Pre-Registration report for Experimental Psychology Lab Study 2019

Research questions

When looking at other people, humans are able to process perceived faces and encode those information in order to form an opinion, or assumption, about the other's personality. Basically, there could be two different forms of processing faces: *featural processing* and *configural processing.*

Featural processing takes only the facial features (e.g. mouth, nose, eyes) into account, while configural processing takes the configuration of those features into account as well (e.g. eyes over nose over mouth).

We want to replicate the study described in the paper "Configural processing and social judgments: Face inversion particularly disrupts inferences of human-relevant traits" by John Paul Wilson, Steven G. Young, Nicholas O. Rule and Kurt Hugenberg from 2018. Our goal is to disrupt the possible configural processing of faces by turning them upside down and see if this has an effect on the perception of different character traits.

All our data and code can be found in our GitHub repository:
https://github.com/annassommer/XPLabFinal

Hypotheses

1. We predict that "trustworthiness" as a character trait is more often appointed to humans, in comparison with animals, than dominance. This hypothesis is based on the findings of Wilson et al. (2018) in Study 4a of their paper.

2. We hypothesize that character traits, which are perceived as being exclusively human, namely "trustworthiness", according to hypothesis 1, rely primarily on the configural processing of faces. Character traits appointed to both humans and animals, namely "dominance", according to hypothesis 1, are processed featurally.

Under the assumption that trustworthiness is attributed stronger to humans (higher ratings in part one of the experiment), we predict that the perception of the "trustworthiness" of faces differs more significantly between upside down faces and upright faces than the perception of "dominance". Upright faces are perceived as more trustworthy than inverted faces, because turning a face around disrupts possible *configural processing*. On the other hand, the perception of "dominance" should not vary much between upright and inverted faces, as the featural processing of faces is maintained across both conditions. This hypothesis is based on the findings of Wilson et al. (2018) in Study 1 of their paper.

Study plan

The study was a web-based experiment which participants could do wherever they chose and without supervision. It was divided into two parts: In the first part, we wanted to take a look at how a certain character trait is classified as being specifically human rather than being equally attributable to humans and non-humans. Specifically we looked at the ratings of trustworthiness ("Vertrauenswürdigkeit") and dominance ("Dominanz"). To mitigate the ratings influencing each other we asked for ratings of several more traits i.e. confidence, aggressiveness, patience,

By: Anna Sophia Sommer, Daniel Pietschke, Esther Chevalier, Fabienne Kock, Sarah Greb

conscientiousness, friendliness, extroversion, politeness, sensibility ("Selbstsicherheit", "Aggressivität", "Geduld", "Gewissenhaftigkeit", "Freundlichkeit", "Extrovertiertheit", "Höflichkeit", "Sensibilität"). This part required a within-subject analysis.

In the second part, we explored the effect inversion has on the judgment of faces regarding two character traits: trustworthiness and dominance. This part required a between-subject analysis.

Participant recruitment

We started recruiting participants as soon as we finish the pre-registration. The experiment was accessible through a web link for one week, from the 16th of July 2019, 21:00 CEST until the 23th of July 2019, 21:00 CEST. We recruited our participants through the e-mail list of the study course "Cognitive Science" at the University of Osnabrück. Furthermore, we recruited participants through social media (What's App, Telegram, Facebook, Reddit).

The study was conducted in German because most of the people that we are going to reach are German native speakers.

Data set

We used the first 36 black and 36 white male face images from the Chicago face database (version 2.0.3) as material plus the character traits "trustworthiness" and "dominance" with some more as fillers for the first part of the study. We resized the pictures to 450 × 316 pixels. We only used male faces because the original study only used male faces as well, in order to avoid social bias playing a role in the participants' judgment.

The face database is freely available here: https://chicagofaces.org/default/

Study design

The experiment consisted of two parts.
The first part started with an introduction slide, that informed the participants about their upcoming task. The explanatory text reads as follows:

"In diesem Teil des Experiments sehen Sie ein Persönlichkeitsmerkmal.
Es ist Ihre Aufgabe auf einer Skala von 1 bis 9 zu bewerten, inwieweit das Merkmal nur auf Menschen oder gleichermaßen auf Menschen und Tiere zutrifft. 1 steht hier für 'Trifft gleichermaßen auf Menschen und Tiere zu' und 9 steht hier für 'Trifft nur auf Menschen zu'.
Wenn Sie bereit sind, drücken Sie bitte auf Start:"

in English: In this part of the experiment you will see one personality trait. Your task is to rate this trait on a scale from 1 to 9 regarding its assignment to humans exclusively or humans and animals equally. 1 stands for "exclusively human" and 9 stands for "equally human and animal". Press the "start" button once you're ready.

After the introduction slide the actual test phase started. The participants viewed randomly one out of ten character traits ("Selbstsicherheit", "Aggressivität", "Geduld", "Gewissenhaftigkeit", "Freundlichkeit", "Extrovertiertheit", "Höflichkeit", "Sensibilität", "Vertrauenswürdigkeit", "Dominanz") at a time. All ten character traits were shown to all participants. Each character trait was rated by the participant on a 9-point rating scale with labeled end points ranging from 1 ("Trifft gleichermaßen auf Menschen und Tiere zu" – "Equally applicable on humans and animals") to 9 ("Trifft nur auf Menschen zu" – "Only applicable on humans").

By: Anna Sophia Sommer, Daniel Pietschke, Esther Chevalier, Fabienne Kock, Sarah Greb

The second part of the experiment started again with an introduction slide, that informed the participants about their upcoming task. This explanation text of the second part reads as follows:

"In diesem Teil des Experiments werden Ihnen Fotos von Gesichtern gezeigt. Es ist Ihre Aufgabe auf einer Skala von 1-7 zu bewerten, inwieweit das Persönlichkeitsmerkmal "Dominanz" (or if applicable "Vertrauenswürdigkeit") auf das Gesicht zutrifft.
1 steht hier für 'Gar nicht dominant (or if applicable vertrauenswürdig)',
7 steht für 'Sehr dominant (or if applicable vertrauenswürdig)'.
Sie sehen zunächst auf den Kopf gestellte Fotos, und dann aufrechte. /
Sie sehen zunächst aufrechte Fotos, und dann auf den Kopf gestellte.

Wenn Sie bereit sind, drücken Sie bitte auf Start:"

In English: In this part of the experiment you will see pictures of faces. Your task is to rate them for "Trustworthiness" *(or "Dominance", which trait is to be rated is decided by coin flip)* on a scale from 1 to 7. 1 stands for "Not trustworthy" *("not dominant" respectively),* 7 stands for "very trustworthy" (*"very dominant" respectively*).

You will firstly see faces which have been turned upside down and upright faces after that. / You will firstly see upright faces and faces which have been turned upside down after that.

(It is decided by coin flip which sentence is displayed. Pictures appear as announced, either turned around and then upright (First version of the sentence) or upright and then turned around (second version of the sentence))

The test phase of the second part consisted of two blocks (counterbalanced). In the first block participants saw 72 upright or inverted facial stimuli one after another, that have been rated on "dominance" or "trustworthiness" (decided by coin flip beforehand and did not change during experiment). The rating took place on a 7-point rating scale ranging from 1 ("not trustworthy/dominant") to 7 ("very trustworthy/dominant") with labeled end points. The second block is identical to the first besides that the facial stimuliwere shown in an inverted (or upright depending on the first block) manner.

The exact wording in German was "Wie dominant/vertrauenswürdig ist dieses Gesicht?" for the question and for the scales "Gar nicht dominant/vertrauenswürdig" (1) and "Sehr dominant/vertrauenswürdig" (7).
In both parts of the experiment, the next trial started immediately after participants selected their answer.

After the second part, participants were asked to give their age, gender and mother tongue and were offered the chance to leave some further comments. The experiment ended with a screen saying "Vielen Dank, dass sie an dem Experiment teilgenommen haben." (in English: "Thank you for participating in this experiment.").

Measured variables

For the first part we collected the ratings of the traits. The independent variable were the different character traits and the dependent variable the rating of the given trait. The character traits are scaled nominal and their ratings are of an ordinal scale.

In the second part we collected the ratings of character traits being ascribed to either upright or inverted pictures. Here the independent variables were the currently rated character trait (either

By: Anna Sophia Sommer, Daniel Pietschke, Esther Chevalier, Fabienne Kock, Sarah Greb

trustworthiness or dominance), the photos and the condition of the photos (either upright or inverted). The dependent variable were again the rating. The independent variables are of nominal scale and the dependent variable of an ordinal scale.

Moreover we kept track of the reaction time for every stimulus, meaning the ratings of the character traits in the first part as well as the ratings of the faces in the second part of the study. Furthermore we collected the age and gender of all participants. The response time will be treated as an interval scale in milliseconds, age as an interval scale and gender as a nominal scale.

Exclusion criteria

We excluded data from participants who gave the same answer to every trial and data from participants who answered in the pattern (1, 2, 3, 4, 5, 6, 7) for at least 3 times. As for reaction times, we excluded data from participants if their reaction times are outside the mean reaction time +/- 3 times the standard deviation interval.

Statistical analysis

We used R for statistical analysis with the packages tidyverse, effsize, readR, brms, desktools and Rmisc.

For the first part of the study, we firstly filtered out the necessary data (which is the rating of dominance and trustworthiness). After that, we visualized the data using boxplots and built a regression model which models the responses of participants against the two character traits per participant. We needed to represent a participant dependency because it is a within subject design.

To judge if our hypothesis 1 (trustworthiness is a more human-like trait and dominance a trait equal to humans and animals) was correct, we needed the 95% confidence interval for the not baseline trait to not contain 0. We had dominance as a baseline - we therefore needed a positive estimate for trustworthiness. If that is the case, we accept the hypothesis.

For the second part, we created 3 different regression models. We have no participant dependence for it is a between participant design.
The models are :
  – Model 1 which models the responses against the categories dominance and trustworthiness
  – Model 2 models the responses against the categories and the picture format (inverted and upright)
  – Model 3 models the responses against the categories and the category per picture format (category|pictureFormat)

Once we had all three models, we used the *loo* and *loo_compare* function of R to determine the model fitting the data best.

To judge for acceptance or rejection of the hypothesis 2 based on the model, we take a look at the best model determined by *loo_compare*. Then, we checked the 95% confidence interval for "category". For the hypothesis to be accepted the confidence interval must not contain 0.

By: Anna Sophia Sommer, Daniel Pietschke, Esther Chevalier, Fabienne Kock, Sarah Greb

Participants

According to our exclusion criteria described in the preregistration of this study, we excluded 6 participants. As a result, we analyzed the data of 79 participants. There were 27 male (average$_{age}$ = 28.9 years, median$_{age}$ = 24, SD = 14.5), 51 female (average$_{age}$ = 26.2 years, median$_{age}$ = 21 years, SD = 10.9) and 1 diverse (age = 102) subjects. Participants had the chance to leave comments at the end of the study. One participant mentioned they got uncomfortable rating faces (assuming wrongly the faces had been mirrored). Three other participants mentioned that they found the task to be tiring and too long. One participant reported issues about the loading of the stimuli.

Results

In the first part of our study, we focused on the perception of the character traits "dominance" and "trustworthiness". We expected the first to be significantly being more attributed to both humans and non-humans than the latter. During our analysis, we took the character trait "dominance" as reference category. The analysis can be found as HTML file in the experiment's [GitHub repository]. Our model compared the ratings of "dominance" and "trustworthiness" only and ignored the fillers mentioned in our study design. The 95% confidence interval (CI) of the model is [0.85; 2.28] and does not contain 0. Therefore, we accept our model as significant. The point estimate is 2.00. Since our reference category is dominance and the point estimate is positive, we conclude that the character trait "trustworthiness" was significantly rated higher than was "dominance". We infer that "trustworthiness" is more often appointed to specifically humans than "dominance". Therefore, we accept our first hypothesis.

In the second part of our study, we tested the perception of trustworthiness and dominance in upright and upside down male faces. We predicted the ratings of trustworthiness to be significantly lower in upside down than upright faces. In the same way, we expected the ratings of dominance to be similar in both conditions. As described in the statistical analysis section of this paper, we generated three models by using the *loo* function in R. We compared the models using *loo_compare* and found model 2 to be the most accurate model. The population-level effects table of model 2 shows [0.39; 0.53] as 95%-CI. The CI is strictly positive, we therefore accept the model as being significant. Again, we defined dominance as the reference category. According to our hypothesis, the point estimate of trustworthiness should be negative as we predicted it to be more influenced by the picture format (upside down or upright). The point estimate of "trustworthiness" in model 2 is 0.46 and consequently, we had to reject our hypothesis.

The second part of our analysis does not only show that "trustworthiness" was in fact rated higher across conditions. It also reveals [-0.14; -0.00] as the 95%-CI of the model for the picture format. Hence, there are no significant differences in ratings between upright and upside down faces independently of the character trait condition.

Discussion

We replicated successfully the findings of study 4a from the paper Wilson et al. (2018) i.e. "trustworthiness" is more uniquely applicable to humans than is "dominance". We used it as a premise of our second hypothesis. Despite our true premise, our results do not support the findings of study 1 in Wilson et al. (2018). The disruption of configural processing in faces did not have a significant impact on a character trait likely to be perceived as uniquely human i.e. "trustworthiness".

There were some issues that could have an impact on the result, both in Wilson et al. (2018) and the study we carried out. Some participants noted anonymously that the great number of pictures (72

By: Anna Sophia Sommer, Daniel Pietschke, Esther Chevalier, Fabienne Kock, Sarah Greb

pictures divided in two blocks) was tiring. There is a high chance some participants got less concentrated towards the end of study which may lead to less accurate results. In addition, Wilson et al. (2018) mentioned that the "*p*-values [as in Study 2 of Wilson et al. are] very near the criterion for significance". Accordingly they decided to replicate the study to get a bigger participants and to ground their results. Using Mechanical Turk, they had the opportunity to test 150 participants for the equivalent of the second part of our study. In contrast, our study was carried out with only 80 participants. The effect tested being very small, a greater number of participants might have large consequences on the result. For this reason, we hope that future replication of this particular finding will be done.

Reference

Wilson, John & Young, Steven & O Rule, Nicholas & Hugenberg, Kurt. (2018). Configural Processing and Social Judgments: Face Inversion Particularly Disrupts Inferences of Human-Relevant Traits. *Journal of Experimental Social Psychology. 74.* 1-7. doi:10.1016/j.jesp.2017.07.007.

By: Anna Sophia Sommer, Daniel Pietschke, Esther Chevalier, Fabienne Kock, Sarah Greb