Gender Bias in Trustworthiness

Lina Elkjær Pedersen (LP) | 201905535
Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark.

Thea Pedersen (TP) | 201904704
Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark.

Characters in total: 40,942
Standard pages: 17.07

Cognitive Science
Social and Cultural Dynamics in Cognition
Aarhus University

Thea Pedersen (TP) | 201904704 27/05 - 2021
Lina Elkjær Pedersen (LP) | 201905535

Abstract

Gender inequality has been widely debated for centuries. Today gender differences in safety in the public space is a highly discussed topic. Most recently an Instagram campaign with the words “Text me when you get home” reignited the debate. This debate was the main inspiration for the present study which investigates how the gender and the facial expression of a stranger can influence their perceived trustworthiness in different social settings.

It was found that both female and male participants trusted *female* facial stimuli more than male facial stimuli in scenarios related to safety. In addition to this, it was found that a *joyful* face compared to a neutral face increased trustworthiness in scenarios related to safety.

Other results remained inconclusive, however, they indicated that both female and male participants might trust *male* facial stimuli more than female facial stimuli in relation to capabilities and that *neutral* faces compared to joyful faces could increase trustworthiness in scenarios related to capabilities.

Earlier research on the subject is conflicting. Thus, the findings of this experiment support the results from some earlier studies. However, since results remain conflicting, more research needs to be conducted on the subject in order to make generalizable conclusions.

Introduction

Inequality is a major issue in modern day society. Even though there have been many advancements in terms of neutralizing inequality, it is still present in many different aspects of our society today. Inequality is experienced in relation to aspects such as gender, race, and wealth, and affects many people all over the world on a day to day basis. Therefore, it is no wonder that inequality is such a widely debated subject and that it has been for so many years.

Focusing on gender inequality, for instance, the first women's rights convention in the US was held in 1848 (History, 2019). It took decades of activism before this eventually led to granting women voting rights in 1920 (History, 2010). Today, a century later we have made vast improvements, yet, it remains a well-known fact that men take up more professional positions of power and earn higher wages for the same jobs (Evans, 2017).

Premise and Motivation

(TP) Gender inequality is not only present in a professional context, it is also especially apparent in relation to safety. In an analysis conducted by World Health Organization (WHO) in 2018, it was found that across 161 countries 30% of women have experienced physical and/or sexual violence (WHO, 2021). The issue of women's safety was further displayed earlier this year, through an Instagram message reading "Text me when you get home". The message went viral and sparked debate after a young British woman disappeared as she was walking home at night. Many people, particularly women, shared and commented on this post on social media. One person commented on the fact that women in modern day society are taught to follow certain rules in order to try to stay safe:

> *"Sarah Everard kept to all the 'rules' that society has set out for women to stay safe and she still wasn't able to walk home safely" (Macfarlane, para. 27)*

(LP) It seems clear that there are some social dynamics in modern day society that make women believe they have to take precautions to be safe in the world. This was the motivation for the present study. We believe that there could be looked more into these dynamics and the effects they possibly have on our society. Therefore, we want to explore how this feeling of being unsafe affects people when interacting with strangers in a public space. More precisely, we want to clarify what the gender and the demeanor of the stranger means for our perception

of this person. Thus, the present study aims to investigate how the gender and the facial expression of a person influence their perceived trustworthiness in different scenarios.

Theoretical Background

Trust and its Implications

(TP) Trust is essentially the firm belief in the reliability, truth, or ability of someone or something (Oxford Dictionary, 2016). Moreover, trust is a natural part of being human. This has much to do with the way that we, and especially our brains, are structured. Because of the large size of our fully developed brains, human babies are born physically premature and defenseless. This means that our physical bodies and brains still need years to develop after birth. If human babies were to be born fully formed to take on the world, the head would simply be so large that it would be impossible for women to contain and deliver the baby. Thus, human babies are born very unable to survive on their own with a high dependency on nurturers. Because of this need to rely on other human beings from the minute we are born, it is in our nature to form social relationships. As social beings we are born to engage and communicate with others, and this is a huge part of what trust is all about (Kramer, 2009).

(LP) The ability to trust has been an important link in our evolution as a species. Scientists believe that the bond between nurturer and child, the cooperation, and other social engagements are all critical parts of what made the human brain develop the way it has. Our success as a species can in some ways be explained by the fact that we are social beings at our very core (Kramer, 2009). In the modern world, social trust is said to contribute positively to a wide range of different phenomena, including economic growth, social integration, cooperation and harmony (Newton, 2009).

There are two broad schools of thought when it comes to the understanding of trust. The first views trust as an individual trait. It considers trust to be connected with characteristics of the individual such as personality, or social and demographic aspects such as class, income, age, and gender. The second school of thought believes that social trust is a property of social systems (Newton, 2009). In this study, we attempt to combine these two schools of thought. We believe that social trust is affected by the social systems present in our society. However, we also believe that the level of trust is highly influenced by individual traits such as gender. We

thus investigate trust as seen by the individual, in a social context, in order to reveal structures of trust in society.

(TP) Trust can be approached from many different angles. In this study, we have chosen to look at two specific sides of trust. Trust in relation to safety and trust in relation to capabilities. When examining trust in relation to safety, we examine whether people perceive strangers to have good intentions when interacting with them. In relation to capabilities, we will look at trust in a perspective where a person's competence and skills are mainly in focus.

We constructed two scenarios for each of the described categories. The aim when constructing these scenarios was that they should be believable. If they were unrealistic or created a feeling of too extraordinary circumstances, we would risk the participants dismissing or rejecting the scenario as they read it. People should be able to relate to the scenarios and imagine them as realistic and lifelike possibilities (Selin, 2005).

Gender Bias in Trust

(LP) As aforementioned, work related gender equality has yet to be reached. In a study from 2011, Bevelander and Page investigated the way men and women network in their professional careers. Their results suggest that women have the same scale of social networks as men, however, women tend to trust men more than other women in risky professional environments. We suspect that the patriarchal society we live in can have affected our perception of trustworthiness in regards to capabilities, such that we believe men are more capable in professional contexts. If this is the case, then it seems likely that people would consider males more trustworthy than females in scenarios related to capabilities.
In regards to safety, however, we believe that women are perceived to be more trustworthy than men. Supporting this theory, statistics show that males on average are more violent than females. According to the FBI's official website, males comprised 80.4% of all individuals arrested for violent crimes in the US in 2011. In addition to this, 88.2% of individuals arrested for murder in 2011 were males (FBI Federal Bureau of Investigation, 2011). Accordingly, we believe that the perception of men as being more violent and potentially dangerous, as compared to women, will have an impact on trust in scenarios where safety is in focus.

(TP) Other empirical evidence suggests that a within-gender bias is present (Bonein & Serra, 2009). This means that we tend to trust people of our own gender more than people of other

genders. However, we believe that people will be more influenced by the dynamics of our society. Thus, despite some evidence towards a within-gender bias we expect that people will trust men more than women in scenarios related to capabilities and that people will trust women more than men in safety related scenarios.

Facial Expressions and their Relation to Trust

(LP) In an article from 2015, Ruben et al. found that when it comes to smiling in job interviews less is more. Two studies were performed. In the first study they wanted to see how smiling affected hiring for a newspaper reporting job. Participants were randomly assigned either the role as applicant or interviewer. In this study they found that smiling had a negative effect on hiring, meaning the applicants who smiled more were less likely to be deemed suitable for hiring. In the second study the researchers wanted to see whether the effect was constant across different job types. In order to do so, each video from the first study was randomly assigned one of four job types: newspaper reporter, middle manager, elementary school teacher, or salesperson. Participants were told the job type and asked to watch the interview videos. In this study, participants still deemed applicants who smiled less more suitable for the jobs, however, the size of the effect greatly depended upon the job type; jobs perceived as more serious, showed a larger effect. The findings suggest that people who smile less are more likely to be taken seriously in terms of jobs and capabilities. Based on this we assume that neutral facial expressions will be trusted more than joyful facial expressions in scenarios related to capabilities.

(TP) It has been found that smiling can induce trust. In a study from 2010, Ozono et al. concluded that participants trust smiling faces more than non-smiling faces. In addition to this, Beaupré and Hess (2003) found that smiling individuals were more likely to be judged as an in-group than an out-group member. This distinction between in-group and out-group is associated with a feeling of safety (APA Dictionary of Psychology, 2014). Thus, we believe that these results indicate that people consider smiling people as less of a threat and therefore trust them more in regards to their safety. In a series of experiments Belkin and Rothman (2017) demonstrated that emotional valence affects perception of sociability, morality and competence. In addition to this, their results strongly indicated that expressions of happiness are conducive to safety related trust. Based on these findings, we assume that joyful faces will increase trustworthiness in scenarios related to safety. Contradicting the finding by Ruben et

al. (2015) described in the section above, Belkin and Rothman (2017) also found that expressions of happiness increased the perceived level of competence. Thus according to this, we should expect to find that smiling will increase trustworthiness both in capabilities and in safety scenarios. In spite of this, we have chosen to proceed with the assumption that neutral facial expressions will increase perceived trustworthiness in the capabilities condition.

Hypotheses (LP, TP)

As aforementioned the aim of this study is to investigate the effect of gender and facial expression on perceived trustworthiness. In order to do so, an experiment was conducted. Based on the empirical evidence presented above, it is expected that the gender and facial expression of a person presented in the experiment will have an effect on perceived trustworthiness, more specifically:

- Both female and male participants will trust *male* facial stimuli more in relation to capabilities.
- Both female and male participants will trust *female* facial stimuli more in relation to safety.
- A neutral face compared to a happy face will increase trustworthiness in the *capabilities* condition.
- A happy face compared to a neutral face will increase trustworthiness in the *safety* condition.

Methods

(TP) We used a within-participants design to investigate our hypotheses. We examined which factors can influence how the trustworthiness of a stranger is perceived. The factors that we have chosen to focus on are gender of image stimuli, facial expression of the image stimuli, and trustworthiness category of the written scenario displayed above the image stimuli. These factors and their effect on trustworthiness ratings will be examined.

Participants

(LP) The participants included 58 people, 22 males and 36 females. Ages ranged from 15 to 64 years (M = 26.09, SD = 10.81). All participants were native Danish speakers and were encouraged to participate in the experiment through a post on Facebook. The post stated that an

age limit was set for participation. This limit was set to participants between 15 and 70 years of age. This limit was set in order to focus our analysis by excluding children and elderly people in our investigation of trustworthiness between the genders. The distribution of gender of the participants is somewhat balanced. However, there is a majority of female participants.

Stimuli

(TP) In the experiment, a combination of text and visual stimuli was displayed in each trial. The text stimuli created by the researchers offered descriptions of four different scenarios (see example in Figure 1 below). All scenarios were presented in Danish. Two of the scenarios were related to safety and were thus placed in a category of the same name. The other two scenarios related to capabilities and were placed in a category of this name. With this categorization, the interpretation of trustworthiness becomes twofold as we look at trust in relation to feeling of safety and trust in capabilities. The visual stimuli used were AI generated photos of faces (Photos by Generated Photos). The stimuli consisted of 40 photos divided into four different categories. The categories were created according to the gender and facial expression of the facial stimuli. The four categories were: male joyful, male neutral, female joyful, and female neutral. In order to limit unwanted variance within the experiment, the image stimuli were narrowed to photos of young Caucasian adults with a front facing head pose and a neutral background. With the purpose of excluding any confusion regarding the gender of the stimuli, short haired women and long haired men were excluded. The stimuli were categorized by gender for the purpose of the analysis.

For the full set of text and image stimuli see appendix (pp. 3-7)

Figure 1:

Figure 1 illustrates an example of a trial within the experiment. In this example the scenario category is capabilities. The facial stimulus is of the gender category male, and the emotion category neutral. Translation of scenario example: "You need surgery on your knee while undergoing full anesthesia. The person in this picture is the doctor who will be performing the surgery. How likely are you to trust the person in the picture to perform the surgery?"

Procedure

(LP) As aforementioned, the experiment was advertised via a Facebook post. In the post participants were informed about requirements for participation and the general purpose of the experiment. In addition to this, they were asked to sign a consent form. The form could be accessed through a link in the post. In short, the form stated that the data responsible researchers (Thea Pedersen and Lina Elkjær Pedersen) were allowed to analyze the data provided in the experiment and to use the anonymized data for an exam paper. Furthermore, it stated that participation was completely voluntary and that participants could withdraw consent at any moment by contacting either of the researchers. The Facebook post also contained a link to the experiment, which was built in Psychopy Builder (Peirce et al., 2019) and run online at Pavlovia.org (Pavlovia, 2021). Lastly, participants were asked to contact either of the researchers in case any questions arose before or after the experiment.

(TP, LP) When participants clicked on the link to start the experiment, a fullscreen window was opened. Subsequently a pop-up box appeared, and the participants were asked to fill in their initials, age and gender. After this information was given, the participants were shown written instructions to the experiment. When these instructions were read and understood, the participant had to press the spacebar in order to proceed to the experiment. In each trial the participant was presented with a scenario, a facial stimulus, and a rating scale (see Figure 1 above). Participants were instructed to rate their level of trust towards the facial stimulus in relation to the belonging scenario on a scale from 1 (very unlikely to trust) to 8 (very likely to trust). The participant had to choose a rating by clicking a point on the scale (See figure 1). Hereafter, the participant would see a new facial stimulus and scenario displayed on the screen along with a rating scale, and the experiment would continue like this for 40 trials. There was a total of 40 images, 10 of each category (male_joyful, male_neutral, female_joyful, female_neutral). As it was not possible to randomize which image and scenario was presented together, the four different scenarios were distributed as equally as possible across the four categories of images. Thus, each of the four scenarios would appear 10 times, but across participants each scenario would always be attached to the same image stimulus. The trials were randomized in such a manner that each image stimulus was presented only once in a unique random sequence for each participant.
All ratings were recorded and stored in a logfile. When the participant had gone through all 40 trials (or chosen to abort the experiment prematurely) the data was saved, and the experiment ended.

Analysis:

(LP) The data was downloaded from Pavlovia (Pavlovia, 2021). A few files contained only between 1-9% of completed trials. The rest of the datafiles included 70% or more completed trials. The files containing under 70% of data were excluded from the analysis. This was done in order to ensure that there was enough data on each participant to meaningfully include a random intercept by participant ID in the data analysis.
The analysis was run in RStudio (RStudio Team, 2021). In the preprocessing of the data a function was created. The function automatically transformed the participants' initials to anonymous participant ID's. In addition to this, the function created the new column "Scenario_type" which took two different values, "safety" or "capabilities". Scenario 1 and 2

were categorized as “safety” and scenario 3 and 4 were categorized as “capabilities”. A column for gender of stimuli, “Gender_stim”, was also created. This took the two values “male” or “female”, as we had no participants categorizing themselves as other gender types. Furthermore, a column denoting the facial expression of the stimuli was also created. This was called “Emotion” and took the two values “joyful” or “neutral”. All columns that were deemed unnecessary for the analysis were excluded along with all rows containing one or more missing values. Lastly, the function merged all csv files and thus created one large dataframe. After preprocessing was completed, the data was made up of 2300 observations, collected from 58 participants.

(TP) The analysis was performed using a Bayesian statistical method. When building the models, we specified the family function as gaussian in all our models as we, based on previous findings (Jowell, 2003), expect the data to be somewhat normally distributed. It is important to note that the outcome variable, Trust_ratings, was discrete. Therefore, it might have been beneficial to use a cumulative family function instead, however, we were advised against this as it would complicate the analysis without really changing the results. Choosing to make the model assume a gaussian distribution, which is centered around 0, meant that it was beneficial to scale the outcome variable (Trust_ratings) around 0 as well.

Nine models were created using the brm() function from the package “brms” (Bürkner P, 2018). Out of the nine models, three would not converge. The rest of the models were compared using LOO-criterion with the function loo_compare from the package “loo” (Vehtari et al., 2017). Hereby, two final models were chosen. One of which was used to investigate hypothesis 1 and 2. The other was used to investigate hypothesis 3 and 4. Prior and predictive checks were conducted for both models. This was done in order to ensure that the priors were not too constraining and hereby preventing the models from learning from the data. Both models had learned from the data (for plots, see figure 1- 4 in appendix, pp. 1-2), so the analysis proceeded using these two models. All plots were created using either bayesplot (Gabry, 2021) or ggplot2 (Wickham, 2016).

Model 1 (LP)

In order to explore hypothesis 1 and 2, the following model was created:

Trust_rating_scaled ~ 0 + Gender_stim:Scenario_type + (1| ID)

As can be seen from the pseudo-code above; the model contained a scaled version of the trust ratings as the outcome, and a two-way interaction between gender of stimuli and scenario type as a predictor. Thus, it was possible to explore whether the gender of the stimuli had an effect on trust ratings and if it changed based on which type of scenario was presented. In addition to this, the model was run without intercept in order to avoid a baseline estimate. Lastly, the model included random intercepts for participant ID. This was done in order to limit unwanted variance. In order to make the model converge, the number of iterations was increased to 3000 instead of the default being 2000.

Model 2 (TP)

In order to explore hypothesis 3 and 4, the following model was created:

Trust_rating_scaled ~ 0 + Emotion:Scenario_type + (1|ID)

This model contained a scaled version of trust ratings as the outcome variable, and a two-way interaction between emotion and scenario type as a predictor. Hereby, we were able to investigate whether there was an effect of emotion on trust ratings and if it would change based on which type of scenario was presented. The model was run without an intercept. Lastly, the model included a random intercept for participant ID in order to exclude unwanted variance. The number of iterations was increased to 3000 instead of the default of 2000 in order to make the model converge.

Results:

Model 1 (LP)

For Model 1 there was a total number of 6000 post-warmup samples. In addition to this the Rhat was 1.01 for all beta estimates. This suggests good convergence of the chains.

The estimate for Gender_stimfemale:Scenario_typecapabilities (participants judging female facial stimuli, in the capabilities scenarios) is -0.09 *(Estimated Error = 0.07, 95% CI lower boundary = -0.24, 95% CI upper boundary = 0.05)*.

The estimate for Gender_stimmale:Scenario_typecapabilities (participants judging male facial stimuli in the capabilities scenarios) is 0.09 *(Estimated Error = 0.07, 95% CI lower boundary = -0.05, 95% CI upper boundary = 0.24)*.

The estimate for Gender_stimfemale:Scenario_typesafety (participants judging female facial stimuli in the safety scenarios) is 0.26 *(Estimated Error = 0.07, 95% CI lower boundary = 0.12, 95% CI upper boundary = 0.41)*.

The estimate for Gender_stimmale:Scenario_typesafety (participants judging male facial stimuli in the safety scenarios) is -0.24 *(Estimated Error = 0.07, 95% CI lower boundary = -0.39, 95% CI upper boundary = -0.10)*.

Figure 2:

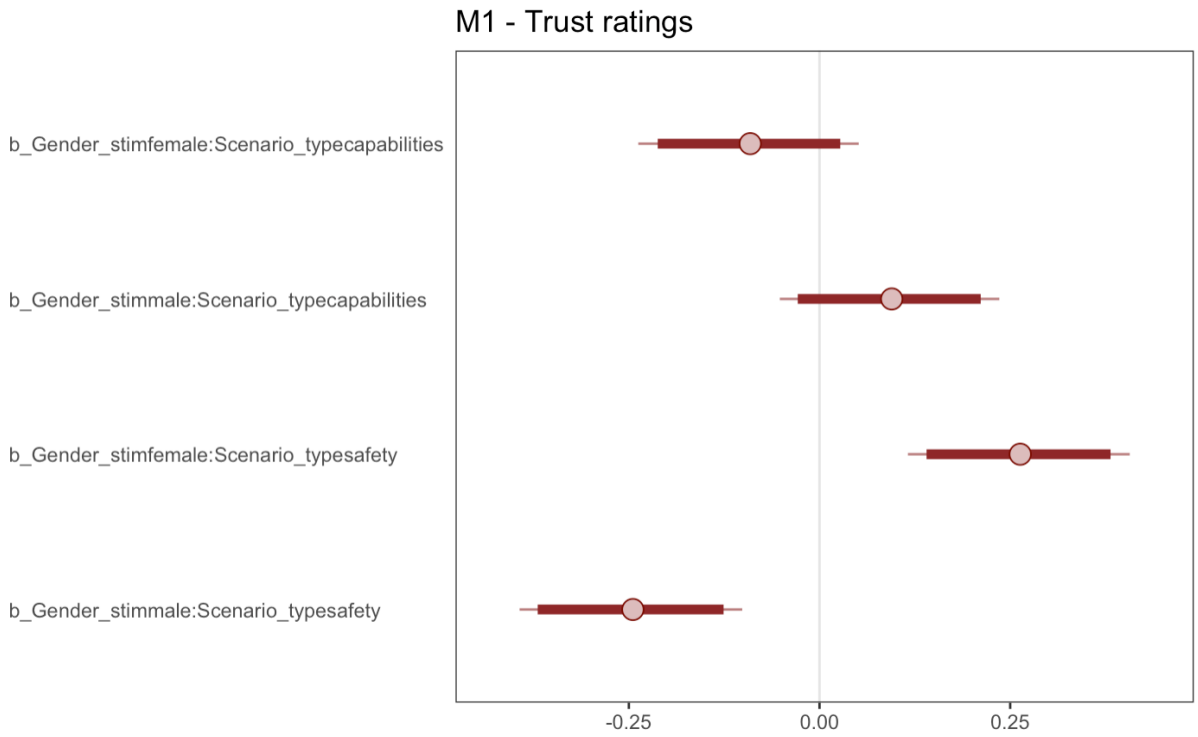


Figure 2 illustrates the estimates of Model 1 and their confidence intervals.

Model 2 (TP)

For Model 2 there were 6000 post-warmup samples. The Rhat is 1.00 for the intercept and all beta estimates. This suggests very good convergence of the chains.

The estimate for Emotionjoyful:Scenario_typecapabilities (participants judging joyful facial stimuli in the capabilities scenarios) is 0.09 *(Estimated Error = 0.07, 95% CI lower boundary = -0.05, 95% CI upper boundary = 0.24)*.

The estimate for Emotionneutral:Scenario_typecapabilities (participants judging neutral facial stimuli in the capabilities scenarios) is -0.06 *(Estimated Error = 0.07, 95% CI lower boundary = -0.20, 95% CI upper boundary = 0.07)*

The estimate for Emotionjoyful:Scenario_typesafety (participants judging joyful facial stimuli in the safety scenarios) is 0.15 *(Estimated Error = 0.07, 95% CI lower boundary = 0.01, 95% CI upper boundary = 0.28)*.

The estimate for Emotionneutral:Scenario_typesafety (participants judging neutral facial stimuli in the safety scenarios) is -0.20 *(Estimated Error = 0.07, 95% CI lower boundary = -0.35, 95% CI upper boundary = -0.06)*.

Figure 3:

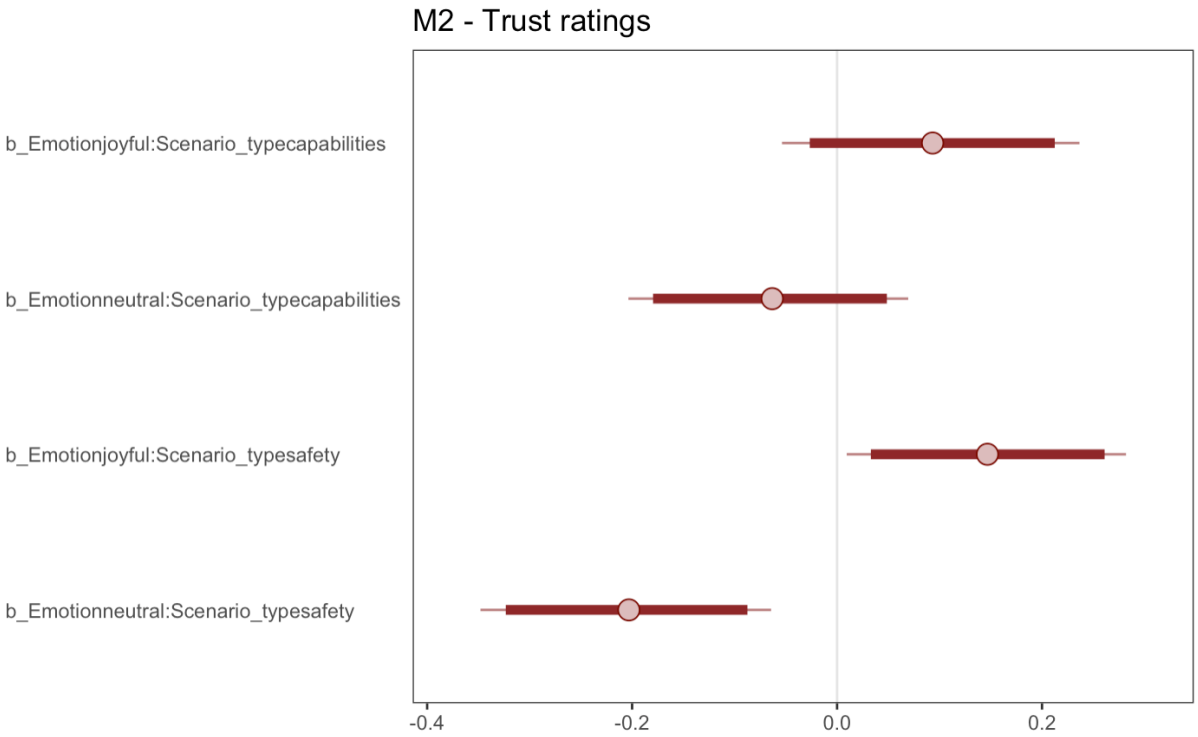


Figure 3 illustrates the estimates of Model 2 and their confidence intervals.

Discussion:

Main Results

(LP) As reported in the results paragraph, the estimate for Gender_stimmale:Scenario_type-capabilities is positive. In comparison, the estimate for Gender_stimfemale:Scenario_type-capabilities is negative. This suggests that male facial stimuli are rated higher on the trustworthiness scale than the female facial stimuli in the capabilities scenarios. However, since the confidence intervals for both estimates span from positive to negative, the model is not confident in determining whether there actually is a difference.
Thus, the results seem to support hypothesis 1, but they remain inconclusive.

(TP) The estimate GenderstimfemaleScenario_typesafety is positive. Since the lower as well as the upper boundaries of the confidence intervals are positive, the model is pretty confident of this result. In comparison, the estimate for Gender_stimmale:Scenario_typesafety is negative. The model is pretty confident of this result, as both the lower and upper boundaries of the confidence intervals are negative. This suggests that there is a positive effect on trust ratings when female facial stimuli are presented in the safety scenarios.
Thus, in line with hypothesis 2; female facial stimuli are rated higher than male facial stimuli in the safety scenarios.

Figure 4

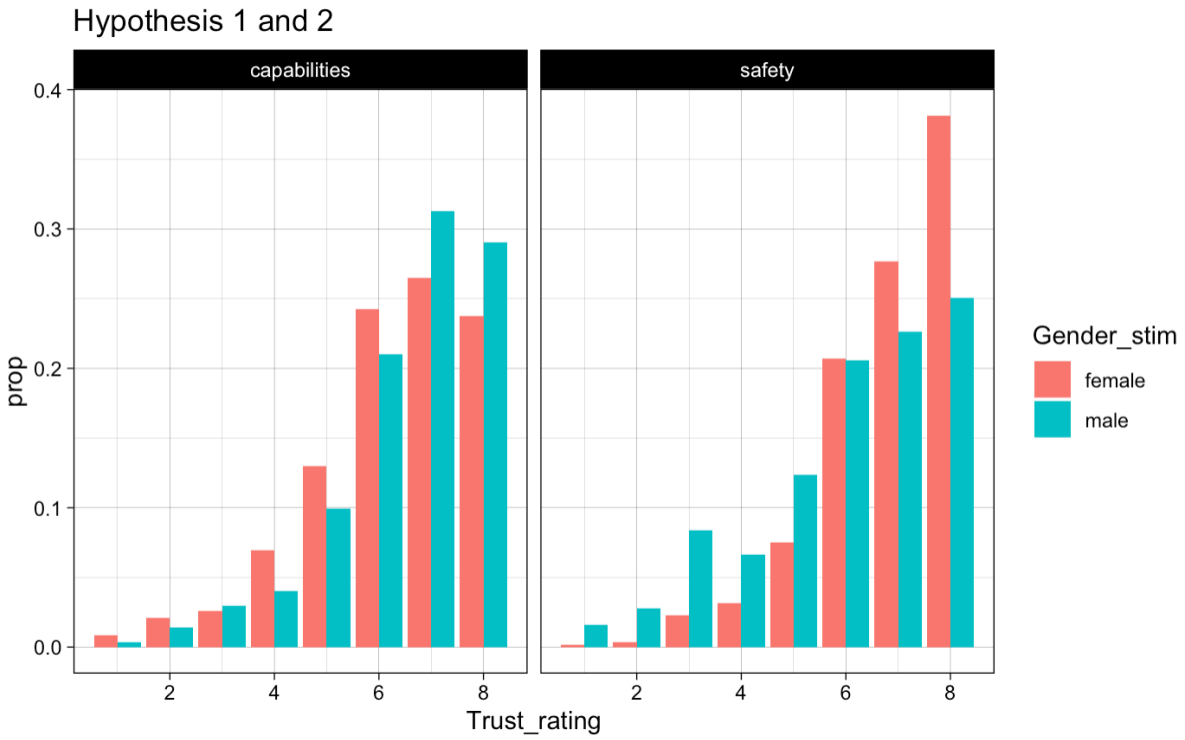


Figure 4 shows two plots. The left plot shows the ratings of trust in the capabilities scenarios. The right plot shows the ratings of trust in the safety scenarios.

(LP) In figure 4, both plots have trust ratings on the x-axis and the number of ratings on the y-axis. Trust ratings span from 1 (very unlikely to trust) to 8 (very likely to trust).
The plot on the left shows that male facial stimuli are rated slightly higher than females in the capabilities condition as consistent with hypothesis 1. However, as reported above, this effect is not conclusive.
Supporting hypothesis 2, the plot on the right shows that female facial stimuli are rated higher than males in the safety condition. According to the statistical analysis, this effect is reliable.

(TP) The estimate for Emotionjoyful:Scenario_typecapabilities is positive. In comparison, the estimate for Emotionneutral:Scenario_typecapabilities is negative. This contradicts hypothesis 3, as it suggests that the neutral facial expressions have a negative effect on trust ratings, as compared to the joyful facial expressions in the capabilities scenarios. However, since the confidence intervals span across both the positive and negative end of the scale, the model is not confident enough to determine whether there actually is a difference.
Thus, the results seem to contradict hypothesis 3, however, they are inconclusive.

(LP) The estimate Emotionjoyful:Scenario_typesafety is positive. As the lower and upper boundaries are both positive, the model is pretty confident of this result. In comparison, the estimate for Gender_stimmale:Scenario_typesafety is negative. Both the lower and the upper boundaries are negative, meaning the model is pretty confident of this result. This suggests that there is a positive effect on trust ratings, when joyful facial expressions are presented in the safety scenarios.

Thus, the results support hypothesis 4; joyful facial stimuli are rated higher than neutral facial stimuli in the safety scenarios.

Figure 5:

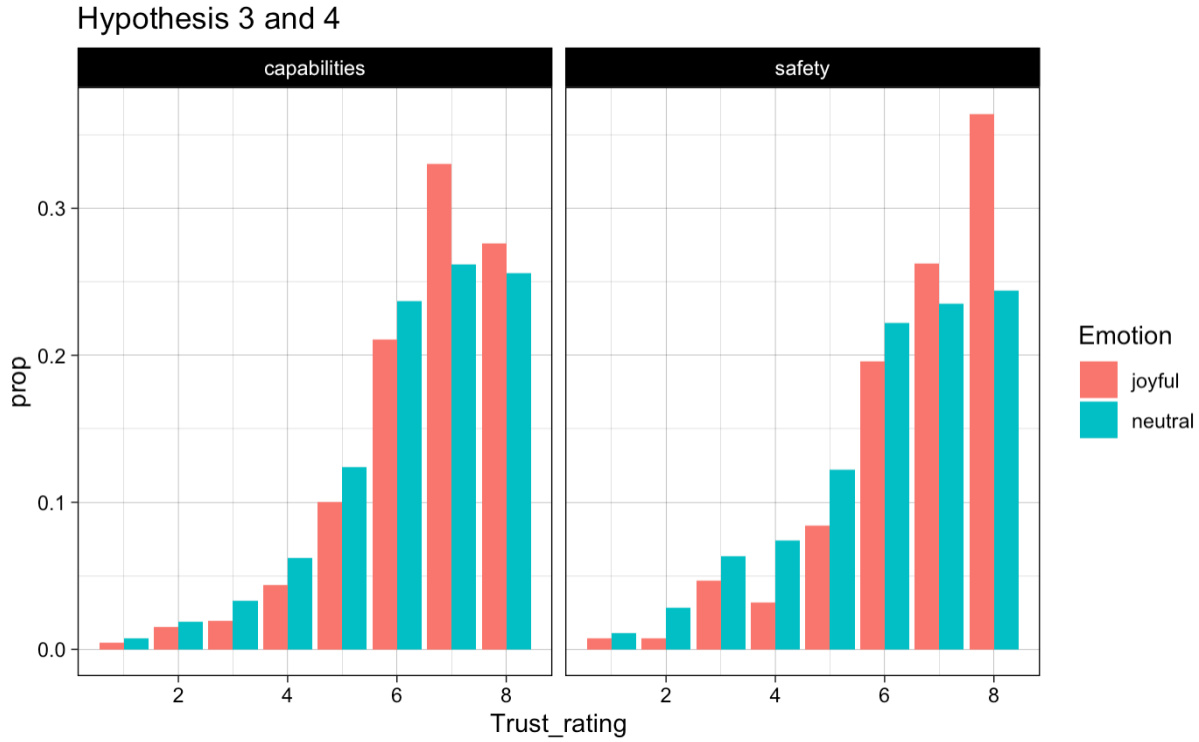


Figure 5 shows trust ratings for the two different emotion categories in the capabilities and safety scenarios.

(TP) The left side plot on figure 5 contains the capabilities scenarios, and the right side plot contains the safety scenarios. As can be seen from the plots, joyful faces receive higher ratings in the capabilities scenarios. This is the opposite effect of what we expected in hypothesis 3. However, this effect was not conclusive according to the statistical analysis.

Joyful faces receive a larger number of high ratings in the safety category, supporting hypothesis 4. As can be seen from the model results, there was a clear positive effect of this. Thus, it seems that the scenario type is not important for the perception of facial expressions in relation to trustworthiness. Accordingly, it seems that joyful facial expressions yield more trustworthiness in general.

Secondary Results

(LP) As mentioned above, no conclusive results were found for hypothesis 3.
In line with the findings by Ruben et al. (2015), we expected that neutral facial expressions would yield higher trust ratings in the capabilities scenarios. Instead, a negative effect was found although this effect was not conclusive. This negative effect can be seen in the plot above; it seems that joyful facial expressions receive higher ratings in both scenario types. This suggests that joyful facial expressions might increase trustworthiness regardless of the type of scenario presented. Thus, our results contradict the findings by Ruben et al. (2015), and instead support the findings by Belkin and Rothman (2017) which suggest that smiling increases trust both in terms of sociability, morality, and competence. However, as mentioned before the effect of joyful facial expressions in the capabilities scenario was not conclusive.

(TP) As stated in the introduction, some empirical evidence has suggested that there is a within-gender bias when it comes to trusting other people. Our results suggest that the trust ratings for male and female stimuli might depend on which type of scenario is presented. While it is still possible that there is an effect of within-gender trust, our results indicate that this effect is overshadowed by the effect of scenario type, particularly in the safety category. However, in order to explore this properly we would have needed to hypothesize this possible effect and then created a model where gender of the participant would interact with gender of the stimuli. This could be explored in future research as we cannot conclude anything about within-gender trust from our specific hypotheses and models.

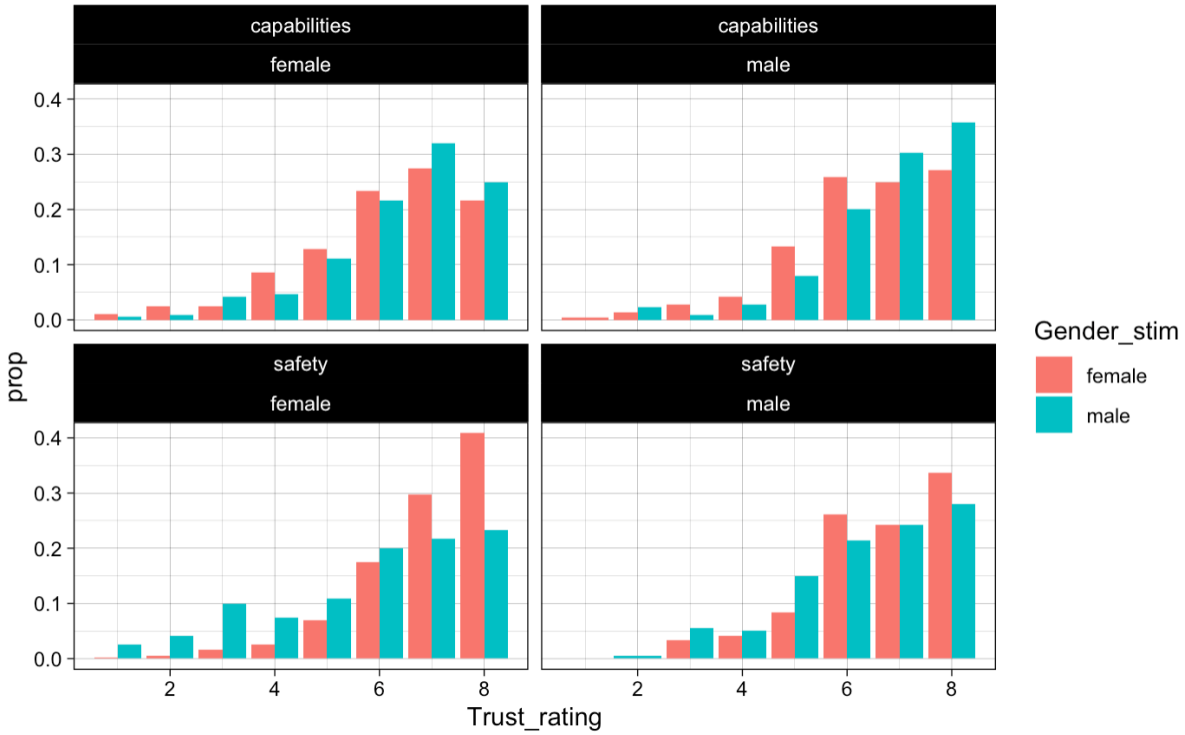
(LP) As a postdoc exploration we wanted to review the effect of gender of the participant in relation to the first two hypotheses. Overall, the plots in figure 6 show that male participants rate female and male facial stimuli somewhat evenly in both scenario types. However, there is a small difference showing that the male participants give female facial stimuli slightly

higher ratings in the safety scenarios whereas they give male facial stimuli slightly higher ratings in the capabilities scenarios.

Female participants rate female facial stimuli higher in the safety scenarios, as opposed to the male facial stimuli, which receive a greater amount of the lower ratings. In the capabilities scenarios, female participants give male facial stimuli a slightly larger amount of high ratings. Taking the "Text me when you get home" debate into consideration it makes a lot of sense that female participants are less trusting towards males in regards to safety.

In order to properly explore whether gender of the participant interacts with gender of the stimuli and scenario type, it would be necessary to build a three-way interaction model and further research this effect.

Figure 6



The plots in figure 6 illustrate the trustworthiness ratings for gender of stimulus in the two scenario types, as seen in relation to the gender of the participant that is giving the ratings.

Limitations and Future Research

(TP) It is important to note that we have created the scenarios ourselves, and that we only have two categories with two scenarios in each. This means that the study is very specific and it is probably too narrow to generalize and establish overall conclusions about trust in modern

day society. A more empirically based and nuanced categorization of scenarios and types of scenarios would be beneficial in future research on the subject. To really understand the implications of trust between genders in society, we would need more scenarios and possibly more subcategories of trust to examine.

Importantly, each of the four specific scenarios that we created could have a different effect on trust ratings. Since the scenarios all represent different situations, it would for example be plausible that scenario 1 seems more unsafe than scenario 2. You might feel more vulnerable and uneasy being approached on the street by a stranger, as compared to a situation where you have an agreement that entails going to a stranger's house to buy a piece of furniture.

Figure 7:

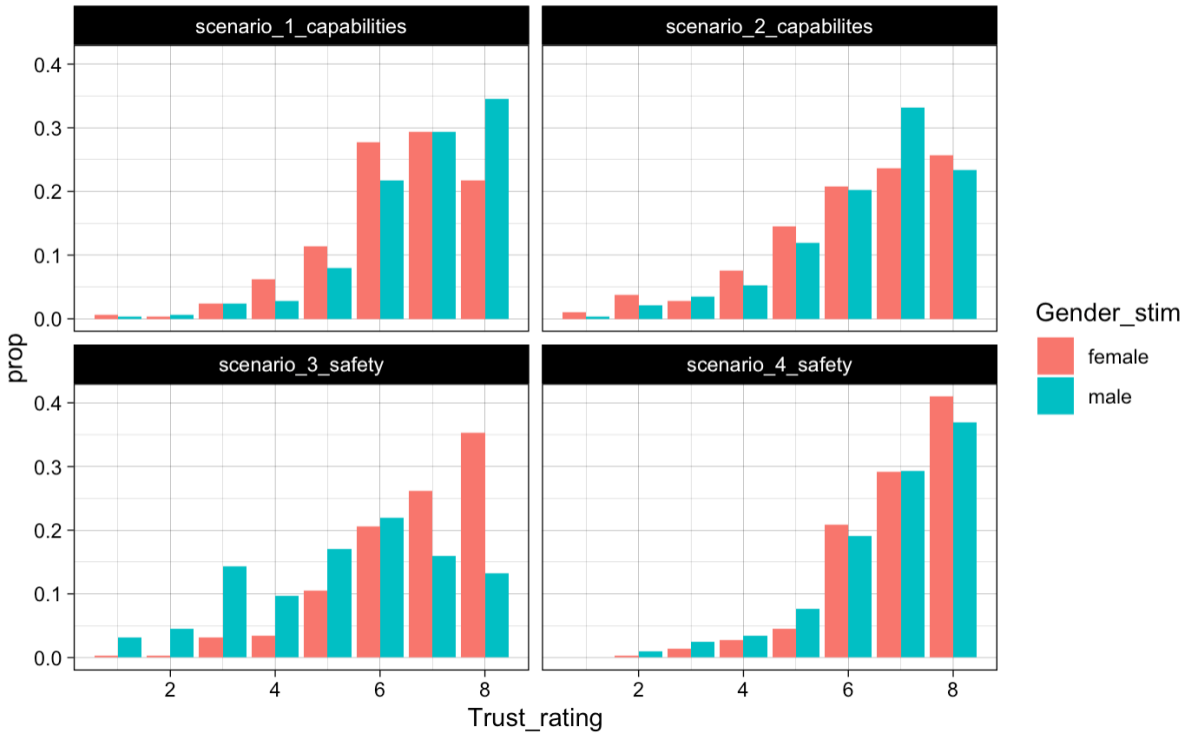


Figure 7 illustrates the effect of gender of stimuli across the specific scenarios.

(LP) Based on the plots in figure 7, it appears that there is a difference between trust ratings based not only on the category of the scenario, but also on the specific scenario presented. Thus, it seems that the scenarios within the two categories are perceived differently from each other. Nonetheless, in order to properly explore a possible effect of the specific scenario type in relation to gender of the stimuli or facial expression, we would need to incorporate this variable in our models.

In further research, it would also be interesting to include a more homogeneous span of age groups in the analysis. We mainly had participants roughly between the ages of 20-30 in this study, even though we had set broad limits for age inclusion. In order to see how trust between genders might be different across generations, we would need to reach more participants across all age groups.

(TP) A limitation of the study is that we were not able to randomize which image stimulus and which scenario were presented together. This was due to issues with uploading the PsychopyBuilder script (Peirce et al., 2019) to Pavlovia (Pavlovia, 2021). Thus initially, we randomized which scenario would appear with which image stimulus. Each of the four scenarios appeared 10 times and thus matched the four categories of images that were also presented 10 times. In sum, this meant that every category was represented evenly across participants. However, since this complete randomization was not possible to upload to Pavlovia (Pavlovia, 2021), we distributed the scenarios as evenly as possible across the four categories of images. Moreover, the order in which the stimulus would appear was randomized for each participant. The consequence of this is that the effects found in this study could be a result of the perception of the individual image stimuli in relation to the specific scenario. Thus, the effect might not be a reflection of the category of image stimuli as a whole.

(LP) Another limitation is that we assumed a Gaussian distribution as a basis for our model even though technically it would have been more correct to use a cumulative approach. In spite of this, we proceeded with the Gaussian family, as we believed the trust ratings would be normally distributed based on a survey from 2003 by European Social Survey (ESS), that reported ratings of trust which were normally distributed (Jowell, 2003). However, during the analysis we observed that the participants of the present study had a clear tendency to rate on the higher end of the trust scale. We suspect that this might relate to the cultural dynamics which are present in Denmark in relation to safety, and thus that it is due to the fact that all participants were Danish. In the survey by ESS the participants were from all over Europe. (TP) In a report from 2012 by the OECD (Organisation for Economic Co-operation and Development), individuals' feeling of safety when walking alone at night in the city or area where they live has been studied across different countries. The report shows that the feeling of being safe when walking alone at night is generally strong in the Nordic countries. More specifically, 83% of Danes feel safe walking alone at night, as compared to 72% as the average for all countries included in the study (OECD, 2014). Moreover, in a paper from 2014,

Mewes found that women from countries with high employment equality, such as Denmark, are more trustful than to other European women.

Conclusion

Trust can be understood in various different ways. The present study sought to investigate trust in relation to safety and capabilities. More specifically, the aim was to examine how gender and facial expressions play a role in our perception of trust across safety and capability scenarios. In order to investigate these parameters, an experiment was conducted. The results from the experiment suggest that females are trusted more than males in safety scenarios. Moreover, the results indicated that males are trusted more in capabilities scenarios, however, this effect was inconclusive. Furthermore, the results suggest that a joyful facial expression will increase trust in a scenario related to safety. In addition to this, the results indicate that a joyful facial expression will increase trust in relation to capabilities, however, this result remains inconclusive.

In conclusion, this study has shown that a joyful demeanor can induce trust, specifically in regards to safety. Moreover, this study shows that trust can be influenced by gender as an individual trait in certain social contexts. This implies dynamics of trust in society where men are possibly perceived as less trustworthy in regards to safety. Such an effect might be due to the fact that men are more often associated with violence and crime, as compared to women. However, more research is needed in order to make decisive conclusions about trust dynamics in society.

References:

APA Dictionary of Psychology. (2014). APA Dictionary of Psychology. Retrieved May 26, 2021, from Apa.org website: https://dictionary.apa.org/ingroup-bias

Beaupré, M., & Hess, U. (2003). In my mind, we all smile: A case of in-group favoritism. Journal of Experimental Social Psychology, 39(4), 371–377. https://doi.org/10.1016/S0022-1031(03)00012-X

Bonein, A., & Serra, D. (2009). Gender pairing bias in trustworthiness. *The Journal of Socio-Economics*, *38*(5), 779–789. https://doi.org/10.1016/j.socec.2009.03.003

Bürkner, P. "Advanced Bayesian Multilevel Modeling with the R Package brms" (2018). *The R Journal*, **10**(1), 395–411. doi: 10.32614/RJ-2018-017.

Evans, M. (2017) *The Persistence of Gender Inequality* (1. ed.). Cambridge, UK: Polity Press

Federal Bureau of Investigation (2011). Table 42. Retrieved May 20, 2021, from FBI website: https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/tables/table-42

Gabry J, Mahr T (2021). "bayesplot: Plotting for Bayesian Models." R package version 1.8.0, https://mc-stan.org/bayesplot/.

Gallery of AI Generated Faces | Generated.photos. (2021). Link to website: https://generated.photos/. Photos retrieved April 28, 2021 from a dropbox link provided by the website holders.

History.com Editors. (2019, February 26). Women's History Milestones: A Timeline. Retrieved May 20, 2021, from HISTORY website: https://www.history.com/topics/womens-history/womens-history-us-timeline

History.com Editors. (2010, March 5). 19th Amendment. Retrieved May 20, 2021, from HISTORY website: https://www.history.com/topics/womens-history/19th-amendment-1

Jowell, R and the Central Coordinating Team, European Social Survey 2002/2003: Technical Report, London: Centre for Comparative Social Surveys, City University (2003). Retrieved 24th of May from: http://essedunet.nsd.uib.no/cms/topics/2/1/1.html#note1

Kramer, R. M (2009). "Rethinking Trust". Harvard Business Review. Retrieved 25th of May from: https://hbr.org/2009/06/rethinking-trust

Macfarlane, J (2021, March 12). *Text me when you get home: meaning of Instagram message that went viral following Sarah Everard disappearance.* Retrieved 23rd of May from Edinburgh News: https://www.edinburghnews.scotsman.com/news/people/text-me-when-you-get-home-meaning-of-instagram-message-that-went-viral-following-sarah-everard-disappearance-3164168

Mewes, J. (2014). Gen(d)eralized Trust: Women, Work, and Trust in Strangers. *European Sociological Review*, *30*(3), 373–386. https://doi.org/10.1093/esr/jcu049

Newton, K (2009). Social and Political Trust. Chapter 1: Social trust and its origin. Retrieved May 24, 2021, from the European Social Survey website: http://essedunet.nsd.uib.no/cms/topics/2/1/

OECD (2014), "Safety and crime", in Society at a Glance 2014: OECD Social Indicators, OECD Publishing, Paris. DOI: https://doi.org/10.1787/soc_glance-2014-30-en

Oxford Dictionary (2016) . TRUST | Definition of TRUST. Retrieved May 25, 2021, from Lexico Dictionaries | English website: https://www.lexico.com/definition/trust

Pavlovia (2021). Retrieved May 25, 2021, from Pavlovia.org website: https://pavlovia.org/

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Ruben, M., Hall, J., & Schmid Mast, M. (2015). Smiling in a Job Interview: When Less Is More. The Journal of Social Psychology, 155(2), 107–126. https://doi.org/10.1080/00224545.2014.972312

Selin, C (2005). Trust and the illusive force of scenarios. https://doi.org/10.1016/j.futures.2005.04.001

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. 27(5), 1413--1432. doi:10.1007/s11222-016-9696-4 (journal version, preprint arXiv:1507.04544).

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

World Health Organization: WHO. (2021, March 9). Violence against women. Retrieved May 20, 2021, from Who.int website: https://www.who.int/news-room/fact-sheets/detail/violence-against-women