Evaluation of User Data acquired in 'When Bias Backfires' (Kuhl & Bush, 2025)

Contents

Introduction

This is an analysis of data acquired in the "When Bias Backfires" study (Kuhl & Bush, 2025) run on Prolific in January 2025.

In this study, our aim is to explore how people make decisions when interacting with biased AI systems. Specifically, aimed to contribute three key insights to the XAI community:

1. Understanding of how AI bias can impact human decision-making through repeated interaction.
2. Examining XAI's role in either facilitating or preventing bias transmission.
3. Highlighting implications for AI systems that support human decision-making while protecting against bias adoption.

In this study, naive users were asked to take on the role of a hiring manager and had to compare candidates for a position in a newly founded department in different phases:

In phase 1 (baseline assessment), participants establish their baseline decision- making patterns by reviewing 20 pairs of candidate profiles without (X)AI as- sistance. For each pair, participants are asked to make a hiring decision for one of the candidates based on the information presented. This phase serves as a crucial baseline measure of any pre-existing biases.

Phase 2 – (X)AI interaction – introduces participants to an (X)AI recruit- ment assistant and represents the core experimental manipulation. Participants review 20 new pairs of candidates, this time with (X)AI recommendations. Par- ticipants in the AI condition receive only the AI's recommendations, while those in the XAI condition receive the same recommendations supple- mented with a CE. In addition to the (X)AI manipulation, the recom- mendations participants encountered were systematically biased either against female or male candidates.

In phase 3 (post-interaction decisions), participants return to making de- cisions without (X)AI assistance, reviewing another set of 20 candidate pairs. This phase is critical to measuring any persistent effects of (X)AI exposure on decision-making patterns.

At the end of each of the decision phases 1-3, participants indicate their confidence in that decision on a 5-point Likert scale ranging from "Not at all confident" to "Extremely confident".

The experiment concludes with phase 4 (post-study assessment), where par- ticipants complete a brief questionnaire about their demographics (age and gen- der identity) and their trust in the (X)AI assistant. Additionally, they respond to two open-ended questions about if they noticed anything about the (X)AI tool and their understanding of the purpose of the study.

First things first: rough data cleaning

Let's first just look at the data we have. Excluding all users that had e.g., incomplete datasets, what is the turnout?

General infos after removal of incomplete datasets

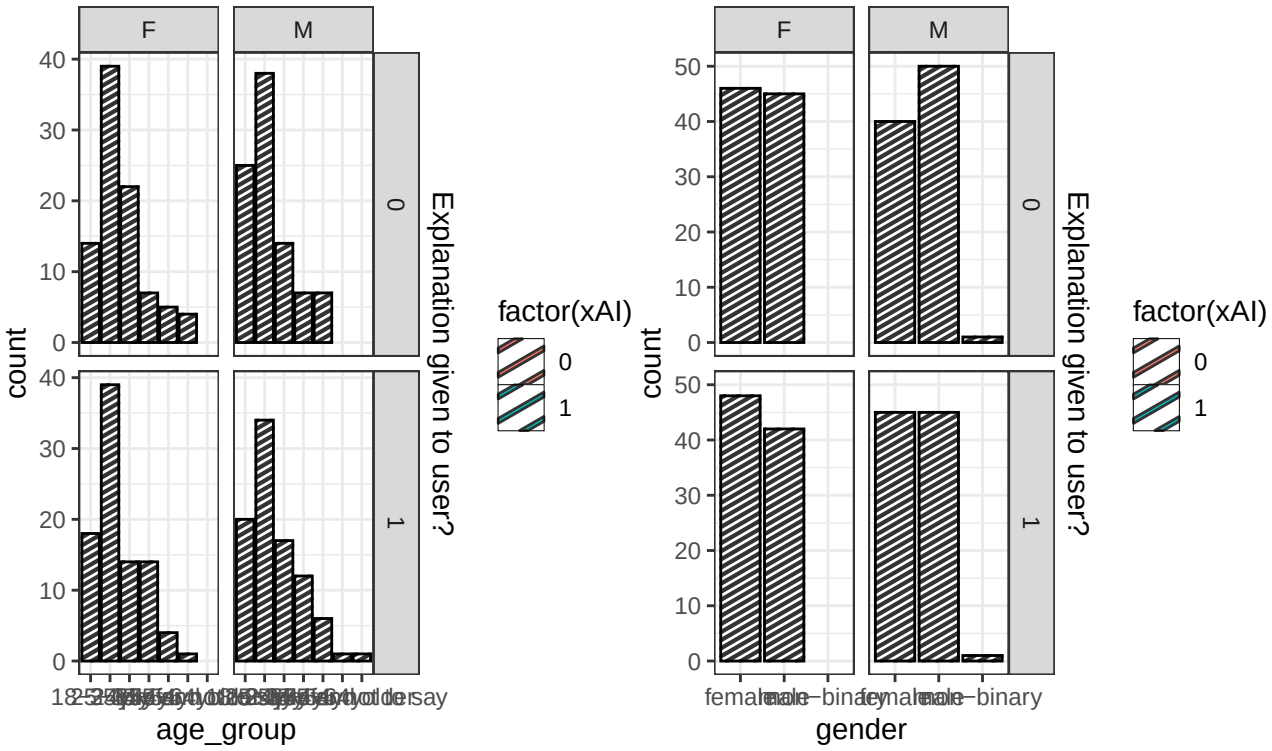
How many users entered the study and inserted their Prolific-ID (i.e., also including users with incomplete datasets or those failing the comprehension checks)? 363

After removal of incomplete datasets, we have 363 participants. Of those,

- 91 participants were in the FemBias-noEx condition
- 91 participants were in the MalBias-noEx condition
- 90 participants were in the FemBias-XAI condition
- 91 participants were in the MalBias-XAI condition

Check covariates across groups

Additionally to assessing performance, we also acquire age and gender information of participants. How do our groups look like? Are the groups comparable?



Let's run a statistical comparison between our four groups. For age and gender, we have ordinal data (in age bands), so we will use a non-parametric statistical test for ordinal data for more than 2 groups, that's the Chi-Square Test of Independence.

We acquired data from 363 participants, with

91 users in the "FB-AI" group (46 female, 45 male, median age group is 25-34y), and

91 users in the "MB-AI" group (40 female, 50 male, 1 non-binary, median age group is 25-34y), and

90 users in the "FB-XAI" group (48 female, 42 male, median age group is 25-34y), and

91 users in the "MB-XAI" group (45 female, 45 male, 1 non-binary, median age group is 25-34y).

The analysis showed for *Age*:

- We have age information for all but 1 participant overall.
- Is there a significant difference in terms of age between the groups? We compared distribution of age of participants between conditions using a Chiˆ2 test. This showed: Chi2=16.9114163, p=0.3241875, Cramer's V = 0.0416679663853576, 0.95, 0, 1

The analysis showed for *Gender*:

- Is there a significant difference in terms of gender distribution between the groups? We compared gender distribution for users in explanation condition and users in the control condition using a Chiˆ2 test. This showed: Chi2=3.4950192, p=0.7446323, Cramer's V = 0, 0.95, 0, 1

Quality criteria

Before going into the hypotheses, we should apply some quality criteria to our data. Sub-quality data should be removed. The following subsections take care of such cases.

Identify "speeders" and "dawdlers"

Speeders are people clicking through the study way too quickly to do the task properly. Dawdlers are people taking unusually long. This might indicate them leaving the screen / being distracted.

Aim: identify IDs being faster / slower than specified values. This part will tag users that deviate less or more than 3*SD from the population's mean.

Identify participants failing the survey attention check

In addition to comprehensive comprehension questions in the beginning, we include 1 attention check during survey (item no 6).

Aim: Identify IDs of users getting this check wrong.

Identify "straight-liners" in decision phases

Identify users who keep selecting either the right or the left image. This can be sign that the task was not done attentively.

Aim: identify IDs of users "straight-lining" in at least one of the non-interaction phases (note: due to AI recommendations, one side might have been favored - we do not want to punish following the AI assistant). How many repeated choices count as "straightlining"? This is a design decision. We decide that a "straightliner" is someone who keeps selecting one side in more than 10 consecutive trials (straightlining_max_task = 10).

Identify "straight-liners" in survey part, and "inconsistent" users

Identify users who always give very uniform answers in the survey part.

Aim: identify IDs of users "straight-lining", i.e. giving only responses with either positive or negative valence. Also identify those giving replies with the same valence to the reworded item and its original counterpart.

Remove data from problematic users

As we have identified users that seem to have unreliable data, we want to remove them.

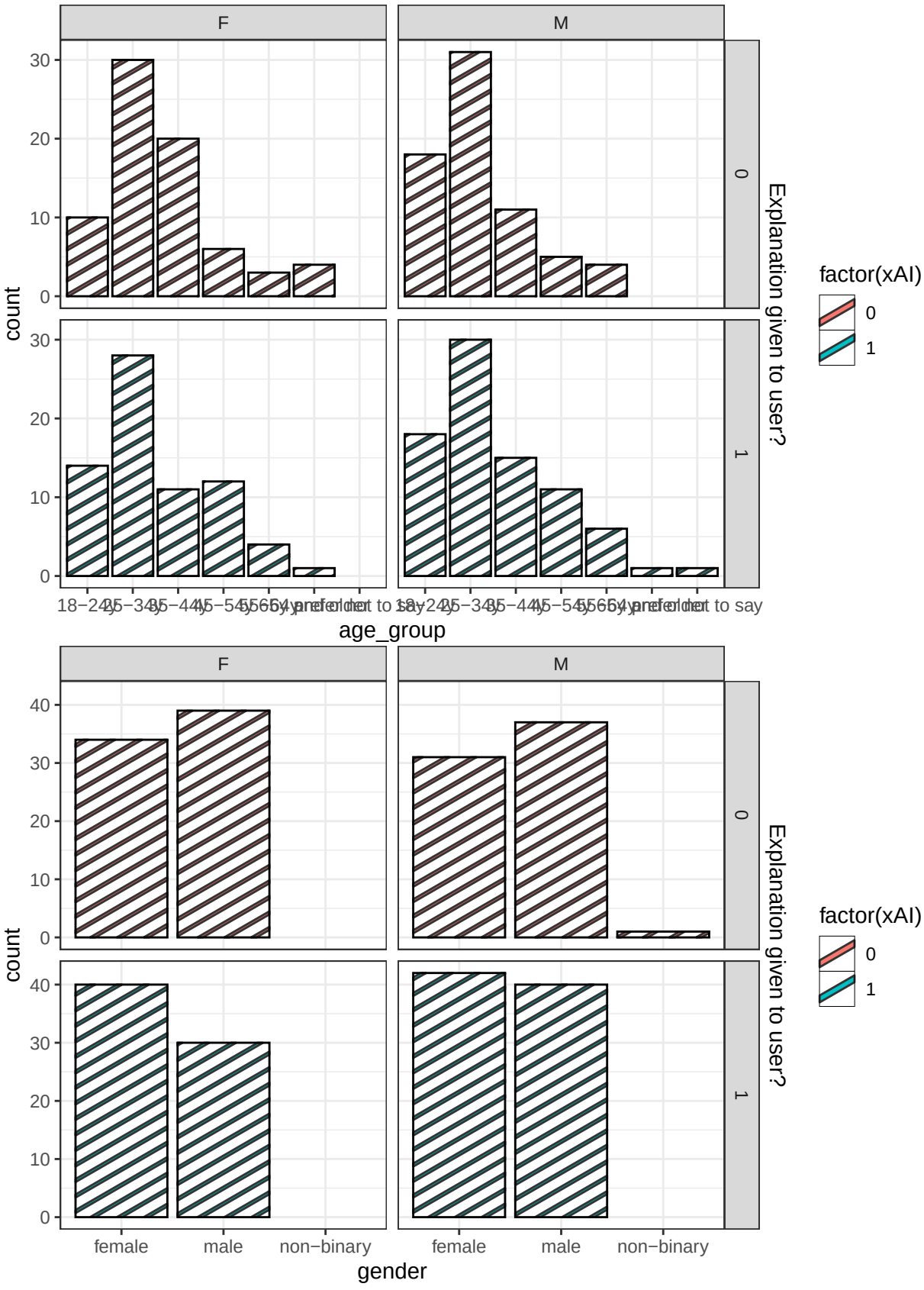
Before it gets serious, do some pre-processing of the response data

So to summarize:

- we have 363 users to begin with
- we remove 0 speeders and 3 dawdlers
- of the remaining participants, we remove 3 users that failed the attention question during the survey
- of the remaining participants, we remove 9 users that consistently clicked on the image on one side
- of the remaining participants, we remove remove 2 users that straightlined in the survey
- of the remaining participants, we remove remove 52 users give contradictory responses in the survey

Finally: How many users do we have in our clean performance df? 294

Final, clean dataset (N=294)



To

sum up, in our final data we have 294 participants, with

73 users in the "FB-AI" group (34 female, 39 male, median age group is 25-34y), and

69 users in the "MB-AI" group (31 female, 37 male, 1 non-binary, median age group is 25-34y), and

70 users in the "FB-XAI" group (40 female, 30 male, median age group is 25-34y), and

82 users in the "MB-XAI" group (42 female, 40 male, median age group is 25-34y).

The analysis showed for *Age*:

- Is there a significant difference in terms of age between the groups? We compared ages of users in explanation condition and users in the control condition using a Chiˆ2 test. This showed: Chi2=17.4512658, p=0.2926108, Cramer's V = 0.0525223643172281, 0.95, 0, 1

The analysis showed for *Gender*:

- Is there a significant difference in terms of gender between the groups? We compared gender distribution for users in explanation condition and users in the control condition using a Chiˆ2 test. This showed: Chi2=5.6175429, p=0.4673654, Cramer's V = 0, 0.95, 0, 1

Hypotheses

Our main hypotheses are:

*H1) Participants receiving XAI CEs will show different rates of agreement with AI recommendations compared to those receiving black-box AI recommendations.

*H2) Interaction with biased (X)AI recommendations will shift participants' gender-based decision patterns in subsequent independent evaluations, increasing alignment with the (X)AI's bias direction.

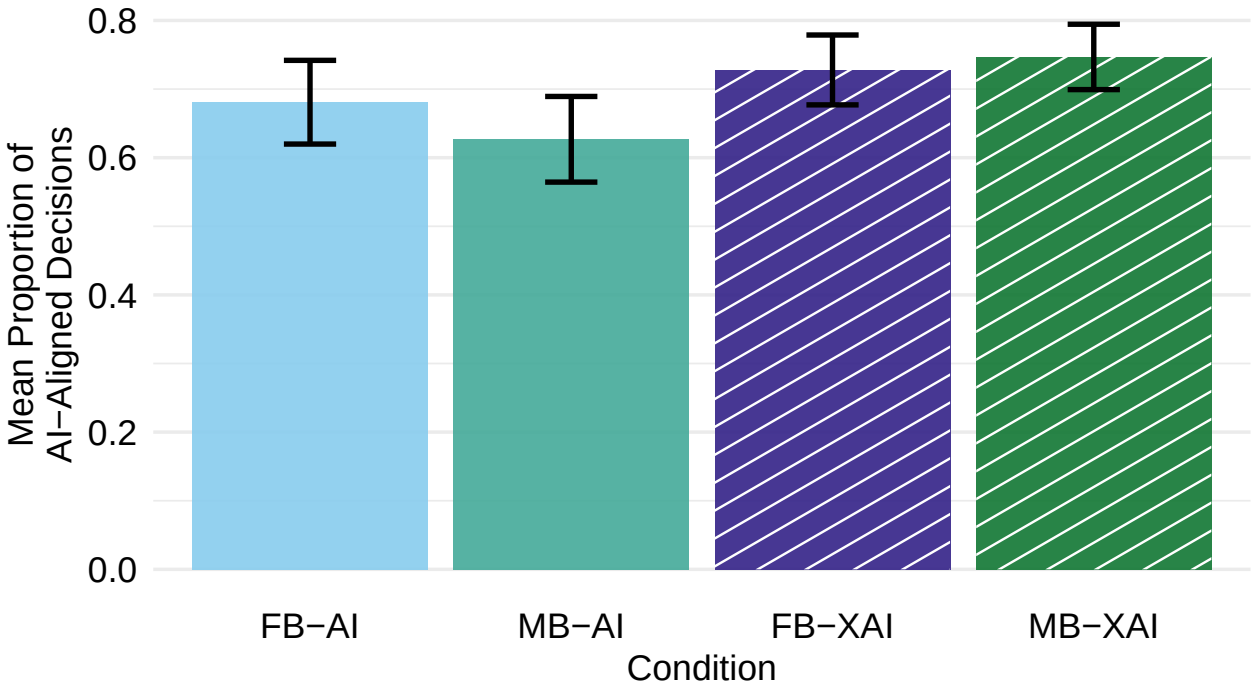
*H3) Decision confidence will vary across experimental phases, depending on the experimental manipulation.

*H4) Participants receiving XAI CEs will show higher rates of trust in the AI recommendations compared to those receiving black-box AI recommendations.

Statistical assessment

We perform all statistical analyses with the experimental condition - Female Bias + AI recommendations (FB-AI), Male Bias + AI recommendations (MB-AI), Female Bias + XAI recommendations (FB-XAI), and Male Bias + XAI recommendations (MB-XAI) - serving as the independent variable. - Distributional differences in demographic covariates were evaluated using χ^2 tests. - The bias shift (BS) in participant behavior is computed as: $BS = MB_{post} - MBpre$, where MB_{post} is the male bias in the post-(X)AI-interaction phase, and MB_{pre} is the male bias in the pre-(X)AI-interaction phase. Positive values indicate a shift toward male bias, while negative values indicate a shift toward female bias. To evaluate the bias shift, the proportion of (X)AI alignment during the interaction phase, and the accumulated trust scores derived from the post-study assessment, we fitted separate 2×2 linear models with factors biased gender (FB vs. MB) and XAI (AI vs. XAI). - For the longitudinal analysis of confidence, measured after each of the three decision phases, we employ a linear mixed-effects model. This model incorporates fixed effects for group, phase, and their interaction, and included a by-subject random intercept to account for within-participant correlations.

H1) Participants receiving XAI CEs will show different rates of agreement with AI recommendations compared to those receiving black-box AI recommendations.



Now on to the statistics: are there systematic group differences in terms of AI alignment?

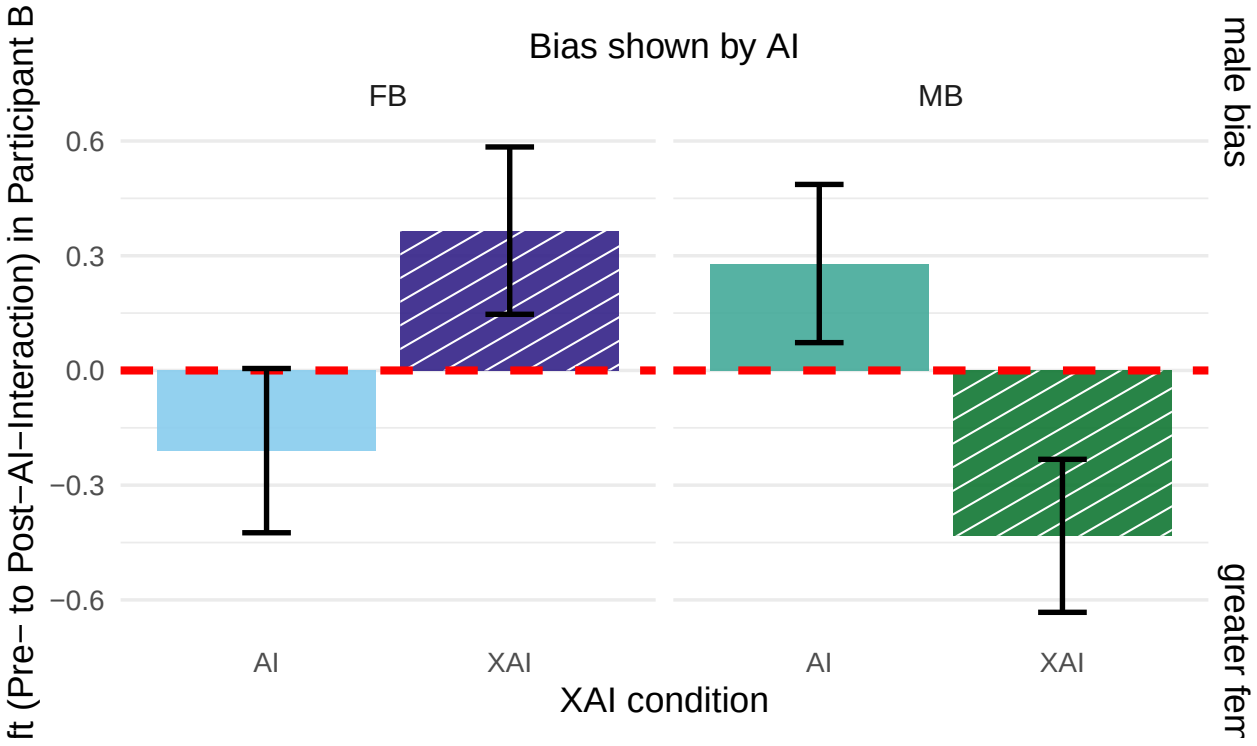
```
## [1] "ANOVA table:"
```

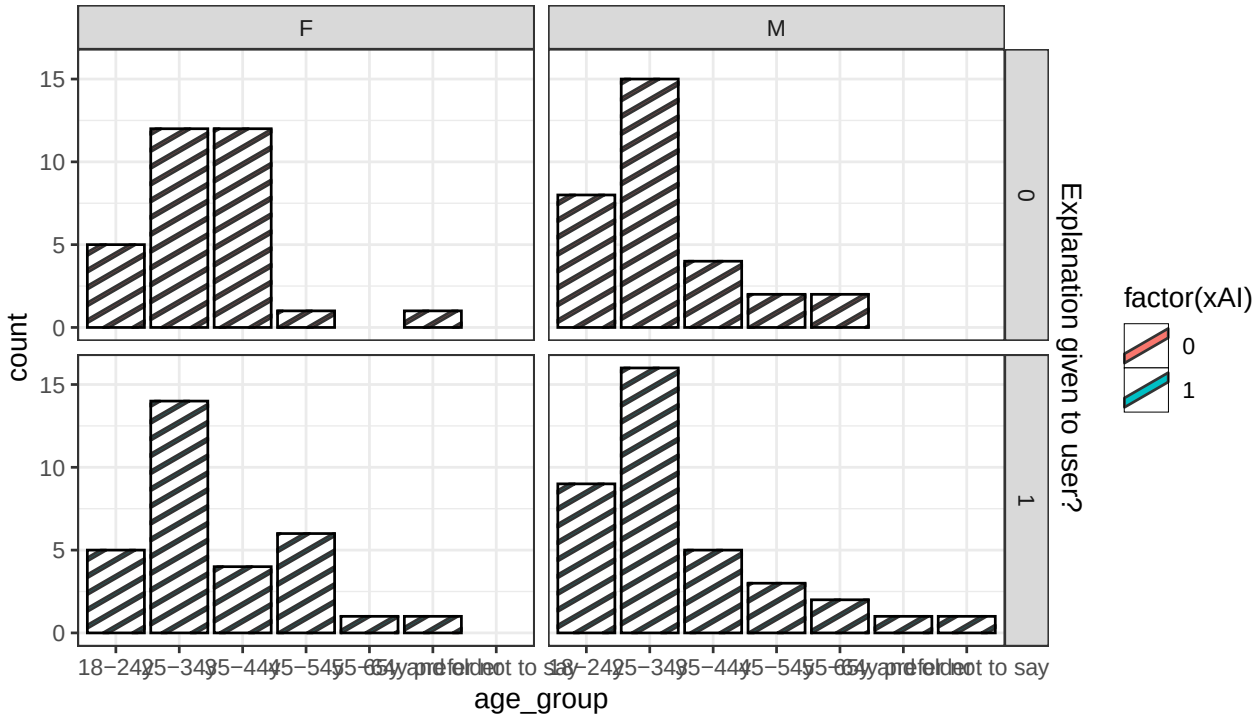
```
## Analysis of Variance Table
##
## Response: percentage_AI_aligned
##                     DF   SumSq  MeanSq  Fvalue  Pvalue
## biased_gender        1  0.0034 0.00344 0.02218 0.88177
## xAI                  1  0.3714 0.37143 2.39298 0.12345
## biased_gender:xAI    1  0.0675 0.06746 0.43460 0.51049
## Residuals          202 31.3537 0.15522
```

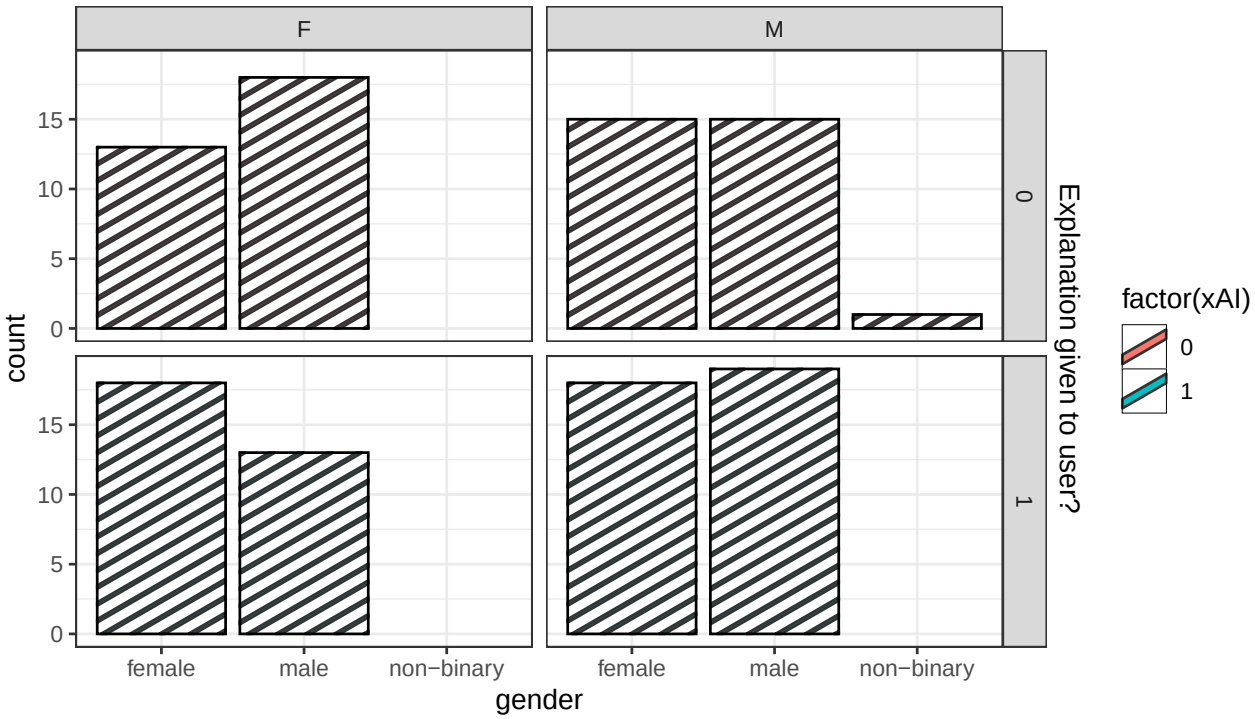
Neither main effect, nor the interaction seems to be statistically significant. Thus, we cannot confirm H1 here.

H2) Interaction with biased (X)AI recommendations will shift participants' gender-based decision patterns in subsequent independent evaluations, increasing alignment with the (X)AI's bias direction.

First, we need to compute and visualize the independent measure 'bias_shift' (from pre-post).


Bias shown by AI
FB
MB
ft (Pre− to Post−AI−Interaction) in Participant B
male bias
greater fem
0.6
0.3
0.0
−0.3
−0.6
AI
XAI
AI
XAI
XAI condition


F
M
0
1
count
15
10
5
0
18−24
25−34
35−44
45−54
55−64
65+
prefer not to say
age_group
Explanation given to user?
factor(xAI)
0
1



IMPORTANT NOTE! For the bias shift, we look at a further sub-sample of participants. We are reduced to the "equal aptitude set".

In our equal aptitude data we have 130 participants, with

31 users in the "FB-AI" group (13 female, 18 male, median age group is 25-34y), and

31 users in the "MB-AI" group (15 female, 15 male, median age group is 25-34y), and

31 users in the "FB-XAI" group (18 female, 13 male, median age group is 25-34y), and

37 users in the "MB-XAI" group (18 female, 19 male, median age group is 25-34y).

The analysis showed for *Age*:

- Is there a significant difference in terms of age between the groups? We compared ages of users in explanation condition and users in the control condition using a Chi^2 test. This showed: Chi2=17.4419916, p=0.2931356, Cramer's V = 0.0784309865313729, 0.95, 0, 1

The analysis showed for *Gender*:

- Is there a significant difference in terms of gender between the groups? We compared gender distribution for users in explanation condition and users in the control condition using a Chi^2 test. This showed: Chi2=4.8631757, p=0.561478, Cramer's V = 0, 0.95, 0, 1

Now on to the statistics, looking at the bias shift shown by participants in individual conditions, ending up with a relatively simple mixed 2x2 model (biased_gender X xAI).

```
## [1] "ANOVA table:"
```

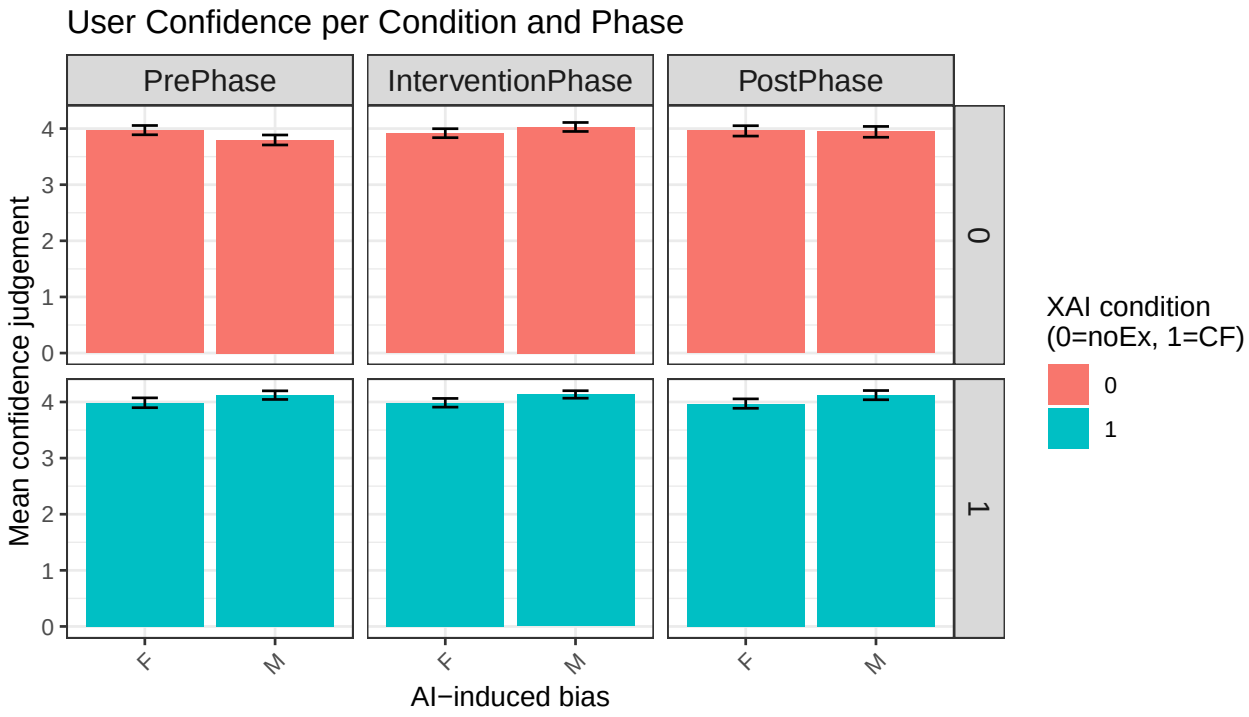
```
## Analysis of Variance Table
##
## Response: bias_shift
##                     DF   SumSq  MeanSq Fvalue  Pvalue
## biased_gender        1   1.120  1.1196 0.7815 0.37838
## xAI                  1   0.296  0.2956 0.2063 0.65045
## biased_gender:xAI    1  13.385 13.3849 9.3428 0.00273
```

```
## Residuals          126 180.513  1.4326
```

These results show something striking: participants who did not receive explanations shifted their inherent biases in the same direction as the bias displayed by the AI, whereas those who received CEs shifted their bias in the opposite direction of the XAI bias!

H3) Decision confidence will vary across experimental phases, depending on the experimental manipulation.

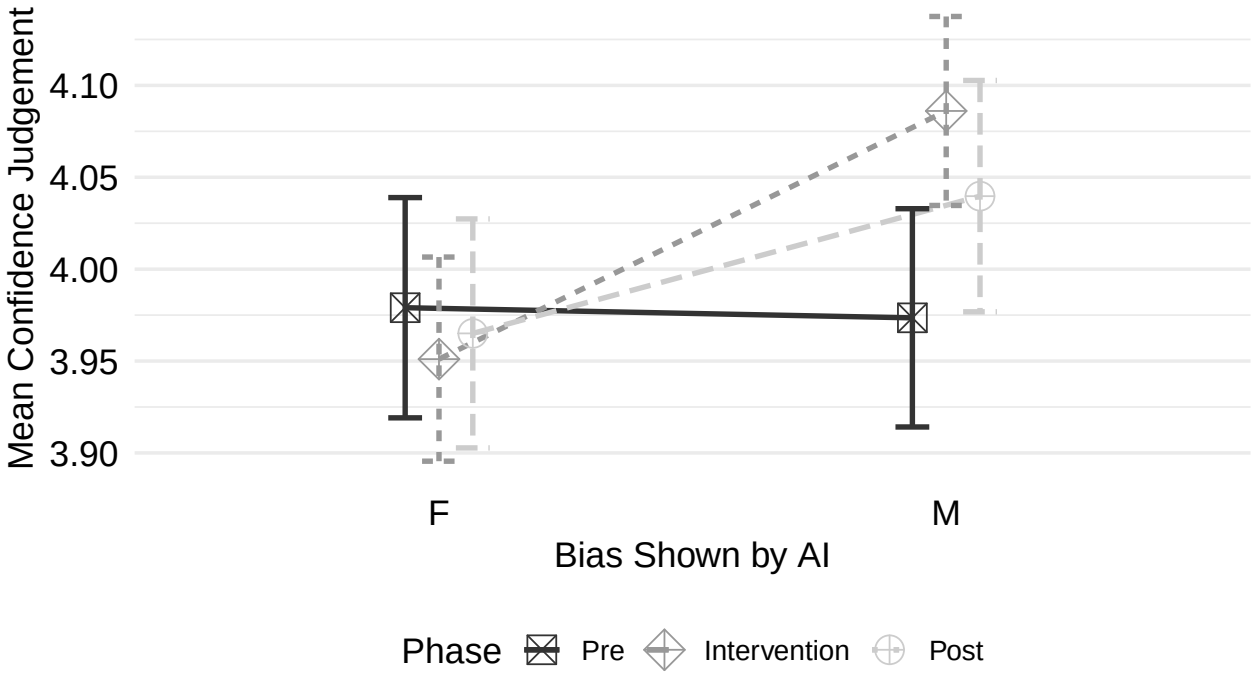
Did the experimental manipulation elicit different levels of confidence about one's own decisions?



Next: On to the statistics on confidence: We are looking at the full data across phases with a relatively complex mixed 2x2x3 model.

```
## [1] "ANOVA table:"
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                                SumSq  MeanSq NumDF DenDF Fvalue  Pvalue
## biased_gender                0.01152 0.01152     1   290 0.0668 0.79617
## xAI                          0.44590 0.44590     1   290 2.5870 0.10883
## phase_name                   0.59930 0.29965     2   580 1.7385 0.17670
## biased_gender:xAI            0.24027 0.24027     1   290 1.3940 0.23870
## biased_gender:phase_name     1.46315 0.73157     2   580 4.2444 0.01479
## xAI:phase_name               0.29717 0.14858     2   580 0.8620 0.42284
## biased_gender:xAI:phase_name 0.68964 0.34482     2   580 2.0006 0.13619
```

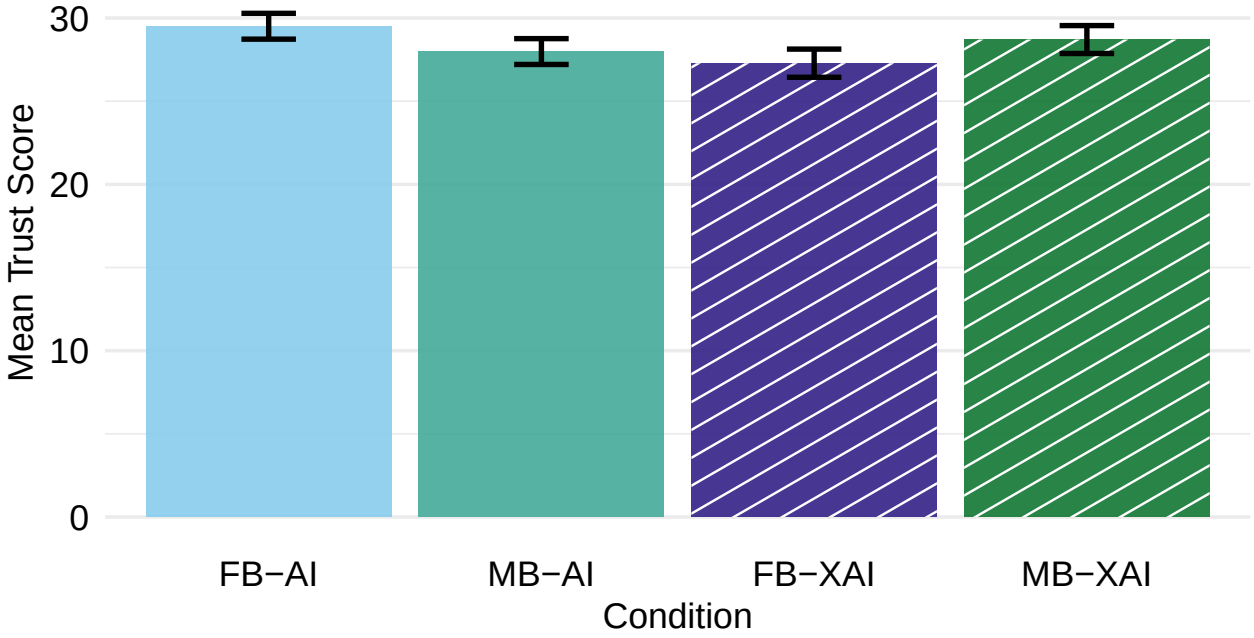


The analysis revealed a significant two-way interaction (biased_gender x phase_name).

These results suggest that while overall confidence did not differ significantly as a function of XAI or phase alone, the impact of phase on confidence was contingent upon the gender bias condition.

H4) Participants receiving XAI CEs will show higher rates of trust in the AI recommendations compared to those receiving black-box AI recommendations.

We will be running a 2x2 model to evaluate this (XAI x bias direction). The hypothesis expects a significant effect of XAI factor on the accumulated trust measure.



Now on to the stats.

```
## [1] "ANOVA table:"
```

```
## Analysis of Variance Table
##
## Response: summed_trust
##                     DF   SumSq  MeanSq Fvalue  Pvalue
## biased_gender        1     0.1   0.130 0.0027 0.95881
## xAI                  1    37.4  37.400 0.7679 0.38160
## biased_gender:xAI    1   158.4 158.415 3.2525 0.07235
## Residuals          290 14124.5  48.705
```

Neither main effect, nor the interaction seems to be statistically reliable. Thus, we have limited evidence for H4 here.

Final post-hoc analysis of candidate matchings: How balanced were candidate pairs on average?

This shows: the difference in total feature scores between the paired candidates was to be kept small (mean difference = -0.0715742, SD = 5.0101292 in final study).