Appendix: What People Think AI Should Infer From Faces

SEVERIN ENGELMANN*, Technical University of Munich, Chair of Cyber Trust, Germany
CHIARA ULLSTEIN*, Technical University of Munich, Chair of Cyber Trust, Germany
ORESTIS PAPAKYRIAKOPOULOS, Princeton University, Center for Information Technology Policy, USA
JENS GROSSKLAGS, Technical University of Munich, Chair of Cyber Trust, Germany

ACM Reference Format:
Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. Appendix: What People Think AI Should Infer From Faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3531146.3533080

*Denotes equal contribution.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Copyright held by the owner/author(s).
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea*, https://doi.org/10.1145/3531146.3533080.

1 VIGNETTE SCENARIOS

a) Advertisement Scenario

A company developed a software that uses **artificial intelligence** to analyze images.

The software analyzes portraits of **users** uploaded to a social media platform in order to show these users suitable advertisements for products. How does that work? The artificial intelligence is presented with a portrait of a user showing only the user's face but nothing else. The software 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 particular advertising material on the social media platform.

Which statement best describes the scenario presented above?

- ○ Product advertisements will be recommended to a user based on inferences by an artificial intelligence on his or her profile picture.
- ○ Recommended product advertisements are based on inferences by a company's employees, who assess the portraits of users.

b) Hiring Scenario

A company developed a software that uses **artificial intelligence** to analyze images.

The software will analyze portraits of **applicants** in order to select suitable candidates during hiring procedures. How does that work? The artificial intelligence is presented with a portrait of an applicant showing only the applicant's face but nothing else. The software scans the applicant's face and makes a variety of inferences about the applicant.

Based on these and other inferences an applicant will be selected or rejected for a job position.

Which statement best describes the scenario presented above?

- ○ The selection of candidates is based on inferences by a company's employees, who assess the portraits of applicants.
- ○ Candidates will be selected based on inferences by an artificial intelligence on the applicant's profile picture.

Fig. 1. Vignette description of the hypothetical advertising scenario a) and hiring scenario b).

2 PRIMARY TASK

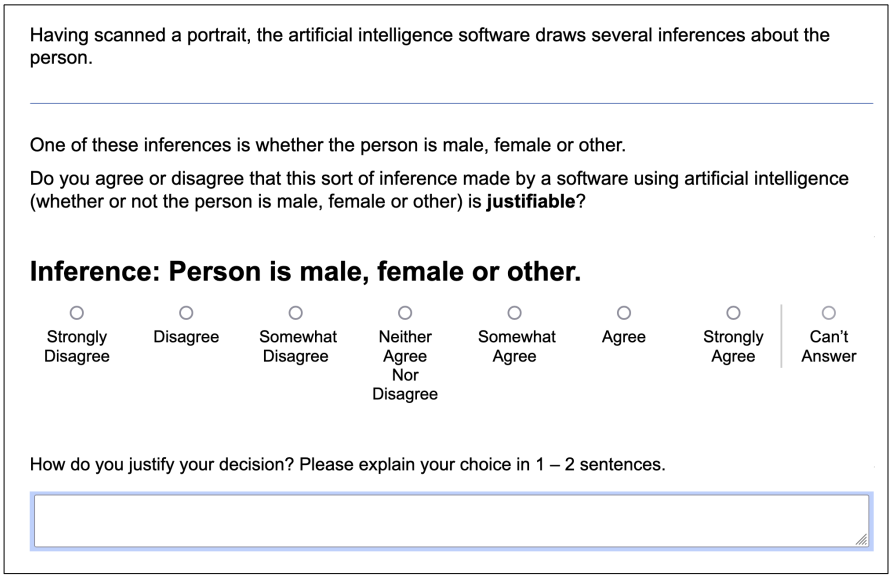

Fig. 2. Example interface of the primary rating task and the prompt to provide a written response. Example does not show treatment with the presentation of a definition of the evaluative term.

3 GENERIC DEFINITIONS OF EVALUATIVE TERMS

Table 1. Generic definitions of the six evaluative adjectives presented to half of the participants. All definitions were based on the Cambridge Dictionary, some formulations were slightly adapted to fit our context.

inference	definition
reasonable	What do we mean by **reasonable**? Something is reasonable if it's based on good sense and/or in accordance with reason.
fair	What do we mean by **fair**? Something is fair if it's based on equality without favoritism or discrimination.
justifiable	What do we mean by **justifiable**? Something is justifiable if it can be marked by a good or legitimate reason.
responsible	What do we mean by **responsible**? Something is responsible if it can answer for its conduct and obligations.
appropriate	What do we mean by **appropriate**? Something is appropriate if it's suitable or compatible in the circumstances.
acceptable	What do we mean by **acceptable**? Something is acceptable if it can be agreed on and is worthy of being accepted.

4 DATA CLEANING

The data was cleaned based on the criteria presented in Table 2, which gives an overview on the measures taken and a count of identified cases per measure. The SoSci Survey online survey tool provides a relative speed index (RSI) that identifies fast responding participants. This index indicates how much faster a participant has completed the experiment than the typical participant (median). As recommended by SoSci, all respondents with an RSI >= 2 (n = 418) are removed. All samples with duration time between 2 minutes and 4 minutes, cases that rated all inferences with the same rating, and cases with a RSI value above 1.75 were manually checked. Cases identified as problematical were discussed with a second researcher and removed in case of agreement.

Table 2. Summary of measures to clean data and number of removed cases

description	removed cases	N
Original N		4752
Time_RSI > 2	418	4334
< 18 years old	1	4333
Attention Check AD	245	4088
Attention Check HR	208	3880
Duration < 120	0	3880
Duration > 120 & < 240	9	3871
Straightliners	52	3819
TIME_RSI > 1.75 & < 2	67	3752
Double Turkers	4	3748
Nonsense Samples	3	3745

5 TWO-SIDED WELCH TWO-SAMPLE T-TEST

Participants rated the inferences gender (*mean* AD=2.66, *mean* HR=3.82; t(3513.1)=-18.536; P<0.001; 95% *CI*: (-1.28, -1.04); d=0.62), skin color (*mean* AD=2.88, mean HR=4.19; t(3513.1)=-18.536; P<0.001; 95% *CI*: (-1.44, -1.17); d=0.61), emotion expression (*mean* AD=2.97, *mean* HR=3.62; t(3654.7)=-11.079; P<0.001; 95% *CI*: (-0.75, -0.52); d=0.36), and wearing glasses (*mean* AD=2.03, *mean* HR=3.16; t(3147.2)=-18.082; P<0.001; 95% *CI*: (-1.26, -1.01); d=0.59) significantly more positively in the low-stake advertisement than in the high-stake hiring scenario.

Subjects rejected inferences intelligent, trustworthy, assertive, and likable regardless of the decision context: The inference ratings for intelligent (*mean* AD=5.25, *mean* HR=5.34; t(3662.2)=-1.425; P=1; 95% *CI*: (-0.21, 0.03); d=0.05), trustworthy (*mean* AD=5.29, *mean* HR=5.18; t(3637.5) = 1.685; P=0.74; 95% *CI*: (-0.02, 0.23); d=0.06), and likable (*mean* AD=5.04, *mean* HR=5.16; t(3695.7)=-2.059; P=0.32; 95% *CI*: (-0.24, -0.006); d=0.06) did not show a significant difference between the two scenarios. Only ratings for the inference assertive (*mean* AD=4.69, *mean* HR=4.89; t(3668.3) = -3.219; P=0.01; 95% *CI*: (-0.32, -0.078); d=0.11) were significantly different between the two scenarios, but the effect was negligible.

6 EXPLORATORY FACTOR ANALYSIS (EFA)

Prior to the computation of the exploratory factor analysis (EFA), several assumptions were tested.

6.1 Assumptions

Missing Data for Inference Ratings. Missing values appeared to be random and were less than 2% per variable (max. n=71 for the variable *assertive*, accounting for 1.9%; min n=31 for the variable *wearing glasses*, accounting for 0.83%). For EFA, all samples with missing values for the inference ratings were removed (in total 208). The sample size was reduced to 3537.

Normality and Linearity. Table 3 lists statistics for each of the dependent inference variables, including skewness and kurtosis. The deviations from normal skewness and kurtosis are within an acceptable range. Additionally, given the large sample size, the impact of departures from normal skewness and kurtosis is negligible.

Table 3. Statistics for each dependent variable

	mean	sd	median	trimmed	skew	kurtosis	se
gender	3.26	1.96	3.00	3.07	0.68	-0.80	0.03
emotion expression	3.30	1.80	3.00	3.16	0.67	-0.64	0.03
wearing glasses	2.59	2.00	2.00	2.26	1.13	-0.12	0.03
skin color	3.53	2.25	3.00	3.41	0.46	-1.36	0.04
intelligent	5.32	1.92	6.00	5.58	-0.95	-0.46	0.03
trustworthy	5.25	1.93	6.00	5.52	-0.95	-0.44	0.03
assertive	4.80	1.88	5.00	4.94	-0.46	-1.06	0.03
likable	5.12	1.85	6.00	5.33	-0.73	-0.72	0.03

Absence of Multicollinearity and Singularity. None of the correlation coefficients displayed in Fig. 2 of the main article are greater than .8. This suggested there is no multicollinearity or singularity. Additionally, the determinant of the R-matrix was 0.031 and greater than the heuristic of 0.00001. [2, p. 771]

Factorability of the Correlation Matrix. The correlation coefficient matrix in Fig. 2 of the main article displayed several correlations above .3. An alternative measure is the Kaiser-Mayer-Olkin (KMO) measure of sampling adequacy [6]. A factor analysis is said to yield reliable and distinct factors, if values are close to 1, which suggests that correlation patterns are relatively compact [2, p. 769]. We used the KMO criteria based on [5]. The KMO values for all inference ratings were above .71 and fell within the range of middling values. The overall MSA value was .82, falling in the range of meritorious values [4, 6].

6.2 Number of Factors

Given the result from the parallel analysis and scree plot in Fig. 3 and other criteria such as the Velicer's MAP test, Very Simple Structure test of complexity 1, and Kaiser's criterion, first a two-factor solution was computed and compared to the results of a three-factor solution and a four-factor solution.

6.3 Test Specifications

It was reasonable to assume that the constructs underlying the measured dependent variables correlated, because we measured the agreement to inferences made from the facial region. Therefore, we first applied oblimin as oblique

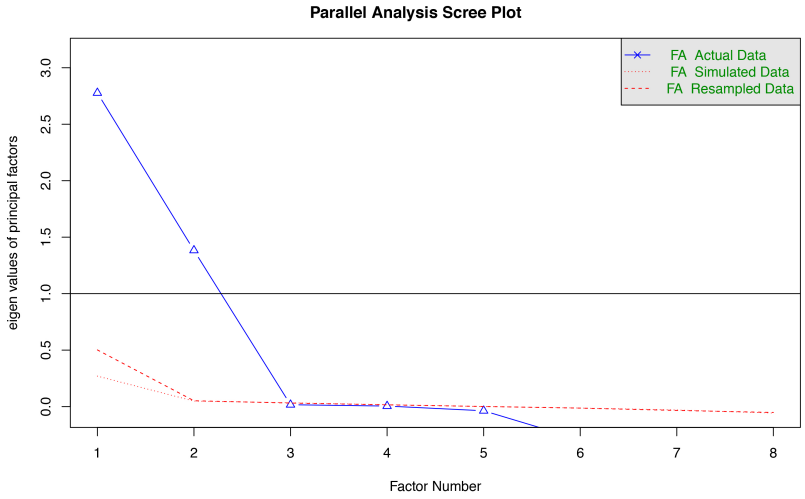

Fig. 3. Graphical analysis for the number of factors using parallel analysis scree plot.

rotation and estimated factor scores using tenBerge for preserving correlations. Supporting this decision, [1, 2] points out that in practice there are many reasons to believe that orthogonal rotation is not appropriate for data involving people, because any construct of psychological nature is correlated in some way with another psychological construct. However, for two factors, oblique rotation resulted in two factors with no correlation. This indicates that the two factors were independent. For correlations of factors below 0.32, [7] suggest orthogonal rotation. Therefore, we applied varimax for orthogonal rotation. Minimum residual (minres) was retained as factoring method, because multivariate normality does not have to be assumed [8]. Factor scores were estimated using regression. To compute the exploratory factor analysis, the R psych package and the GPArotation package were used.

6.4 Factor analysis model with 2 factors

Fig. 4 a) displays the structure of the factor analysis with two factors and indicates the rounded loadings. MR1 represents the first factor labeled *second-order inferences* and MR2 the second factor labeled *first-order inferences*. Fig. 4 b) is a graphical representation of the item's grouping based on their loadings on both of the factors.

There were no residuals > 0.05. The root-mean-square residual was 0.014. The residuals appeared to be approximately normally distributed. Regarding the factor scores, no outliers were identified.

We validated the results by randomly splitting the data in half and running the factor analysis on both subsets. This procedure was repeated three times. For each validation procedure, both factor analyses on the two subsets of the data set resulted in the variables having the same patterns of the factor loadings as with the complete sample. Additionally, the communalities were similar. This validated the factor solution previously obtained on the full dataset.

Both sub-scales had high reliability, the overall α is 0.89 for the factor labeled *second-order inferences* and 0.77 for the factor labeled *first-order inferences*.

Table 4 displays all solutions with two, three and four factors.

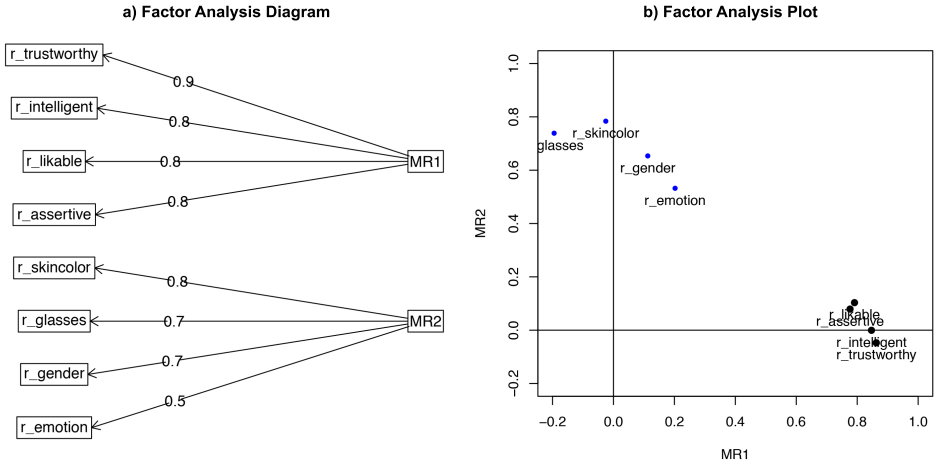

Fig. 4. Summary of two-factor solution with factor diagram and factor plots.

Table 4. Overview of Exploratory Factor Analysis Solutions with 2, 3 and 4 Factors.

	Two Factors		Three Factors			Four Factors			
	MR1	MR2	MR1	MR2	MR3	MR1	MR2	MR3	MR4
gender	0.11	**0.65**	0.14	**0.65**	0.01	0.07	**0.66**	-0.01	0.09
emotion expression	0.20	**0.53**	0.08	0.09	**0.62**	0.01	-0.00	**1.00**	-0.00
wearing glasses	-0.19	**0.74**	-0.21	**0.60**	0.17	-0.19	**0.67**	0.07	0.01
skin color	-0.03	**0.78**	0.01	**0.83**	-0.03	0.06	**0.82**	-0.01	-0.05
intelligent	**0.85**	-0.00	**0.87**	0.05	-0.08	**0.86**	0.01	-0.02	0.00
trustworthy	**0.86**	-0.05	**0.87**	-0.04	-0.03	**0.87**	-0.05	0.00	-0.00
assertive	**0.78**	0.08	**0.75**	-0.04	0.14	0.01	-0.00	-0.00	**0.99**
likable	**0.79**	0.10	**0.77**	0.03	0.08	**0.73**	0.06	0.05	0.06
eigenvalues	2.78	1.89	2.73	1.48	0.45	2.07	1.57	1.00	1.00
proportion variance	0.35	0.24	0.34	0.17	0.06	0.26	0.20	0.13	0.13
cumulative variance	0.35	0.58	0.34	0.53	0.58	0.26	0.46	0.58	0.71
α	0.89	0.77	0.89	–	0.76	0.87	0.76	–	–

6.5 Factor analysis for 3 and 4 factor solutions

The factor analyses with three and four factors resulted in one and two factors with only one indicator variable respectively (see Table 4). This is opposed to the general idea of a factor analysis identifying latent constructs by forming factors out of a combination of at least two variables [3]. Additionally, for the three-factor solution, the cumulative variance was equal to the cumulative variance for a two-factor solution. The third factor had an eigenvalue of < 1. The composition of the three factors was not robust when computing the factor analysis on randomly sampled subsets of the complete data. While the cumulative variance explained by a factor analysis for four factors was the greatest among all tested factor analysis models, this solution was also not robust. Running the factor analysis on two randomly sampled subsets resulted in different patterns of the loadings on the factors. Altering the random sampling produced different patterns of loadings once again.

Although the fit based upon off diagonal values equaled 1 in each of the models, the solutions with three and four factors were neither appropriate in terms of variables per factor nor robust across subsets of the data. Hence, exploratory factor analysis of the eight items measured in this study revealed that two factors were sufficient to explain the underlying structure of common inferences from faces.

6.6 Distribution of EFA factor scores and original ratings

The global means for all variables that load on the first factor and all variables that load on the second factor are highlighted by the horizontal lines in Fig. 5 a) and b). The bold lines in panels a) and b) indicate the means for the individual groups. By using the factor scores as dependent variables for further analysis, the interpretation of the dependent variables depicted in panels c) and d) changes compared to the original inference ratings. A factor score of approximately 0 indicates that a participant's mean rating of all variables that load on this factor is close to the global mean of these variables (horizontal lines in panels a) and b)). A negative factor score indicates this subject gave lower than average ratings. A factor score close to 1 indicates that the subject's ratings for the variables loading on this specific factor are about one standard deviation above the average rating.

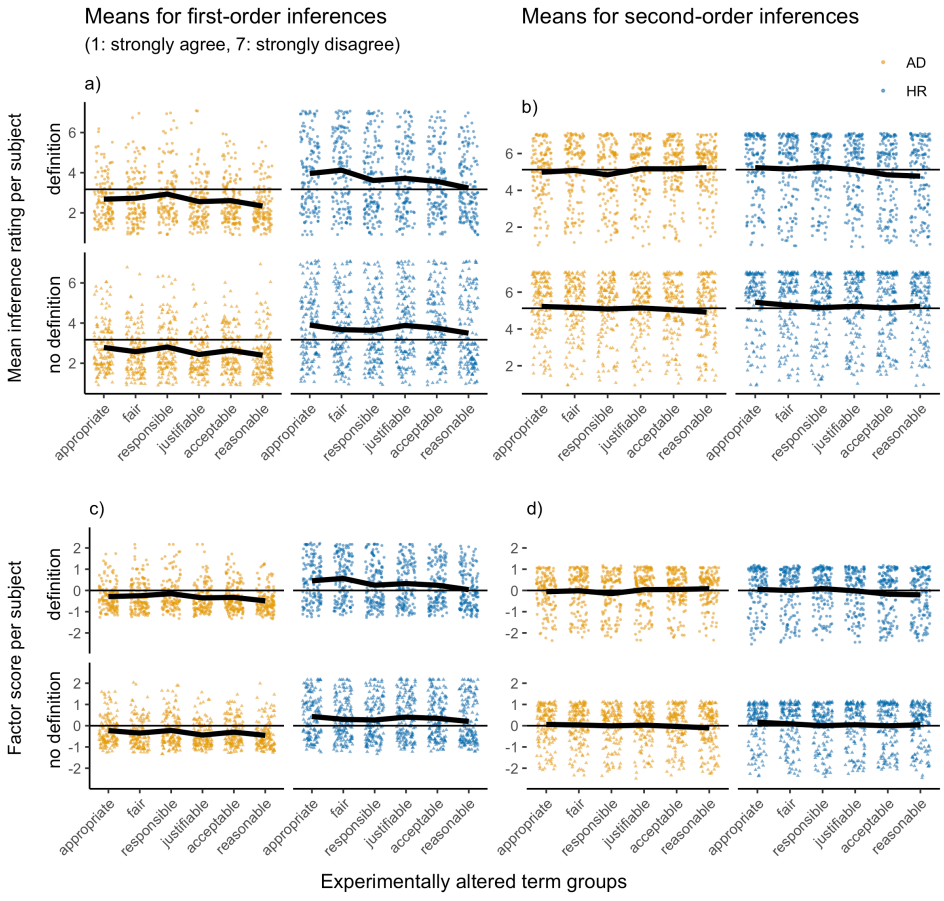

Fig. 5. Distribution of participants' ratings and distribution of the factor scores extracted from the exploratory factor analysis.

7 MANOVA

We performed a multi-factorial MANOVA to statistically test the differences in group means. The two factors identified by performing exploratory factor analysis served as dependent variables. We included three experimentally altered independent variables (context, adjective terms, definition), all measured control variables (AI knowledge, gender, age, education and occupation) and the main justification types for first-order and second-order inferences from the classification. All predictors were included as categorical variables. For the MANOVA and ANOVA analysis, the R car package was used.

7.1 Assumption tests and fitting the model

Assumption tests prior to fitting the model

Although the exploratory factor analysis produced uncorrelated factor scores, we first computed a MANOVA to obtain an overview of patterns between first-order ratings and second-order ratings as dependent variables. Given the lack of correlation and thus no further information from the correlation structure of the dependent variables, we expected a diffused structure of results. Running the MANOVA based on factor scores from the factor analysis with oblique rotation did not change the results. Nine further cases with missing data, i.e., no justification provided for their ratings, were additionally removed.

The following assumptions were tested prior to computing the MANOVA. **Adequate Sample Size**. We applied the one-in-ten-rule for adequate sample size. Our sample size of 3,528 with at least 133 subjects per group based on the experimentally altered independent variables exceeded the threshold of 100 subjects (ten times the number of independent variables: Context, Adjective Terms, Definition, AI Knowledge, Age, Gender, Education, Occupation, Main Justification First-Order, Main Justification Second-Order). **Independent Observations**. Given the randomization, all observations were independent. **Outliers Based on Raw Data**. Neither univariate extreme outliers based on the boxplot method with observations being three interquartile ranges far from the first or third quartile nor multivariate outliers based on Mahalanobis distance were identified. **No Multicollinearity**. There was no multicollinearity.

Model Fitting 1: Testing for Interaction Effects

To test the other assumptions based on residual analysis, we fitted a model with interaction terms first. There were no significant interaction effects. All partial η^2 were calculated using the etasq function from the R heplots package.

Model Fitting 2: Residual Analyses

Because none of the interaction effects were significant at α =0.01, they were removed and a new model without interaction effects was fitted. Residual analyses were conducted on the linear model of this MANOVA.

The following assumptions were tested after fitting the MANOVA. **Linearity of Data**. The residuals vs. fitted values plot indicates that the linearity assumption is met. The line is approximately horizontal at zero. **Homogeneity of Variances of Residuals**. The spread-location plot shows that the residuals have an equal variance above and below the line, which is approximately horizontal across the plot. This indicates that the spread of the residuals is approximately equal at all fitted values and that the assumption of homoscedasticity is satisfied. **Normality of Residuals**. The histogram of residuals indicates that the residuals are approximately normally distributed. However, in the Q-Q plot of residuals, the points in the lower left and upper right corner of the plot deviate somewhat from the reference line. A further analysis of outliers and influential cases could help identify cases that might cause the deviations.

Observations having extreme residuals (> 3.5, < -3.5), extreme Cook's Distance values (> 0.0056), extreme hat values (> 0.062, < -0.062), or extreme dffits values (> 0.5, < -0.5) were identified and inspected. These thresholds are based on graphical analysis and are all less strict than common thresholds such as the > 2(p+1)/n for hat values (with p being the number of predictors and n the sample size). Model results for the removal of varying sets of outliers and influential cases were compared. Finally, 36 cases having either extreme residuals (> 3.5, < -3.5) or extreme Cook's Distance values (> 0.0057) were removed. Removing more of the previously identified cases did not improve the results.

Model Fitting 3: Final Multivariate Assumption Check

Table 5 presents the output for the model after removing the identified 36 cases. Significant effects are highlighted in bold. The panels in Fig. 6 indicate that linearity of data, homogeneity of variances of residuals as well as normality of residuals are now met.

Table 5. Final MANOVA without interaction effects and with outliers and influential cases removed

	Df	test stat	approx F	num Df	den Df	Pr(>F)	Bonferroni	partial η^2
(Intercept)	1	0.01	21.43	2	3445	0.000	0.000	0.012
first-order justification	6	0.50	190.76	12	6892	0.000	**0.000**	**0.249**
second-order justification	6	0.45	164.60	12	6892	0.000	**0.000**	**0.223**
AI knowledge	4	0.03	13.43	8	6892	0.000	**0.000**	**0.015**
age	5	0.01	4.50	10	6892	0.000	0.000	0.006
gender	2	0.00	1.97	4	6892	0.097	1.000	0.001
occupation	8	0.01	2.38	16	6892	0.001	0.016	0.006
education	7	0.00	0.94	14	6892	0.519	1.000	0.002
context	1	0.04	73.68	2	3445	0.000	**0.000**	**0.041**
terms	5	0.01	3.58	10	6892	0.000	0.001	0.005
definition	1	0.00	0.61	2	3445	0.543	1.000	0.000

Comparison of final model with model based on an equalized dataset

The results of the final model from Table 5 were compared to the results of a model for an equalized dataset based on the three experimentally altered independent variables (context, adjective terms, definition). The same outliers and influential cases as in the previous model were removed. After equalization, this dataset contained 3,168 subjects. Because the assumptions based on the graphical analysis did not differ and the results were similar to the previous results of Table 5, this model was discarded in favor of retaining more observations in a sample without equalized groups.

7.2 Follow-up analysis

To identify which individual predictors had a significant effect on which dependent variable, we conducted univariate analyses.

Univariate Analysis: ANOVA for First-Order Dependent Variable

Graphical analysis served to test the model assumptions. While the assumptions of normality and linearity seemed to be approximately met, heterogeneity of variances was questionable. However, the removal of 13 identified extreme

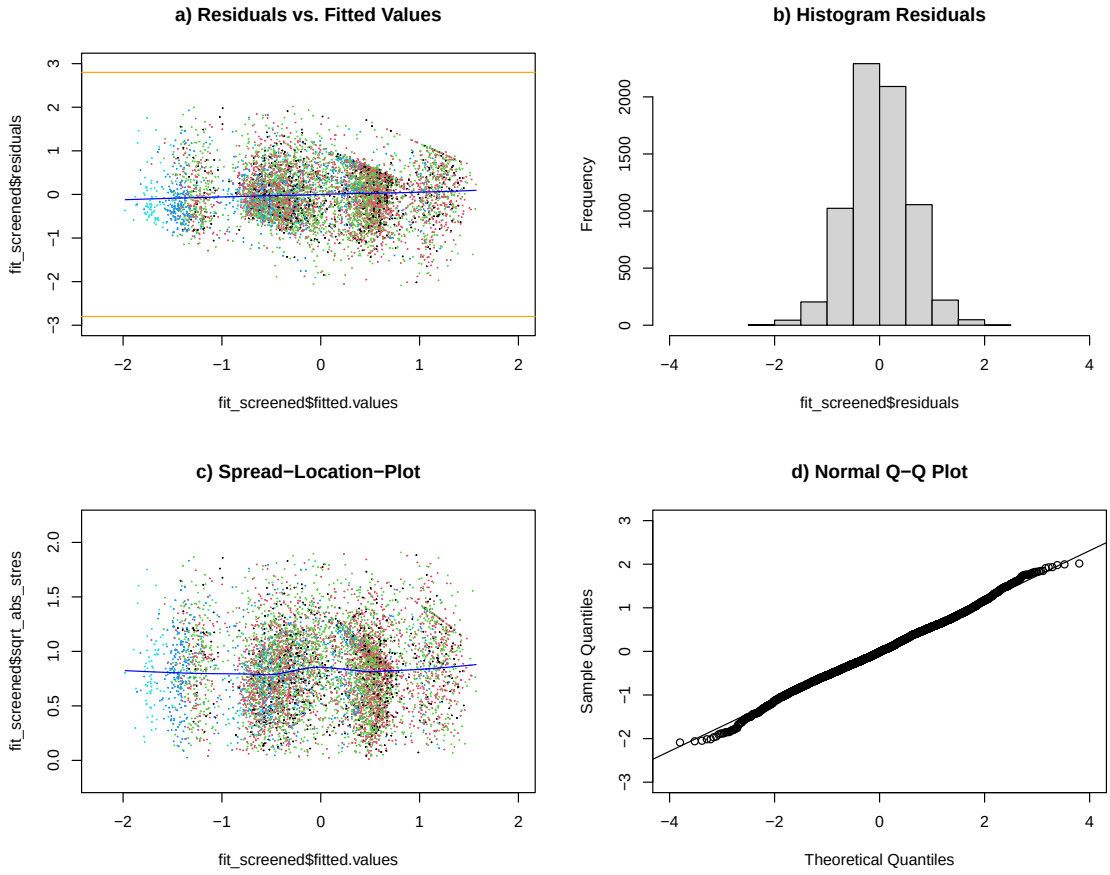

Fig. 6. Graphical analysis of MANOVA test assumptions after removing 36 identified cases.

outliers and influential cases did not improve the homogeneity of variances. To control for the family-wise error rate, we applied a Bonferroni correction to adjust the P values for multiple comparisons of a multiway ANOVA. Additionally, the P values were compared to a Bonferroni-corrected α-level = 0.005 (= 0.01/2) for two ANOVAs.

Univariate Analysis: ANOVA for Second-Order Dependent Variable

Graphical analysis served to test the model assumptions. While the assumptions of normality and linearity seemed to be approximately met, heterogeneity of variances was questionable. However, the removal of twelve extreme outliers and influential cases did not improve homogeneity of variances. As we did for the ANOVA for the first-order dependent variable, we applied a Bonferroni correction to adjust the P values for multiple comparisons of a multiway ANOVA. In addition, the P values were compared to a Bonferroni-corrected α-level = 0.005 (= 0.01/2) for two ANOVAs.

7.3 Pairwise comparisons

For first-order inferences, pairwise comparisons for the variable *adjective terms* and the significant experimental variable *context* based on estimated marginal means revealed significant group differences between the advertisement and the hiring context at each level of the variable *adjective terms* (see Table 6, rows 1-6). These differences could not be observed for second-order inferences. All groups differed significantly between first-order and second-order inferences (see Table 6, rows 7-18). These results are in line with the rating behavior depicted in Fig. 5 and the ANOVA results (see Appendix 7.2 and for ANOVA outputs Table 1 of the main text), i.e., the assignment to a context, either advertisement or hiring, had a significant effect on the rating behaviors of participants for first-order inferences. Also, the rating behaviors on first- and second-order inferences within one context differed significantly.

Table 6. All significant pairwise tests for context and adjective terms based on estimated marginal means for the complete model

terms	variety	context	contrast	estimate	SE	df	t.ratio	p.value
acceptable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
appropriate	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
fair	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
justifiable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
reasonable	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
responsible	factor1st	.	HR - AD	0.26	0.02	3454.00	12.11	0.00
acceptable	.	AD	factor2nd - factor1st	-0.55	0.11	3454.00	-5.08	0.00
acceptable	.	HR	factor2nd - factor1st	-0.75	0.11	3454.00	-6.89	0.00
appropriate	.	AD	factor2nd - factor1st	-0.54	0.11	3454.00	-5.06	0.00
appropriate	.	HR	factor2nd - factor1st	-0.74	0.11	3454.00	-6.88	0.00
fair	.	AD	factor2nd - factor1st	-0.64	0.11	3454.00	-5.87	0.00
fair	.	HR	factor2nd - factor1st	-0.84	0.11	3454.00	-7.67	0.00
justifiable	.	AD	factor2nd - factor1st	-0.55	0.11	3454.00	-5.01	0.00
justifiable	.	HR	factor2nd - factor1st	-0.75	0.11	3454.00	-6.81	0.00
reasonable	.	AD	factor2nd - factor1st	-0.58	0.11	3454.00	-5.30	0.00
reasonable	.	HR	factor2nd - factor1st	-0.78	0.11	3454.00	-7.12	0.00
responsible	.	AD	factor2nd - factor1st	-0.71	0.11	3454.00	-6.55	0.00
responsible	.	HR	factor2nd - factor1st	-0.91	0.11	3454.00	-8.36	0.00

The influence of the *justification* variables becomes apparent when computing estimated marginal means for a model without the *justification* variables. When controlling for the *justifications*, the effect of the variable *context* decreases. Nevertheless, the same significant differences of main interest are identified between the AD and HR context.

8 SUBJECTS' JUSTIFICATIONS

8.1 Documentation of category classes and F1 scores

Table 7. Generated category classes for participants' justifications, together with example comments of classified observations per class and test set F-1 score for each class.

	Category classes	Examples	F1 score
1	AI can tell	"You should be able to determine the race of a person with a picture of their face."	0.94
2	AI cannot tell	"You can not tell if a person is likable or not in a photo."	0.96
3	Inference is relevant for the decision making	"Some positions require emotion, or at least sympathy or empathy."	0.96
4	Inference is not relevant for the decision making	"it does not matter if a person is black or white when the AI is recommending products and services"	0.95
5	Inference creates harm (e.g., illegal, discrimination).	"This is unacceptable, as it may be discriminatory against the transgender population."	0.97
6	AI has human biases	"Artificial intelligence is no less susceptible to bias than humans are. Especially considering that humans pick the training data and that affects how AI forms it's models.."	0.97
7	Incomprehensible & nonsensical responses	"this person is not fully trustworthy", "Not very like"	0.95

8.2 Categories

Table 8 defines all categories, provides application descriptions, and differentiates the category to related ones. More examples comments are provided.

Table 8. Definition of categories and examples (Code book).

Category	Description	Example
AI can tell (e.g. "easy to tell")	**Definition:** The AI/software is able to/can make an inference because the portrait image provides sufficient evidence for the inference. Alternatively, the data basis on which the AI was trained and/or the data used for the analysis in the given context and/or the physical nature of the trait to be inferred are suitable/good/sufficient for the AI to make the inference. **Application:** The category is assigned when someone *agrees* that an AI is able to make the inference based on sufficient evidence. Sometimes a *specific reference* to the photograph, portrait, image, picture, or visual data type is made. The word "obvious" can be an indicator to use this category.	*Very easy to tell. All you need is a picture and a database.* (P635/2575) *Can always tell this from a color pic.* (P1329/4565) *AI can determine this easily. It can see if you wear glasses or not.* (P557/2327) *Also extremely obvious and superficial.* (P1257/4338)
AI cannot tell (e.g. "not easy to tell")	**Definition:** The AI/software is not able to/cannot make an inference because the evidence in the portrait image is insufficient for the inference. Alternatively, the data basis on which the AI was trained and/or the data used for the analysis in the given context and/or the physical nature of the trait to be inferred are not suitable/good/sufficient for the AI to make the inference. **Application:** The category is assigned when someone *disagrees* that an AI is able to make the inference. In some cases, it is *specifically highlighted* that a facial image or visual data type is not correct/insufficient to make a certain inference.	*AI cannot determine whether a person is trustworthy or not.* (P333/1605) *Intelligence is not a physical trait and cannot be determined from a photograph by an AI.* (P220/1207) *You cannot determine whether someone is intelligent based on the way that they look.* (P1362/4610)
Inference is relevant for the decision making	**Definition:** The inference is relevant/important and/or useful for the purpose of application. **Application:** This category is assigned if someone explains why/that a certain inference is relevant for making a decision for a specific application.	*[...] this piece of information is needed for better predictions.* (P260/1339) *[...] I think having emotions is a crucial part of an interview.* (P3515/5661)
Inference is not relevant for the decision making	**Definition:** The inference is not relevant/important/appropriate and/or not useful for the purpose of the application. **Application:** This category is assigned if someone explains why/that a certain inference is not relevant for making a decision for a specific application.	*It does not matter whether a person is assertive or not.* (P46/550) *A sex does not define a person.* (P1109/3856)

Inference creates harm (e.g. illegal, discrimination)	**Definition:** An AI inference is considered discriminatory and/or violates personal rights. **Application:** This category is assigned when drawing an inference would lead to a discriminatory outcome or harm a person in any other way.	*this form of racism should be unacceptable. you cannot infer such a thing on skin color alone.* (610/2491) *Trying to determine a user's personality and trustworthiness is a pretty massive breach of privacy.* (P133/894)
AI has human bias	**Definition:** Inference is affected by human bias; the inference cannot be made without human bias. **Application:** This category is assigned if someone highlights the dependency of AI on humans and hence the implicit integration of human bias, for example, into the data and ultimately into the decision made by an AI.	*I do not see how an AI could make such a determination without relying on human biases to be programmed into it. [...]* (P1862/1966) *Artificial intelligence is no less susceptible to bias than humans are. Especially considering that humans pick the training data and that affects how AI forms it's models.* (P1708/1272)
Incomprehensible responses	**Definition:** The comment is unrelated to the task and/or contains text copied from the instructions or nonsensical text. **Application:** This category is assigned if the comment is not a justification for the rating. Additionally, this category is applied if it becomes apparent from the comments that a participant did not understand the task. If one comment of a respondent can clearly be assigned to this category, all comments by this same respondent have to be assigned to this category, because it cannot be assumed that the person trustfully filled out the questionnaire.	*ok a so like in* (P1419/4830) *they are intelligent* (P607/2486) *I agree that person is or is not wearing glasses. because it is useful to portrait a person.* (P928/3352)

8.3 Justifications results for the "Glasses" inference

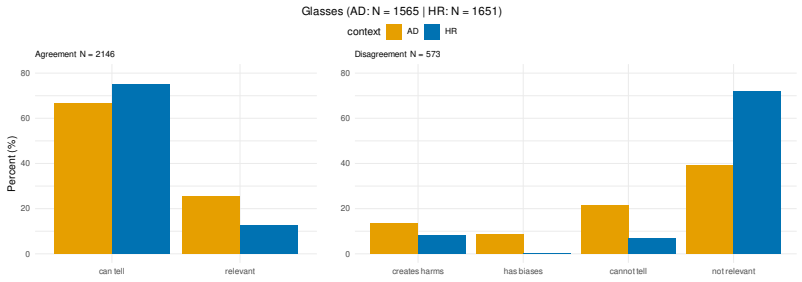

Fig. 7. Justifications results for the "Glasses" inference.

REFERENCES

[1] Anna B. Costello and Jason Osborne. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation* 10, 1 (2005), 7. https://doi.org/10.7275/jyj1-4868

[2] Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering statistics using R.* Sage.

[3] Robin K. Henson and J. Kyle Roberts. 2006. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement* 66, 3 (2006), 393–416. https://doi.org/10.1177/0013164405282485

[4] Matt C. Howard. 2016. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction* 32, 1 (2016), 51–62. https://doi.org/10.1080/10447318.2015.1087664

[5] Graeme D. Hutcheson and Nick Sofroniou. 1999. *The multivariate social scientist: Introductory statistics using generalized linear models.* Sage.

[6] Henry F. Kaiser. 1970. A second generation little jiffy. *Psychometrika* 35, 4 (1970), 401–415. https://doi.org/10.1007/BF02291817

[7] Barbara G. Tabachnick and Linda S. Fidell. 2013. *Using multivariate statistics.* Pearson Education.

[8] Conrad Zygmont and Mario R. Smith. 2014. Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *The Quantitative Methods for Psychology* 10, 1 (2014), 40–55. https://doi.org/10.20982/tqmp.10.1.p040