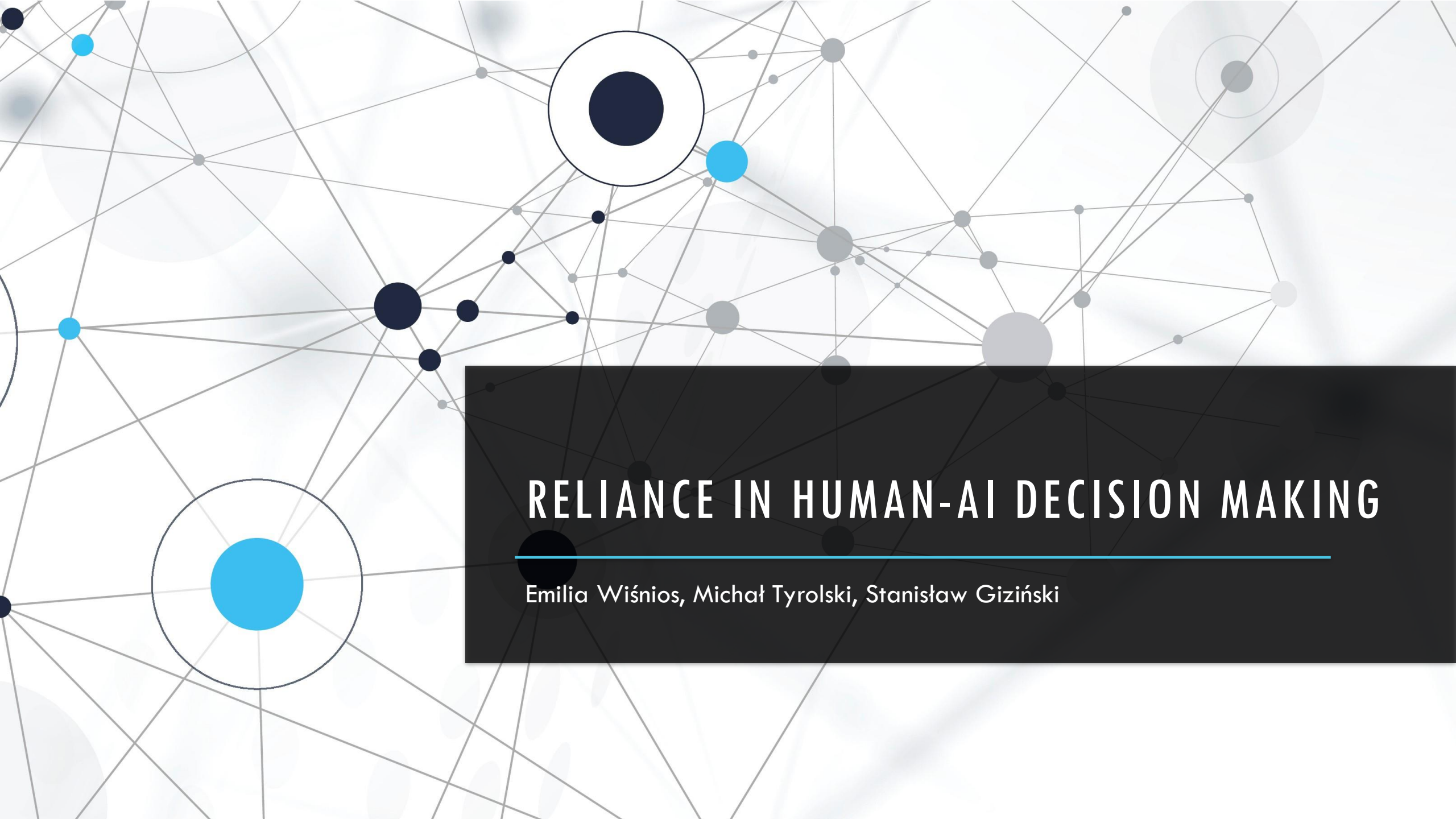


# RELIANCE IN HUMAN-AI DECISION MAKING

Emilia Wiśnios, Michał Tyrolski, Stanisław Giziński

# MOTIVATION & BACKGROUND

Nowadays ML models are more and more broadly used across different domains

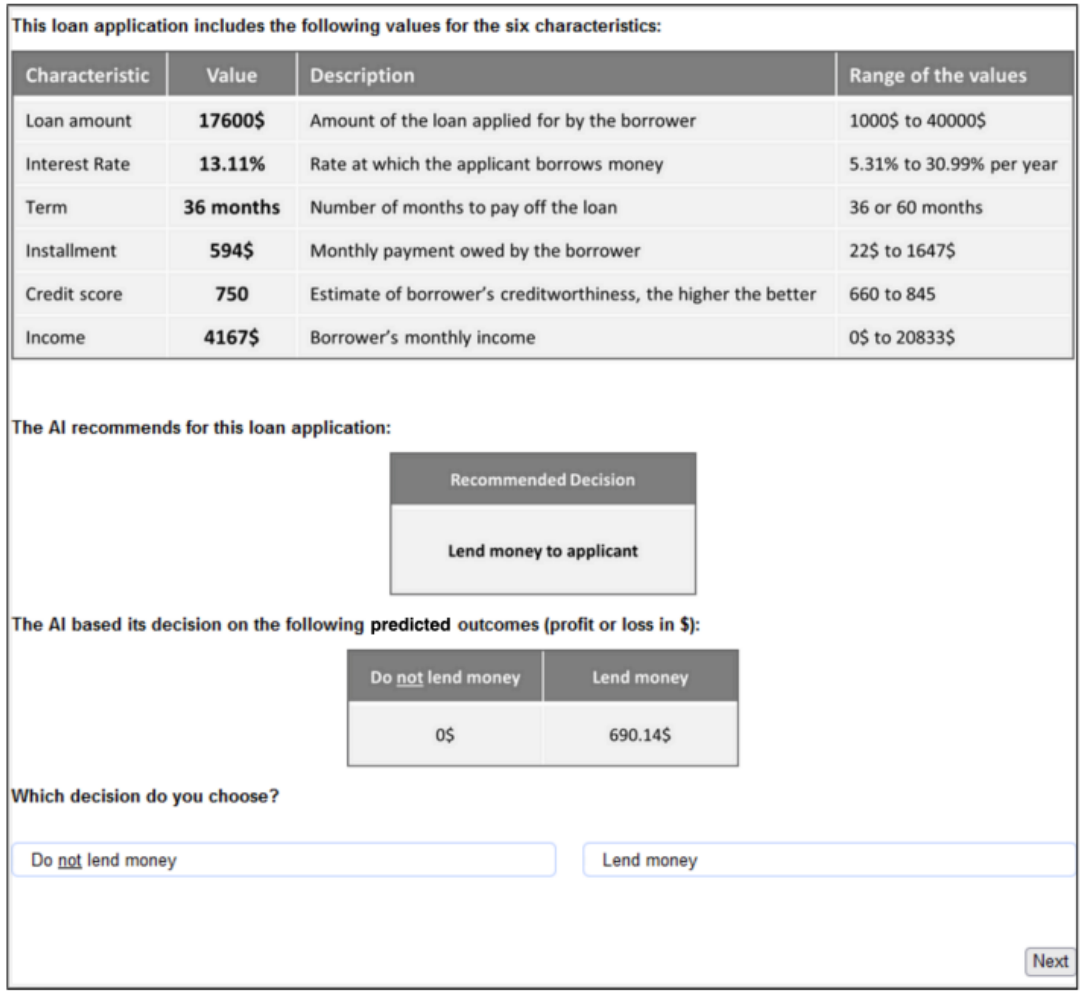
Models are not perfect, thus we can't fully rely on tchem

Explanations should increase reliability of cooperation of models with humans

…do they?

# RELATED WORK

Jakubik et al. An Empirical Evaluation of Predicted Outcomes as Explanations in Human-AI Decision-Making. ECML XKDD Workshop 2022



This loan application includes the following values for the six characteristics:

| Characteristic | Value | Description | Range of the values |
|---|---|---|---|
| Loan amount | **17600$** | Amount of the loan applied for by the borrower | 1000$ to 40000$ |
| Interest Rate | **13.11%** | Rate at which the applicant borrows money | 5.31% to 30.99% per year |
| Term | **36 months** | Number of months to pay off the loan | 36 or 60 months |
| Installment | **594$** | Monthly payment owed by the borrower | 22$ to 1647$ |
| Credit score | **750** | Estimate of borrower's creditworthiness, the higher the better | 660 to 845 |
| Income | **4167$** | Borrower's monthly income | 0$ to 20833$ |

The AI recommends for this loan application:

| Recommended Decision |
|---|
| **Lend money to applicant** |

The AI based its decision on the following **predicted** outcomes (profit or loss in $):

| Do not lend money | Lend money |
|---|---|
| 0$ | 690.14$ |

Which decision do you choose?

Do not lend money | Lend money

Next

Fig. 1: Exemplary trial from our study presenting the task and relevant information in the *AI with predicted outcomes* condition.

# RELATED WORK TAKEAWAYS

General idea: Binary Predictions vs Regression Predictions vs Binary Predictions + Global Explanation

People tends to follow AI recommendations more often when supplemented with predictions

This effect is particularly pronounced when AI recommendations are incorrect, leading to over-reliance

Predictions determine a reduced ability to distinguish between correct and incorrect AI recommendations.

# RELATED WORK CONT'D

Baniecki H, Parzych D, Biecek P. The grammar of interactive explanatory model analysis. arXiv preprint arXiv:2005.00497. 2020 May 1.

# RELATED WORK TAKEAWAYS

- Break Down vs +=Ceteris Paribus vs +=Shapley Values
- Accuracy & Confidence increases for last case (BD+CP+SV)
- "I don't know" decreases for last case (BD+CP+SV)
- Impact in decision: BD+CP+SV > BD+CP

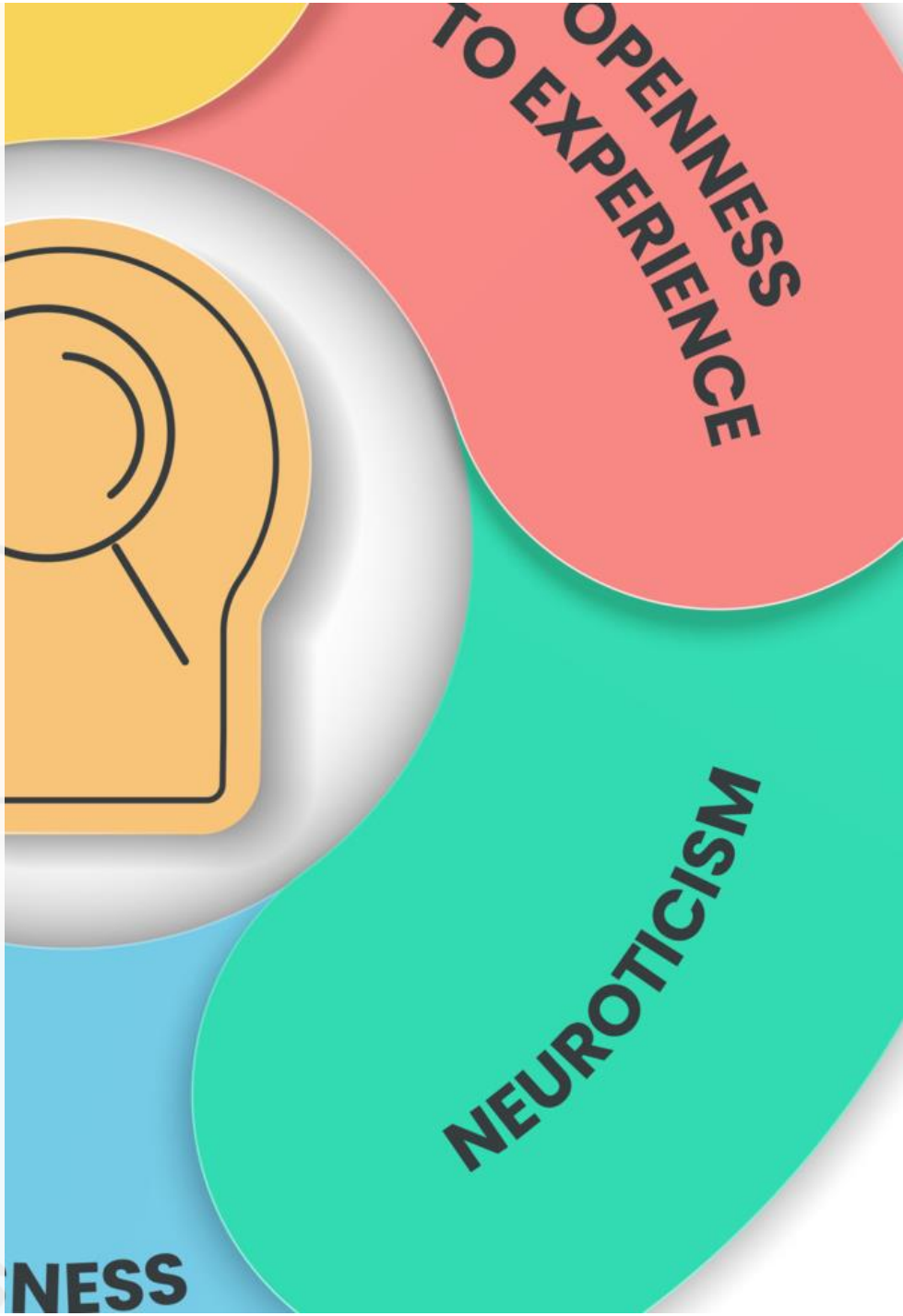
# OUR APPROACH

Census bureau database by Ronny Kohavi and Barry Becker

Determine whether a person makes over $50K a year

Analize data over 3 groups

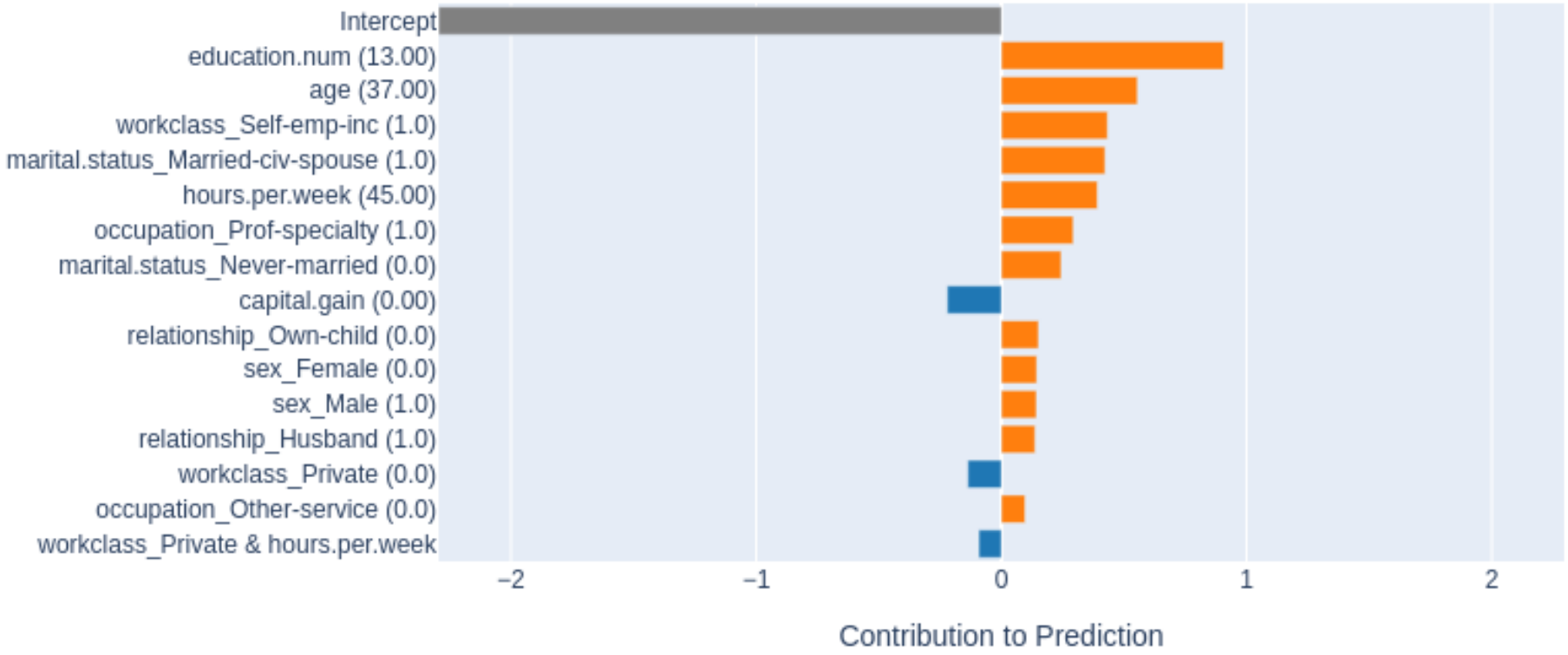
Contributions: Big Five Test, Evaluation using EBMs



# MODEL

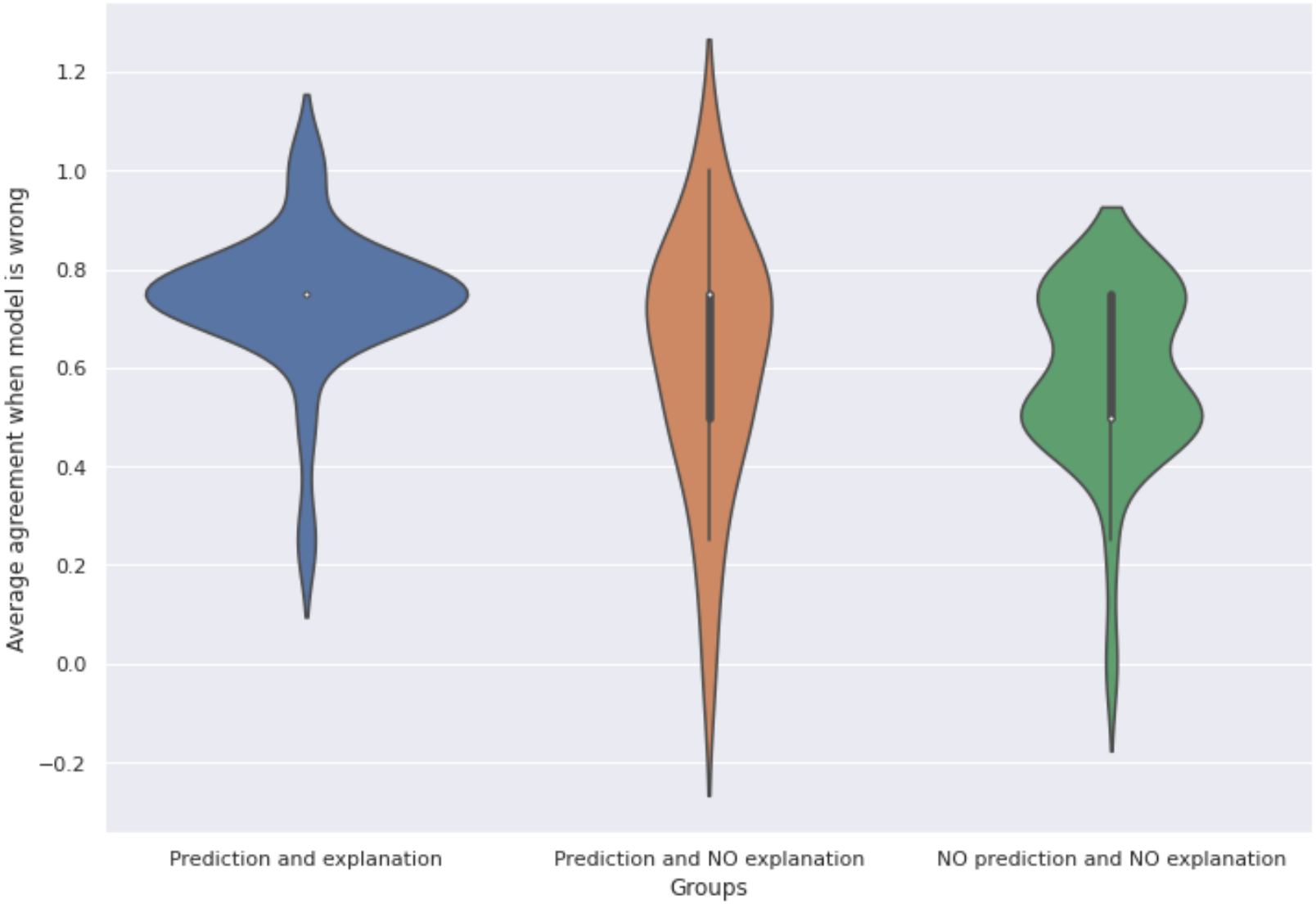
Explainable Boosting Machines

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

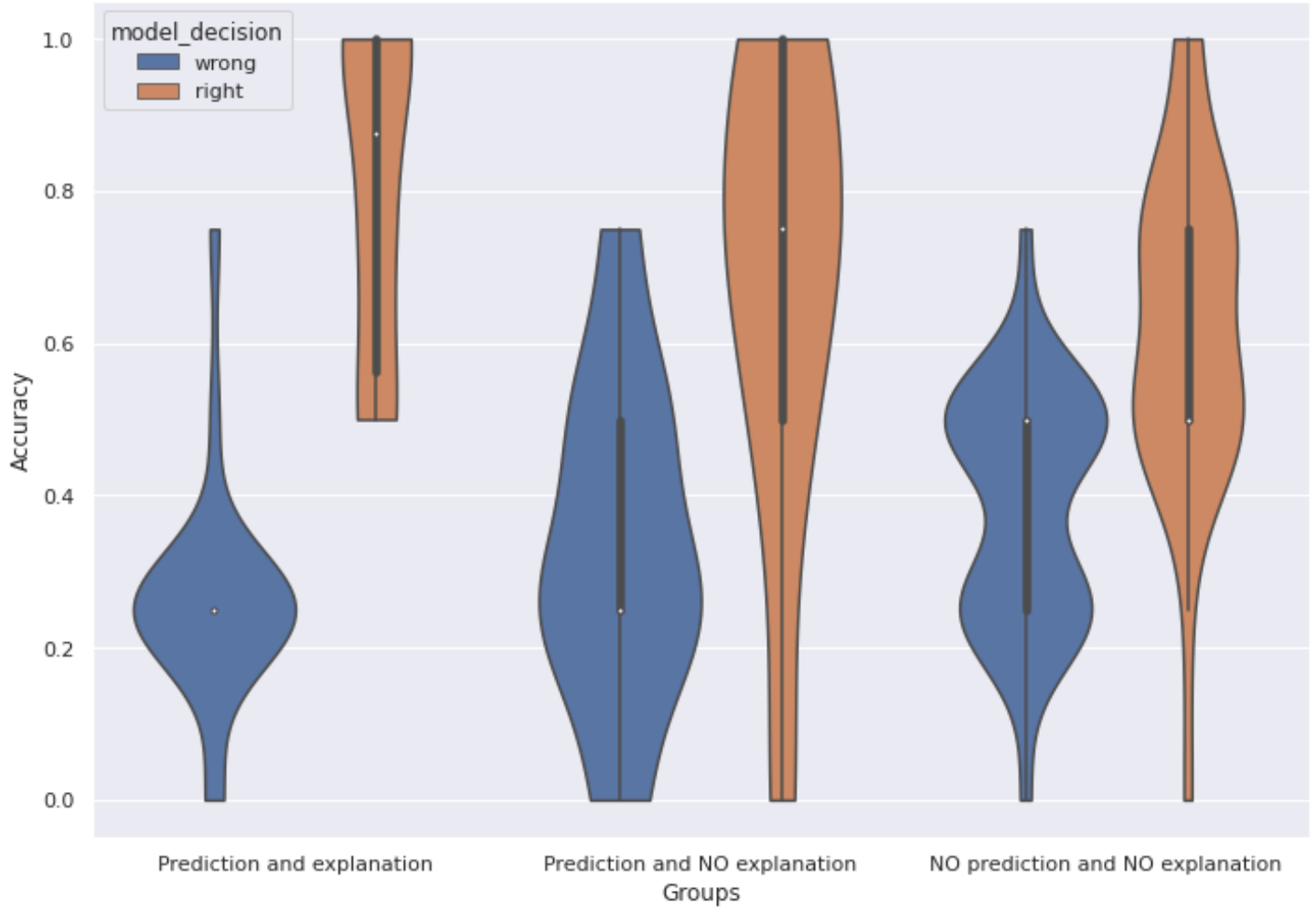


# RESULTS

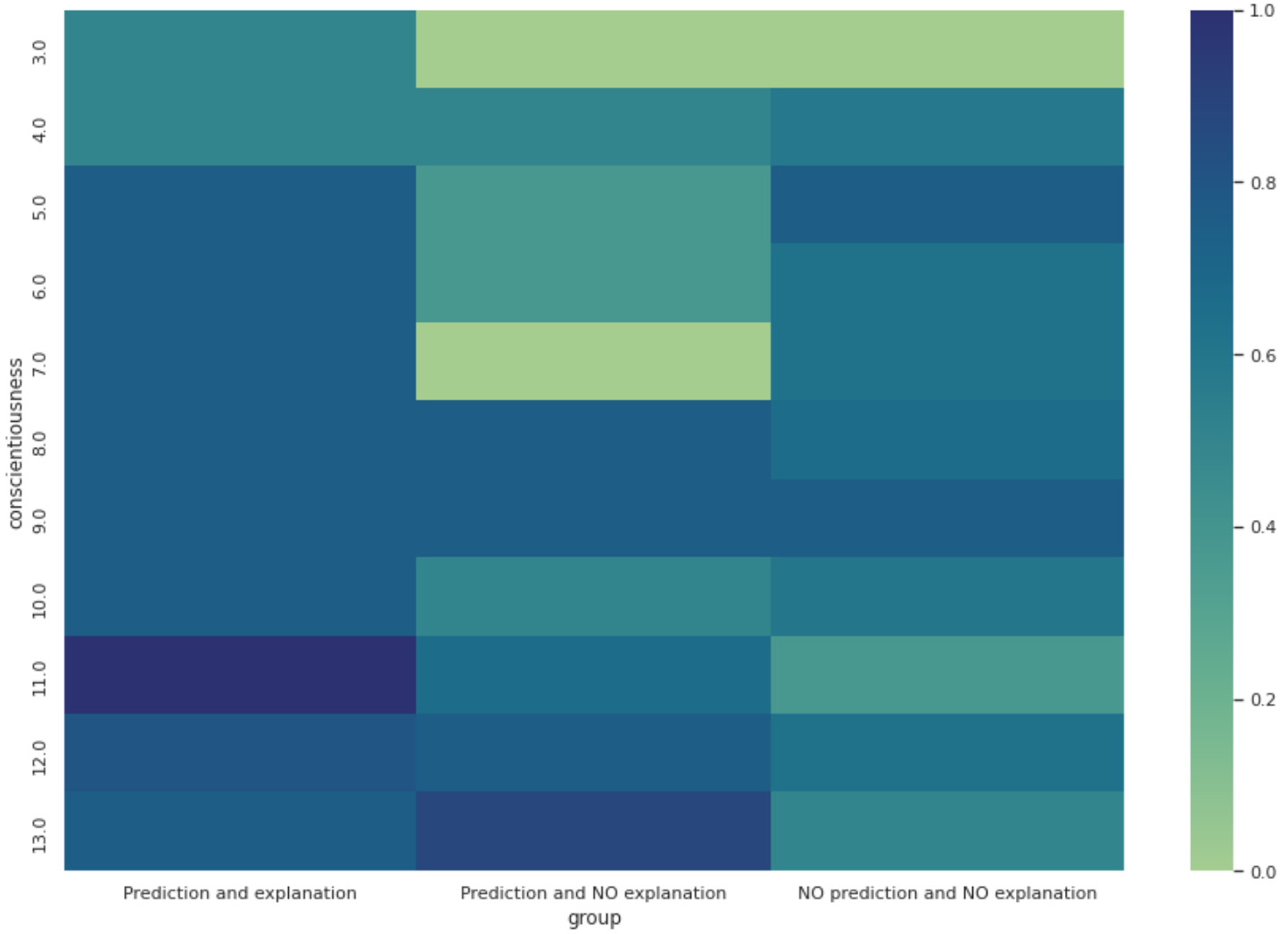
# AVERAGE AGREEMENT WHEN MODEL IS WRONG (PER PERSON)



# ACCURACY VS MODEL DECISION



# AVERAGE AGREEMENT WHEN THE MODEL IS WRONG WRT CONSCIENTIOUSNESS



# MODEL ANSWERS VS GROUP

| Group | Model Response | Can reject Chi^2 test |
|---|---|---|
| Explanation | TP | No |
| Explanation | TN | Yes |
| Explanation | FP | No |
| Explanation | FN | +/- (1/2) → No |
| Prediction only | TP | +/- (1/2) → No |
| Prediction only | TN | No |
| Prediction only | FP | No |
| Prediction only | FN | No |

# RELIANCE

| Model Response | Accuracy | Reliance |
|---|---|---|
| FP | 5% | 55% |
| FN | 50% | 27% |
| TP | 91% | 70% |
| TN | 75% | 61% |

# KEY TAKEAWAYS

People are strongly inspired by model responses, even if they say otherwise

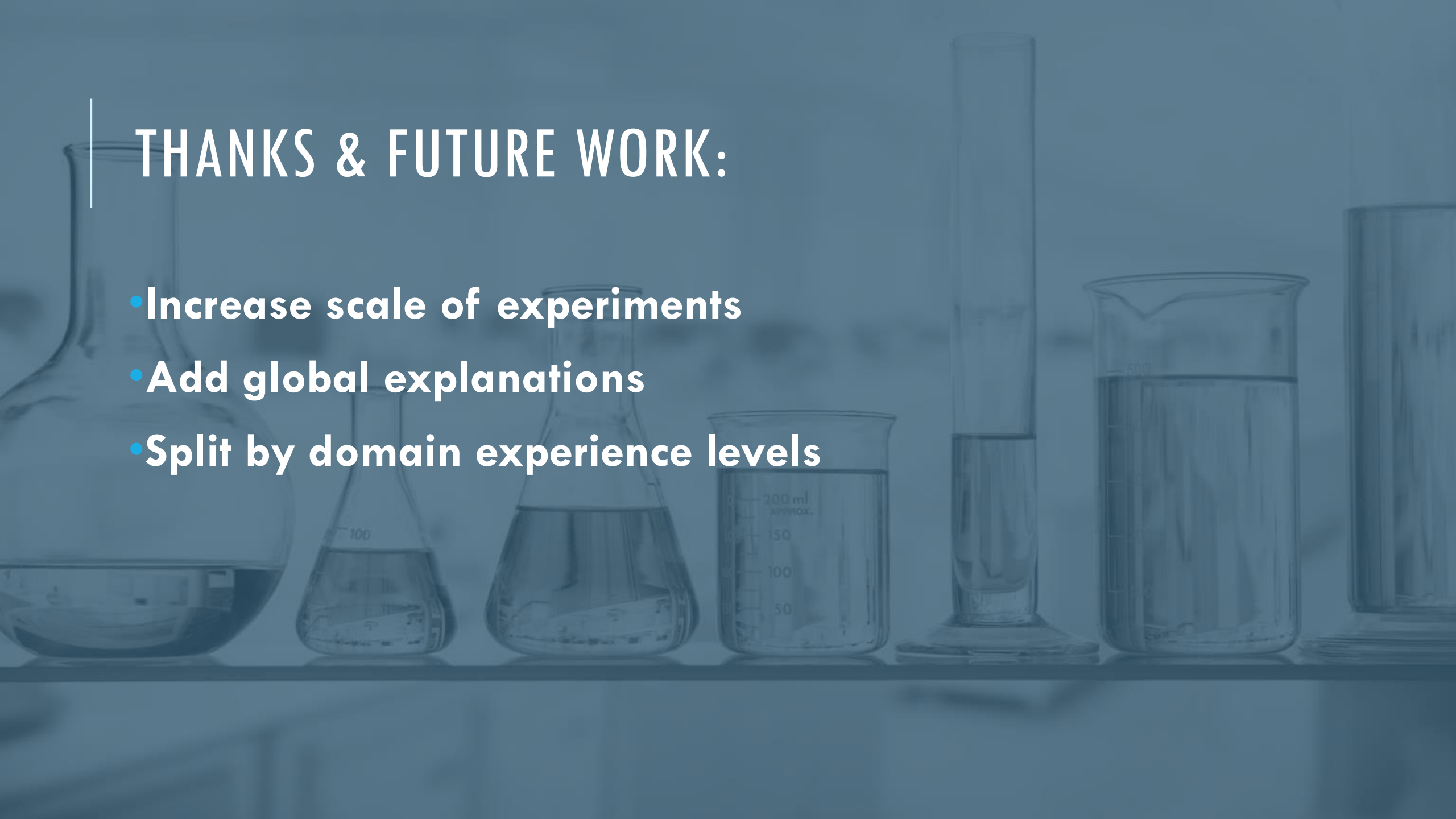
Models don't increase the reliability of Human-AI cooperation in all cases

Several groups of personalities, (ex. Conscientiousness), are more susceptible to ML explanations

There is a strong need for constructing a better „Reliance" definition

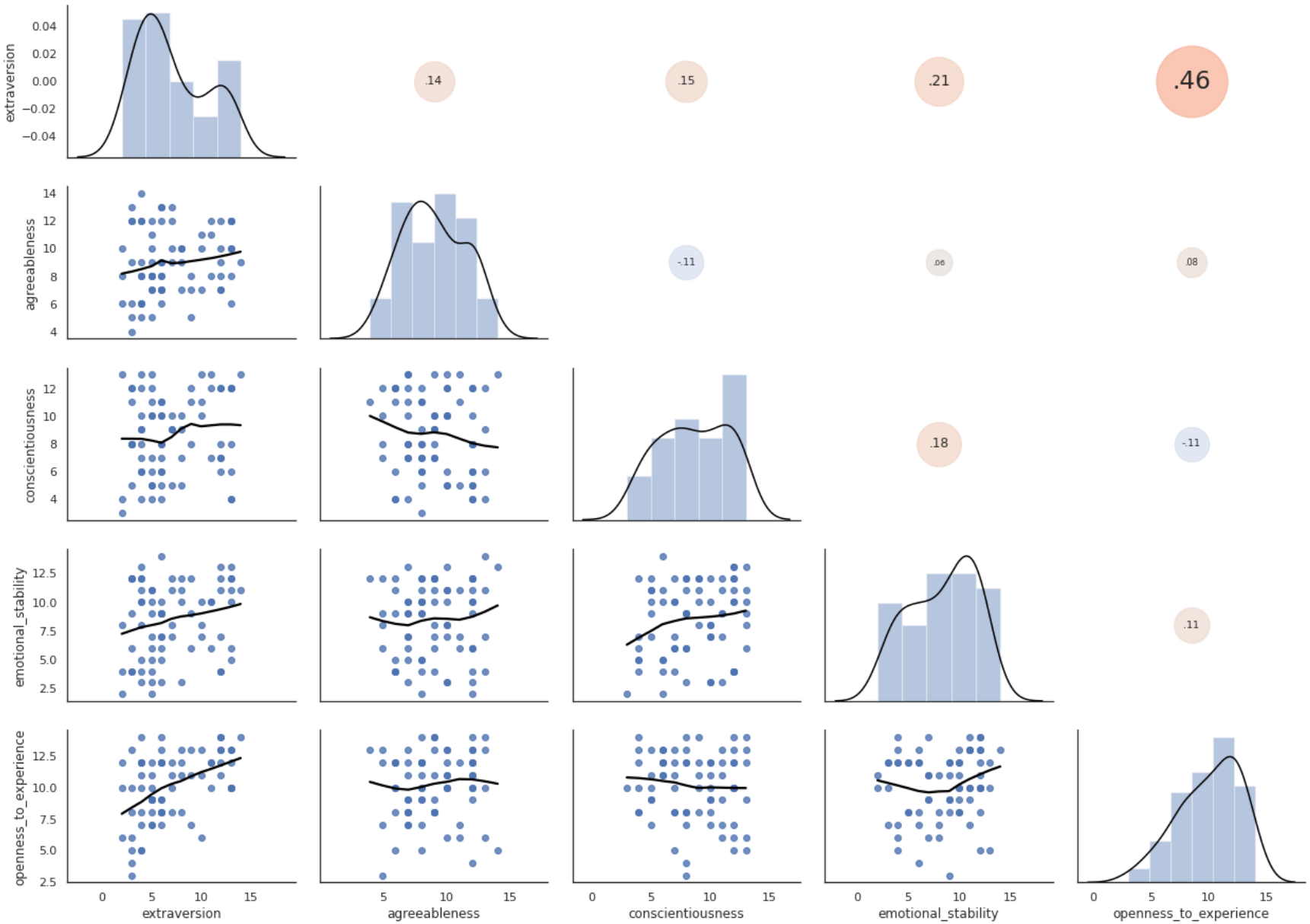
# THANKS & FUTURE WORK:

- Increase scale of experiments
- Add global explanations
- Split by domain experience levels



# APPENDIX

# BIG FIVE


extraversion
agreeableness
conscientiousness
emotional_stability
openness_to_experience
.14
.15
.21
.46
-.11
.06
.08
.18
-.11
.11