Reliance in human-AI decision-making

Stanisław Giziński S.GIZINSKI@STUDENT.UW.EDU.PL
Michał Tyrolski M.TYROLSKI@STUDENT.UW.EDU.PL
Emilia Wiśnios E.WISNIOS@STUDENT.UW.EDU.PL
University of Warsaw, Poland

Abstract

1. Introduction

The current era is characterized by the widespread use of machine learning across various aspects of life, particularly in critical domains such as medicine (e.g. (Reynisson et al., 2020)), law (e.g. Surden (2014)) or finance (e.g. Huang et al. (2020)). To improve the reliability of AI systems in these fields, it is often proposed that AI recommendations should support human decision-making, as this solution is thought to be both safe and efficient. However, recent studies, among others (Jakubik et al., 2022), have revealed otherwise, with incorrect predictions stemming from a model's inability to perform well with data outside its training distribution, as well as the over-reliance on human-AI decision-making. According to that study, participants were more likely to follow AI recommendations when they were supplemented with predicted outcomes, compared to situations with no explanation or feature-based explanations. However, this increased trust in AI led to the phenomenon of over-reliance, especially when the AI recommendation was incorrect. The use of predicted outcomes as explanations also reduced participants' ability to distinguish between correct and incorrect AI recommendations. These findings underscore the importance of carefully considering the type of explanation in the design of human-AI decision-making systems.
In our study, investigated the influence of personality type on over-reliance on human-AI decision-making. Our main contributions are:

1. To the best of our knowledge, we as first investigated the influence of personalities on the human-AI decision-making process.

2. Within introducing the metric of "self-reliance", we show that participants unconsciously use models' suggestions, even if they say otherwise.

2. Methodology

Dataset In our study, we utilized the Adult Census Income dataset obtained from Kaggle. The data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Dua and Graff, 2017), and the objective of the prediction task is to determine whether an individual earns an annual income of over $\$50,000$. We selected the following features for our analysis: age, race, gender, native country, marital status, relationship, work class, occupation, years of education, hours of work per week, capital gain, and capital loss. Categorical features were then transformed through one-hot-encoding. Univariate feature selection with mutual information as a metric was performed, resulting in the selection of 30 out of 85 features for the final dataset.

Model We employed the use of Explainable Boosting Machines (EBMs) (Caruana et al., 2015) from the InterpretML (Nori et al., 2019) package for model training. EBMs are a variant of gradient boosting algorithms that incorporate feature importance and local interpretability methods to provide explanations for individual predictions. This approach allows for the identification of the most important features in the model, as well as the understanding of how these features influence the model's predictions. EBMs have been shown to achieve comparable or even superior performance compared to traditional gradient-boosting methods. Given the class imbalance in the dataset, we chose to optimize the model using the F1 metric. The final model achieved a score of 85.9% accuracy and 68.7% F1, demonstrating an effective balance between precision and recall.

Experimental Design To perform the study, we carried out a scenario-based online experiment. Participants were shown a description of an individual and asked to determine if their yearly income exceeded $\$50,000$. We divided the participants into three groups: a control group and two groups that received AI predictions, with one group also receiving a local explanation from the EBM model. The group compositions are listed in Table 5. The features and range of values were introduced at the start of the study, with a detailed description provided in Appendix B. To ensure a balanced distribution of answers, the groups were randomly assigned. However, one group tend to quit the survey before the end, and thus had a higher dropout rate, so only the least represented group was included in the final days of the experiment. All cases were consistent across groups (see Appendix A), but with varying levels of information provided. Additionally, all participants completed common demographic and personality questions, including age, gender, education level, and a shortened version of the Big Five personality test (10-Item Personality Inventory; (Rammstedt and John, 2007)) obtained from the following website (last accessed: 30.01.2023). Moreover, we asked in order participants to score their own machine-learning knowledge to check if this statistically influences results. The study also included a consent form and an introduction to the task.

Subject Recruitment A total of 74 subjects were recruited for the study. Dissemination of information regarding the study was accomplished through various channels, including students' mailing lists, LinkedIn and Twitter profiles, as well as slack channels of various

Condition	Explanation
No prediction and no explanation (group 3)	Study participants were provided only with features' values.
Prediction and no explanation (group 2)	Study participants were provided with features' values, the model's prediction, and its confidence.
Prediction and local explanation (group 1)	Study participants were provided with features' values, the model's prediction, its confidence, and local explanations produced by the model.

Table 1: Experimental conditions of our study design.

research groups. Despite efforts to obtain a diverse participant pool, limitations in time and resources resulted in a majority of participants having mathematical and computer science backgrounds. All subjects participated voluntarily and received no remuneration for their involvement. Final participants distribution was 22 for group with prediction and explanation, 21 for group with prediction-only group and 31 for the control group.

3. Experimental results

3.1 Big Five

Our first series of experiments cover the investigation of the influence of big five personality types on participations' responses. Based on survey answers, we assign to each person values of binary features, namely: agreeableness, conscientiousness, extraversion, openness to experience, and emotional stability. Due to sharing common features, those parts can partially overlap. We calculate the correlation in answers between each of the features and average responses from all parts. Additionally, we set hypothesis

$$H_0 = \textbf{There is no difference between responses of all participants vs group G}$$

$$H_1 = \textbf{There is a significant difference between responses}$$

For each case, we calculated the p-value. We set a threshold of 0.05. Results are shown in Table 2. We see that participants with conscientiousness attributes tend to behave differently than the rest, both for group 1 and group 2. Additionally, participants which are open to experience tend to behave differently in case the model is wrong.

3.2 Reliance and Self-Reliance

Definition 1 (Self-reliance) *Let us consider the number of cases where a study participant stated that the model was useful in decision-making to be represented by the variable n, and the total number of decisions made to be represented by N. The concept of self-reliance can then be defined as $1 - \frac{n}{N}$ which represents the proportion of people who relied on their own judgment, as opposed to relying on the model.*

	AGR		CONSC		EXTR		OPEXP		EMSTAB	
Group	corr	p-val	corr	p-val	corr	p-val	corr	p-val	corr	p-val
1 avg	-0.259	0.244	0.141	0.528	-0.102	0.650	0.002	0.991	0.145	0.518
1 fp/fn	0.139	0.535	0.532	0.010	0.084	0.708	-0.441	0.039	0.025	0.909
2 avg	-0.387	0.091	0.352	0.127	-0.138	0.561	-0.367	0.110	0.115	0.629
2 fp/fn	-0.290	0.2137	0.559	0.010	-0.026	0.912	-0.226	0.337	0.115	0.629

Table 2: Correlations and p-values between average reliance (avg) or average reliance when model wrong (fp/fn) in groups where model prediction was shown. Table legend: AGR - agreeableness, CONSC - conscientiousness, EXTR - extraversion, OP EXP - openness to experience, EWM STAB - emotional stability. P-values < 0.05 are marked on the green, P-values < 0.1 on yellow. Fp/fn indicates that model prediction differs from ground truth.

In order to investigate the relationship between average self-reliance and average performance of study participants in Group 1 (which received model predictions and explanations), we divided the calculations into four separate groups, considering that the results may vary significantly in the case of incorrect predictions. Our study showed that the lowest level of self-reliance was associated with the highest accuracy, while the highest accuracy was associated with self-reliance at a level of 45% (see: Table 3).

	Accuracy	**Self-reliance**
FP	0.05	0.45
FN	0.5	0.73
TP	0.91	0.3
TN	0.75	0.39

Table 3: Average accuracy and self-reliance for group 1

4. Conclusions and future work

We study the influence of the big five personality types on models. We introduce the self-reliance metric along with emphasizing its influence on study participants, especially in cases when the model is wrong. We see the following directions for future work:

- **Fundamental analysis on the influence of different explanation types on participants**. This refers to among others introducing global explanations or hiding model scores by leaving only binary predictions.
- **Increase scale of experiment**. More people would allow us to establish more reliable results along with introducing more groups.
- **Incorporate additional group of domain experts**. This would verify which explanations and how influence participations' choices.

References

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2788613. URL https://doi.org/10.1145/2783258.2788613.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

Jian Huang, Junyi Chai, and Stella Cho. Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):1–24, 2020.

Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing, and Niklas Kühl. An empirical evaluation of predicted outcomes as explanations in human-ai decision-making, 2022. URL https://arxiv.org/abs/2208.04181.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *CoRR*, abs/1909.09223, 2019. URL http://arxiv.org/abs/1909.09223.

Beatrice Rammstedt and Oliver P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212, 2007. ISSN 0092-6566. doi: https://doi.org/10.1016/j.jrp.2006.02.001. URL https://www.sciencedirect.com/science/article/pii/S0092656606000195.

Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. *Nucleic acids research*, 48(W1):W449–W454, 2020.

Harry Surden. Machine learning and law. *Wash. L. Rev.*, 89:87, 2014.

Appendix A. Detailed description of cases from the study

age	race	gender	country	martial Status	rship.	work class	occ.	yrs. edu.	hrs of work	cap. gain	cap. loss	ground truth	model's pred.	conf.
60	white	male	USA	married civ spouse	husband	other	exec. mgmt.	10	40	0	0	below 50k	above 50k	59.2%
32	other	male	other	married civ spouse	husband	private	prof speciality	14	40	0	0	below 50k	above 50k	57.6%
38	other	male	USA	married civ spouse	husband	private	prof speciality	13	70	0	0	above 50k	below 50k	50.8%
44	white	male	USA	married civ spouse	husband	private	other	10	60	0	0	above 50k	below 50k	52.3%
58	white	male	other	married civ spouse	husband	private	exec. mgmt.	11	40	0	0	above 50k	above 50k	61.4%
21	white	female	USA	never married	not in family	private	exec. mgmt.	10	40	99999	0	above 50k	above 50k	78.7%
59	white	male	USA	divorced	not in family	self emp	exec. mgmt.	9	60	0	0	below 50k	below 50k	67.9%
27	other	male	USA	never married	not in family	private	exec. mgmt.	10	40	0	0	below 50k	below 50k	98.3%

Table 4: Detailed descriptions of cases used in the study with ground truth and model prediction. Additional legend for abbreviations in the table: relationship (rship.), occupation (occ.), years of education (yrs. edu.), confidence of the prediction (conf.), executive managerial (exec. mgmt.).

Appendix B. Detailed description of features

Feature	**Range or possible values**
Age	17 to 90
Race	White or other
Gender	Female or Male
Native Country	Few most popular like USA, France, etc.
Martial status	Married civilian spouse, Divorced, Never-married, Separated, etc.
Relationship	Represents what this individual is relative to others. For example, an individual could be a husband. Each entry only has one relationship attribute and is somewhat redundant with marital status.
Work class	Private, Self-employed, Other
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Years of Education	1 to 16
Hours of work per week	Average Weekly Hours in the United States averaged 34.40 Hours from 2006 until 2022
Capital gain	
Capital loss	

Table 5: Features and its description provided in the study.

Appendix C. Local explanations of cases used in the study

The enumeration of people is consistent with the row's numeration in Appendix A.

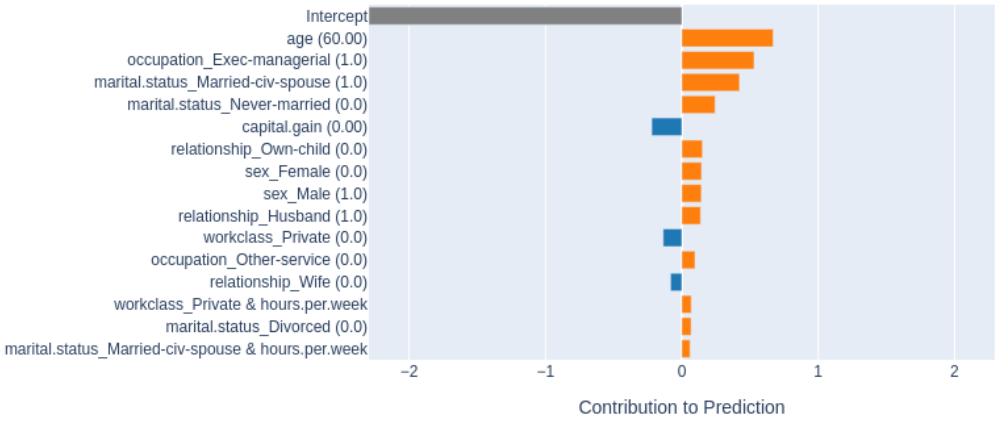

Figure 1: Local explanation of prediction for person 1.

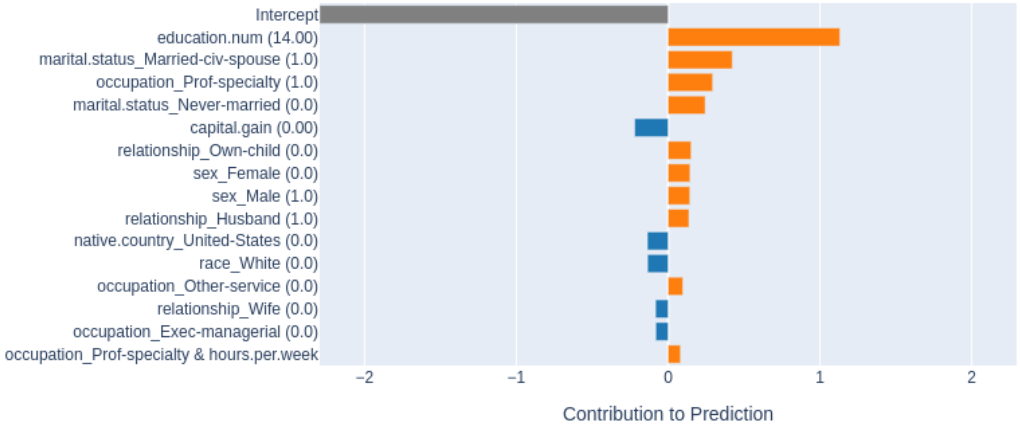

Figure 2: Local explanation of prediction for person 2.

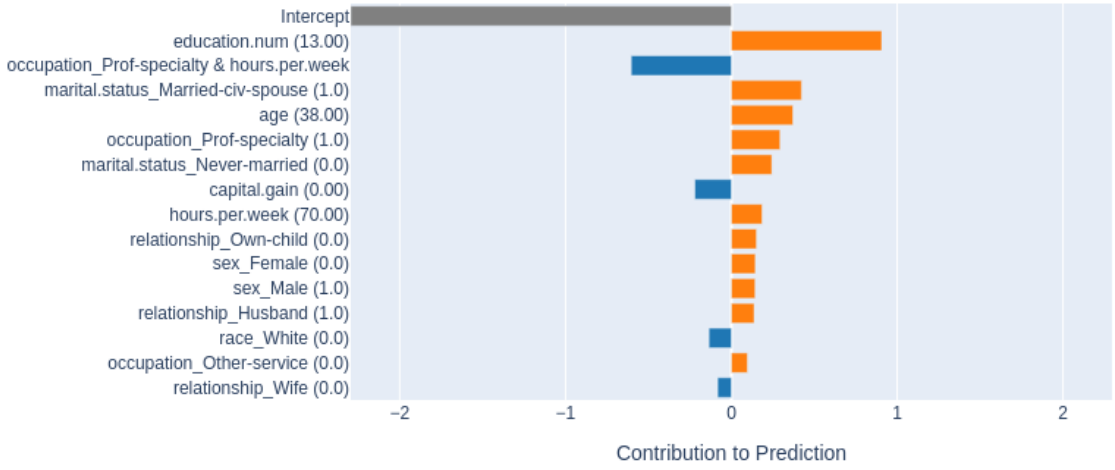

Figure 3: Local explanation of prediction for person 3.

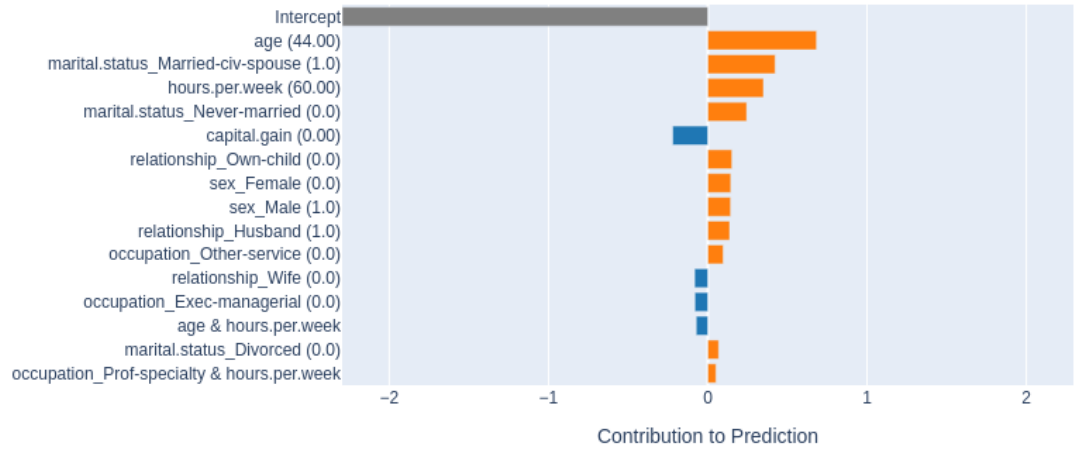

Figure 4: Local explanation of prediction for person 4.

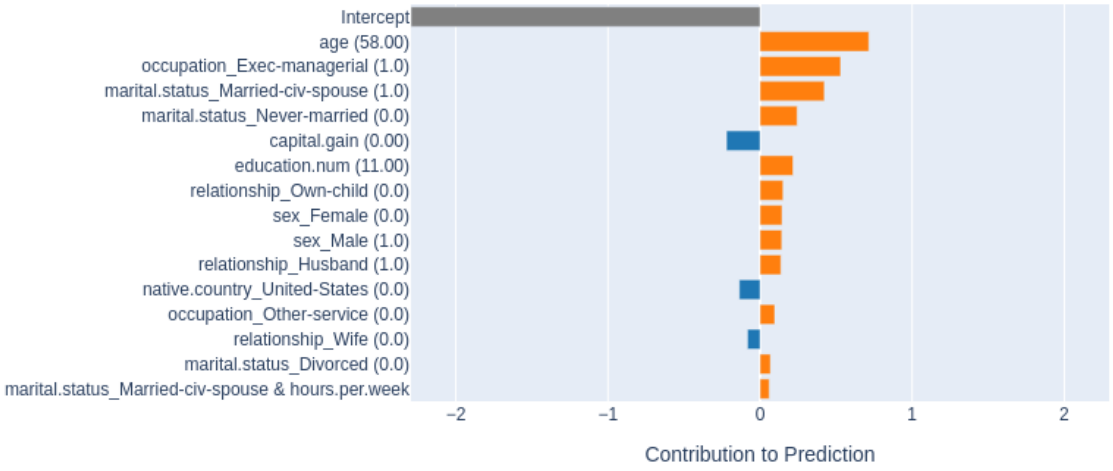

Figure 5: Local explanation of prediction for person 5.

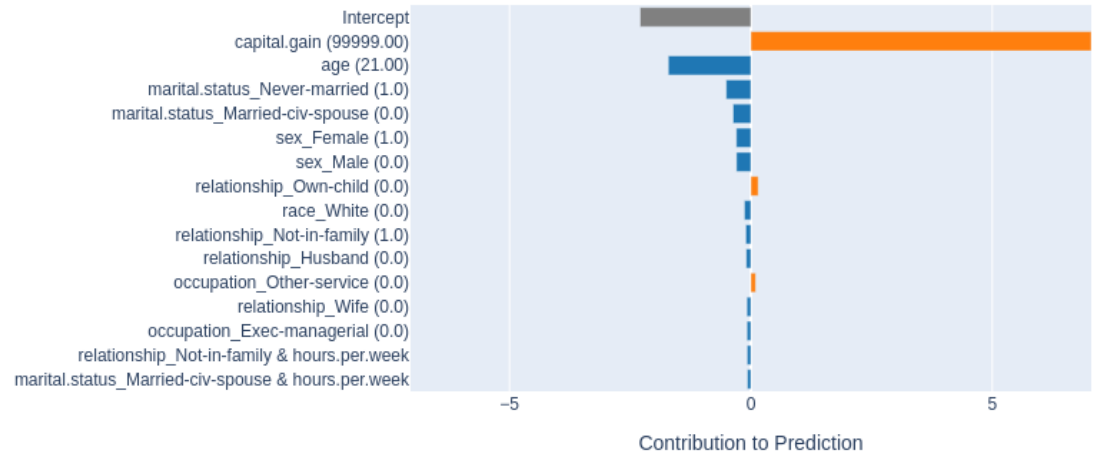

Figure 6: Local explanation of prediction for person 6.

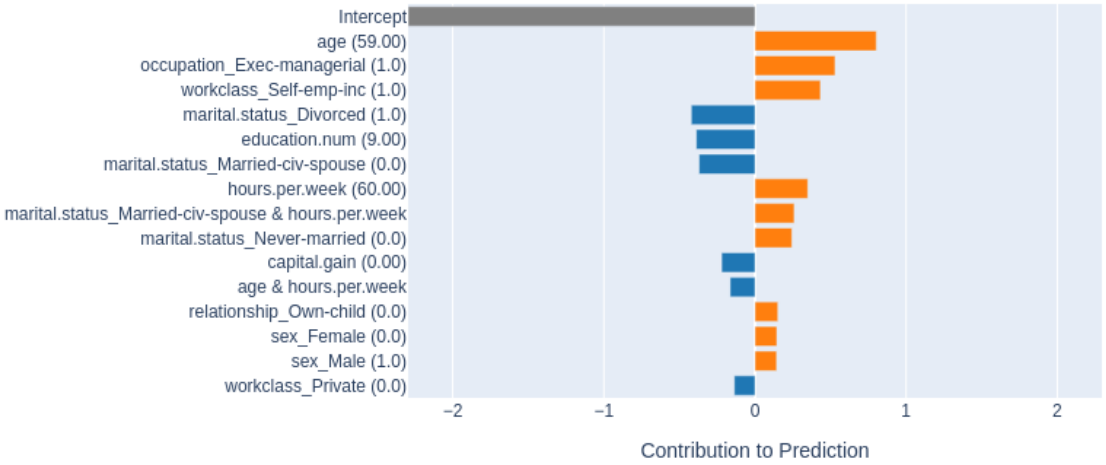

Figure 7: Local explanation of prediction for person 7.

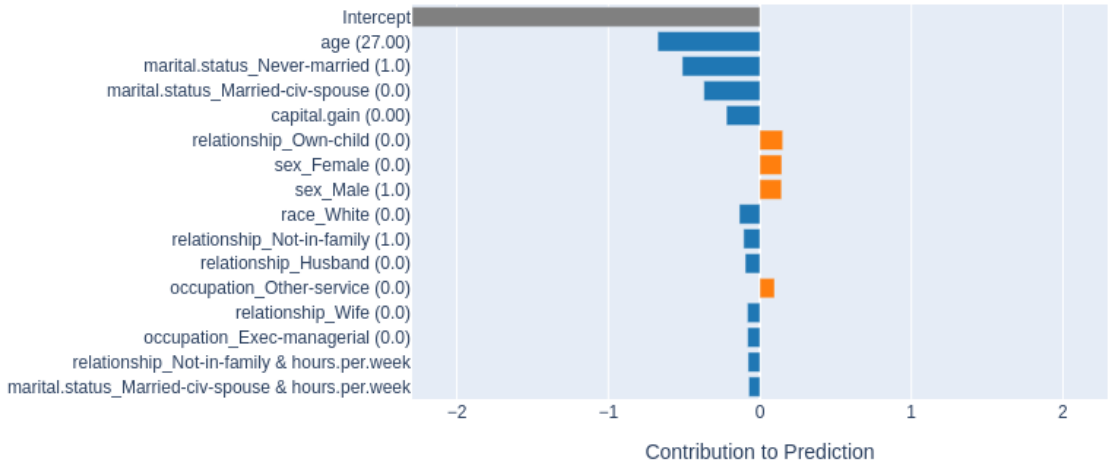

Figure 8: Local explanation of prediction for person 8.

Appendix D. Relations between personality types

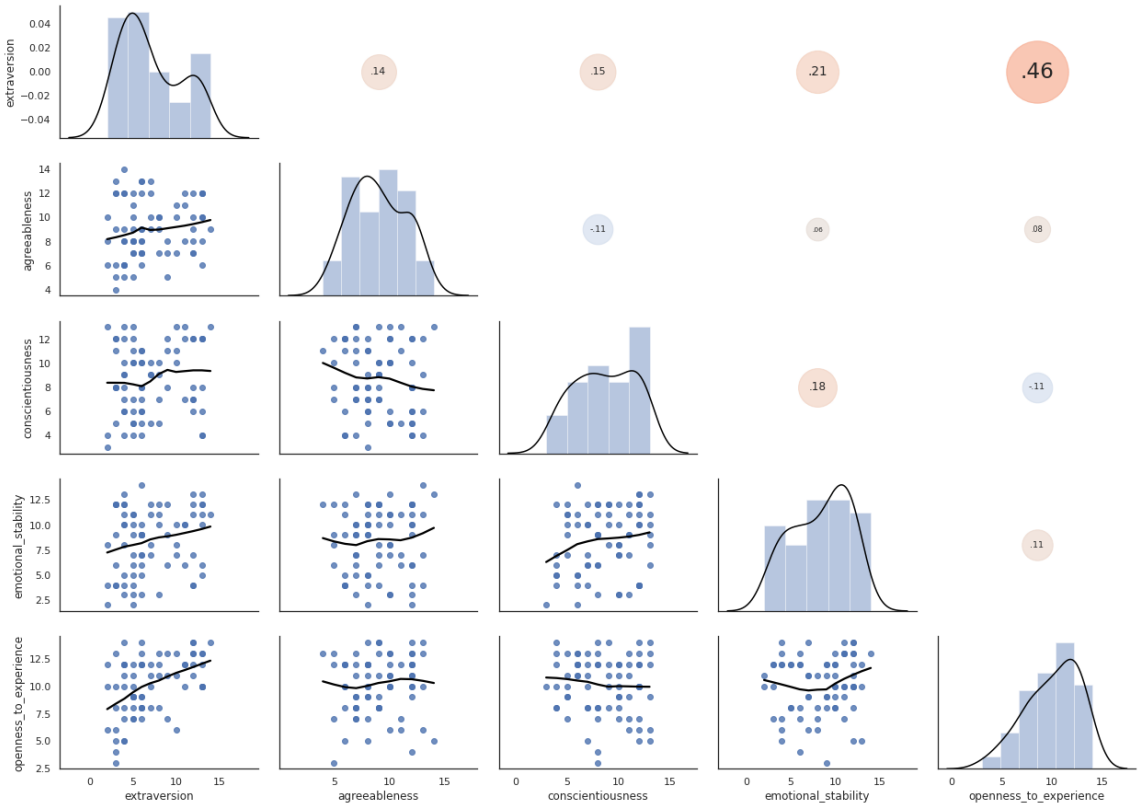

Figure 9: Correlation plot between personality types

In the above plot:

- The distribution of each variable is shown on the diagonal.
- On the bottom of the diagonal : the bivariate scatter plots with a fitted line are displayed.
- On the top of the diagonal : the value of the correlation plus the significance level circle size.

Appendix E. Additional plots with statistics

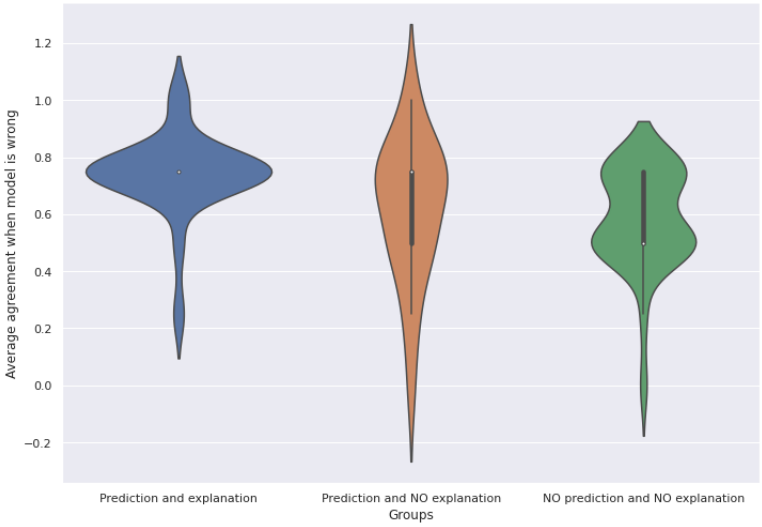

Figure 10: Average reliance when model is wrong.

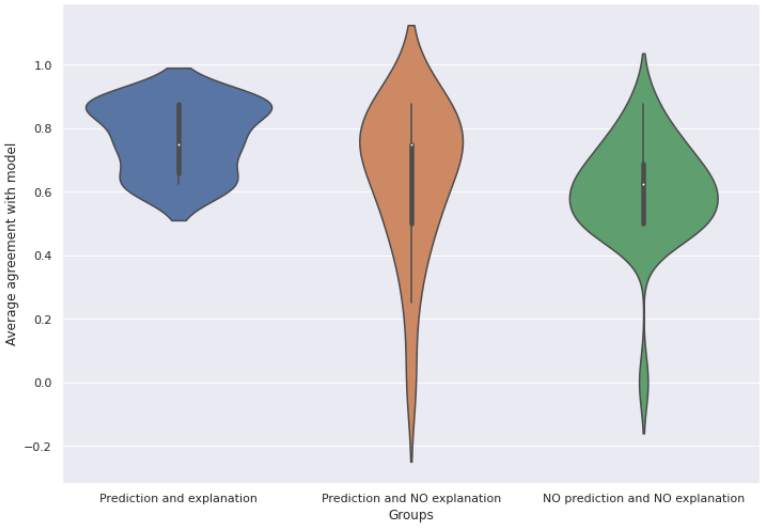

Figure 11: Average reliance when model is right.

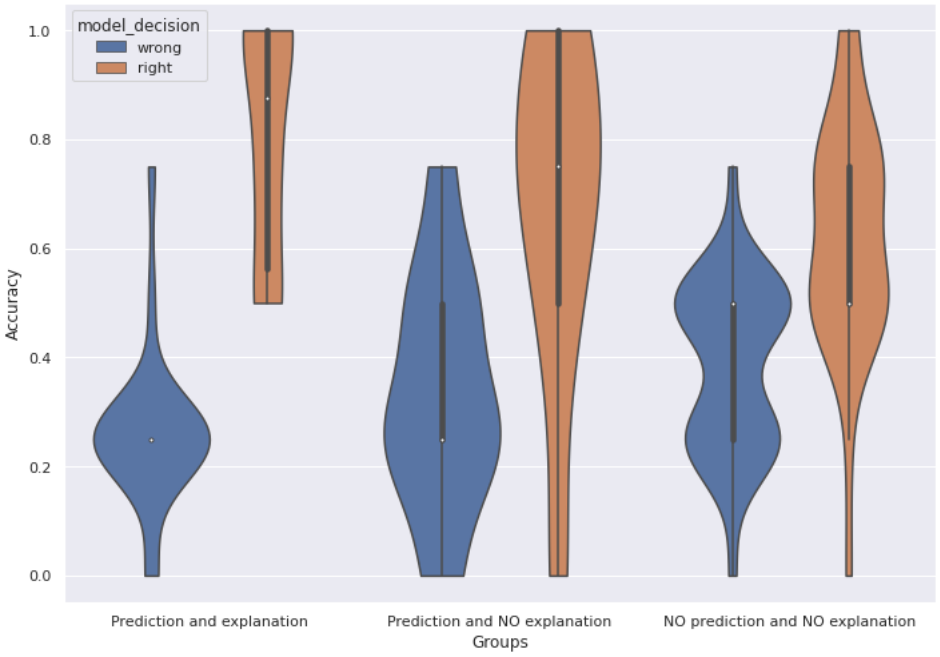

Figure 12: Accuracy of participants when the model prediction was right and wrong.

Appendix F. Participant Info

Gender	Age	Education level	ML Level	Group
Male	25	Bachelor's or equivalent	2	2
Male	22	Bachelor's or equivalent	3	2
Male	22	Bachelor's or equivalent	2	2
Male	27	Master's or equivalent	1	2
Female	21	Bachelor's or equivalent	1	2
Male	18	Bachelor's or equivalent	2	2
Male	19	Bachelor's or equivalent	1	2
Male	23	Master's or equivalent	3	2
Male	21	Bachelor's or equivalent	1	2
Male	24	Master's or equivalent	3	2
Female	25	Doctorate or equivalent	2	2
Male	22	Bachelor's or equivalent	0	2
Male	27	Doctorate or equivalent	4	2
Male	25	Master's or equivalent	2	2
Other	35	Bachelor's or equivalent	1	2
Male	20	Secondary education (High School or equivalent)	1	2
Male	23	Master's or equivalent	2	2
Male	22	Master's or equivalent	1	2
Female	21	Bachelor's or equivalent	1	2
Male	19	Doctorate or equivalent	4	2
Male	25	Master's or equivalent	2	3
Male	22	Other	2	3
Female	21	Bachelor's or equivalent	0	3
Male	21	Bachelor's or equivalent	0	3
Male	21	Bachelor's or equivalent	3	3
Male	23	Master's or equivalent	3	3
Male	21	Bachelor's or equivalent	1	3
Female	21	Secondary education (High School or equivalent)	0	3
Male	24	Master's or equivalent	3	3
Male	21	Bachelor's or equivalent	1	3
Female	19	Bachelor's or equivalent	0	3
Male	25	Master's or equivalent	0	3
Prefer not to say	24	Bachelor's or equivalent	2	3
Male	21	Bachelor's or equivalent	0	3
Male	24	Master's or equivalent	3	3
Male	23	Master's or equivalent	1	3
Female	21	Secondary education (High School or equivalent)	1	3
Female	23	Bachelor's or equivalent	1	3
Female	20	Bachelor's or equivalent	0	3
Male	23	Master's or equivalent	3	3
Male	23	Bachelor's or equivalent	1	3
Male	21	Bachelor's or equivalent	0	3
Female	19	Bachelor's or equivalent	0	3

Gender	Age	Education level	ML Level	Group
Male	19	Secondary education (High School or equivalent)	0	3
Male	24	Master's or equivalent	2	3
Female	21	Bachelor's or equivalent	2	3
Female	20	Bachelor's or equivalent	0	3
Male	21	Bachelor's or equivalent	1	3
Female	19	Bachelor's or equivalent	1	3
Male	21	Bachelor's or equivalent	1	3
Female	21	Bachelor's or equivalent	2	1
Male	20	Bachelor's or equivalent	1	1
Male	19	Bachelor's or equivalent	0	1
Male	23	Bachelor's or equivalent	1	1
Male	23	Master's or equivalent	2	1
Male	23	Master's or equivalent	0	1
Male	23	Master's or equivalent	1	1
Female	23	Master's or equivalent	3	1
Male	31	Master's or equivalent	4	1
Male	24	Bachelor's or equivalent	2	1
Female	25	Master's or equivalent	3	1
Female	21	Bachelor's or equivalent	1	1
Female	20	Bachelor's or equivalent	1	1
Male	20	Bachelor's or equivalent	0	1
Male	24	Master's or equivalent	1	1
Male	23	Master's or equivalent	4	1
Male	23	Bachelor's or equivalent	2	1
Female	20	Bachelor's or equivalent	0	1
Male	20	Secondary education (High School or equivalent)	2	1
Male	20	Bachelor's or equivalent	1	1
Male	27	Doctorate or equivalent	1	1
Male	22	Bachelor's or equivalent	1	1

Table 6: Table of participant information. For ML knowledge, participants rated themselves using the following scale: 0 - "No experience", 2 - "Project/Comptetition", 4 - "ML publications".