and traits such as intelligence [e.g., 37].

We asked participants to rate their agreement or disagreement with each inference on a 7-point Likert scale[3] in the context of the advertising or hiring decision-making scenario (1: "strongly agree" to 7: "strongly disagree"). After each rating, we asked participants to provide a brief justification of their rating. In this paper, we focus on analyzing the open-text responses to understand what justification patterns participants provide when arguing for or against specific AI inferences from faces. The survey concluded by asking participants to indicate their demographics.

2.3 Data collection and participant sample

We contacted various online survey companies to collect data from Argentina, Japan, Kenya, and the USA. However, several of these companies either were unable or unwilling to gather data in all four countries. Reasons included a lack of panels or partners in each country and the required sample size for the requested quotas exceeded the capabilities of their available panels. In the end, we contracted the online survey company Talk Online Data Collection AG[4], which is an online-panel company with partners in Argentina, Japan, Kenya, and the USA. The company recruited participants based on country-specific target quotas[5] for gender and age (see

[3]Prior research [32] has shown that study participants are not primed by asking whether they agree with an inferred trait being, for example, justifiable, reasonable, or fair. For this reason, when translating the English survey into Spanish for Argentinan subjects, and into Japanese for Japanese subjects, we allowed some freedom in the translation, intending to make the survey as comprehensible as possible in each of the three languages. In the English and Japanese survey, the term "justifiable" and the corresponding translation "正当である" was used, while in Spanish the term "razonable" was perceived as more adequate.

[4]https://talk-group.com

[5]'Target quota' refers to the company's attempted but not guaranteed participant representation due to online-panel size constraints in some countries.

Table 3 in the Appendix). We also established a target quota for two levels of education, given that it is known that individuals with higher education levels are more likely to engage in online surveys.

The institution of the first author provided the resources to run the study. This institution does not mandate ethics approval for online questionnaire studies. In the execution and data analysis of the study, we adhered to established ethical research norms, including providing a thorough explanation of the study procedures to participants, securing participants' informed consent, and refraining from gathering either personal or device-specific data. We employed the survey platform Sosci Survey[6], which adheres to the GDPR enforced by the EU. Our study was free of any misleading components, and participants were free to withdraw from the study at any time. Privacy was a priority; all collected data were anonymized, and the confidentiality of participation was preserved throughout the research process.

We conducted a power analysis using the software G*Power [38] to estimate the appropriate sample size for the data collection. The sample size calculation reflects our vignette design described above. Based on standard values of α=0.05 and power (1-β)=0.9, the minimum required sample size for achieving small effect sizes for a comparative analysis of all collected variables was a total of 952 survey participants per country (see Appendix A.3 for a detailed explanation). Before the full launch of the survey, we performed a pre-survey with 200 participants (50 participants per country) in December 2022. The research team manually checked the pre-survey data. In particular, we thoroughly read the responses to the open text fields to investigate whether participants fully comprehended the survey vignettes and whether their responses were sensible. We made minor adjustments to the survey and then launched it.

We collected the data from December 2022 to March 2023. Our final sample consists of data from 4,468 participants (see Table 3 in the Appendix for detailed demographics). This dataset includes 1,123 responses from participants from Argentina (50.3% female, 48.9% male; 18-34: 37.4%, 35-54: 36.8%, 55+: 25.8%), 1,068 from Japan (50.4% female, 49.1% male; 18-34: 21.2%, 35-54: 34.1%, 55+: 44.7%), 1,107 from Kenya (47.7% female, 51.6% male; 18-34: 51.6%, 35-54: 34.4%, 55+: 13.9%), and 1,170 from the USA (52.8% female, 46.8% male; 18-34: 30.1%, 35-54: 33.1%, 55+: 35.8%). The dataset is representative[7] for each country for the variables gender (female, male) and age groups (three levels).

We applied several measures to ensure data quality. We included two comprehension checks in the survey to ensure participants were aware of the described hypothetical scenario. We manually examined all survey data of those participants who wrongly answered the comprehension checks. We also checked that no participant's age was younger than 18 years. We removed participants if their open-text responses appeared to be nonsensical, or linguistically indecipherable. We identified such cases as "garbage" in a coarse data cleaning together with the online-panel provider and during the process of manually labeling data for automatic classification. This process included the following steps: We first identified all cases with repeated entries for open-text responses and with identical inference ratings for all eight inferences. These cases were manually screened for reasonable justifications. Second, we manually checked all cases that deviated by two standard deviations from the average duration of taking the study. Third, we manually selected 100 *garbage* inference justifications to train a classification model and used the model to screen the entire data corpus for further garbage cases. After the final automatic classification of all responses, we set a threshold rule of 75% (i.e., cases where six or more responses were classified as "garbage") and manually checked these responses.

2.4 Qualitative coding procedure

An interdisciplinary team of researchers contributed to the analysis of open-text responses. The research team focused on coding participants' responses to the survey task, *"Please justify your rating in 1 - 2 sentences."* for each of the presented eight inferences. In total, the qualitative data analysis procedure involved the coding[8] of 35,579 open-text responses. We followed best practices to produce reliable qualitative results. Specifically, we utilized an iterative [54] data analysis process and transparently documented all analytical procedures and workshop session outcomes in this section and in Appendix B [66]. To ensure the quality of our analysis, we applied analyst triangulation [71] by forming four country teams of two researchers each (see Section 4.3 for researcher positionality). We used triangulation of sources [71] by collecting data not from one but four countries with the aim of analyzing consistency and variation across data sources (i.e., countries).

In order to analyze the open-text responses, we developed an analysis process that combined both manual and automatic text classification. First, the manual text classification served to construct a coding scheme that encompassed all justification types. Second, we created a database of examples for each of the identified codes. This database formed the basis for training and validating an automatic text classifier based on three variants of Bert [23].

Developing a coding scheme. We developed the coding scheme over a period of six months, through four coding phases and workshops – ideating, testing, refining, validating – iteratively [54] until full consensus was reached (Figure 1). Each of the four country teams performed the same coding tasks on their country-specific dataset. To develop the coding scheme, we applied both deductive and inductive content analysis following best practices based on [54]. We use deductive content analysis based on our previous study [32, 98]. The following base codes from [32, 98] served as a deductive starting point to build our category scheme representative of the participants' responses: *AI (in)ability, inference task, reference to data, reference to (ir)relevance of inference for purpose of AI system, ethics and norms, comparison to human, miscellaneous.* However, as a research team, we agreed on the importance of inductive content analysis to complement or replace the base codes. This is important to build a coding scheme that best reflects the themes that emerge from analyzing participants' justifications. We decided on a case-wise analysis (i.e., coding all responses from one participant at a time in order to take contextual information into account). Because responses often contained multiple justification themes, we agreed

[6]https://www.soscisurvey.de/

[7]Please note that post-data cleaning, achieved quotas slightly differ from the targeted representative quotas.

[8]We use the verb 'coding' and the noun 'code' in correspondence with practices in qualitative content analysis. The terms refer to the building of classes or categories of themes that are represented in the textual data. Typical synonyms are 'annotations', 'classifications', 'classes', and 'categories'.

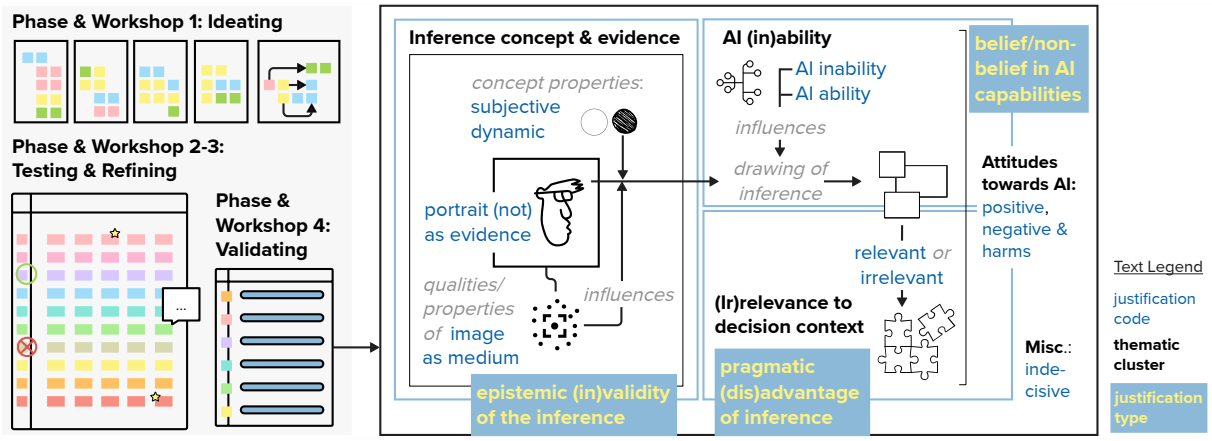

Figure 1: Workshop process to reach a final coding scheme (here visualized). See Appendix B for workshop documentation.

to assign multiple codes to a participant's response when appropriate. This decision to apply multi-class, multi-label classification [16] increased the complexity of a classification model, but allowed for a better representation of participants' arguments.

In preparation for our first coding workshop, entitled "ideating," each researcher individually worked through eight open-text responses from 50 participants to produce an ideation of codes based on apparent themes in the data. The country teams met individually and then collectively to discuss the themes identified, and then to structure them on a Mural board[9] into clusters of themes (coding phase 1), reflecting on the initial base codes provided. During the first workshop, all country teams presented the identified themes. Commonalities and differences were discussed. The workshop resulted in a draft coding scheme shown in Figure 7 in Appendix B.1.1.

In preparation for our second workshop, entitled "testing," each researcher tested the draft coding scheme by applying it to the previous sample data of 50 participants (coding phase 2). We collected examples of each code on a new Mural board. Prior to the joint workshop with all researchers, all country teams met individually and discussed their observations. The aim of this preparatory exercise was to identify which codes of the draft scheme did not represent the data well and whether country-specific themes were sufficiently reflected. During the workshop, changes to the draft coding scheme suggested by the researchers were discussed on the basis of the examples collected. Sub-codes were rearranged, and clusters were renamed. This second workshop ended with a revised coding scheme. We repeated this process of applying the revised coding scheme and discussing proposed changes in coding phase 3 and during the third workshop entitled "refining." The basis for our discussions was established by a jointly developed document with working definitions of the codes, sample responses for each code, and sample keywords that were likely to indicate a code. The final version of this document eventually became our codebook. The whiteboard documenting our work in workshops 2 and 3 is shown in Figure 8 in Appendix B.1.2.

[9]Mural is a collaborative online whiteboard platform allowing simultaneous document access by multiple users. See https://app.mural.co.

Building a multilingual classified database. After establishing a coding scheme, we categorized random samples of the dataset according to the process shown in Figure 2 to build a training and validation dataset for the automatic text classifier. Team members individually coded subsets of the dataset, 19 survey participants in each round case-wise [54]. The two members in each country team discussed their disagreements (in terms of different codes assigned) and then repeated this process twice (coding round A to C in Figure 2). Thus, all coded responses were double-coded by the country team members and, in the case of different assigned codes, discussed until the team was able to unanimously assign one or more codes. The discussion of assigned codes allowed us to ensure that the manual coding was not prone to the subjectivity of individual researchers. After the first round (coding round A), we held a final coding workshop ("validation") and made final changes to the coding scheme. The workshop concluded with the finalization of the codebook (Appendix B.2), which all researchers followed when coding the data in the subsequent categorization rounds. This step completed the manual coding process.

With the aim of obtaining enough exemplary responses per code for the subsequent training of an automated classifier, each country team filled up codes with fewer than 33 examples (see top right plot in Figure 10 in Appendix B.3 for details). We do not report an intercoder reliability score (IRR), as all of the responses were reviewed and discussed extensively among the coders throughout the workshops. We note the perspective of McDonald et al. [61] that reaching informal consensus through discussion meetings suffices to communicate reliability, particularly in the context of unstructured data and multi-label coding [61]. This perspective aligns with our methodology, emphasizing the importance of discussion and consensus-building in ensuring the reliability of our qualitative analysis.

We designed a classification approach to address country-specific coding biases. Instead of training a classifier for each language using examples from that language, we followed an inter-lingual ensemble approach. We first translated all responses into all languages. To realize the translations of participants' responses, we automatically translated sample responses using the deepL API.

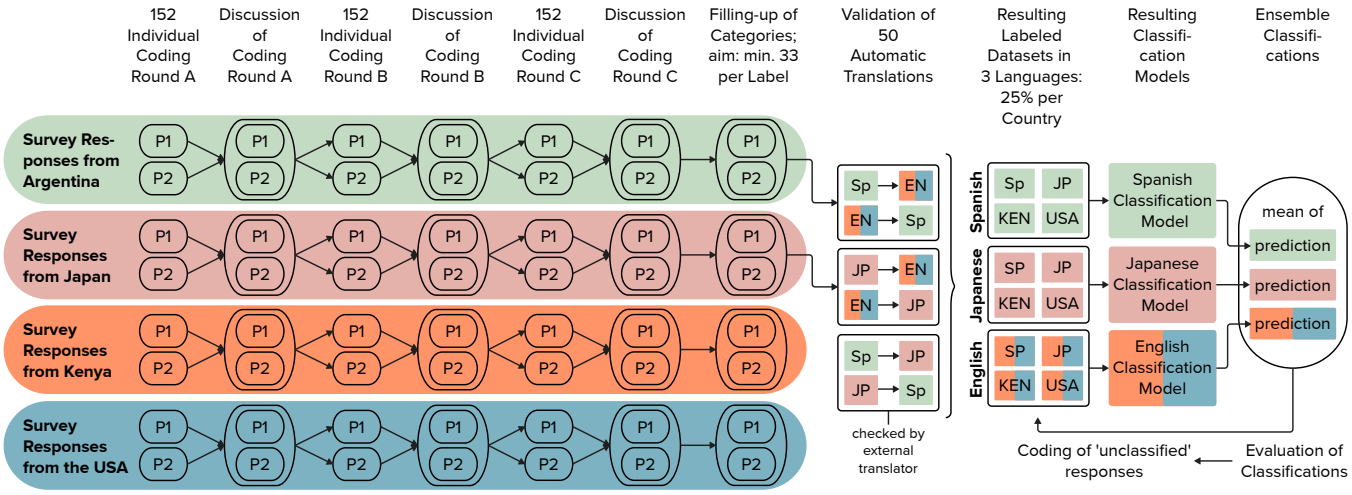

Figure 2: Process from manually coding the data for a multilingual database to developing an automatic classifier.

The Spanish and Japanese country teams validated the automatic translations for the language pairs Spanish-English and Japanese-English. We hired an external translator to validate the Spanish-Japanese response translations. While we obtained low error rates for Spanish-English and Japanese-English translations, error rates were too high for the Spanish-Japanese language pair (see error rates in Appendix B.3.2). For Spanish-Japanese translations, we decided to have text responses translated into English first, before translating them into the target language. We translated each of the labeled responses into each of the other languages, leading to three identical datasets in Japanese, Spanish, and English, with an average translation accuracy of 90%.

Developing the classifier. We then developed three classifiers (English, Spanish, Japanese) for the original and translated responses. For each response to be classified, we obtained a prediction from each classifier. The final labels assigned to each response included any label that was assigned a probability score of 0.5 or higher by any of the three models. We used three variants of Bert [23] to train our models, each fine-tuned on data of the respective language, on the multilabel classification task. We did not use assistant-based LLMs such as ChatGPT, as they fail to compete with classical fine-tuned LLMs in classification tasks [106]. We split the data (original and translated) to train (80%), test (10%), and evaluation (10%) sets, with each model's accuracy having at least an F-1 score of 0.70 on the evaluation set. The final ensemble model achieved an F-1 score of 0.75. We used the ensemble model to obtain codes for the totality of the responses (N=35,579).

Evaluating the classifier. Four researchers manually checked the classification of 130 responses over two independent validation rounds and discussed disagreements. These produced 85.38% absolute agreement or partial agreement with the classifier. While satisfied with the accuracy, the model produced 6677 unclassified responses, which amounts to lost information for the data analysis. Therefore, four researchers manually classified 300 of the unclassified responses over two independent coding rounds and discussed disagreements. The coded responses were added to the multilingual classified database. The resulting final annotated dataset consisted of 9164 responses, 4553 in English (USA and Kenya), 2302 in Spanish (Argentina), and 2309 in Japanese. The models were retrained. The accuracy of each model achieved at least an F-1 score of 0.70 on the evaluation set. The final ensemble model achieved an F-1 score of 0.71. We used the ensemble model to obtain codes for the totality of the responses (N=35,579). The number of unclassified responses was reduced to 1226.

2.5 Quantitative analysis process

We analyzed the resulting justification classifications using frequency analysis [54] to understand what justification patterns participants apply to argue for or against AI inference-making. We computed Welch two-sample t-tests for unequal variances and Kruskal Wallis tests to analyze differences in participant groups, both across countries and across contexts (advertisement vs. hiring). We supplement this analysis with Welch two-sample t-tests on the participants' inference ratings across the advertisement and hiring contexts. This mixed-methods approach [54] allows us to obtain a more in-depth understanding of people's perceptions of AI inference-making from faces. We report all statistical analysis results in Appendix C.

3 RESULTS

3.1 Qualitative justifications: How do people reason about facial AI inferences?

Overview. We identified three main types of justifications that participants apply when evaluating facial AI inferences. First, they consider if an AI inference is epistemically sound. Second, they turn to AI as an epistemic technology, reasoning about whether they believe AI can or cannot perform certain inferences. Third, participants evaluate AI inferences according to their relevance for the decision context, taking a pragmatic justification.

Specifically, we identified twelve different justifications (i.e., codes) that people provided when explaining their agreement or disagreement with AI inferences from people's portraits. Four of the justifications have a positive connotation (*ability of AI, can be*

Table 1: Examples of participants' responses for largest justification clusters. See more examples in the codebook, Appendix B.2.

Cluster abbrev.	Inference concept & evidence	AI (in)ability	(Ir)relevance
affirmative connotation	*This is easy to identify by looking at photos*	*AI can tell a person's age | AI can identify*	*Necessary to segment advertising*
negative connotation	*IQ cannot be seen just through a scan.*	*AI cannot deduce the intellectual ability*	*Irrelevant to personality assessment*
justification type	(1) epistemic (in)validity of the inference; 43.37% of responses*	(2) belief/non-belief in AI capabilities; 18.34% of responses	(3) pragmatic (dis)advantage of inference; 16.78% of responses

*The percentages do not amount to 100%, because the justification codes belonging to the cluster 'general attitudes toward AI and ethical reflection' are not represented in this table. They fall into the justification type *belief/non-belief in AI capabilities* or *pragmatic (dis)advantage of inference.*

inferred from image, inference relevant, positive attitude and ethics), while six have a negative connotation (*inability of AI, cannot be inferred from image, dynamic concept, subjectivity, inference not relevant, negative attitude and discrimination or harms*). Two justifications are neutral (*the medium image, indecisive*). Taking a different perspective, the twelve justifications can also be clustered into five major thematic areas (see Table 1 for the three largest clusters). For example, the cluster AI (in)ability represents justifications for stated beliefs and non-beliefs in the capabilities of AI. In the following, we describe each justification by cluster allocation (frequencies in brackets and Appendix C.1).

Cluster: Inference concept and evidence for inference-making. The largest cluster includes five different justification codes, which each refer to different notions of the inference concept or the evidence used for inference-making. The most used justification expresses people's affirmative perception that an inference **can be inferred from a portrait** or that the inference is **obvious or clear** (N=8252, 16.40% of all assigned classifications). According to this line of justification, the concept of the inference to be drawn is unambiguous and drawing the inference from the portrait "would be something obvious". They reasoned that the face "shows" the inference. It "can be seen" and "is easy to identify by looking at photos". Thus, a portrait is perceived as providing sufficient evidence for an AI system to draw a particular inference.

On the contrary, the second most common justification is used to explain that the **inference cannot be inferred from a portrait** or, more generally, that there are problems with the **measurement** (N=7236, 14.38% of all assigned classifications). This justification is most often used to express disagreement with the inferences *trustworthy* and *intelligent*. The argument here is that the portrait (alone) does not provide sufficient evidence for the inference to be drawn. At least other non-visual input parameters or alternative tests are required for measurement. For example, one participant explained that "IQ cannot be seen just through a scan. It can only be measured through tests." Focusing on the image as inadequate evidence for drawing the inference, others argue that "[o]ne cannot decipher someone's intelligence just from looking at an image" or that "[a]ppearance does not imply a person's gender."

With these two commonly used arguments, the participants reflected on whether an image provides a sufficient epistemic foundation to draw an inference. The normative evaluation is thus based on an assessment of the proportionality of the image as evidence and the inference as an informational goal. We refer to this type of justification as discussing the *epistemic (in)validity of the inference.*

We identified three additional lines of reasoning within this thematic cluster and justification type. Although less frequently used, these justifications were typically employed to justify rejection or to limit their approval of AI inference-making. Participants highlighted that the concept of an inference or the face as a basis for the inference has a **dynamic** component, meaning that it is not constant over time or can be manipulated (N=3197, 6.35%). This justification was often made with the inferences *age*, *emotion expression*, *wearing glasses*, and *gender*. For example, participants described that "the emotion can only be a thing of the moment", that "[p]eople apply makeup" or that in "today's day and time it is more difficult to determine whether a person is male, female or other." Closely related is the notion of **subjectivity** (N=2224, 4.42%), which participants perceived as being at odds with AI inference-making, in particular, when drawing the inference *beautiful*. Participants highlighted that a conclusion is "relative to different values" or varies with "different perceptions." A few participants contextualized their responses by pointing to the required qualities of **the medium image** (N=916, 1.82%). The image or photo as a file to be analyzed may have been manipulated by "effects" or "filters" or may be of poor quality if "lighting conditions" were unfavorable.

Cluster: AI (in)ability. The second largest cluster of justifications includes two justification codes. They refer to the **ability of AI** (N=4819, 9.58%) and the **impossibility or difficulty for AI** to draw an inference (N=4408, 8.76%), in general. Justifications in this cluster are less specific in their reasoning. Participants highlighted that the inference task "is easy for the AI" to solve or that the AI system is "accurate". Some participants made a comparison with a human being who is able to infer information. Hence, AI systems can do this, too. Other participants were convinced of the opposite and explained that the inference task is "difficult", "not accurate" or "impossible" for an AI to solve, e.g., "AI cannot deduce the intellectual ability of an applicant." Again, some participants made a comparison with a human, stating that a human may or may not be able to make the inference, but in any case, the inference task is not solvable for an AI, e.g., "It is impossible, even for humans." With these two arguments, participants justify their perception of whether AI can solve a facial AI inference task, which we will hereafter refer to as *belief/non-belief in AI capabilities.*

Cluster: (Ir)relevance. In the third largest cluster consisting of two justifications, participants discuss the **irrelevance** (N=4546, 9.03%) or **relevance** (N=3900, 7.75%) of an AI inference for informing a decision in the advertisement or hiring context. Participants stated, for example, that an inference is "[i]rrelevant [for] personality assessment" or for "determining qualified candidates." Affirming

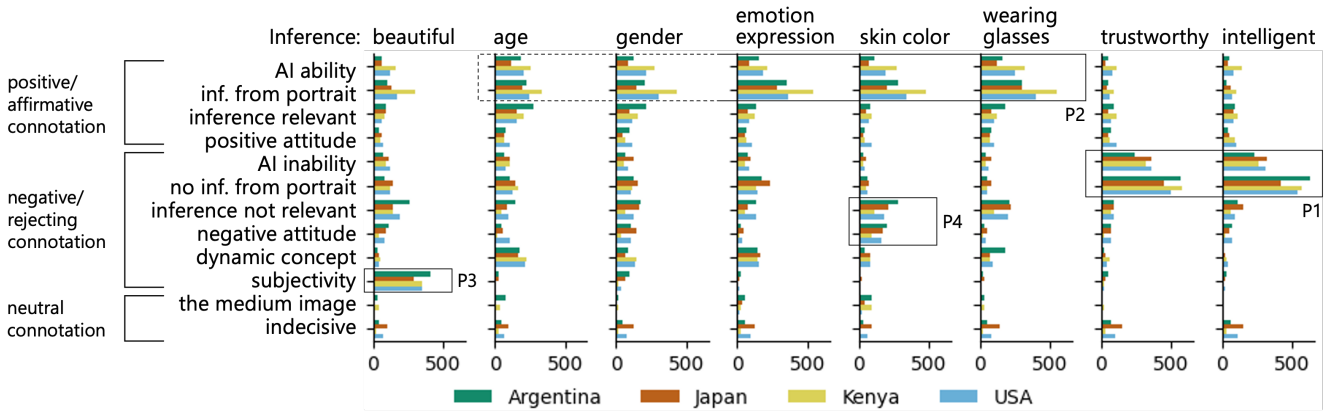

Figure 3: Comparison of frequencies of justifications used by participants from different countries to justify (dis)agreement.

facial AI inference-making, participants explained that an inference is relevant, such as for "personaliz[ing] ads" or "select[ing] applicants". Participants leverage these two arguments to anchor their normative reasoning about whether the inferences could lead to effects that are useful or positive or, conversely, useless or negative – a type of justification that we refer to as *pragmatic (dis)advantage of the inference*.

Cluster: General attitudes toward AI and ethical reflections. The fourth largest cluster with two justifications contains responses that conveyed participants' general attitudes toward AI and ethical reflections. Some participants presented a generally **positive attitude** toward AI technology, such as that "AI is a good invention" (N=2448, 4.86%). Others conveyed a generally **negative attitude** toward AI technology (e.g., "I do not trust AI") or highlighted **harmful consequences** resulting from drawing the inference from portraits (N=2528, 5.02%). Subjects provided examples such as privacy harms or discrimination, including but not limited to racism and sexism. Hence, in this thematic cluster, participants' normative evaluation is based on their *belief/non-belief in AI capabilities* and/or the *pragmatic (dis)advantage of the inference*.

Cluster: Misc. In the last cluster, we collected responses that reflected participants' **indecisiveness** (N=2265, 4.87%). Some participants seemed not to have an opinion on facial analysis AI or to be unsure how to think of it, e.g., "Can't judge", "I do not feel qualified", "I don't know."

In summary, we identified twelve justification codes clustered into five thematic areas across three types of justifications: Epistemic (in)validity of the inference (43.37%), belief/non-belief in AI capabilities (>18.34%), and finally pragmatic (dis)advantage of the inference (>16.78%). The most frequently used justifications describe that an inference can(not) be drawn from a facial portrait.

3.2 Inference differences: How do people argue for or against specific AI inferences?

We identified similar lines of reasoning for or against certain inference groups, some of which apply across all countries. In combination with the numerical inference ratings, we distinguish some justification particularities for different groups of inferences (P1-P4 in Figure 3; see Table 7 in the Appendix for absolute frequencies and Figure 12 for distribution of ratings).

Participants from all countries expressed disagreement with inferences of the traits *trustworthy* (M_{AD}=5.0, sd_{AD}=1.8, M_{HR}=4.8, sd_{HR}=1.9) and *intelligent* (M_{AD}=5.1, sd_{AD}=1.8, M_{HR}=4.9, sd_{HR}=1.9; rating 5="rather disagree" on a 7-point scale). This opposition was primarily supported by the contention that these attributes cannot be deduced from a single portrait or, more broadly, from an AI system (P1). These two inferences stand apart from the remaining six presented to participants as they reflect character traits. Participants from all countries generally agreed that making assumptions about character traits based on facial images is rather not justifiable, and they expressed skepticism about the ability of AI systems to accomplish this inference task based on a facial image.

The two justifications that AI can make an inference in general or from a portrait (P2) were most frequently used by subjects across all countries in the context of the inferences *wearing glasses*, *skin color*, and *emotion expression*. To a lesser extent, this also applies to the inferences *gender* and *age*. It should be noted, however, that these two justifications (AI ability and inference from portrait) were used less frequently in absolute terms per inference than their negative counterarguments (AI inability and no inference from portrait) for the inferences *trustworthy* and *intelligent*. We can, therefore, conclude that the participants' perception of the inferability of the former inferences is less strong than their perception that AI cannot infer in general or from a portrait whether a person is *intelligent* or *trustworthy*. This is also reflected in the participants' numerical ratings, which tend to show agreement (*wearing glasses*: M_{AD}=2.7, sd_{AD}=1.7, M_{HR}=3.5, sd_{HR}=2.1; *skin color*: M_{AD}=3.3, sd_{AD}=1.9, M_{HR}=4.1, sd_{HR}=2.2; *emotion expression*: M_{AD}=3.3, sd_{AD}=1.6, M_{HR}=3.7, sd_{AD}=1.8).

Participants were divided in their opinion about the inference *beautiful* with high variation in inference rating within and between countries (M_{AD}=4.2, sd_{AD}=1.9, M_{HR}=4.5, sd_{HR}=2.0). Participants predominantly argued that the inference is subjective (P3), for example, by mentioning that "[b]eauty is in the eye of the beholder" or that "[i]t would be based only off the programmers' view of beauty." Subjects' ratings for the inference *skin color* presented the highest rating variation within and between countries. While

saying that skin color can be inferred from an image (e.g., "easy to identify"), many also highlighted the irrelevance of skin color for the decision context (e.g., skin color "is not a factor for good work") or expressed concerns often related to discrimination (e.g., "It is discriminatory to take into account skin color"; P4).

Besides a concentration on a few justifications for some inferences, we also observed a wide variety of justifications used for other inferences. Significant variation in justifications exist for the inferences *age*, *gender*, *emotion expression*, and *beautiful*. These differences in perceptions within and across countries highlight the epistemic complexity of these inferences.

Summarizing, for some inferences, we observed a concentration on a few justifications, whereas, for other inferences, participants used a variety of justifications to argue for or against the inference. On the one hand, there is a strong agreement among participants that the personality traits *intelligent* and *trustworthy* are not justifiable and cannot be inferred, both confirmed by ratings and arguments. In particular, participants stress the semantic ambiguity of facial portraits and question the use of a portrait as evidence. On the other hand, there is agreement, albeit less strong, that the inferences *wearing glasses*, *skin color*, and *emotion expression* can be inferred, in general, or from a portrait. For the latter three inferences, and even more for the inferences *gender*, *age*, and *beautiful*, the classification results also show that the justifications can be multifaceted: while they warrant some epistemic justifiability, they are perceived as subjective, irrelevant to the decision context, or associated with negative consequences. The two main argumentation particularities (P1 and P2) are each composed of justifications that discuss the epistemic (in)validity of the inference and the participants' belief/non-belief in AI capabilities.

3.3 Country and context differences: What differences exist in lines of reasoning?

Figure 4 comparatively displays how often different justification arguments were put forward by study participants in the advertisement (solid bar) and in the hiring context (dashed bar; see Appendix C for frequencies and for statistical analysis of group differences). We observed different argumentation patterns across countries and across the advertising and hiring context.

First, positive justifications (top right corner in the radial graphs in Figure 4; Table 2), for example, that AI can draw an inference in general or from a portrait, were the most used by participants from Kenya. Overall, 53.1% of the Kenyan subjects' arguments have a positive connotation. Participants from Kenya used the justification that an inference can be drawn from an image the most frequently. The share of positive arguments is lowest among participants from Japan, with 30.8%. The relevance argument was used most often by participants from Argentina.

Second, arguments with a negative connotation (left half of the radial graphs in Figure 4; Table 2), for example, that AI cannot draw an inference in general or from a portrait, were the most used by participants from Japan. Overall, 58.4% of the Japanese subjects' arguments have a negative connotation. The second most negative are arguments by Argentinean subjects with 55.3%. The least negative are arguments by participants from Kenya, with

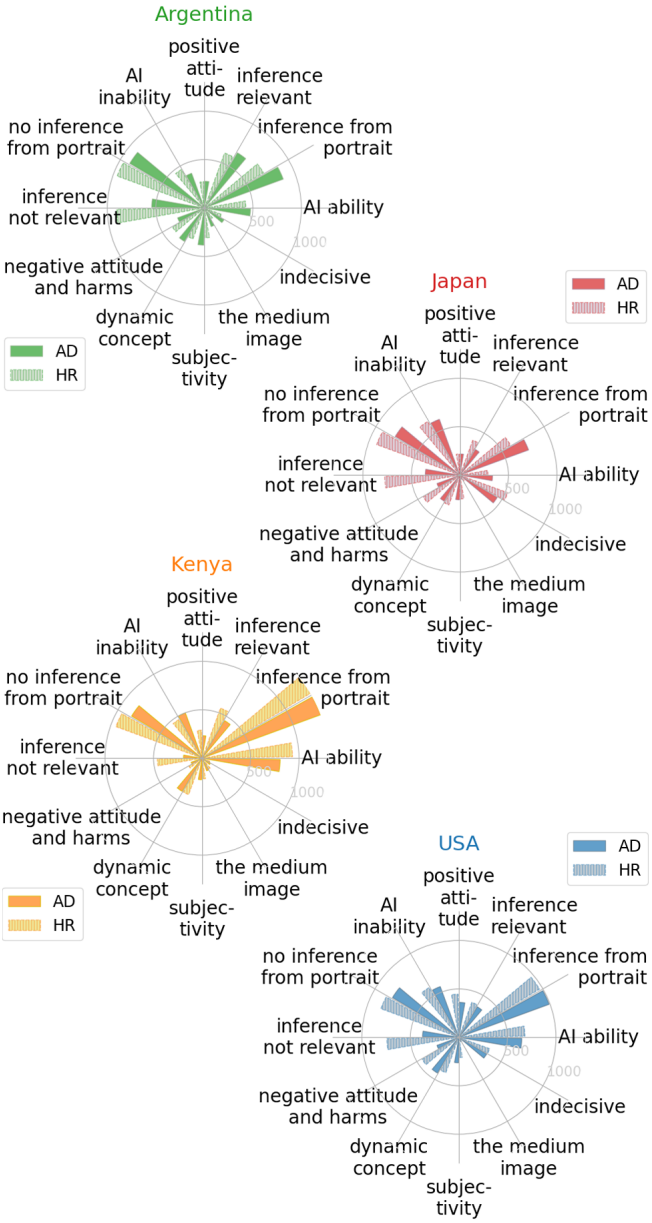

Figure 4: Comparison of frequencies of justifications used by country per context. The four justification codes in the upper right quarter of each radial graph have a positive or affirmative connotation. The six justification codes on the left half of each graph have a negative or rejecting connotation. The justifications *indecisive* and *the medium image* are of neutral nature.

42.9%. The argument of an inference being irrelevant was used most often by participants from Argentina.

Third, the decision context influences certain lines of argumentation. In all countries, the argument describing the irrelevance of the inference was used significantly more often by participants in the

Table 2: Sum of positive/negative/neutral justifications.*

country	∑ positive	∑ negative	∑ neutral	*sum*
Argent.	4298 (38.0%)	6255 (55.3%)	761 (6.7%)	11314
Japan	3121 (30.8%)	5920 (58.4%)	1092 (10.8%)	10133
Kenya	6099 (53.1%)	4924 (42.9%)	462 (4.0%)	11485
USA	4849 (42.1%)	5906 (51.1%)	771 (6.7%)	11526
sum	18367 (41.2%)	23005 (51.8%)	3086 (7.0%)	44458

*The numbers refer to all classified justification labels. A participant's response might have received more than one label. Hence, the sum of classified labels is larger than the sum of participants' responses.

hiring context than by participants in the advertising context (Argentina: t(8769)=-10.0, p<.001, Japan: t(7751)=-10.6, p<.001, Kenya: t(8295)=-9.2, p<.001, USA: t(8684)=-9.7, p<.001; see Table 11 in the Appendix). Participants in the hiring context argued significantly more often for the irrelevance of an inference than for its relevance, except for participants from Kenya (see Table 10 in the Appendix). In all countries except Kenya, justifications describing negative attitudes and harm are used significantly more often by participants in the hiring context than in the advertising context (Argentina: t(8769)=-3.2, p=.001164, Japan: t(7751)=-5.2, p<.001, USA: t(8684)=-6.4, p<.001). Justifications describing that an inference can be drawn from a portrait are used significantly more often in the advertising context than in the hiring context (Argentina: t(8769)=4.8, p<.001, Japan: t(7751)=7.6, p<.001, USA: t(8684)=3.7, p<.001)). There are no context-related differences in the reasoning regarding positive attitudes, the relevance of an inference, and no inference from the portrait.

In summary, participants from Japan, followed by participants from Argentina, were the most skeptical about the justifiability of AI inference-making. In contrast, participants from Kenya applied positive arguments most frequently. We observe more rejecting perceptions (both in inference ratings and arguments) from subjects in the hiring compared to the advertising context, and this across all countries.

4 DISCUSSION

4.1 Discussion of results

We surveyed participants from Argentina, Japan, Kenya, and the USA whether and why they agree or disagree with AI inferring specific attributes from faces. Documenting participants' written evaluations of AI inferences, we conclude that there are different types of justifications (epistemic validity, AI capabilities, pragmatic (dis)advantage), each with different arguments. Different inferences are subject to particular justification profiles. For example, the perception that an AI is incapable and that an image is not sufficient evidence were the most prominent justifications against the justifiability of inferring whether a person is intelligent or trustworthy. The context in which the inference is used is also a critical factor in participants' justifications. For example, the participants in the hiring context argue more frequently than participants in the advertising context for the irrelevance of an inference and highlight the negative consequences that can result from inferring an inference. Our results support recent policy developments. They also indicate the power of corporations' rhetoric on positive and useful use cases that seem to have high relevance to some of our study participants.

During the negotiations on the EU AI Act proposal [33], strong voices from the scientific community and civil society contributed to, for example, the ban or more restrictive use of facial recognition and emotion recognition technologies in specific contexts [45, 79, 89]. However, many of these voices are not satisfied with the result achieved [e.g., 78]. Concerning the inference of emotion, amendments made by Members of the European Parliament to the EU AI Act Proposal explicitly addressed the lack of validity by adding to the recitals that there "are serious concerns about the scientific basis of AI systems aiming to detect emotions [...]" [34, Recital 44]. While it is evident that the scientific concerns are known to EU policymakers, the agreements between the European Parliament and the Council of the European Union to ban AI systems that infer emotion only in some contexts and to introduce several exemptions and limitations on the ban of biometric categorization, as outlined in Section 1, highlights the power of involved interest groups. While our study results confirm that people perceive AI inference-making in the hiring context as more consequential than in the advertising context, the lack of validity of the practice of inferring emotions applies equally to all contexts. Our participants recognized the epistemic invalidity, in particular, for the inferences *intelligent* and *trustworthy*. Discussions on the reasonableness and justifiability of facial analysis AI in different contexts and concerning different inferences must continue. And this needs to happen across national borders.

Our results indicate that subjects perceived the inferences *trustworthy* and *intelligent* as rather unjustifiable. This replicates the findings of previous studies [32, 98]. However, most arguments justifying the inferences *skin color*, *wearing glasses*, and *emotion expression* concern the ability of AI to make these inferences in general or from portraits. Fewer participants criticize that the inference is not relevant, potentially causing harm, or not stable as a target variable. There is more variation in the use of different justifications for the inferences *beautiful*, *age*, and *gender*. In line with Miceli et al.'s [62] findings that target variables are hardly ever questioned, we perceive such questioning to occur primarily for the inferences *trustworthy* and *intelligent*, and to be of less intensity for the other inferences. Equally, there appeared to be some participant groups that were more critical across all inferences than other participants. For example, referring to the concept of the inference *beautiful*, one participant argued that the conclusion "would be based only off the programmers view of beauty" [sic]. Others recognized that the target variables to be inferred are societal constructs, as stressed by researchers [e.g., 83, 84]. But this was only of concern to a minority of participants.

Participants' affirming justifications, which portray their perceptions of pragmatic advantages of AI inference-making, could be illustrative of corporations' rhetoric on positive and useful use cases of facial analysis AI. The affirmative responses emphasizing participants' positive attitude and their belief in the capabilities of AI could suggest that some participants trust companies to deploy facial analysis AI in certain application contexts. This, in turn, would underline the near-unlimited epistemic authority of organizations – both corporate and research – in shaping the interpretation of

human faces in visual data by determining target variables they presume to be useful for the product. An example of such communicative corporate practice is a facial analysis AI company that analyzes customers' faces to improve their aesthetics [74]. They advertise their service by citing academic studies that show that attractive individuals often have better career opportunities [102] or that having an attractive partner typically leads to more contentment [9]. Not only is the generalizability of such studies up for debate, but it is also questionable if these are the inferences and facial analysis AI use cases that society actually perceives as justifiable or reasonable.

Considering participants' ratings, a considerable participant group seems not to have a clear opinion on facial analysis AI (indicated by many inference ratings close to or at 4: "neither agree nor disagree"). Additionally, opinions vary significantly across and within countries, pointing to disagreements on the justifiability of AI inference-making (see Figure 12 in the Appendix). This emphasizes the need for public debate over these types of AI systems. Public debate should provide societies with opportunities to become informed about the scientific concerns on facial analysis AI, including their conceptual invalidity and risks of harm. These perspectives are probably not as prevalent in the everyday lives of laypeople as the positive rhetoric from companies about the benefits and power of AI for facial analysis.

4.2 Future research and limitations

Future research could replicate this study and provide participants with more information on facial analysis AI, including critical accounts. A comparison with the results of this study, which measured participants' spontaneous reactions to facial analysis AI promoted by the fictitious company ImageInsights, could help to understand what factors influence participants' beliefs in AI capabilities for facial analysis AI tasks. Such future analysis could also further elaborate on country-specific or cultural factors that influence people's perceptions of technology [43, 46, 47].

Reflecting on our methodological approach to analyzing participants' qualitative text responses, some limitations should be highlighted. First, we cannot be sure to have identified all themes raised by participants. A pure qualitative analysis approach would have allowed extract even more justification types. This was infeasible given the large number of data and shows the limits of automatic text classification. It also required reducing the coding scheme to a reasonable number of justification codes so that we were realistically able to prepare a training, testing, and validation dataset for the classification model. However, with our research team set-up and iterative approach, we have taken measures to find a set of justifications representing participants' responses across countries and to counteract subjective interpretation. Second, the F-1 score of 0.75 for the ensemble model indicates that the predictions still contain false positives and false negatives. Future studies should address this issue, for example, by including the user ratings for each of the inferences as additional information in the classification models to better contextualize participants' responses.

4.3 Researcher positionality

We acknowledge our positionality relative to this study. Our study critically complements predominant research on the ethical perceptions and judgments of AI with participants from WEIRD (Western, Educated, Industrialized, Rich, Democratic) backgrounds. To operationalize this effort, we established a cross-national research. It took six months to form a research team representing researchers from all studied countries. Our final research team consists of Japanese, Kenyan, and US-based researchers, as well as a researcher from the same linguistic area as Argentina and a country bordering it. We had at least two researchers from each linguistic community manually code participants' justifications – only researchers with the specific national experience or background interpreted the written justifications of their linguistic community.

Besides this multinational makeup, our team is multi-gender, varied in socioeconomic status, and multi-disciplinary. It represents computer sciences with a focus on NLP and software development, science and technology studies, political science, philosophy, and privacy economics. Our team has complementary expertise in qualitative and quantitative research. It includes undergraduate, graduate, post-graduate, and faculty researchers. At the start of the collaboration, all team members had a university affiliation. Undergraduates were either compensated for their contributions or received course credit. The composition of this research team minimized the potential bias resulting from a nation- and discipline-specific interpretation of participants' ethical justifications.

5 FINAL REMARKS

To conclude, our study presents cross-country differences and commonalities in agreements and disagreements with facial analysis AI from visual data. We identified twelve themes of justification categorized into five coherent thematic clusters and three justification types: 44% of the participants' responses discussed the epistemic (in)validity of the inference. In the other responses, participants expressed their belief or non-belief in AI capabilities or highlighted pragmatic (dis)advantages of AI inference-making. For the inferences *trustworthy* and *intelligent*, we observed the most rejection, both in inference ratings and justifications. These findings support politicians advocating for a ban or restrictions on these types of classification tasks. We also found significant variations in perceptions within and across countries, both in ratings (e.g., *beautiful*, *skin color*) and justifications (e.g., *age*, *gender*). The decision context had an effect both on inference ratings (more agreement in the advertising context) and on the normative judgment for some inferences. In the hiring context, participants raised concerns about adverse consequences as well as highlighted the irrelevance of an inference more frequently than participants in the advertising context. This suggests that participants evaluate the consequences of facial analysis differently across decision contexts. The diversity of laypeople's opinions makes it clear that there is no universal "common sense" that supports AI inference-making. These findings underscore the importance for critical data scientists and civil society organizations to persist in raising awareness, offering societies a platform to learn about the scientific validity of facial analysis AI and develop an informed stance.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments that improved the paper. We thank Prof. Arisa Ema and Miriam Rateike for their support in connecting us with researchers in their network. We thank Emmanuel Syrmoudis and Prof. Sameer Patil for their support in improving the survey experience. We thank the participants of the *2023 European Workshop on Algorithmic Fairness* as well as the participants of the *2023 Many Worlds of AI: Intercultural Approaches to the Ethics of Artificial Intelligence Conference* for their constructive comments and feedback on our project presentations. This research is partially supported by the TUM Institute for Ethics in Artificial Intelligence. Our early work on this project was supported by a Volkswagen Foundation Planning Grant.

REFERENCES

[1] Glenn Adams and Hazel Rose Markus. 2003. Toward a conception of culture suitable for a social psychology of culture. In *The Psychological Foundations of Culture*, Mark Schaller and Christian S. Crandall (Eds.). Lawrence Erlbaum Associates, Mahwah, NJ, 344–369. https://doi.org/10.4324/9781410608994

[2] James Agarwal, Naresh K. Malhotra, and Ruth N. Bolton. 2010. A cross-national and cross-cultural approach to global market segmentation: An application using consumers' perceived service quality. *Journal of International Marketing* 18, 3 (2010), 18–40. https://doi.org/10.1509/jimk.18.3.18

[3] Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. 2017. Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition* 72 (2017), 15–26. https://doi.org/10.1016/j.patcog.2017.06.031

[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. https://doi.org/10.1027/1614-2241/a000014

[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] 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] BeautyScoreTest. 2024. Beauty Score Calculator. https://www.beautyscoretest.com/

[8] 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://doi.org/10.1109/CVPR.2016.600

[9] Ellen Berscheid, Karen Dion, Elaine Walster, and G. William Walster. 1971. Physical attractiveness and dating choice: A test of the matching hypothesis. *Journal of Experimental Social Psychology* 7, 2 (1971), 173–189. https://doi.org/10.1016/0022-1031(71)90065-5

[10] Betaface. 2021. Betaface API. https://www.betafaceapi.com/wpa/

[11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184. https://doi.org/10.1145/3531146.3533083

[12] 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

[13] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

[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 (2018), 185–203. https://doi.org/10.1007/s42235-018-0015-y

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

[17] Clarifai. 2023. Appearance multicultural classifier (ethnicity-demographics-recognition). https://clarifai.com/clarifai/main/models/ethnicity-demographics-recognition

[18] Richard Cook, Adam Eggleston, and Harriet Over. 2022. The cultural learning account of first impressions. *Trends in Cognitive Sciences* 26, 8 (2022), 656–668. https://doi.org/10.1016/j.tics.2022.05.007

[19] Alan S. Cowen, Dacher Keltner, Florian Schroff, Brendan Jou, Hartwig Adam, and Gautam Prasad. 2021. Sixteen facial expressions occur in similar contexts worldwide. *Nature* 589, 7841 (2021), 251–257. https://doi.org/10.1038/s41586-020-3037-7

[20] Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, New Haven, CT. https://doi.org/10.12987/9780300252392

[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. AI Now Institute. https://ainowinstitute.org/publication/ai-now-2019-report-2

[22] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & Society* 36, 4 (2021), 1105–1116. https://doi.org/10.1007/s00146-021-01162-8

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[24] 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

[25] 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

[26] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press.

[27] Charles Efferson and Sonja Vogt. 2013. Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports* 3, Article 1047 (2013), 7 pages. 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] 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] Severin Engelmann and Orestis Papakyriakopoulos. 2023. Social media as classification systems: procedural normative choices in user profiling. In *Handbook of Critical Studies of Artificial Intelligence*. Edward Elgar Publishing, 619–630.

[31] Severin Engelmann, Valentin Scheibe, Fiorella Battaglia, and Jens Grossklags. 2022. Social Media Profiling Continues to Partake in the Development of Formalistic Self-Concepts. Social Media Users Think So, Too.. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 238–252.

[32] 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*. 128–141. https://doi.org/10.1145/3531146.3533080

[33] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://tinyurl.com/43eap7bz

[34] European Parliament. 2024. Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts - CORRIGENDUM to the position of the European Parliament adopted at first reading on 13 March 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

[35] Eyeris. n.d.. Artificially Intelligent Emotion Recognition Technology that reads facial micro-expressions. http://www.emovu.com/

[36] Face++. n.d.. Beauty Score. https://www.faceplusplus.com/beauty/

[37] Faception. 2021. Our technology. https://www.faception.com/our-technology

[38] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. https://doi.org/10.3758/BF03193146

[39] Lauren Fichten. 2023. 'This is a problem': A new hyper-realistic TikTok beauty filter is freaking people out. *Vice* (February 2023). https://www.vice.com/en/article/pkg747/tiktok-beauty-filter-bold-glamor-problem

[40] 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

[41] 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

[42] Srinivas Gutta, Harry Wechsler, and P. Jonathon Phillips. 1998. Gender and ethnic classification of face images. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. 194–199. https://doi.org/10.1109/AFGR.1998.670948

[43] Tilman Hartwig, Yuko Ikkatai, Naohiro Takanashi, and Hiromi M Yokoyama. 2023. Artificial intelligence ELSI score for science and technology: a comparison between Japan and the US. *AI & SOCIETY* 38, 4 (2023), 1609–1626. https://doi.org/10.1007/s00146-021-01323-9

[44] Aya Hassouneh, A.M. Mutawa, and M. Murugappan. 2020. Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked* 20, Article 100372 (2020), 9 pages. https://doi.org/10.1016/j.imu.2020.100372

[45] Merve Hickok and Marc Rotenberg. 2023. *Making the AI Act Work: How Civil Society Can Ensure Europe's New Regulation Serves People & Society*. Report. European Artificial Intelligence & Society Fund. https://europeanaifund.org/wp-content/uploads/2023/10/041023-FINAL-for-publication-EAISF-AIA-implmentation-report.pdf

[46] Michel Hohendanner, Chiara Ullstein, Yosuke Buchmeier, and Jens Grossklags. 2023. Exploring the Reflective Space of AI Narratives Through Speculative Design in Japan and Germany. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good* (Lisbon, Portugal) *(GoodIT '23)*. Association for Computing Machinery, New York, NY, USA, 351–362. https://doi.org/10.1145/3582515.3609554

[47] Michel Hohendanner, Chiara Ullstein, Dohjin Miyamoto, Emma Fukuwatari Huffman, Gudrun Socher, Jens Grossklags, and Hirotaka Osawa. 2024. Metaverse Perspectives from Japan: A Participatory Speculative Design Case Study. arXiv:2401.17428 [cs.CY]

[48] Bastian Jaeger, Bastiaan Oud, Tony Williams, Eva G. Krumhuber, Ernst Fehr, and Jan B. Engelmann. 2022. Can people detect the trustworthiness of strangers based on their facial appearance? *Evolution and Human Behavior* 43, 4 (2022), 296–303. https://doi.org/10.1016/j.evolhumbehav.2022.04.004

[49] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. 2020. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision* 12, 1–3 (2020), 1–308. https://doi.org/10.1561/0600000079

[50] 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, Article 8487 (2020), 11 pages. https://doi.org/10.1038/s41598-020-65358-6

[51] 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

[52] 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.

[53] Michal Kosinski. 2021. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports* 11, Article 100 (2021), 7 pages. https://doi.org/10.1038/s41598-020-79310-1

[54] Udo Kuckartz. 2014. *Mixed methods: Methodologie, Forschungsdesigns und Analyseverfahren*. Springer Fachmedien, Wiesbaden, Germany. https://doi.org/10.1007/978-3-531-93267-5

[55] Tuan Le Mau, Katie Hoemann, Sam H. Lyons, Jennifer M.B. Fugate, Emery N. Brown, Maria Gendron, and Lisa Feldman Barrett. 2021. Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nature Communications* 12, 1, Article 5037 (2021), 13 pages. https://doi.org/10.1038/s41467-021-25352-6

[56] Howard Lee and Yi-Ping Phoebe Chen. 2015. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications* 42, 12 (2015), 5356–5365. https://doi.org/10.1016/j.eswa.2015.02.005

[57] 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

[58] John Leuner. 2019. *A Replication Study: Machine Learning Models are Capable of Predicting Sexual Orientation from Facial Images*. Master's thesis. University of Pretoria. https://doi.org/10.48550/arXiv.1902.10739

[59] Peter Lewinski, Tim M. Den Uyl, and Crystal Butler. 2014. Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics* 7, 4 (2014), 227–236. https://doi.org/10.1037/npe0000028

[60] VCloudX PTE. LTD. 2023. EnableX Face AI Transforming Conversations with Intelligence. https://www.enablex.io/cpaas/faceai/

[61] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW, Article 72 (2019), 23 pages. https://doi.org/10.1145/3359174

[62] 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, Article 115 (2020), 25 pages. https://doi.org/10.1145/3415186

[63] 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

[64] Microsoft. 2023. Perceived Emotion Recognition Using the Face API. https://learn.microsoft.com/en-us/xamarin/xamarin-forms/data-cloud/azure-cognitive-services/emotion-recognition

[65] Sophie J. Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8, Article e2120481119 (2022), 3 pages. https://doi.org/10.1073/pnas.2120481119

[66] Helen Noble and Joanna Smith. 2015. Issues of validity and reliability in qualitative research. *Evidence-based Nursing* 18, 2 (2015), 34–35. https://doi.org/10.1136/eb-2015-102054

[67] Marwa Obayya, Saud S. Alotaibi, Sami Dhahb, Rana Alabdan, Mesfer Al Duhayyim, Manar Ahmed Hamza, Mohammed Rizwanullah, and Abdelwahed Motwakel. 2022. Optimal deep transfer learning based ethnicity recognition on face images. *Image and Vision Computing* 128, Article 104584 (2022), 11 pages. https://doi.org/10.1016/j.imavis.2022.104584

[68] 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

[69] 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

[70] 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

[71] Michael Quinn Patton. 1999. Enhancing the quality and credibility of qualitative analysis. *Health Services Research* 34, 5 (1999), 1189–1208. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089059/

[72] Mariana Pinto da Costa. 2021. Conducting cross-cultural, multi-lingual and multi-country focus groups: guidance for researchers. *International Journal of Qualitative Methods* 20 (2021), 1–6. https://doi.org/10.1177/16094069211049929

[73] QOVES. 2023. Facial Assessment Tool. https://qoves.com/facial-assessment-tool/

[74] QOVES. 2023. Home. https://qoves.com/#citation

[75] 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*. 429–435. https://doi.org/10.1145/3306618.3314244

[76] 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

[77] Stig Hebbelstrup Rye Rasmussen, Steven G. Ludeke, and Robert Klemmensen. 2023. Using deep learning to predict ideology from facial photographs: Expressions, beauty, and extra-facial information. *Scientific Reports* 13, 1, Article 5257 (2023), 8 pages. https://doi.org/10.1038/s41598-023-31796-1

[78] European Digital Rights. 2023. EU AI Act Trilogues: Status of Fundamental Rights Recommendations. https://edri.org/our-work/eu-ai-act-trilogues-status-of-fundamental-rights-recommendations/

[79] European Digital Rights. 2023. Reclaim Your Face. https://reclaimyourface.eu/

[80] Stein Rokkan. 1993. Cross-cultural, cross-societal and cross-national research. *Historical Social Research / Historische Sozialforschung* 18, 2 (1993), 6–54. https://doi.org/10.12759/hsr.18.1993.2.6-54

[81] 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

[82] 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, Article 317 (2021), 37 pages. https://doi.org/10.1145/3476058

[83] 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), 20539517211053712. https://doi.org/10.1177/20539517211053712

[84] 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, Article 58 (2020), 35 pages. https://doi.org/10.1145/3392866

[85] 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

[86] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How western, educated, industrialized, rich, and democratic is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 160–171. https://doi.org/10.1145/3593013.3593985

[87] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. arXiv:2404.17293

[88] 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

[89] Pia Sombetzki. 2023. Biometrische Überwachung: Wie biometrische Erkennungssysteme Grundrechte beschneiden können. https://algorithmwatch.org/de/wp-content/uploads/2023/08/9_v2.pdf

[90] 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

[91] 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

[92] Visage Technologies. 2023. Age detection. https://visagetechnologies.com/age-detection

[93] Visage Technologies. 2023. Arbelle - Virtual Makeup SDK. https://visagetechnologies.com/makeup-sdk/

[94] Visage Technologies. 2023. Gender detection. https://visagetechnologies.com/gender-detection

[95] Alexander Todorov. 2017. *Face Value: The Irresistible Influence of First Impressions*. Princeton University Press, Princeton, NJ. https://doi.org/10.1515/9781400885725

[96] 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

[97] Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, and Jens Grossklags. 2023. A reflection on how cross-cultural perspectives on the ethics of facial analysis AI can inform EU policymaking. In *Workshop Proceedings of the 2023 European Workshop on Algorithmic Fairness*. Article 40, 5 pages. https://ceur-ws.org/Vol-3442/paper-40.pdf

[98] Chiara Ullstein, Severin Engelmann, Orestis Papakyriakopoulos, Michel Hohendanner, and Jens Grossklags. 2022. AI-competent individuals and laypeople tend to oppose facial analysis AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. Article 9, 12 pages. https://doi.org/10.1145/3551624.3555294

[99] Fons J.R. Van De Vijver. 2009. Types of comparative studies in cross-cultural psychology. *Online Readings in Psychology and Culture* 2, 2, Article 2 (2009), 12 pages. https://doi.org/10.9707/2307-0919.1017

[100] Konstantina Vemou and Anna Horvath. 2020. *Facial Emotion Recognition*. EDPS TechDispatch. Technology and Privacy Unit of the European Data Protection Supervisor (EDPS). https://edps.europa.eu/system/files/2021-05/21-05-26_techdispatch-facial-emotion-recognition_ref_en.pdf

[101] 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

[102] Chris Warhurst, Diane Van den Broek, Richard Hall, and Dennis Nickson. 2009. Lookism: The new frontier of employment discrimination? *Journal of Industrial Relations* 51, 1 (2009), 131–136. https://doi.org/10.1177/0022185608096808

[103] 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

[104] 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

[105] 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

[106] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291. https://doi.org/10.1162/coli_a_00502

APPENDIX

These appendices document and support our analysis process.

A STUDY DESIGN

A.1 Vignette design

	BASE STUDY			Follow-up 1: PRAGMATISM		Follow-up 2: ACCURACY	
	Control [BASE]	Control [PRAG-MATISM]	Control [ACCU-RACY]	LOW	HIGH	LOW	HIGH
AD	A company called ImageInsight has developed software that uses artificial intelligence (AI) to analyze images. The software analyzes portraits of users on social media in order to show social media users suitable advertisements for products. How does that work? ImageInsight's deep learning AI is presented with a portrait of a user showing only the user's face but nothing else. The AI scans the user's face and makes a variety of inferences about the user. Based on these and other inferences a user will be shown a personalized advertising material on the social media platform. [*PAGE BREAK*] To recap: ImageInsight's AI scans and analyzes the portrait of a social media user. Along with other evaluation metrics, the analysis is then used to show the social media user suitable product advertisement on social media.						
	-	ImageInsight states that users are more likely to engage with personalized advertising when its AI is used. In other words, ImageInsights claims that using its AI leads to more user engagement with personalized advertisement.	ImageInsight states that its AI achieves accuracy for all facial inferences that are used for personalized advertising. In other words, ImageInsights claims that its AI can correctly predict inferences from faces that it scans and analyses.	ImageInsight states that users are 20% more likely to engage with personalized advertising when its AI is used. In other words, ImageInsights claims that users are 0.2 times more likely to engage in personalized advertising when its AI is used.	ImageInsight states that users are 300% more likely to engage with personalized advertising when its AI is used. In other words, ImageInsights claims that users are 3 times more likely to engage in personalized advertising when its AI is used.	ImageInsight states that its AI achieves 60% accuracy for all facial inferences that are used for personalized advertising. In other words, ImageInsights claims that its AI correctly predicts an inference for 60 out of 100 faces that it scans and analyses.	ImageInsight states that its AI achieves 95% accuracy for all facial inferences that are used for personalized advertising. In other words, ImageInsights claims that its AI correctly predicts an inference for 95 out of 100 faces that it scans and analyses.
	A social media user just uploaded a portrait image to a social media platform.						
HR	A company called ImageInsight has developed software that uses artificial intelligence (AI) to analyze images. The software analyzes portraits of job applicants in order to select suitable candidates during hiring procedures. How does that work? ImageInsight's deep learning AI is presented with a portrait of an applicant showing only the applicant's face but nothing else. The AI scans the applicant's face and makes a variety of inferences about the applicant. Based on these and other inferences an applicant will be **invited** for a job interview. [*PAGE BREAK*] To recap: ImageInsight's AI scans and analyzes the portrait of an applicant. Along with other evaluation metrics, the analysis is then used to decide whether the applicant is invited for a job interview.						
	-	ImageInsight states that companies are able to hire more top performing employees when its AI is used. In other words, ImageInsights claims that using its AI leads to companies hiring more successful employees.	ImageInsight states that its AI achieves accuracy for all facial inferences that are used when hiring job candidates. In other words, ImageInsights claims that its AI can correctly predict inferences from faces that it scans and analyses.	ImageInsight states that companies are able to hire 20% more top performing employees when its AI is used. In other words, ImageInsights claims that using its AI leads to companies hiring 0.2 times more top performing employees.	ImageInsight states that companies are able to hire 300% more top performing employees when its AI is used. In other words, ImageInsights claims that using its AI leads to companies hiring 3 times more top performing employees.	ImageInsight states that its AI achieves 60% accuracy for all facial inferences that are used when hiring job candidates. In other words, ImageInsights claims that its AI correctly predicts an inference for 60 out of 100 faces that it scans and analyses.	ImageInsight states that its AI achieves 95% accuracy for all facial inferences that are used when hiring job candidates. In other words, ImageInsights claims that its AI correctly predicts an inference for 95 out of 100 faces that it scans and analyses.
	An applicant just submitted an application with a portrait image to a prospective employer.						
	ImageInsight's AI scans the portrait and then draws several inferences about the person. One of these inferences is whether or not the person is of a certain age, is beautiful, is intelligent, is trustworthy, is of a certain gender, is wearing glasses, is of a certain skin color, is expressing an emotion. [*one inference is individually shown on one a separate survey page*]						
	Do you agree or disagree that this sort of inference made by ImageInsight's AI (whether or not a person is <inference>) is justifiable? [RATING] **Please justify your rating in 1 - 2 sentences? [JUSTIFICATION]**						

Figure 5: Vignette Design: Wording of text in survey for all experimental groups.

A.2 Comprehension checks

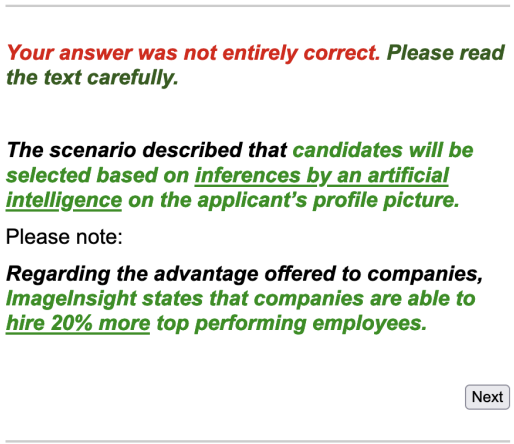

Figure 6: Comprehension Check: Information displayed to participants who wrongly answered the comprehension checks. The exemplary text in the image was displayed to a participant in the advertisement context and the pragmatism low scenario.

A.3 Sample size calculation

We conducted a power analysis using the software G*Power [38] to estimate the appropriate sample size for the data collection. The sample size calculations reflect our vignette design with two different contexts, three different informational settings (comprising seven more detailed information scenarios), and four countries, resulting in 24 experimental groups overall. Intending to calculate multivariate statistics, we base our assumptions of achievable effect sizes on prior studies [32] that achieved a Pillai V = 0.04 at a small effect size (η^2 = 0.041) for the differentiation of two contexts in an experimental setting with 24 groups. Based on standard values of α = 0.05 and power (1-β) = 0.9, the minimum required sample size for achieving small effect sizes is a total of 952 survey participants per country. For the sample size calculation, the following considerations were taken:

- 2 contexts * 3 studies (meaning: 3 information settings) * 4 countries = 24 groups
- sample size calculations results in 1632 participants in total
- 1632 participants / 3 studies = 544 participants per study, i.e., per information setting
- 544 * 7 detailed information scenarios to build the 3 information settings = 3808 participants
- 3808 participants / 4 countries = 952 participants per country.

A.4 Demographics

Table 3: Participant sampling and target quotas

Level	Argentina	Japan	Kenya	USA
Gender				
female	565 (0.51)	538 (0.51)	528 (0.5)	618 (0.51)
male	549 (0.49)	524 (0.49)	571 (0.5)	548 (0.49)
non-binary	7	1	3	3
no answer	2	5	5	1
Age				
18-34	420 (0.4)	226 (0.21)	571 (0.54)	360 (0.31)
35-54	413 (0.36)	364 (0.33)	381 (0.32)	387 (0.33)
55+	290 (0.26)	477 (0.46)	154 (0.14)	419 (0.37)
no answer	0	1	1	0
University				
with degree	584 (0.5)	751 (0.58)	734 (0.5)	738 (0.65)
without degree	527 (0.5)	312 (0.42)	350 (0.5)	419 (0.35)
no answer	12	5	23	13
total N	1123	1068	1107	1170

Numbers in brackets indicate targeted quotas based on representative samples for each country.

B QUALITATIVE CATEGORIZATION PROCESS

B.1 Building a category scheme

B.1.1 Coding workshop 1. The first 2h coding workshop "ideating" concluded with the construction of a preliminary category scheme shown in Figure 7. During this first workshop, we discussed the categories previously identified by the country teams during the coding phase 1. We agreed on the following clusters of categories that were likely to be relevant to data from all countries: AI (in)ability, general opinion on technology, face as evidence, ir/relevance, human comparison, ethics, and miscellaneous. This set of category clusters served as a draft category scheme for coding phase 2.

B.1.2 Coding workshop 2 and 3. Figure 8 presents the results of workshop 2 and 3.

B.1.3 Coding workshop 4. Figure 9 presents the results of workshop 4.

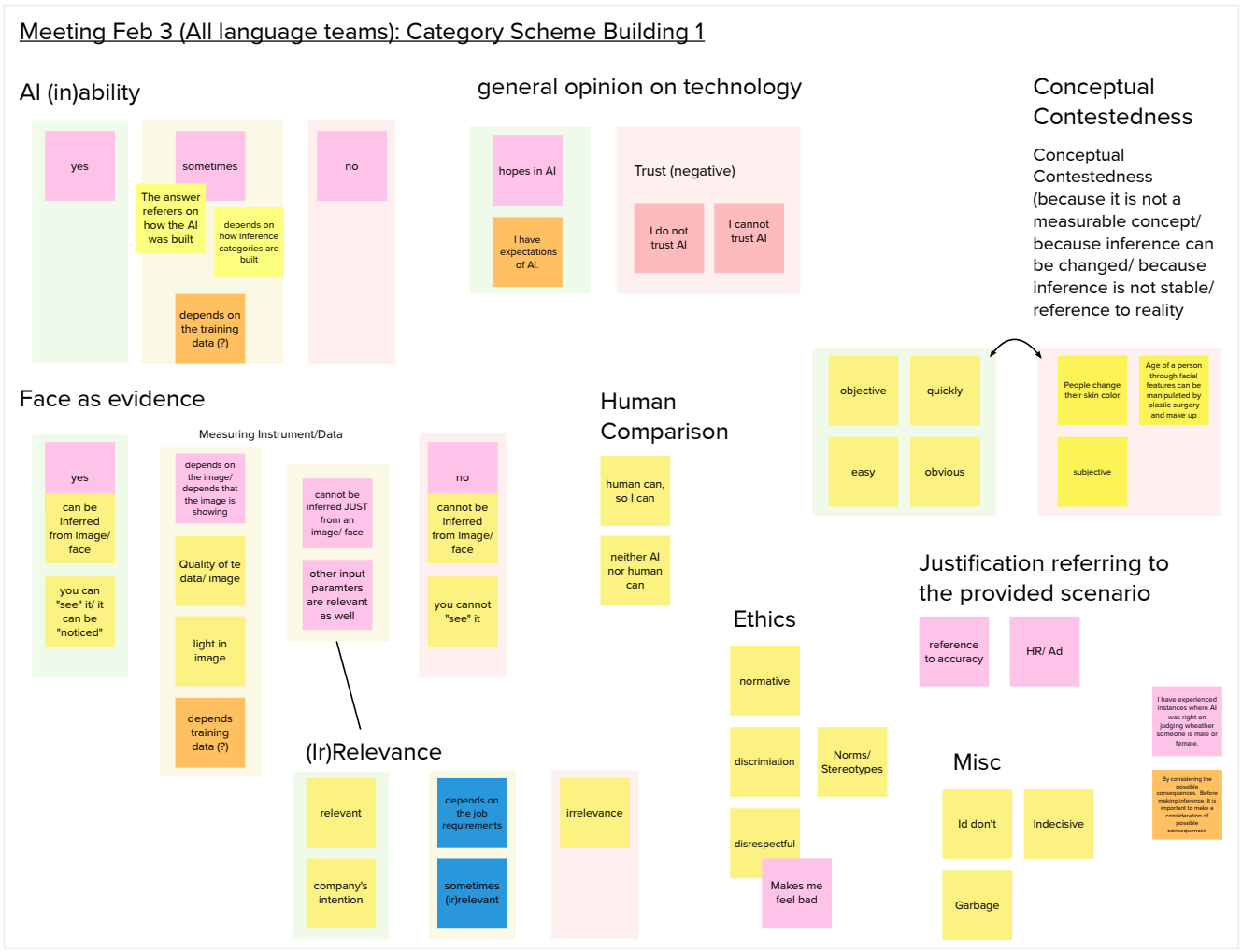

Figure 7: Result of Category Ideation Workshop 1 (zoom in to see details).

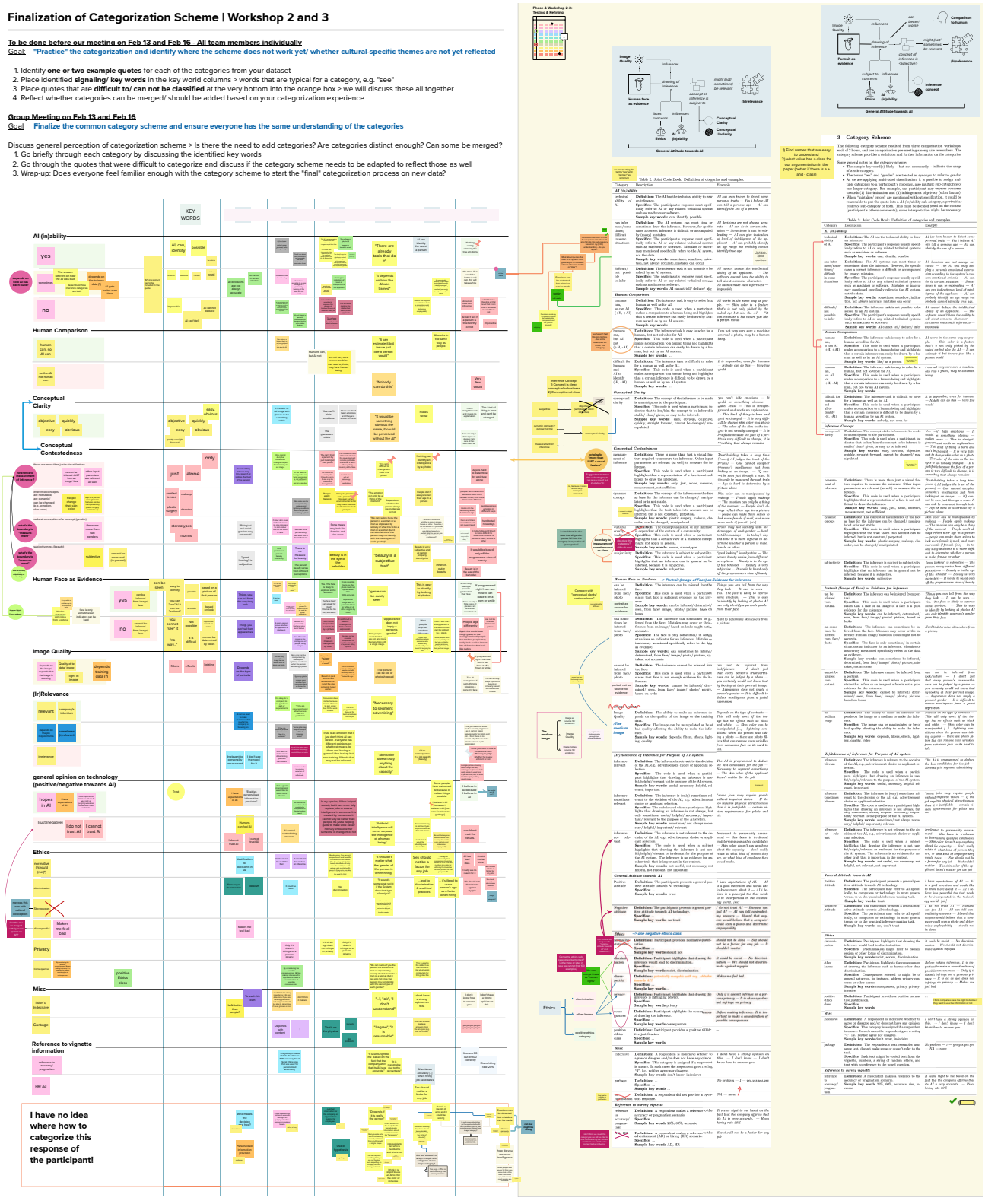

Figure 8: Result of Category Workshop 2 and 3 (zoom in to see details).

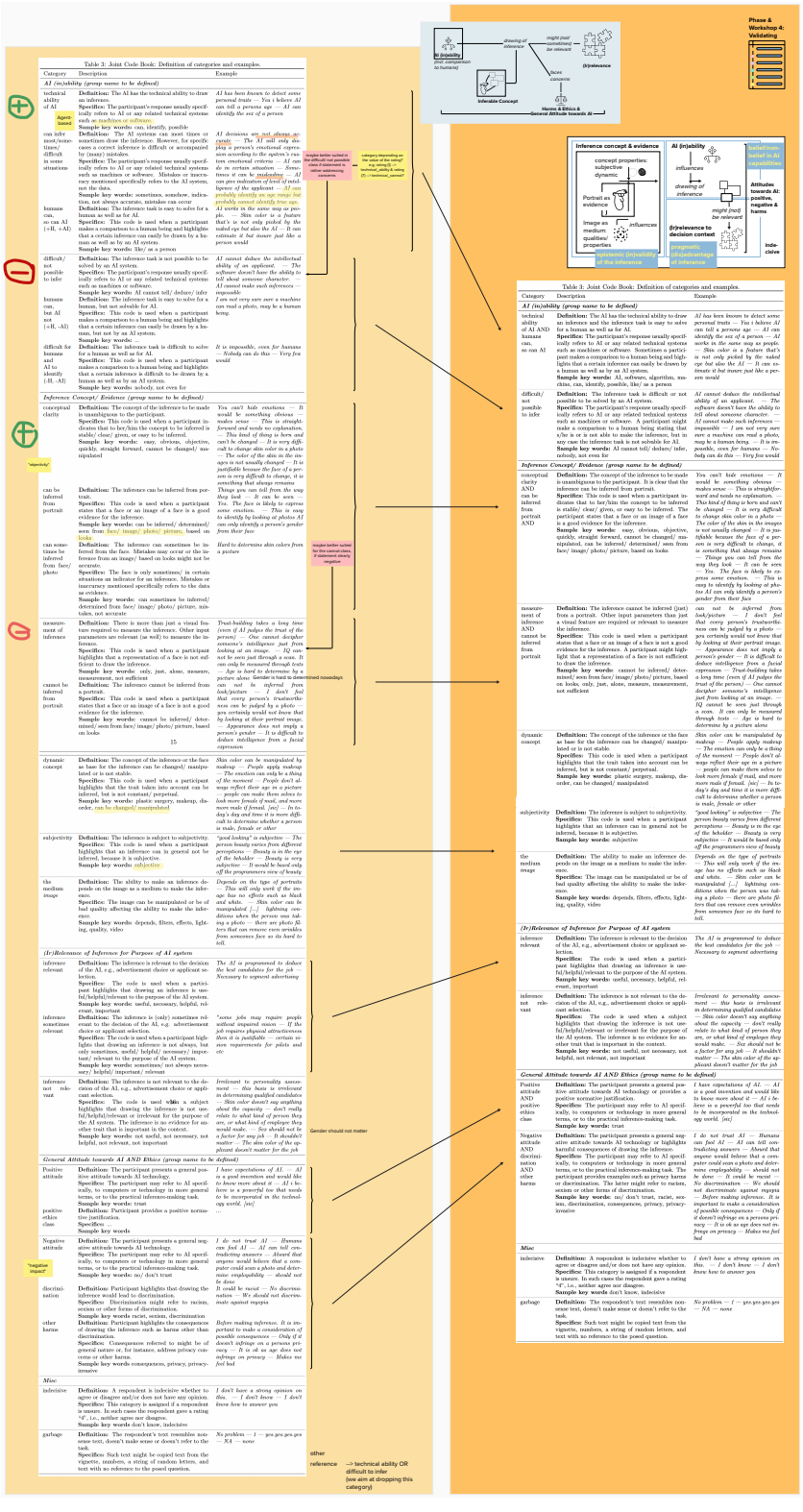

Figure 9: Result of Category Workshop 4 (zoom in to see details).

B.2 Codebook

The following coding scheme resulted from one categorization pre-meeting and four categorization workshops, each of 2 to 3 hours, among the research team. The coding scheme provides a definition and further information on each of the code. The coding scheme is furthermore based on experiences gathered while applying previous draft versions of the coding scheme by each of the researchers. The experiences were discussed and codes with similar application use cases were merged.

Some general notes on the category scheme:

- The sample keyword(s) likely – but not necessarily – indicate the usage of a sub-category.
- The terms "sex" and "gender" are treated as synonym to refer to *gender*.
- As we are applying multi-label classification, it is possible to assign multiple codes to a participant's response.
- When "mistakes/ errors" are mentioned without specification, it could be reasonable to put the quote into a *AI (in)ability* sub-category, a *portrait as evidence* sub-category, or both. This must be decided based on the participant's other comments and the ratings.
- For comments that mention that a system can "sometimes" draw an inference (from a portrait) or that the inference is "sometimes" relevant, the attribution to its respective positive or negative label has to be decided based on the rating and the context (participant's others comments).

The following tables define the justification codes and provide examples.

Table 4: Joint Code Book: Definition of categories and examples.

Category	Description	Examples
AI (in)ability *Note: agent-based (The AI or sometimes also "you" in general can or cannot make the inference.)*		
technical ability of AI, humans can and so can AI	**Definition:** The AI has the technical ability to draw an inference and the inference task is easy to solve for an AI. The AI system is accurate. **Specifics:** The participant's response usually specifically refers to AI or any related technical systems such as machines or software. Sometimes a participant makes a comparison to a human being and highlights that a certain inference can easily be drawn by a human as well as by an AI system. **Sample key words:** AI, software, algorithm, machine, can, identify, possible, like/ as a person	*AI has been known to detect some personal traits | Yea i believe AI can tell a persons age | AI can identify the sex of a person | AI works in the same way as people. | Skin color is a feature that's is not only picked by the naked eye but also the AI | It can estimate it but insure just like a person would*
difficult/ not possible to infer, difficult for human and AI, human can and AI cannot	**Definition:** The inference task is difficult or not possible to be solved by an AI system. **Specifics:** The participant's response usually specifically refers to AI or any related technical systems such as machines or software. A participant might make a comparison to a human being stating that s/he is or is not able to make the inference, but in any case the inference task is not solvable for AI. **Sample key words:** AI cannot tell/ deduce/ infer, nobody, not even for	*AI cannot deduce the intellectual ability of an applicant. | The software doesn't have the ability to tell about someone character. | AI cannot make such inferences | impossible | I am not very sure sure a machine can read a photo, may be a human being. | It is impossible, even for humans | Nobody can do this | Very few would*
Inference Concept/ Evidence *Note: inference-concept-based*		
conceptual clarity, can be inferred from portrait	**Definition:** The concept of the inference to be made is unambiguous to the participant. The inference concept is clear. It is obvious that the inference concept can be inferred from a portrait. The face "shows" the inference. The inference concept can be seen. **Specifics:** This code is used when a participant indicates that to her/him the concept to be inferred is stable/ clear/ given, or easy to be inferred. The participant states that a face or an image of a face is a good evidence for the inference. **Sample key words:** easy, obvious, objective, quickly, straight forward, cannot be changed/ manipulated, can be inferred/ determined/ seen from face/ image/ photo/ picture, based on looks	*You can't hide emotions | It would be something obvious | makes sense | This is straightforward and needs no explanation. | This kind of thing is born and can't be changed | It is very difficult to change skin color in a photo | The color of the skin in the images is not usually changed | It is justifiable because the face of a person is very difficult to change, it is something that always remains | Things you can tell from the way they look | It can be seen | Yes. The face is likely to express some emotion. | This is easy to identify by looking at photos*
measurement of inference, cannot be inferred from portrait	**Definition:** The inference cannot be inferred (just) from a portrait. Other input parameters than just a visual feature are required or relevant to measure the inference. **Specifics:** This code is used when a participant states that a face or an image of a face is not a good evidence for the inference. A participant might highlight that a representation of a face is not sufficient to draw the inference. **Sample key words:** cannot be inferred/ determined/ seen from face/ image/ photo/ picture, based on looks, only, just, alone, measure, measurement, not sufficient	*can not be inferred from look/picture | I don't feel that every person's trustworthiness can be judged by a photo | you certainly would not know that by looking at their portrait image. | Appearance does not imply a person's gender | It is difficult to deduce intelligence from a facial expression | Trust-building takes a long time (even if AI judges the trust of the person) | One cannot decipher someone's intelligence just from looking at an image. | IQ cannot be seen just through a scan. It can only be measured through tests | Age is hard to determine by a picture alone*
dynamic concept	**Definition:** The concept of the inference or the face as base for the inference can be changed/ manipulated or is not stable. **Specifics:** This code is used when a participant highlights that the trait taken into account can be inferred, but is not constant/ perpetual. **Sample key words:** plastic surgery, makeup, disorder, can be changed/ manipulated	*Skin color can be manipulated by makeup | People apply makeup | The emotion can only be a thing of the moment | People don't always reflect their age in a picture | people can make them selves to look more female if mail, and more more male if femail. [sic] | In today's day and time it is more difficult to determine whether a person is male, female or other*

Category	Description	Examples
subjectivity	**Definition:** The inference is subject to subjectivity. **Specifics:** This code is used when a participant highlights that an inference can in general not be inferred, because it is subjective. **Sample key words:** subjective	*"good looking" is subjective | The person beauty varies from different perceptions | Beauty is in the eye of the beholder | Beauty is very subjective | It would be based only off the programmers view of beauty*
the medium *image*	**Definition:** The ability to make an inference depends on the image as a medium to make the inference. **Specifics:** The image as a medium can be manipulated or be of bad quality affecting the ability to make the inference. **Sample key words:** depends, filters, effects, lighting, quality, video	*Depends on the type of portraits | This will only work if the image has no effects such as black and white. | lighting conditions when the person was taking a photo | there are photo filters that can remove even wrinkles from someones face so its hard to tell.*
(Ir)Relevance of Inference for Purpose of AI system		
inference relevant	**Definition:** The inference is relevant to the decision of the AI, e.g., advertisement choice or applicant selection. **Specifics:** The code is used when a participant highlights that drawing an inference is useful/helpful/relevant to the purpose of the AI system. **Sample key words:** useful, necessary, helpful, relevant, important	*The AI is programmed to deduce the best candidates for the job | Necessary to segment advertising | Useful to offer e.g. certain skin care products*
inference not relevant	**Definition:** The inference is not relevant to the decision of the AI, e.g., advertisement choice or applicant selection. **Specifics:** The code is used when a subject highlights that drawing the inference is not useful/helpful/relevant or irrelevant for the purpose of the AI system. The inference is no evidence for another trait that is important in the context. **Sample key words:** not useful, not necessary, not helpful, not relevant, not important	*Irrelevant to personality assessment | this basis is irrelevant in determining qualified candidates | Skin color doesn't say anything about the capacity | don't really relate to what kind of person they are, or what kind of employee they would make. | Sex should not be a factor for any job | It shouldn't matter | The skin color of the applicant doesn't matter for the job*
General Attitude Toward AI and Ethics Reflections		
Positive attitude, positive ethics	**Definition:** The participant presents a general positive attitude towards AI technology or provides a positive normative justification. **Specifics:** The participant may refer to AI specifically, to computers or technology in more general terms, or to the practical inference-making task. **Sample key words:** trust	*I have expectations of AI. | AI is a good invention and would like to know more about it | AI i believe is a powerful too that needs to be incorporated in the technology world. [sic]*
Negative attitude, discrimination, and other harms	**Definition:** The participant presents a general negative attitude towards AI technology or highlights harmful consequences of drawing the inference. **Specifics:** The participant may refer to AI specifically, to computers or technology in more general terms, or to the practical inference-making task. The participant provides examples such as privacy harms or discrimination. The latter might refer to racism, sexism or other forms of discrimination. **Sample key words:** no/ don't trust, racist, sexism, discrimination, consequences, privacy, privacy-invasive	*I do not trust AI | Humans can fool AI | AI can tell contradicting answers | Absurd that anyone would believe that a computer could scan a photo and determine employability | should not be done | It could be racist | No discrimination | We should not discriminate against myopia | Before making inference. It is important to make a consideration of possible consequences | Only if it doesn't infringe on a persons privacy | It is ok as age does not infringe on privacy | Makes me feel bad | It is discriminatory to take into account skin color | That is outright racism. It's disgusting*
Misc		
indecisive	**Definition:** A respondent is indecisive whether to agree or disagree and/or does not have any opinion. **Specifics:** This category is assigned if a respondent is unsure. In such cases the respondent gave a rating "4", i.e., neither agree nor disagree. **Sample key words** don't know, indecisive	*I don't have a strong opinion on this. | I don't know | I don't know how to answer you*
[garbage]*	**Definition:** The respondent's text resembles nonsense text, doesn't make sense or doesn't refer to the task. **Specifics:** Such text might be copied text from the vignette, numbers, a string of random letters, and text with no reference to the posed question.	*No problem | 1 | yes.yes.yes.yes | NA | none*

* Garbage class used for cleaning of data.

B.3 Building an automatic classification model

B.3.1 Workshop for finalizing sample datasets for training automatic classifiers. Figure 10 presents the workshop instructions for the data-filling-up workshop. This workshop aimed at finalizing sample datasets for training automatic classifiers. After each team labeled 456 comments (57 survey participants x 8 responses) based on our final coding scheme, we met for another workshop, this time focusing on filling up the categories with insufficient exemplary responses to train a classifier model. The top right plot in Figure 10 helped identify the categories that required finding further exemplary comments to have sufficient data to train a classifier model. We aimed at 33 examples per language per category.

B.3.2 Automatic translation: Documentation of error rates. The following table documents our error rates for the automatic translations using the deepL API. The language pairs Spanish-English and Japanese-English were validated by team members. The language pair Spanish-Japanese was validated by a hired translator.

Table 5: Error rates for automatic translations.

	source language	target language 1	error rate	target language 2	cumulative error rate
round 1	English	Spanish	6%		
	Spanish	English	6%		
	English	Japanese	8%		
	Japanese	English	10%		
	Japanese	Spanish	30%		
	Spanish	Japanese	28%		
round 2	Spanish	English		Japanese	14%
	Japanese	English		Spanish	14%

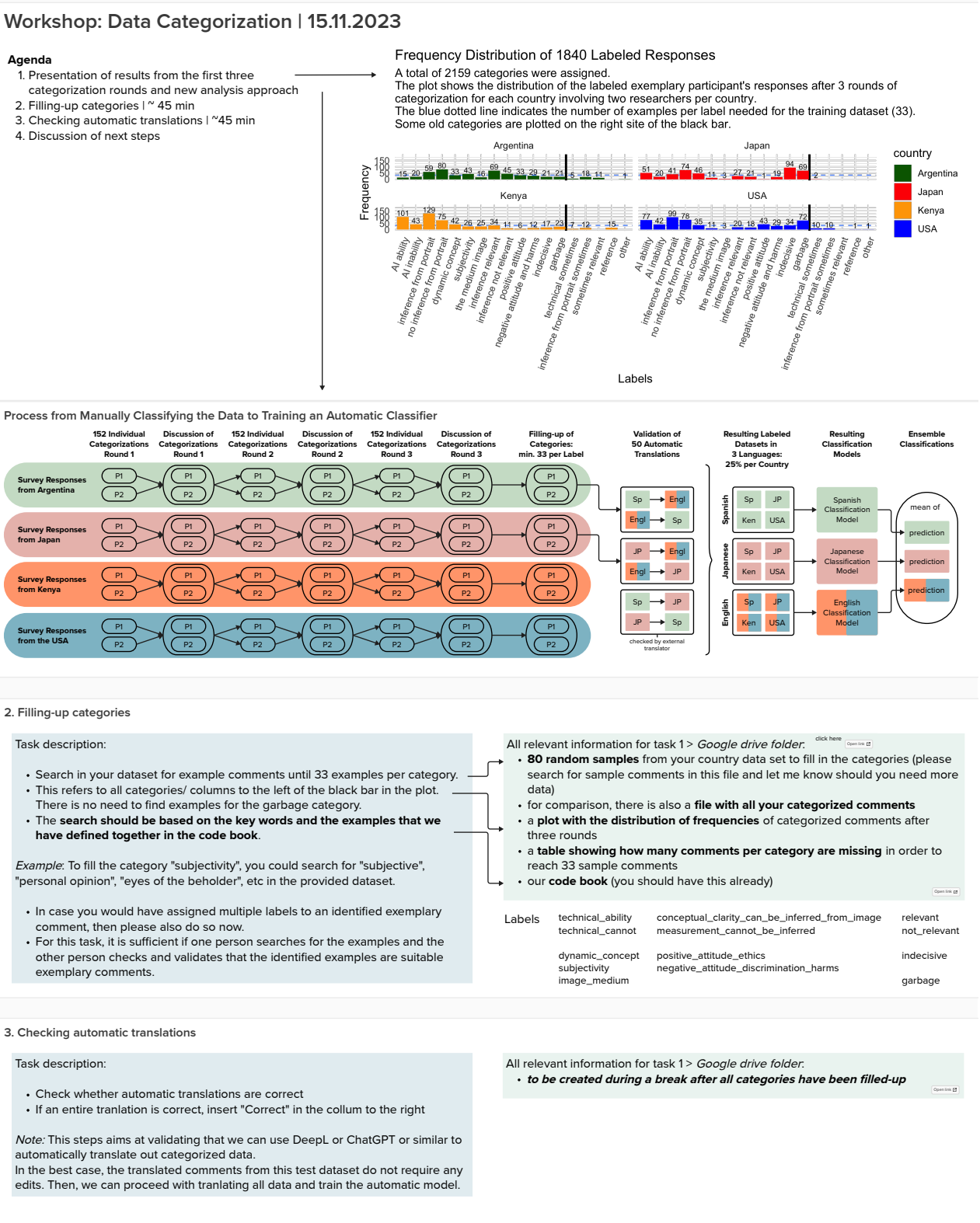

Figure 10: Workshop instructions for the final data-filling-up workshop (zoom in to see details).

C RESULTS DOCUMENTATION

C.1 Frequencies of qualitative justifications

The following tables document the absolute frequencies for all identified justification codes. The frequencies refer to the number of participants' comments to which we assigned a specific justification code.

Table 6: Frequencies of identified justification codes in participants' responses. (sorted by size)

Justification code	Absolute frequency	Percentage
inference from portrait	8252	16.40
no inference from portrait	7236	14.38
AI ability	4819	9.58
inference not relevant	4546	9.03
AI inability	4408	8.76
inference relevant	3900	7.75
dynamic concept	3197	6.35
negative attitude and harms	2528	5.02
indecisive	2507	4.98
positive attitude	2448	4.86
subjectivity	2224	4.42
the medium image	916	1.82
garbage*	3347	6.65

*If all of a participant's comments only received the label "garbage", then we did not consider this response for analysis.

Table 7: Frequencies of identified justification codes in participants' responses per country and per context.

country	case	AI ability	inference from portrait	inference relevant	positive attitude
Argentina	AD	476	866	664	276
Argentina	HR	428	711	605	272
Japan	AD	337	754	310	219
Japan	HR	290	608	384	219
Kenya	AD	806	1302	453	233
Kenya	HR	933	1531	551	290
USA	AD	648	990	368	361
USA	HR	681	969	386	446

Table 7 continued

country	case	AI inability	no inference from portrait	inference not relevant	negative attitude and harms
Argentina	AD	382	896	543	294
Argentina	HR	463	957	905	383
Japan	AD	614	777	357	252
Japan	HR	646	915	776	440
Kenya	AD	499	846	192	120
Kenya	HR	466	946	461	159
USA	AD	558	798	375	241
USA	HR	599	858	744	442

Table 7 continued

country	case	dynamic concept	subjectivity	the medium image	indecisive
Argentina	AD	401	383	204	235
Argentina	HR	340	308	133	189
Japan	AD	314	255	47	448
Japan	HR	328	246	70	527
Kenya	AD	399	222	138	100
Kenya	HR	400	214	133	91
USA	AD	430	263	47	334
USA	HR	389	209	51	339

C.2 Statistical analysis of group differences in the use of specific justifications

We calculated two sample Welch t-tests for unequal variances and Kruskal Wallis Tests for identifying statistical differences across groups. For the t-tests, a positive t-statistic suggests that the sample mean of the first group is higher than the sample mean of the second group, while a negative t-statistic indicates the opposite. We indicate the groups in the table captions. We report the significance levels as follows: p-value < 0.05 (*), p-value < 0.01 (**), p-value < 0.001 (***).

C.2.1 Differences: 'AI ability' vs. 'AI inability'.

Table 8: Welch t-tests: Differences in use of the justification (1) AI ability vs. (2) AI inability.

Country	Case	T-Stat	P-Val	CI Low	CI Upp	df	sig
Arg.	HR	-1.24	0.22	-0.02	0.00	8861.20	
Arg.	AD	3.38	0.00	0.01	0.03	8584.15	***
Japan	HR	-12.45	0.00	-0.10	-0.07	7562.94	***
Japan	AD	-9.72	0.00	-0.09	-0.06	6656.29	***
Kenya	HR	13.73	0.00	0.09	0.12	8305.58	***
Kenya	AD	9.38	0.00	0.06	0.10	7380.40	***
USA	HR	2.47	0.01	0.00	0.03	9211.34	*
USA	AD	2.81	0.00	0.01	0.04	8099.18	**

C.2.2 Differences: 'inference from portrait' vs. 'no inference from portrait'.

Table 9: Welch t-tests: Differences in use of the justification (1) inference from portrait vs. (2) no inference from portrait.

Country	Case	T-Stat	P-Val	CI Lo	CI Upp	df	sig
Arg.	HR	-6.70	0.00	-0.07	-0.04	8758.34	***
Arg.	AD	-0.80	0.42	-0.02	0.01	8664.61	
Japan	HR	-8.73	0.00	-0.09	-0.06	8225.17	***
Japan	AD	-0.66	0.51	-0.03	0.01	7069.16	
Kenya	HR	13.97	0.00	0.11	0.15	8751.83	***
Kenya	AD	11.70	0.00	0.10	0.14	7513.74	***
USA	HR	2.90	0.00	0.01	0.04	9218.64	**
USA	AD	5.15	0.00	0.03	0.07	8081.62	***

C.2.3 Differences: 'inference relevant' vs. 'inference not relevant'.

Table 10: Welch t-tests: Differences in use of the justification (1) inference relevant vs. (2) inference not relevant.

Country	Case	T-Stat	P-Val	CI Lo	CI Upp	df	sig
Arg.	HR	-8.51	0.00	-0.08	-0.05	8652.63	***
Arg.	AD	3.76	0.00	0.01	0.04	8605.04	***
Japan	HR	-12.51	0.00	-0.11	-0.08	7780.75	***
Japan	AD	-1.91	0.06	-0.03	0.00	7041.91	
Kenya	HR	3.01	0.00	0.01	0.03	8892.36	**
Kenya	AD	10.82	0.00	0.06	0.08	6710.40	***
USA	HR	-11.45	0.00	-0.09	-0.06	8581.72	***
USA	AD	-0.27	0.79	-0.01	0.01	8129.42	

C.2.4 Differences in justification usage: advertisement vs. hiring context.

Table 11: Welch t-tests: Differences in use of justifications between participants in the (1) advertisement and the (2) hiring context.

Justification Code	Country	T-Stat	P-Val	df	sig.
AI ability	Argentina	2.06	0.04	8769	*
AI ability	Japan	4.22	0.00	7751	***
AI ability	Kenya	0.26	0.80	8295	
AI ability	USA	1.54	0.12	8684	
AI inability	Argentina	-2.58	0.01	8769	*
AI inability	Japan	2.42	0.02	7751	*
AI inability	Kenya	3.71	0.00	8295	***
AI inability	USA	1.04	0.30	8684	
inference from portrait	Argentina	4.83	0.00	8769	***
inference from portrait	Japan	7.88	0.00	7751	***
inference from portrait	Kenya	-0.16	0.88	8295	
inference from portrait	USA	3.74	0.00	8684	***
no inference from portrait	Argentina	-1.03	0.30	8769	
no inference from portrait	Japan	0.29	0.77	7751	
no inference from portrait	Kenya	1.09	0.28	8295	
no inference from portrait	USA	1.25	0.21	8684	
dynamic concept	Argentina	2.67	0.01	8769	**
dynamic concept	Japan	1.74	0.08	7751	
dynamic concept	Kenya	2.29	0.02	8295	*
dynamic concept	USA	3.41	0.00	8684	***
subjectivity	Argentina	3.29	0.00	8769	***
subjectivity	Japan	2.44	0.01	7751	*
subjectivity	Kenya	2.07	0.04	8295	*
subjectivity	USA	3.95	0.00	8684	***
the medium image	Argentina	4.16	0.00	8769	***
the medium image	Japan	-1.20	0.23	7751	
the medium image	Kenya	1.62	0.11	8295	
the medium image	USA	0.23	0.82	8684	
inference relevant	Argentina	2.24	0.02	8769	*
inference relevant	Japan	-0.52	0.60	7751	
inference relevant	Kenya	-0.65	0.52	8295	
inference relevant	USA	1.15	0.25	8684	
inference not relevant	Argentina	-10.00	0.00	8769	***
inference not relevant	Japan	-10.61	0.00	7751	***
inference not relevant	Kenya	-9.17	0.00	8295	***
inference not relevant	USA	-9.75	0.00	8684	***
positive attitude	Argentina	0.46	0.65	8769	
positive attitude	Japan	1.89	0.06	7751	
positive attitude	Kenya	-0.72	0.47	8295	
positive attitude	USA	-1.24	0.21	8684	
negative attitude and harms	Argentina	-3.25	0.00	8769	**
negative attitude and harms	Japan	-5.17	0.00	7751	***
negative attitude and harms	Kenya	-1.05	0.29	8295	
negative attitude and harms	USA	-6.39	0.00	8684	***
indecisive	Argentina	2.54	0.01	8769	*
indecisive	Japan	0.23	0.82	7751	
indecisive	Kenya	1.74	0.08	8295	
indecisive	USA	1.52	0.13	8684	

C.2.5 Differences in justification usage across all countries for selected inferences and justification codes. The following Table 12 documents that there are statistically significant differences across all countries for the presented inferences. However, due to the large number of responses per justification code, the significant levels should be interpreted with caution. The identified differences cannot necessarily be interpreted to be of real-world significance. Indeed, while we observe statistical significance across the countries for these particular justifications and inference combinations, participants from different countries align in their usage patterns of these justification arguments in comparison to other arguments (see Figure 3 in Section 3.2 in the main text).

Table 12: Kruskal Wallis Tests: Differences across all countries for selected inferences

Inferences	Justification Code	H-Statistic	P-Value	sig
beautiful	subjectivity	20.01	0.00	***
emotion expression	inference from portrait	137.41	0.00	***
gender	inference from portrait	224.96	0.00	***
wearing glasses	inference from portrait	161.00	0.00	***
intelligent	no inference from portrait	68.55	0.00	***
intelligent	AI inability	29.15	0.00	***
trustworthy	no inference from portrait	32.26	0.00	***
trustworthy	AI inability	52.26	0.00	***

C.3 Statistical analysis of rating differences in groups

C.3.1 Differences in ratings across countries (all inferences merged).

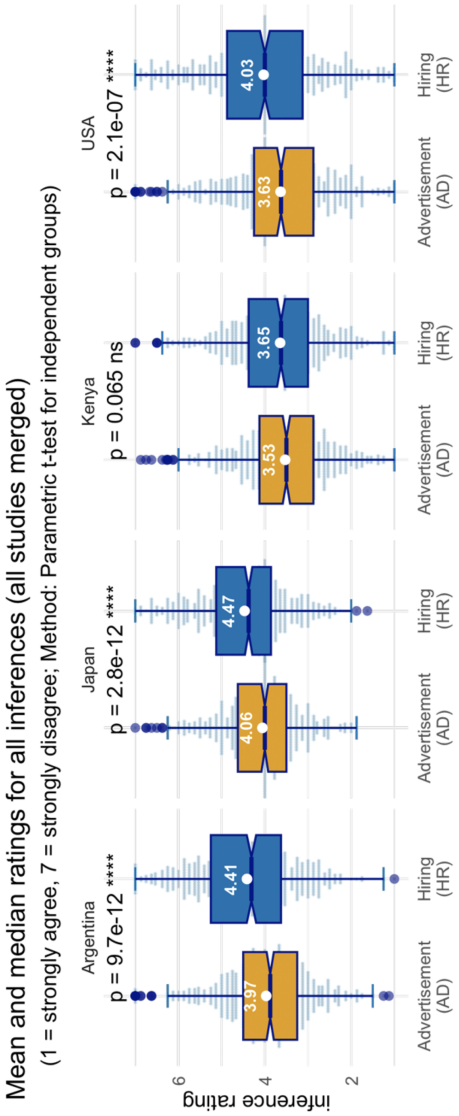

Figure 11: Differences in mean ratings for all countries and contexts (inferences merged). Orange: advertisement context; blue: hiring context. (Method: Parametric t-test for two independent groups, unequal variance)

C.3.2 Differences in ratings across countries for each of the inferences. Please note: the Figure below displays violin plots and boxplots with notches based on medians. The t-tests are calculated for the means of each sub-group. While non-overlapping notches indicate significant differences in groups, some of the t-tests suggest differing conclusions. These special cases might indicate that there is, in particular, a lot of variation in participants' opinions.

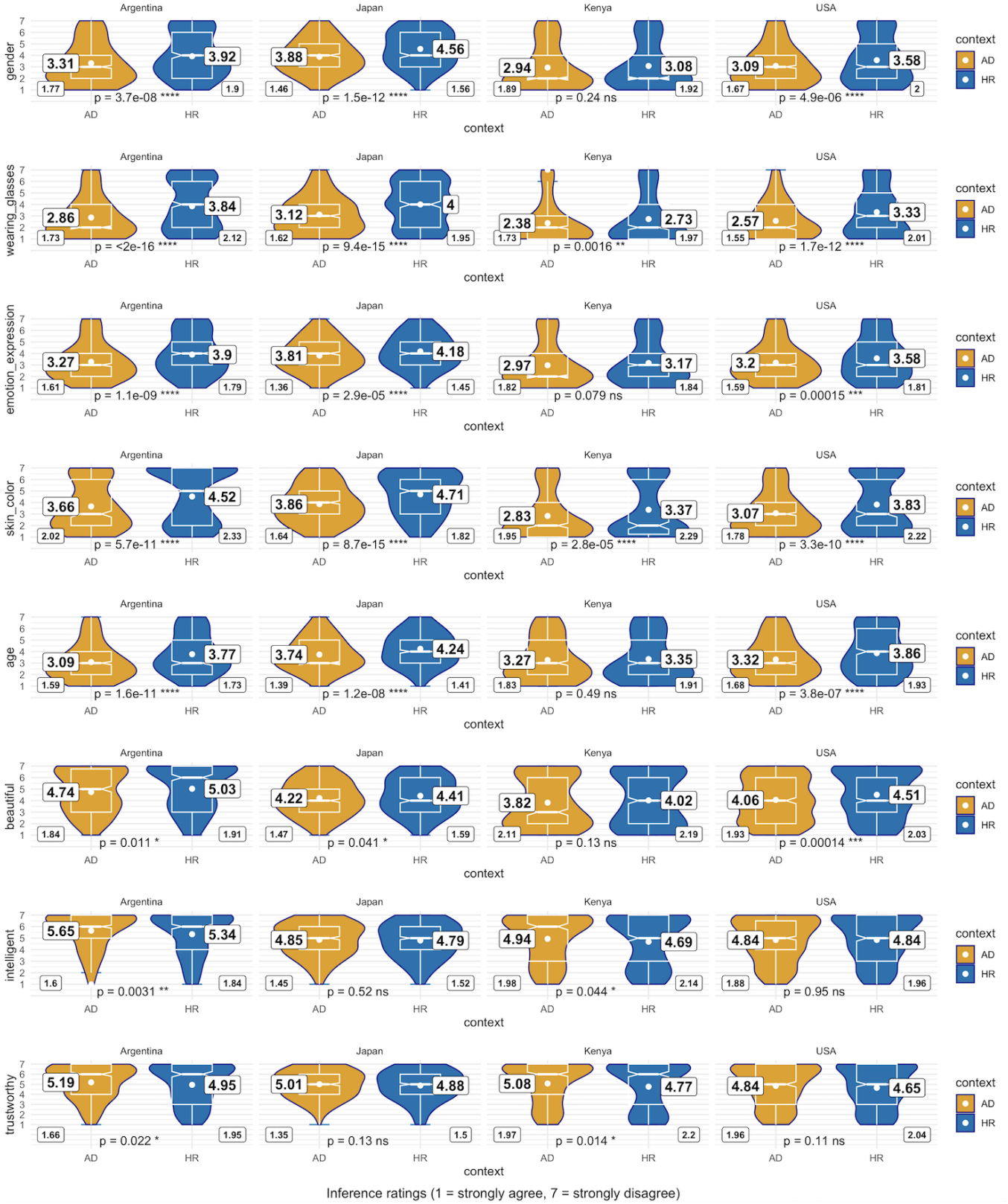

Figure 12: Differences in mean ratings for all inferences across all countries and contexts. Standard deviation plotted to bottom corners. Orange: advertising context; blue: hiring context. (Method: Parametric t-test for two independent groups, unequal variance; Cases with "Can't answer" (NA) removed)