Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias

Philipp Schmidt[1(✉)] and Felix Biessmann[1,2]

[1] Amazon Research, Berlin, Germany
{phschmid,biessman}@amazon.com
[2] Einstein Center Digital Future, Berlin, Germany
felix.biessmann@beuth-hochschule.de

Abstract. Transparent Machine Learning (ML) is often argued to increase trust into predictions of algorithms however the growth of new interpretability approaches is not accompanied by a growth in studies investigating how interaction of humans and Artificial Intelligence (AI) systems benefits from transparency. The right level of transparency can increase trust in an AI system, while inappropriate levels of transparency can lead to algorithmic bias. In this study we demonstrate that depending on certain personality traits, humans exhibit different susceptibilities for algorithmic bias. Our main finding is that susceptibility to algorithmic bias significantly depends on annotators' affinity to risk. These findings help to shed light on the previously underrepresented role of human personality in human-AI interaction. We believe that taking these aspects into account when building transparent AI systems can help to ensure more responsible usage of AI systems.

Keywords: Transparent AI · Machine learning · HCI · Risk affinity

1 Introduction

Decision support by Machine Learning (ML) models has become ubiquitous in everyday life. Responsible usage of such assistive technology requires an appropriate level of trust into ML systems [7]. Trust into this technology is often argued to require interpretability of ML predictions. This is why transparent ML methods have become an active focus of research. A central research question in this field is how methodological advances translate into improvements in human-AI interaction and trust. Transparent ML research aims at better explaining ML predictions and systems to humans, but it is difficult to assess how much human-AI interaction profits from scientific advances in the field. Many studies proposing new interpretable ML methods do report results with human-in-the-loop experiments to test the impact of their proposed method on

P. Schmidt and F. Biessmann—Equal contribution.

Published by Springer Nature Switzerland AG 2020
A. Holzinger et al. (Eds.): CD-MAKE 2020, LNCS 12279, pp. 431–449, 2020.
https://doi.org/10.1007/978-3-030-57321-8_24

human-AI collaboration [26,37]; and there are approaches to compare the quality interpretability methods in psychophysical experiments with humans [43]. Yet there appears to be no consensus in the community on how the impact of interpretability methods on human-AI interaction should be evaluated [14].

So while there has been substantial growth in the research field of transparent AI, studies on the impact of transparency in AI systems on human cognition remain underrepresented. This development is at odds with the increased usage of ML systems as decision support systems. In many societies humans interact more often with computer systems than with other humans and the importance of a better understanding of machine behaviour and human machine interaction is widely recognized [36]. When humans interact with ML systems as often as with machines, studying the interdependence of machine behaviour and human behaviour becomes increasingly important. Everyday work in many professions with high responsibility employs assistive ML technology, including doctors, policemen and judges. To what extent these professions can profit from increased transparency or are negatively affected by algorithmic bias, here referring to humans trusting an AI too much, is an open research question. Algorithmic bias can have severe negative consequences, for instance when policemen trust biased ML predictions too much. Similarly, algorithm aversion can have devastating consequences, for example when experienced doctors ignore correct ML predictions. We argue that the danger of algorithmic bias or algorithm aversion should motivate not only calibrating the transparency level of AI systems. An equally important focus of research is the calibration of humans for responsible usage of transparent AI systems. The goal of this study is to highlight the potential of existing psychological research for a better calibration of human-AI interaction by taking into account personality traits. Our working hypothesis is that adding transparency does not have the same effect on users of an AI system, depending on their personality. We employ concepts from psychological research to investigate how different personality characteristics influence the impact of transparency in AI assisted decision making.

Human decision making in a collaborative context has been studied extensively in the psychological literature. One major focus of this research is on personality traits that are related to how tolerant subjects are with respect to risk and ambiguity. Risk affinity or aversion is studied in decisions in which subjects have access to the odds of a certain outcome. Tolerance to ambiguity is studied in decisions in which subjects to not have access to the probabilities associated with each outcome in a decision. Being optmistic in presence of ambiguity has recently been reported to be an important personality trait for trust. For instance the authors of [50] show that tolerance to ambiguity reliably predicts prosocial behaviour. We hypothesize that such behavioural traits are also involved in human-AI collaboration. We propose to leverage this research to better understand how humans interact with AI systems. Our results indicate that not only in human interaction but also in the context of assistive AI technology, these two factors, tolerance of risky and ambiguous uncertainty, play an important role. After all, most attempts to render AI systems more transparent

can be motivated to reduce these very aspects: Transparency of an AI system is often increased by exposing the model likelihoods for an uncertain decision, or by showing explanations. Users with high or low affinity to risk and ambiguity could benefit differently from these transparency aspects.

Our study combines established methods from experimental psychology to determine personality traits indicative for risk and ambiguity tolerance with experiments that investigate the susceptibility to algorithmic bias of subjects. This combination allows to study the influence of these personality traits on the impact of transparency induced algorithmic bias. More importantly our findings can be directly applied to real world application scenarios of human-AI interaction in which responsible usage of AI systems is of paramount importance, such as policing, jurisdiction, medicine and many others. After illustrating cases of algorithmic bias and the differential susceptibility of humans, depending on their personality traits, we derive guidelines that help to reduce detrimental algorithmic bias and to maximise the benefits of transparent assistive AI.

2 Related Work

The work on interpretability of ML models has become a central topic of research in both theoretical aspects of statistical learning as well as applied ML. Many of the relevant publications at major ML conferences and dedicated workshops can be broadly categorized in more conceptual contributions or position papers and technical contributions to the field of interpretability.

In the category of position papers, an important aspect dealt with in [17] is the question of how we balance our concerns for transparency and ethics with our desire for interpretability. Herman points out the dilemma in interpretability research: there is a tradeoff between explaining a model's decision *faithfully* and *in a way that humans easily understand*. Interpreting ML decisions in an accessible manner for humans is also referred to as *simulatability* [28]. A reasonable working hypothesis is that the subjective value of transparency is an important factor for algorithmic bias. Studying these cases of bias is challenged by the fact that these biases can occur independently of the conscious perception of users. And even worse, these biases are likely to affect humans differently depending on their personality traits.

Intuitive comprehensibility and low cognitive friction of ML prediction explanations that is desired for explanations, and can be used as the basis of quantitative comparisons of interpretability approaches [8,29,32,43], is a two sided sword: Explanations should be comprehensible, but at the same time the negative aspects of algorithmic biases should be avoided. The need for unbiased and automated evaluation metrics for transparency in combination with human-in-the-loop experiments is widely recognized. For instance the authors of [8] highlight the necessity of *understandability* of explanations as well as the lack of consensus when it comes to evaluating interpretability of ML models. They propose an evaluation taxonomy that comprises both automated evaluations but also involves evaluations by human laymen. However, the negative and positive effects of algorithmic biases are not a central focus of that work.

Interpretability Approaches. On the AI side of human-AI interaction studies, there is a large body or work advancing the state of the art in transparent AI. The category of technical contributions can be broadly subdivided into two types of methods. First there are methods that aim at rendering specific models interpretable, such as interpretability methods for linear models [16,54] or interpretability for neural network models [33,44,53]. Second there are interpretability approaches that aim at rendering *any* model interpretable, a popular example are the *Local Interpretable Model-Agnostic Explanations* (LIME) [37]. As these latter interpretability methods work without considering the inner workings of an ML model, they are often referred to as *black box interpretability methods.* Due to the popularity of neural network models especially in the field of computer vision there have been a number of interpretability approaches specialized for that application scenario and the method of choice in this field, deep neural networks. Some prominent examples are *layerwise relevance propagation* (LRP), sensitivity analysis [44] and deconvolutions [53]. For comparing these different approaches the authors of [40] propose a greedy iterative perturbation procedure for comparing LRP, sensitivity analysis and deconvolutions. The idea is to remove features where the perturbation probability is proportional to the relevance score of each feature given by the respective interpretability method. An interesting finding in that study is that the results of interpretability comparisons can be very different depending on the metric: the authors of [13] performed an evaluation of sensitivity analysis and came to a different conclusion than [40].

The idea of using perturbations gave rise to many other interpretability approaches, such as the work on *influence functions* [5,15,24] and methods based on game theoretic insights [30,47]. In [47] evaluations are entirely qualitative; in [30] the authors compare interpretability methods by testing the overlap of explanations with human intuitions. While this approach can be considered quantitative, it is difficult to scale as it requires task specific user studies. Another metric used in that study for comparisons of evaluations is computational efficiency, which is simple to quantify, but is not directly related to interpretability. Other studies also employ user studies and comparisons with human judgements of feature importance. An interesting approach is taken in [38] in which the authors let students of an ML class guess what a model would have predicted for a given instance when provided with an explanation. Similarly the authors of [26] perform user studies in which they present rules generated by an interpretability method and measure how good and how fast students can replicate model predictions. This approach has been taken also in [18] in which the speed and accuracy are measured with which humans can replicate model decisions. Overall many studies investigate the impact of transparent AI on human-AI collaboration. Yet the details of when an explanation leads to detrimental algorithmic bias and the personality traits governing susceptibility to algorithmic bias remain underrepresented. We build our work on the ideas put forward in the above studies, but we place a special focus on the factors determining algorithmic bias in both ML models and human personality.

Decision Making Under Uncertainty. In almost every human-AI collaboration the final decision is made by humans. For this reason it is of utmost importance to understand how humans decide. Human decision making [3] is studied in a number of different fields such as mathematics [12,42], behavioural economics [20,48] and psychology [46]. The existing literature can be divided into work that focuses on normative theory, i.e., how decisions should be made with logical consistency, descriptive models, i.e., how people make decisions, and prescriptive approaches that try to help make people better decisions. Further, decision making problems can be divided by whether a decision maker has access to outcome probabilities, i.e., a decision task associated with risky uncertainty, or is lacking such information, i.e., a decision problem with ambiguous uncertainty. Depending on whether uncertainty is risky or ambiguous [23], humans tend to exhibit different behaviour, generally favoring risky options over ambiguous ones [4,6,45]

Expected utility theory [51] attempts to explain human choice behaviour using utility functions that take into account the objective value of a choice. Prior work has recognized the discrepancy between how people should choose under an expected value maximization principle and their actual choices [39]. It has been found that choices might deviate from the optimal one for a number of reasons, including risk aversion [1,19,34]. A widely known example of this is the St. Petersburg paradox, where players are reluctant to bet on a game with infinite expected value [39]. Further work in this field introduced subjective expected utility (SEU) [41] that allows for subjective variations in decision making behaviour under risk. However the explanatory power of SEU was questioned in [2] and it was later demonstrated by the Ellsberg paradox [10] that the rationality assumptions introduced by SEU did not hold in cases where outcome probabilities are not known, i.e., in cases of ambiguous uncertainty. Ambiguity aversion and its effects were studied thoroughly in [4] and hence it was found that ambiguity sensitive behaviour cannot be explained by SEU. Follow up work then explicitly developed models to explain decision behaviour under ambiguity [9,12,22,42].

In Psychology risk and ambiguity tolerance are considered two distinct personality traits [46]. Attitudes towards risky and ambiguous uncertainty can be estimated using choice models [12], as it was done in [50] to explain pro-social behaviour. In line with these findings, individuals that exihibt ambiguity tolerance were found to be generally optimistic when faced with uncertainty [35,52].

In this work, we employ an utility model that is used to specify subjective value of an option taking into account risky and ambiguous uncertainty [12]. This model has been used in prior work to determine a subjects attitude to risky and ambiguous uncertainty [11,27,50]. We believe that human-AI interaction is inherently associated with risky and ambiguous uncertainty, regardless of whether the interaction occurs during online shopping sessions or when judges draw on algorithmic decision support for bail decisions.

3 Experiment

In order to investigate the impact of levels of transparency and its relationship to risk and ambiguity affinity we ran an annotation task on the crowdworking platform Amazon Mechanical Turk[1]. The annotation task was based on a text classification task, a binary sentiment classification task of IMDb reviews. A ML model was trained on the task and its predictions and explanations thereof were exposed to the annotators. We varied the level of transparency and measured the effect on the agreement between annotators' responses and a) ground truth labels and b) model predictions. In order to examine how risk and ambiguity affinity relates to trust in an AI system with varying levels of transparency we first determined the affinity of annotators to risk and ambiguity in their decisions, using a well established psychological task that involves playing a lottery with known odds (risk) or with unknown odds (ambiguity) [11,27,49,50].

3.1 Risk and Ambiguity Affinity Experiment

The first part of the experiment was an incentivized gambling task for which the experiment participants had to choose between a safe payout and an uncertain monetary option. This type of task is known to yield reliable estimates for subjective risk and ambiguity attitudes. Across trials, participants were exposed to different levels of risk (25%, 50% and 75%) and ambiguity (24%, 50% and 74%). In half of the trials, participants were exposed to risky gambling options, where the winning odds were fully observable. The remaining trials had varying levels of ambiguity associated to them, where the winning probabilities could not be fully observed. See Fig. 1 for a depiction of a risky and ambiguous trial respectively. All of the ambiguous trials had an objective winning probability of 50%. The safe payout option had a monetary value of \$0.25 whereas the uncertain, risky and ambiguous, trials were allocated to equally represent monetary values of \$0.25, \$0.4, \$1.0, \$2.5 and \$6.25.

We have adopted a subjective utility function based on [12], as it has been defined in [50] as

$$\mathrm{SV}\left(p, A, v\right) = \left(p - \beta\frac{A}{2}\right) v^{\alpha} \tag{1}$$

where the subjective value (SV) is calculated as a function of three parameters, the objective winning probability (p), the level of ambiguity (A) and the monetary amount v in that trial. The two free parameters α and β indicate a subjects sensitivity to risk and ambiguity respectively. We modelled choice of the lottery, i.e., whether a subject chose the variable payout option on a given trial as

$$P\left(\text{play lottery}\right) = \frac{1}{1 + e^{\gamma(\mathrm{SV}_F - \mathrm{SV}_V)}}. \tag{2}$$

[1] https://www.mturk.com/.

The fixed monetary option is referred to as SV_F and the variable, risky or ambiguous, option is referred to as SV_V. For each subject we estimated three parameters; α, β and a slope of the logistic function γ using maximum likelihood given choice data that was collected as part of the experiment.

For subjects that are unaffected by ambiguity in their gambling choices, ambiguity sensitivity β will be 0. A positive β indicates ambiguity aversion, i.e., the winning probability is perceived as being less than 50% in ambiguous trials. On the other hand, a negative β indicates that a subject perceives the winning probability to be more than 50% in ambiguous trials. Risk neutral subjects will have an α of 1. A risk averse subject will have an α smaller than 1, and for a risk tolerant subject α is greater than 1.

Overall, the subjective value function explained participants choice behaviour extremely well, in 83.5% of trials across all subjects the model predicted choice correctly. According to the model fit, 41.5% of experiment participants were identified as being risk and ambiguity averse; the second largest group (31.8%) was tolerant to both risky and ambiguous uncertainty.

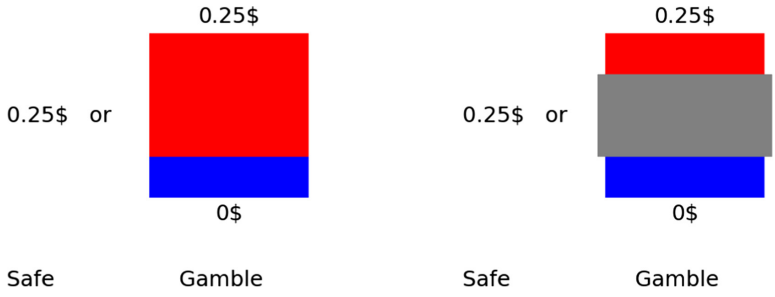

Fig. 1. User interface of the experiment to assess risk affinity and ambiguity tolerance. *Left:* Risk condition, users know the odds of winning the lottery. *Right:* Ambiguity tolerance condition, subjects do not know the odds, the winning probability in these trials was always 50%.

3.2 Annotation Task Experiment

In the second part of the experiment, we asked participants to annotate binary sentiments for 50 movie reviews from IMDb. Depending on the condition, participants were exposed to different levels of algorithmic transparency, see Fig. 2 for an exemplary depiction of the experimental conditions.

3.3 Data Set

The task uses the publicly available IMDb movie review sentiment dataset[2] which was introduced in [31]. The IMDb rating scale is defined from one to

[2] https://www.imdb.com/conditions.

ten stars where all reviews that have less than five stars are considered to have negative sentiment and all reviews that have more than six stars positive. For more controlled experimental conditions we further reduced the complexity of the dataset: We subsampled the full dataset to 50 movie reviews and controlled for various factors in that sample. Reviews were selected such that they were all between 400 and 1000 characters long to ensure comparable cognitive load for all reviews. Positive and negative reviews occurred equally often (25 times each) and had varying degrees of (relative) ease. This assessment of task ease is based on prior work in which subjects classified reviews without support; these values range between 0.5 and 1.0, indicating the fraction of subjects classifying the reviews correctly. In order to avoid learning and adoption effects, the true labels were never revealed throughout the experiment. We selected a set of reviews for which the ML model's classification accuracy was 80%. Moreover, this accuracy was symmetrical across positive and negative reviews. Given this design, all other conventional performance measures such as precision, recall, and specificity amount to 80% as well. Previous research on IMDb review classification found that typical human accuracy ranges between 75% and 80% on similar samples of the same data [43]. All participants were exposed to the same reviews.

Table 1. Held-out per label precision/recall/f1 scores of the ML model used for comparing ML interpretability methods on the full IMDb test dataset. In the annotation task a subsample of this full data set was used, in which several variables were controlled for to ensure comparable cognitive load across data points.

Sentiment	Precision	Recall	f1-score	Support
Negative	0.88	0.87	0.87	12500
Positive	0.87	0.88	0.87	12500
Avg/total	0.87	0.87	0.87	25000

3.4 Machine Learning Model

The ML model was trained on the complete training dataset which consists of 25000 movie reviews and achieved precision/recall/f1 metrics close to 90% on the test dataset, see Table 1. In all experiments we used unigram bag-of-words features that was term-frequency inverse document frequency normalized. English stopwords were removed prior to the feature extraction. Bag-of-words feature vectors $\boldsymbol{x} \in \mathbb{R}^d$, where d denotes the number of unigram features, were used to train an L_2 regularized multinomial logistic regression model. Let $y \in \{1, 2, \ldots, K\}$ be the true label, where K is the total number of labels and $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}^{d \times K}$ is the concatenation of the weight vectors $\boldsymbol{w}_k \in \mathbb{R}^d$ associated with the kth class then

$$p(y = k|\boldsymbol{x}, \boldsymbol{W}) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \qquad \text{with } z_k = \boldsymbol{w}_k^\top \boldsymbol{x} \tag{3}$$

We estimated $\boldsymbol{W}$ using stochastic gradient descent (SGD) using the python library `sklearn` and used a regularization parameter of 0.0001, other values for the regularizer lead to similar model performance.

3.5 Interpretability Method

A simple and efficient approach for rendering linear models interpretable is provided in [16]. While this approach is limited to linear models, it is a special case of the feature importance ranking measure (FIRM) [54] that can be applied to arbitrary non-linear models and there are other non-linear extensions [21]. Following Eq. (6) in [16] we obtain the feature importances for each class k separately by

$$\boldsymbol{a}_k = \boldsymbol{X}^\top \hat{\boldsymbol{y}}_{\boldsymbol{k}} \tag{4}$$

where the matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ denotes the N samples in the held out test data set and the and d denotes the number of unigram features extracted from the training data set. The predictions of the model for class k on the test data are denoted $\hat{\boldsymbol{y}}_{\boldsymbol{k}} \in \mathbb{R}^{N \times 1}$. Each dimension of $\boldsymbol{a}_k \in \mathbb{R}^d$ is associated with a feature, in our case a word in the unigram dictionary. To compute the explanations, i.e. the highlighted words for sample $\boldsymbol{x}_i$, we selected the feature importances $\boldsymbol{a}_k$ associated with the most likely predicted class k under the model and ranked the words in a text according to the element-wise product of features $\boldsymbol{x}_i$ and feature/prediction covariances $\boldsymbol{a}_k$. The highlighted words were those that were present in the text and scored high in terms of their covariance between features and model predictions. The use of this interpretability method was motivated by its general applicability, by its simplicity and by its speed. Also this approach was found to be superior to other interpretability methods including LIME [37] for this particular combination of data set and ML model [43].

3.6 Experimental Setup

All user study experiments were run on Mechanical Turk where we asked annotators to provide the correct sentiment of 50 movie reviews. The annotation user interface (UI) is shown in Fig. 2. In total we collected annotations from 248 distinct workers. For each worker we selected one out of three levels of transparency:

- **control**: only the move review was shown
- **highlights**: movie reviews were shown along with the ML model prediction and the top 3 words according to interpretability score in Eq. 4 were highlighted
- **highlights with confidence**: same as *highlights* condition but with the likelihood for the predicted class as in Eq. 3.

Fig. 2. Annotation user interface for the AI-assisted IMDb movie review sentiment classification experiment. Three transparency levels were examined, shown is only the highest transparency level where the ML prediction with the model confidence was shown along with the explanation in form of highlighted words; the medium transparency level did not show the model confidence and the lowest level of transparency only showed the review text.

Each experimental subject was exposed to a transparency condition drawn uniformly at random. To control for effects related to the number of words highlighted we kept the number of words highlighted fixed to three words in each text, samples with more words highlighted (e.g. due to duplicate words) were discarded. For each annotation we recorded the annotation time, the experimental condition, the true label and the label provided by the annotator.

4 Results

We analysed the effects of increased transparency of ML predictions on human annotators performance. In particular, we investigated the agreement of human annotators with the ground truth labels and the ML predictions. Analysing the effects of transparency on annotators' agreement with the model predictions allows to investigate algorithmic bias of annotators, meaning the overlap of annotators' predictions and the predictions of the ML model. Studying the impact of transparency on human annotators' agreement with ground truth labels allows to differentiate cases of beneficial or detrimental algorithmic bias. These effects were then analysed with respect to the risk and ambiguity affinity of each annotator.

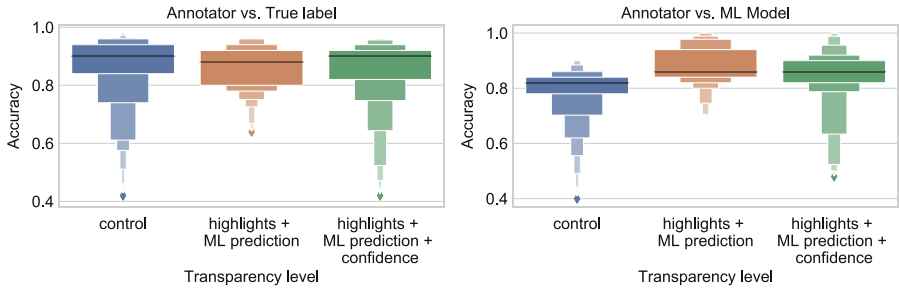

Fig. 3. Agreement between annotations by humans and ground truth (*left*) and ML model predictions (*right*) at increasing levels of transparency. Adding transparency significantly increased overlap between human annotations and ML predictions. But overlap with ground truth labels is not increased.

4.1 Transparency Increases Algorithmic Bias

When increasing transparency of ML predictions the annotators' agreement with the ground truth labels was not affected significantly. As shown in Fig. 3, there seems to be a slight decrease in human annotation accuracy when explanations and the model prediction were shown. In contrast, the annotators' agreement with the ML model predictions are increased significantly when adding model transparency. In particular, compared to the control group, algorithmic bias was increased in the first treatment group where word highlights and the model prediction were shown (Mann-Whitney, $p < 0.001$, bonferroni-adjusted). Similarly, algorithmic bias was also significantly increased in the second treatment group where in addition to the transparency from the first treatment group also model confidence scores were shown (Mann-Whitney, $p < 0.001$, bonferroni-adjusted). These results show that transparency mainly leads to algorithmic bias in our setting, but does not improve the performance of the annotators significantly.

Transparency Biases Annotators to Wrong ML Predictions. In order to better understand these effects we divided the data into cases when the ML prediction was correct and when the ML prediction was wrong. The results shown in Fig. 4 indicate that the effects observed in the overall aggregates in Fig. 3 can be mainly attributed to those cases when the AI was wrong. When the ML prediction was wrong, adding transparency led to a significant decrease in annotators agreement with the ground truth (Kruskal Wallis, $p < 0.001$). In contrast, there was a significant increase in the agreement of annotators with the ML model when the ML prediction was wrong (Kruskal Wallis, $p < 0.001$). When the ML prediction was correct, the effect of adding transparency was not significant.

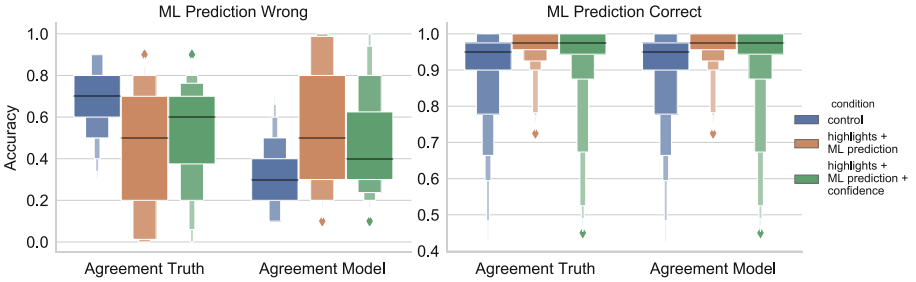

Fig. 4. Agreement between annotations by humans and ground truth or ML model predictions, respectively, when the ML prediction was *wrong* (*left*) and when the ML prediction was correct (*right*) at increasing levels of transparency. When the ML prediction was wrong, transparency decreases annotators' agreement with ground truth but increases their agreement with the ML prediction.

4.2 Algorithmic Bias and Model Uncertainty

While intuitive, these insights from conditioning on whether the model was right or wrong are difficult to translate in real world scenario advices: in most real world applications the true labels are not known. What often is known is the uncertainty of a ML prediction. We investigated how algorithmic bias due to transparency is influenced by model uncertainty. In some cases, when the model has high accuracy and its uncertainty estimates are well calibrated, this quantity is closely related to the accuracy of a model, but that is not necessarily the case. In order to condition on model uncertainty we first computed the entropy of the binary classifier for each data point

$$-p(\text{pos})\log(p(\text{pos})) - p(\text{neg})\log(p(\text{neg})) \tag{5}$$

where $p(\text{pos})$, $p(\text{neg})$ is the predicted likelihood of the ML model for a positive or negative review, respectively. We split the data at the median entropy of all data points into low entropy samples, for which the model was relatively certain about its prediction, and high entropy samples, for which the model was uncertain. Note that only in the condition with the highest transparency degree subjects had access to the probability score, which is in this case directly related to the entropy. In the two other conditions, the entropy of the model was not exposed.

Model Uncertainty Decreases Algorithmic Bias. The agreement of subjects with ground truth labels and model predictions split into low and high entropy samples is shown in Fig. 5. Similar to the overall effect of algorithmic bias in Fig. 3, also in this illustration we observe a pronounced algorithmic bias induced by adding transparency. The agreement between annotators and ML model increases when adding transparency, especially when only explanations and model prediction are shown. However, when conditioning on model entropy we see a strong effect of uncertainty on annotators' bias: their algorithmic bias is

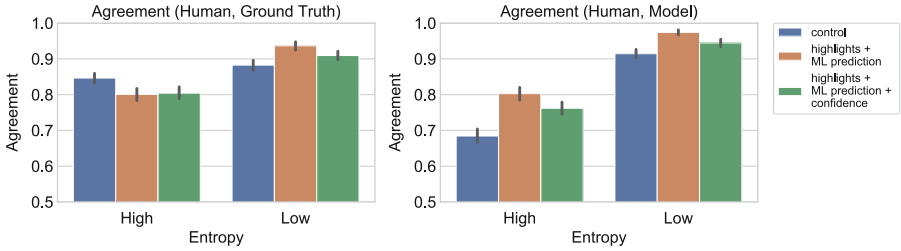

Fig. 5. Effect of model uncertainty on agreement between annotations by humans and ground truth (*left*) and ML model predictions (*right*) at increasing levels of transparency. *Left*: Algorithm transparency is associated with a decrease in annotators' agreement with true labels when predictions are uncertain, but with an increase when the ML model is certain. *Right*: Transparency induces algorithmic bias – especially when the model is certain.

reduced significantly in all transparency conditions when the model is uncertain (chi-square test of independence, $p < 0.001$).

Positive and Negative Effects of Algorithmic Bias. While the overall impact of transparency, shown in Fig. 3, is mainly algorithmic bias, there appears to be no effect on the annotators' agreement with the ground truth. Our results however suggest that when controlling for model uncertainty, transparency does have an effect on annotators' accuracy, but this effect is opposite for low and high entropy predictions of an ML model. In Fig. 5 (*left*) we observe that adding transparency *decreases* annotators' accuracy – but only for high entropy samples (chi-square test of independence, $p = 0.018$). In contrast when the model is certain, annotation accuracy is increased with transparency (chi-square test of independence, $p < 0.001$). These findings have direct implications for the calibration of human interaction with transparent AI systems. As annotators' accuracy becomes worse with added transparency when a model is uncertain, one should consider carefully exposing biasing information in such human-AI interaction when the model is uncertain. Interestingly, this is true also for conditions in which annotators had access to the model entropy. This finding is somewhat in line with the results in [25] in which the authors reported that exposing the probabilities of a prediction did not have a strong positive effect on human subjects.

4.3 Impact of Risk and Ambiguity Tolerance on Transparency Induced Algorithmic Bias

The above results confirm that transparency induces algorithmic bias, meaning annotators are more likely to replicate the ML prediction. In this section we investigate whether annotators exhibit different algorithmic bias susceptibility depending on the personality traits risk aversion and ambiguity tolerance. We partitioned the annotators into risk averse and risk affine as well as ambiguity averse and ambiguity tolerant subjects according to the coefficients in Eq. 1

which were fitted to the data obtained in the gambling experiment. Following [50] annotators with $\alpha > 1$ were classified as risk tolerant and and risk averse otherwise, annotators with $\beta < 0$ were classified as ambiguity tolerant and ambiguity averse otherwise. The histogram of annotators in each segment is shown in Table 2.

Table 2. Histogram of annotators classified as risk and ambiguity averse or tolerant

# annotators	Ambiguity averse	Ambiguity tolerant
Risk averse	103	45
Risk tolerant	21	79

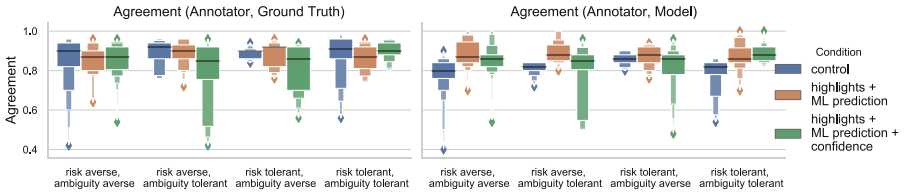

Fig. 6. Same data as in Fig. 3 but here the data is split by risk and ambiguity tolerant behaviour of annotators. When adding transparency, risk and ambiguity averse annotators exhibit higher agreement with the ML model and less agreement with ground truth. Risk tolerant and ambiguity averse annotators show the opposite effect. (Color figure online)

Risk Aversion Correlates with Algorithmic Bias Susceptibilty. The results in Fig. 6 show that annotators with different risk and ambiguity behaviour are impacted differently by adding transparency to the assistive ML system. The most prominent effect is the relationship between risk aversive behaviour and algorithmic bias. Annotators that tended to be risk averse showed the most pronounced algorithmic bias, as shown in Fig. 6, *right*, blue (control) vs. orange (highlights) and green (highlights and confidence) (Mann-Whitney, $p < 0.001$, bonferroni-adjusted). Interestingly, the strongest algorithmic bias was observed for the intermediate level of transparency, when the model prediction confidence was not shown (orange). This suggests that while risk averse subjects are most susceptible to algorithmic bias induced by transparency, this tendency is alleviated when the model confidence is shown. In contrast to risk averse annotators, risk tolerant annotators were not as susceptible to transparency induced algorithmic bias, the agreement of their annotations with the ML model predictions was less affected (Mann-Whitney, $p = 0.003$) by the level of transparency.

Ambiguity Tolerance and Transparency. We also investigated the relationship between ambiguity tolerance and the impact of transparency on algorithmic bias and annotation accuracy. Ambiguity tolerance is generally associated with prosocial behaviour in human interaction [50] and we hypothesized that it could be important for human AI interaction as well. While our results suggest that there is an effect of ambiguity tolerance in combination with risk aversion on the impact on transparency (Fig. 6), these differences are not statistically significant. This suggests, that while ambiguity tolerance is an important personality trait for interactions between humans it is less important for the trust relationship in human-AI interaction. We emphasize however that when conditioning on all combinations of risk aversion and ambiguity tolerance some groups of subjects were too small for detecting a significant effect of ambiguity and risk behaviour on transparency induced algorithmic bias, see also Table 2.

5 Conclusion

Human-Machine interaction has become a central part of everyday life. A large body of literature is dedicated to improving transparency of ML systems for more responsible usage of AI systems. We believe that transparency in itself does not necessarily have positive consequences on human-AI interaction. Both parts, AI systems and human users, should be calibrated well to avoid cases of unjustified algorithmic bias, or cases of ignorance of helpful assistive-AI predictions. Optimal calibration however requires an in-depth understanding of both parties and how they interact. To the best of our knowledge the role of human personality has been underrepresented in the literature on transparent machine learning. In this study we investigated the impact of transparency added to an ML system onto human-AI interaction with a special focus on personality traits that are associated with trust. We analyzed human-AI interaction by conditioning on various aspects of the underlying ML model, such as uncertainty or correctness of a prediction, but also by conditioning on personality traits.

Our results demonstrate that transparency leads to algorithmic bias in human-AI interaction. Extending previous work, we find that both model correctness and model uncertainty have an effect on algorithmic bias. In particular we find that transparency and the induced algorithmic bias can lead to worse annotation accuracy when the ML model prediction is uncertain or wrong – but we also find that algorithmic bias can lead to increased annotation accuracy when the model is certain or correct in its predictions. This finding has direct implications for practical applications: for uncertain model predictions, transparency should be used with care to avoid detrimental algorithmic biases. This is especially important as many cases, when unjustified algorithmic bias can have far reaching negative impact, are cases with high uncertainty, such as time critical decisions in hospitals or policing.

Most importantly however we find that not all subjects were equally susceptible to transparency induced algorithmic bias. Our results show that risk averse annotators were more susceptible to algorithmic bias than risk affine subjects.

When increasing transparency risk aversion was associated with an increase in agreement between annotators and the ML prediction and at the same time with a decrease in annotation accuracy – a sign of blind trust in ML predictions that can be attributed to increased transparency. These findings can also be directly transferred into practical applications. Determining the risk affinity of subjects can help to optimally calibrate the level of transparency for human-AI interaction. In contrast to the effects of risk aversion we did not find a significant effect of ambiguity tolerance on algorithmic bias. This result is different from studies on interaction between humans [50] and suggests that ambiguity tolerance could play a different role in interactions between humans and interactions between humans and an ML system.

Taken together our results highlight the potential of methods from psychology for improving the quality of human-AI interaction. A better understanding of how different personalities use AI can help to design systems that are both easier to use and less prone to algorithmic bias. We hope that this line of research will ultimately help to foster more responsible usage of AI systems.

References

1. Arrow, K.: Aspects of the theory of risk-bearing. Yrjö Jahnsson lectures, Yrjö Jahnssonin Säätiö (1965). https://books.google.de/books?id=hnNEAAAAIAAJ
2. Aumann, R.J.: Agreeing to disagree. Ann. Stat. **4**, 1236–1239 (1976)
3. Bell, D.E., Raiffa, H., Tversky, A.: Decision Making: Descriptive, Normative, and Prescriptive Interactions. Cambridge university Press, Cambridge (1988)
4. Camerer, C., Weber, M.: Recent developments in modeling preferences: uncertainty and ambiguity. J. Risk Uncertainty (1992). https://doi.org/10.1007/BF00122575
5. Cook, R.D.: Detection of influential observation in linear regression. Technometrics **19**(1), 15–18 (1977). http://www.jstor.org/stable/1268249
6. Curley, S.P., Yates, J.F., Abrams, R.A.: Psychological sources of ambiguity avoidance. Organ. Behav. Hum. Decis. Processes **38**(2), 230–256 (1986)
7. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. **144**(1), 114–126 (2015). https://doi.org/10.1037/xge0000033
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
9. Einhorn, H.J., Hogarth, R.M.: Decision making under ambiguity. J. Bus. **59**, S225–S250 (1986)
10. Ellsberg, D.: Risk, ambiguity, and the savage axioms. Q. J. Econ. **75**, 643–669 (1961)
11. FeldmanHall, O., Glimcher, P., Baker, A.L., Phelps, E.A.: Emotion and decision-making under uncertainty: physiological arousal predicts increased gambling during ambiguity but not risk. J. Exp. Psychol. Gen. **145**(10), 1255 (2016)
12. Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. In: Uncertainty in Economic Theory, pp. 141–151. Routledge (2004)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. CoRR abs/1412.6572 (2014). http://arxiv.org/abs/1412.6572

14. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018). https://doi.org/10.1145/3236009. http://dl.acm.org/citation.cfm?doid=3271482.3236009
15. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust Statistics: The Approach Based on Influence Functions, vol. 196. Wiley, Hoboken (2011)
16. Haufe, S., et al.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage **87**, 96–110 (2014)
17. Herman, B.: The promise and peril of human evaluation for model interpretability. CoRR abs/1711.07414 (2017)
18. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. **51**(1), 141–154 (2011). https://doi.org/10.1016/j.dss.2010.12.003, http://www.sciencedirect.com/science/article/pii/S0167923610002368
19. Kahneman, D., Tversky, A.: Choices, values, and frames. In: Handbook of the Fundamentals of Financial Decision Making: Part I, pp. 269–278. World Scientific (2013)
20. Kahneman, D., Tversky, A.: Prospect theory: an analysis of decision under risk. In: Handbook of the Fundamentals of Financial Decision Making: Part I, pp. 99–127. World Scientific (2013)
21. Kindermans, P., Schütt, K.T., Alber, M., Müller, K., Dähne, S.: Patternnet and patternlrp - improving the interpretability of neural networks. CoRR abs/1705.05598 (2017). http://arxiv.org/abs/1705.05598
22. Klibanoff, P., Marinacci, M., Mukerji, S.: A smooth model of decision making under ambiguity. Econometrica **73**(6), 1849–1892 (2005)
23. Knight, F.H.: Risk, Uncertainty and Profit. Courier Corporation, North Chelmsford (2012)
24. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Precup, D., Teh, Y.W. (eds.) ICML. vol. 70, pp. 1885–1894 (2017). http://proceedings.mlr.press/v70/koh17a.html
25. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: a case study on deception detection (2019). https://doi.org/10.1145/3287560.3287590
26. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: a joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1675–1684. ACM (2016)
27. Levy, I., Snell, J., Nelson, A.J., Rustichini, A., Glimcher, P.W.: Neural representation of subjective value under risk and ambiguity. J. Neurophysiol. **103**(2), 1036–1047 (2009)
28. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
29. Lipton, Z.C.: The doctor just won't accept that! arXiv preprint arXiv:1711.08037 (2017)
30. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS, pp. 4768–4777 (2017)
31. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: ACL, pp. 142–150 (2011). http://www.aclweb.org/anthology/P11-1015

32. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. arXiv preprint arXiv:1706.07269 (2017)
33. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recogn. **65**, 211–222 (2017). https://doi.org/10.1016/j.patcog.2016.11.008
34. Pratt, J.W.: Risk aversion in the small and in the large. In: Uncertainty in Economics, pp. 59–79. Elsevier (1978)
35. Pulford, B.D.: Short article: is luck on my side? optimism, pessimism, and ambiguity aversion. Q. J. Exp. Psychol. **62**(6), 1079–1087 (2009)
36. Rahwan, I., et al.: Machine behaviour. Nature **12**(11), 26. https://doi.org/10.1038/s41586-019-1138-y
37. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": explaining the predictions of any classifier. In: SIGKDD, pp. 1135–1144 (2016)
38. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: AAAI Conference on Artificial Intelligence (2018)
39. Rieger, M.O., Wang, M.: Cumulative prospect theory and the St. Petersburg paradox. Econ. Theory **28**(3), 665–679 (2006)
40. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learning Syst. **28**(11), 2660–2673 (2017). https://doi.org/10.1109/TNNLS.2016.2599820
41. Savage, L.J.: The Foundations of Statistics. Courier Corporation, North Chelmsford (1972)
42. Schmeidler, D.: Subjective probability and expected utility without additivity. Econometrica J. Econometric Soc. **57**, 571–587 (1989)
43. Schmidt, P., Bießmann, F.: Quantifying interpretability and trust in machine learning systems. vol. abs/1901.08558 (2019). http://arxiv.org/abs/1901.08558
44. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR abs/1312.6034 (2013). http://arxiv.org/abs/1312.6034
45. Slovic, P., Tversky, A.: Who accepts savage's axiom? Behav. Sci. **19**(6), 368–373 (1974)
46. Stanley Budner, N.: Intolerance of ambiguity as a personality variable 1. J. Pers. **30**(1), 29–50 (1962)
47. Strumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. **11**, 1–18 (2010). https://doi.org/10.1145/1756006.1756007
48. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of uncertainty. J. Risk Uncertainty **5**(4), 297–323 (1992)
49. Tymula, A., et al.: Adolescents' risk-taking behavior is driven by tolerance to ambiguity. Proc. National Acad. Sci. (2012). https://doi.org/10.1073/pnas.1207144109
50. Vives, M.L., Feldmanhall, O.: Tolerance to ambiguous uncertainty predicts prosocial behavior. Nat. Commun. (2018). https://doi.org/10.1038/s41467-018-04631-9
51. Von Neumann, J., Morgenstern, O., Kuhn, H.W.: Theory of Games and Economic Behavior (Commemorative Edition). Princeton University Press, Princeton (2007)
52. Wally, S., Baum, J.R.: Personal and structural determinants of the pace of strategic decision making. Acad. Manag. J. **37**(4), 932–956 (1994)

53. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV, pp. 818–833 (2014)
54. Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G.: The feature importance ranking measure. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5782, pp. 694–709. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04174-7_45