Taylor & Francis
Taylor & Francis Group

Journal of Decision Systems

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjds20

Transparency and trust in artificial intelligence systems

Philipp Schmidt , Felix Biessmann & Timm Teubner

To cite this article: Philipp Schmidt , Felix Biessmann & Timm Teubner (2020): Transparency and trust in artificial intelligence systems, Journal of Decision Systems, DOI: 10.1080/12460125.2020.1819094

To link to this article: https://doi.org/10.1080/12460125.2020.1819094

Published online: 10 Sep 2020.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tjds20

JOURNAL OF DECISION SYSTEMS
https://doi.org/10.1080/12460125.2020.1819094

Taylor & Francis
Taylor & Francis Group

ARTICLE

Check for updates

Transparency and trust in artificial intelligence systems

Philipp Schmidt[a], Felix Biessmann[a,b,c] and Timm Teubner[c,d]

[a]Amazon Research, Berlin, Germany; [b]Informatik und Medien, Beuth University of Applied Sciences, Berlin, Germany; [c]Einstein Center Digital Future (ECDF), Berlin, Germany; [d]Institute of Technology and Management, TU Berlin, Berlin, Germany

ABSTRACT

Assistive technology featuring artificial intelligence (AI) to support human decision-making has become ubiquitous. Assistive AI achieves accuracy comparable to or even surpassing that of human experts. However, often the adoption of assistive AI systems is limited by a lack of trust of humans into an AI's prediction. This is why the AI research community has been focusing on rendering AI decisions more transparent by providing explanations of an AIs decision. To what extent these explanations really help to foster trust into an AI system remains an open question. In this paper, we report the results of a behavioural experiment in which subjects were able to draw on the support of an ML-based decision support tool for text classification. We experimentally varied the information subjects received and show that transparency can actually have a negative impact on trust. We discuss implications for decision makers employing assistive AI technology.

ARTICLE HISTORY
Received 9 October 2019
Accepted 8 July 2020

KEYWORDS
Artificial intelligence; trust; experiment; machine learning; XAI; transparency

"The machine knows!"

– Michael Scott [1]

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have become increasingly important for aiding human decision making in various fields, including medical diagnosis, political predictions, policing, predictive maintenance, and many more (Esteva et al., 2017; Nickerson & Rogers, 2014; Susto et al., 2015). For a growing number of applications, ML-trained algorithms achieve performance comparable to or even surpassing that of humans (Szegedy et al., 2017). In light of the rapid progress of this 'cognitive automation,' the role of human decision makers, however, represents an often-overlooked factor (Ribiero et al., 2016). Specifically, recent cases show that many people dislike relying on AI and prefer to trust human experts, even if these experts are likely to be wrong (Polonski, 2018). In light of this apparent clash, and if AI-based decision support systems[2] are truly meant to benefit people, understanding human users' trust in such systems is key. It should be considered for at least three reasons.

CONTACT Felix Biessmann felix.biessmann@beuth-hochschule.de

© 2020 Informa UK Limited, trading as Taylor & Francis Group

First, trust functions as an important prerequisite of technology acceptance, adoption, and use in general (Venkatesh et al., 2016) and for AI in particular (Siau & Wang, 2018). It is not surprising that people hesitate to put major decisions into the hands of an AI assistant, 'especially when that assistant makes decisions without providing a transparent reasoning for choosing one solution over a set of alternatives' (Polonski, 2016, p. 1). Second, even once professionals have *adapted* AI-based support systems to inform their decisions, trust into such systems will be a prerequisite for them to *actually base their decisions* on the system's predictions, classifications, and recommendations. After all, humans still have the final say on how to deal with the AI's output in many cases. Third, next to questions related to the adoption of AI systems, understanding the processes governing human trust in AI is crucial to counteract the potential ramifications and side-effects of 1) mistakenly denied and 2) unfounded trust. If humans increasingly leverage AI to inform, derive, and justify decisions, it also becomes important to quantify when, how, why, and under which conditions they tend to overly trust or mistrust those systems.

Most recently, the Machine Learning community as well as the public debate have turned towards the comprehensibility of an AI's decisions, including aspects such as transparency, traceability, and hence interpretability (Grzymek & Puntschuh, 2019; Koene et al., 2019; Rohde, 2018). The underlying idea behind this stream of research is often to foster trust in AI systems by rendering them more transparent, often referred to as explainable AI, or XAI. While a substantial body of literature is dedicated to new methods of interpretability (Guidotti et al., 2018; Samek, 2019), the relationship between trust and the transparency of an AI's decision remains underrepresented in this research. This is why the driving and inhibiting factors of trust into AI are yet to be better understood. Specifically, we argue that there are cases in which transparency can actually have a detrimental impact on trust into AI. This can, in turn, lead to suboptimal usage of AI, with potential ramification for decision makers employing such technology but, at least equally importantly, also for persons affected by the outcomes (e.g. patients, defendants in trial; (Yong, 2018)). In this paper, we thus consider the following overarching research question:

RQ: *How does insight into a ML-based decision support tool affect human decision makers' trust in its predictions?*

We report the results of a behavioural experiment in which subjects were able to draw on the support of an ML-based decision support tool for text classification. We deliberately focus on a task that does not require expert domain knowledge. The motivation for this approach is that it allows for drawing conclusions that are not restricted to a given domain or profession. We experimentally varied the information subjects received on the tool's internal process in two dimensions. First, an explanation of the AI's prediction (by highlighting *decisive words* in the texts) was either shown or not. Second, a score on the tool's classification *confidence* was either provided or not. Our results challenge the common and popular narrative of providing highest possible transparency in order to build trust. Quite to the contrary, our results show that providing more insights into how an ML system arrives at its decision can have a *negative* effect on trusting behaviour. Importantly, this effect occurs predominantly for cases in which the ML system's predictions are correct, showing that improvident use of transparency within assistive AI tools can in fact impair human performance. Our findings have important implications for the

design and provision of algorithmic transparency where more insight may not always be preferable.

The remainder of this paper is organised as follows. In Section 2, we review related work on trust and interpretability of AI and motivate the focus variables of the present study. Next, Section 3 presents our experimental design, including task description and treatment structure. Section 4 reports our results which we discuss in Section 5. Last, Section 6 concludes.

2. Related work

Research on interpretability and user acceptance of AI can roughly be divided into three streams. First, there is a recent body of literature that explores *how* to make an AI's inner decision logic visible from a mostly technical perspective. Second, research on *technology acceptance* considers AI and algorithmic decision support from a general user perspective. Third, few recent studies have looked into how specific methods for making an AI's prediction transparent affect user behaviour (e.g. in terms of trust in those predictions) on a single-decision level.

2.1. Technical approaches to algorithm transparency

Algorithm transparency and interpretability can be established through various means, depending on the application domain, the ML model used, and the type of the model's input and output data. The most straightforward approach would be to first understand the modelling problem, then define a generative parametric model, fit its parameters to the data, and then interpret these parameters. This is the classical approach in engineering and other disciplines. Unfortunately, for complex data sets, it can be difficult to define a generative model of the data. If one is primarily interested in making predictions given some input data, it is much easier to use predictive ML models, such as Neural Networks. These models can achieve impressive predictive performance but their inner workings are usually neither easy to access nor to interpret (Ren et al., 2015). The research on interpretability methods aims at rendering these predictive models more transparent. One line of work in this context is the derivation of *importance scores for each feature* used in the prediction of a model. These scores can then be used to interpret the prediction in the feature space, for instance, one can inspect the top ranking features by highlighting words in text classification (Horn et al., 2017; Ribiero et al., 2016) or mask the lowest ranking pixels in image classification (Alber et al., 2018; Lundberg & Lee, 2017; Montavon et al., 2018; Ribiero et al., 2016). Such techniques have successfully been applied in algorithmic diagnosis and have uncovered sample-induced flaws such as the recognition of horses based on watermarks within the image (Lapuschkin et al., 2019). These feature scoring methods can be broadly categorised into two groups. The first group consists of model-agnostic explanations (Guidotti et al., 2018; Lundberg & Lee, 2017; Pacaci et al., 2019; Ribiero et al., 2016), also referred to as black-box explanation methods. The second group of feature scoring methods are explanations that are tailored to specific model classes, for instance, explanations for neural networks (Alber et al., 2018; Montavon et al., 2018; Simonyan et al., 2013; Zeiler & Fergus, 2014) or linear models (Haufe et al., 2014). Here, we will focus on the latter explanation method as it is efficient to compute and was

shown to be superior in terms of explanation quality as compared to more expensive model-agnostic approaches (Schmidt & Biessmann, 2019).

2.2. AI, trust, and technology acceptance

The literature on organisational behaviour and psychology suggests that individuals' intention to use a new technology is determined by perceptions and beliefs about the technology (Ajzen, 2014). In this regard, the Technology Acceptance Model (and its various descendants) represent suitable and often-applied frameworks to capture the behavioural aspects of technology adoption and acceptance. While the original model focussed on perceived usefulness and ease of use (Davis, 1989; Venkatesh & Davis, 2000), numerous extensions and variations have considered additional drivers, factors, moderators, and outcomes within the broader area of technology acceptance (Venkatesh et al., 2016). Among these, trust has emerged as one key driver of technology acceptance (Gefen et al., 2003).

Trust in technology should be differentiated from trust towards humans. In fact, depending on the specific definition of trust, in particular when it includes the willingness to accept *strategic uncertainty*, that is, when facing an actor with agency and intentions, it can be argued that there is no such thing as trust into machines. One of the most commonly adopted definitions refers to trust as the 'willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party' (Mayer et al., 1995, p. 712). This understanding of trust is somewhat less strict in the sense that it does not explicitly require agency or intentions and would hence allow for an AI as the trustee. The notion of trust is typically conceptualised along the dimensions *ability, benevolence*, and *integrity* (Mayer et al., 1995). For technology, for instance, Söllner et al. (2012) propose to differentiate trusting beliefs along with the dimensions performance, process, and purpose – while Lippert and Swiercz (2005) argue for the dimensions utility, reliability, and predictability. Compared to trust towards humans, prior research has argued that people tend to have less trust towards AI by default, where 'people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster' (B. J. Dietvorst et al., 2014, p. 1). This innate scepticism towards AI was termed as 'algorithm aversion'. In follow-up work, B. Dietvorst et al. (2018) showed that people were more than twice as likely to rely on algorithmic forecasts if they were allowed to change a fraction of the AIs predictions compared to when they were bound what the AI predicted. Moreover, trust in ML is found to build up slower and to decrease faster than trust in humans (Dzindolet et al., 2003). This may partly be explained by the high media attention on instances in which AI went wrong, including a Google algorithm that classified Afro-Americans as gorillas,[3] self-driving vehicle accidents and near misses,[4] Amazon's Alexa device offering adult content to children,[5] or Microsoft's Twitter chatbot that developed into a white supremacist within only one day.[6]

To create trust in AI-assisted tools, Hind et al. (2018) propose to include a supplier declaration of conformity (SDoC) which is a long-established procedure in other sectors such as transportation and telecommunication. Hengstler et al. (2016) present a comprehensive overview of existing case studies in transportation and medical sectors that allow insights into how firms systematically increase trust in applied AI. Especially in

safety-critical applications, for instance, for self-driving vehicles, it was proposed to integrate physiological (Hutchins & Hook, 2017) and ethical factors (Adnan et al., 2018) as part of technology acceptance models.

2.3. Algorithm transparency and trusting behaviour

A representative EU survey tested for knowledge and participants' perception of algorithms (Grzymek & Puntschuh, 2019). Overall, while respondents see more advantages than disadvantages when it comes to algorithms, yet 78% responded that algorithms needed more rigorous control. At the same time, around half of the sample reported little to no knowledge about algorithms. Multiple studies found that trust into assistive AI technology can be increased most effectively by confronting users with an easy to understand explanation of why a specific prediction was made (Herlocker et al., 2000; Poursabzi-Sangdeh et al., 2018; Zhao et al., 2019). The reported findings have implications for advice-giving-systems which are increasingly used on e-commerce platforms such as Amazon.com and Google Shopping. The degree of *simulatability* (Lipton, 2016), that is, how accessible model predictions are to humans, aligns well with the aforementioned effects on trust.

When it comes to quantification, it is important to note that trust is measured differently in different research fields. A straightforward approach used in many fields is to ask subjects about their *trusting beliefs* or *trusting intentions* towards a certain entity, typically by means of standardised survey instruments (e.g. Gefen & Straub, 2003). Naturally, such survey scales can only partially assess trust as per the definition of *making oneself vulnerable to the action of another party*, as ticking boxes in a survey do not require the actual critical behaviour and hence is not associated with vulnerability. Self-assessments must thus be considered as a somewhat *weak* indicator. Behavioural trust, in contrast, infers trust by considering actual behaviour and can hence considered to be a stronger indicator. For instance, leaving one's kids with a certain baby sitter suggests the parents to have trust towards that person. However, this indicator also comes with some caveats as well. Behaviour is complex and driven by a multitude of factors (e.g. monetary constraints, lack of alternatives, …). Behaviour will hence always represent a *proxy* for trust, where researchers must be aware of further potential drivers that determine behaviour.

Now, following this thought, trust towards an AI can be operationalised based on behaviour. Specifically, decision makers that draw on the assistance of an AI will ultimately depend on their decisions, for instance, financially or with regard to their reputation. At the same time, they have some uncertainty about the AI's accuracy. In this sense, trust towards an assistive AI tool can be operationalised as the probability with which the decision maker follows the model's prediction (Poursabzi-Sangdeh et al., 2018; Schmidt & Biessmann, 2019). As stated above, there will of course exist other, additional drivers of behaviour. The decision maker may, for instance, be convinced of a certain diagnosis a priori and the AI system incidentally makes the same prediction. While a single decision can hence not be considered a reliable clue, observing and comparing the frequency of multiple decisions – especially in relative terms to other groups or treatment conditions – will very well function as a behavioural proxy of trust.

In behavioural decision making, other measures such as the *weight of advice* (Gino & Moore, 2007; Yaniv, 2004) is used to measure trust and when and how advice is accepted. It was found that several factors drive acceptance of advice, such as one's own knowledge and confidence, and whether tasks are perceived as difficult or easy (Yaniv, 2004). In the mentioned studies, it was found that participants generally had a bias towards self-reliance in decision making, therefore discarding advice even in situations where it would have been beneficial. The self-reliance bias could be corrected for when, over time, participants were able to learn that the AI assistance is indeed superior (Dzindolet et al., 2003).

To summarise, the majority of the literature suggests that explanations and transparency of AI systems should lead to more trust into an AI's prediction. Yet there is also evidence for algorithm aversion. In those cases of unjustified mistrust in AI, should we expect that transparency helps to rebuild trust? Or should we expect the opposite, that is, would explanations decrease trust further in these cases? There is evidence that the mechanisms underlying the build-up and decay of human trust into AI systems are different from those governing trust between humans or from human trust in classical technology that does not learn from data. But the exact conditions under which transparency affects trust in AI systems are not well represented in the existing literature. This work explores several aspects of transparency in AI systems and probes the hypothesis that transparency increases trust in AI systems.

3. Methods

To address the outlined research question, we conducted an online experiment in which human participants take the role of decision makers in a series of classification tasks. Given their high practicality and applicability for general audiences, we use text classification tasks.

3.1. Classification task

In the classification task, participants were asked to classify movie reviews' sentiment from the IMDb database as either positive or negative. The data used is publicly available and is described in more detail in (Maas et al., 2011). Subjects were incentivised to correctly classify as many reviews as possible by the prospect of a bonus payment. Importantly, participants had the prediction of an ML-based decision support tool at their disposal which predicted the movie review either to be positive or negative. A detailed description of the training procedure and the training data preparation is provided in Appendix B.

Participants engaged in 50 consecutive classification tasks. All participants were exposed to the same reviews. Positive and negative reviews occurred equally often (25 times each) and had varying degrees of (relative) ease. Task ease is based on the frequency of correct (unaided) classifications from prior work (Schmidt & Biessmann, 2019). These values range between 0.5 (chance level performance) and 1.0. In order to avoid learning and adoption effects, the true sentiment values were not revealed throughout the experiment.

We selected a set of reviews for which the ML-based decision support tool'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 also amount to 80%. Previous research on IMDb review classification found that typical human accuracy ranges between 75% and 80% (Schmidt

& Biessmann, 2019). Participants were hence informed that the AI's performance was similar to that of humans, while not perfect. This information was important to avoid participants employing strategies in which they simply copy the prediction of the AI system. We provide the experimental instructions in Appendix A.

Figure 1 shows a sample screenshot of the experiment's UI. Depending on the condition, participants either saw the highlights and/or confidence score. In this case, both word highlighting and confidence were displayed.

3.2. Treatment structure

We considered two means by which the ML-based decision support tool 'explained' its predictions to participants. First, we considered the display of *word highlighting* within the movie reviews. Specifically, words that had a particular impact on the tool's classification were highlighted, typically highly positive or highly negative expressions. For each movie review, the three most relevant words were highlighted. The technical details on how these explanations were computed are described in Appendix C. Depending on the treatment condition, these highlights were either present or not. Second, we considered the display of a *confidence score*. With this score, the tool provided some insight into how certain it was about its prediction. The score is derived directly from the logistic regression estimates for positive (p_{pos}) and negative (p_{neg}) sentiment and is encoded on a scale from 0% (purely guessing) to 100% (absolute confidence). This score was either displayed or not displayed, depending on the treatment condition.

We considered each of the resulting $2 \times 2 = 4$ combinations of these two treatment variables (i.e. full-factorial design). To avoid cross-over and sequence effects, each participant engaged in exactly one of these conditions (i.e. between-subjects design). Figure 2 summarises the treatment design. Subjects were allocated to treatments at random; there occurred no significant differences across treatments with regard to age, gender, or individual disposition to trust.

3.3. Procedure and sample

Participants were recruited via the online platform Prolific.ac (Palan & Schitter, 2018). They received a payoff comprising an unconditional part and a performance-contingent part.

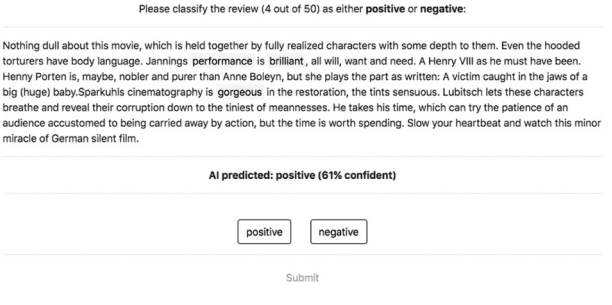

Figure 1. Screenshot of the labelling UI. The to-be-classified text is shown on top with the AI prediction, in this case including its confidence, below it. Participants were asked to select either a positive or negative sentiment for the given movie review and were then able to submit the answer.

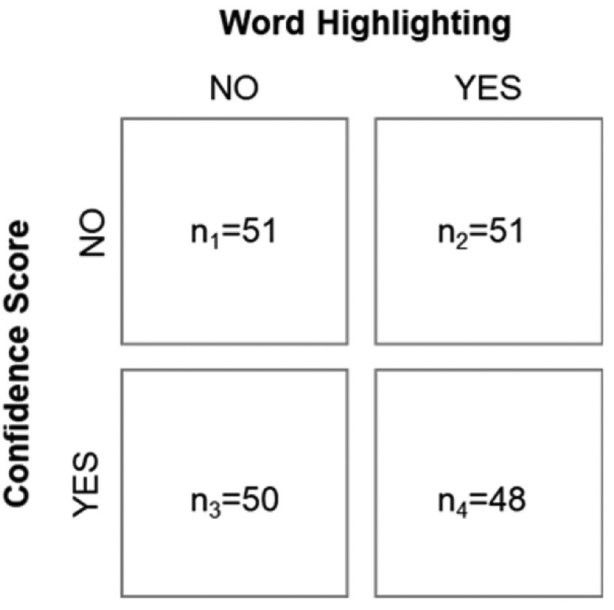

Figure 2. Treatment design.

Specifically, they received an unconditional base payoff of GBP 1.50 and an additional payoff of GBP 0.80 if their classification accuracy was equal to or above 85%. Importantly, participants were informed that the AI's accuracy was below this threshold so that simply adopting the AI's predictions would fail to qualify for the bonus payment.

After completing the classification tasks, we surveyed participants' on their general disposition to trust using validated survey instruments (Gefen, 2000). Overall, 200 participants took part in the study, 103 of which were female. Participants' age ranged between 18 and 71 years, with a mean of 29.1 and a median of 27 years. The great majority of subjects was from the UK (35%), other European (48%), or English-speaking countries such as the US, Canada, or Australia (12%). Average classification accuracy was 87% (minimum: 40%, maximum: 98%, median: 90%), where roughly 3 out of 4 participants qualified for the bonus payment.

4. Results

As a behavioural proxy for trust, we here consider how often subjects followed the AI's prediction across the 50 tasks, that is, their average willingness to depend on the AI. As a first step of analysis, we consider the *overall treatment effects* of showing highlights and/or the confidence score on trusting behaviour. Figure 3 (left) illustrates trusting behaviour for the different treatment conditions. We observe marked *negative* differences for both features. While showing highlights decreases trusting behaviour of about 1 to 1.5 percentage points, showing the confidence score has an effect of roughly 2.5 to 3 percentage points. This reduction of trust based on the display of highlights and confidence is also reflected in decreases in accuracy (Figure 3, right), where the drop in accuracy is particularly strong when both elements are displayed.

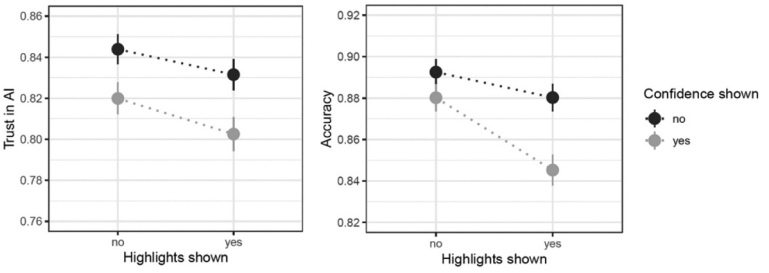

Figure 3. Overall treatment effects on Trust in AI (left) and Accuracy (right); standard errors indicated.

To better understand the observed overall treatment effects and to back up this first visual assessment statistically, we conduct a series of two-way random effects logit panel regressions with repeated measures per subject. The basic regression equation is given as

$$y_{it} = \alpha + \sum_{n=1}^{4} \beta^{n} x_{i}^{n} + \sum_{m=1}^{2} \gamma^{m} x_{t}^{m} + \in_{i} + \xi_{t} + \mu_{it}$$

where y_{it} denotes whether or not subject i followed the AI's recommendation for review t (and whether or not subject i classified review t correctly, respectively). The variables in x_i represent subject-specific factors that only depend on i, including the treatment dummies (Highlights, Confidence Score), as well as gender and individual disposition to trust. Moreover, x_t contains the task-specific variables (i.e. Task Ease, AI's prediction correct/incorrect). Last, ε, ξ, and μ denote subject- and task-specific errors, as well as the residual model error. These basic models (I(a) and II(a)) are then extended by interaction effects (I(b) and II(b)). All results are summarised in Table 1.

Model I(a) confirms the negative effect of displaying the confidence score ($\beta_1 = -.259$, $p < .05$), while – overall – displaying highlights does not have a significant effect ($\beta_2 = -.136$, $p = .173$). To fully understand the effects of the tested features, we control for the respective interactions with the AI's correctness in Model I(b). Here, we see that the main treatment effects are driven by those tasks in which the AI's prediction is actually correct. For these cases, the transparency features have strong negative effects (highlights: $\beta_1 = -.358$, $p < .01$; confidence: $\beta_2 = -.513$, $p < .001$). The positive and significant interaction effects between the treatment variables and the dummy variable 'AI is incorrect' indicate that these effects are close to zero in case the AI's prediction is incorrect anyway (Highlights: −.358 + .436 = .078, Confidence Score: −.513 + .494 = −.019). In other words, for incorrect AI predictions, subjects' trusting behaviour does not (or only hardly) depend on the tested transparency features, while these features do make a substantial difference for correct predictions.

Controlling for task-specific properties, we see that Task Ease does not significantly affect trusting behaviour ($\gamma_1 = .106$, $p = .659$). In contrast, the AI's correctness has a strong effect, where expectedly, trusting behaviour is consistently lower if the AI is incorrect ($\gamma_2 = -3.863$, $p < .001$). Last, none of the considered subject-specific control variables exerts any significant effects on trusting behaviour (females: $\beta_3 = -.053$, $p = .602$, disposition to trust: $\beta_4 = .083$, $p = .128$).

Table 1. Two-way random effects logit panel regression results (*** p<.001; ** p< .01; * p< .05).

Dependent Variable	Trust in AI				Accuracy			
Model	I(a)		I(b)		II(a)		II(b)	
	Coef. (SE)	Sig.	Coef. (SE)	Sig.	Coef. (SE)	Sig.	Coef. (SE)	Sig.
Highlights	−.136 (.100)		**−.358** (.122)	**	−.217 (.132)		**−.321** (.152)	*
Confidence Score	**−.259** (.101)	*	**−.513** (.124)	***	−.198 (.133)		**−.501** (.154)	**
AI is incorrect	**−3.387** (.076)	***	**−3.863** (.131)	***	**−2.199** (.075)	***	**−2.612** (.133)	***
Highlights × Incorrect			**.436** (.139)	**			.209 (.146)	
Confidence × Incorrect			**.494** (.140)	***			**.583** (.147)	***
Gender: Female	−.047 (.103)		−.053 (.102)		.052 (.136)		.046 (.135)	
Disposition to Trust	.091 (.056)		.083 (.055)		.135 (.074)		.127 (.073)	
Task Ease	.107 (.241)		.106 (.241)		**3.148** (.270)	***	**3.151** (.270)	***
Intercept	2.648 (.297)	***	2.940 (.302)	***	.356 (.360)		.612 (.366)	
N	9,672		9,672		9,672		9,672	

Overall, we observe very similar results for subjects' classification accuracy which are summarised in Models II(a) and II(b). Here, again, accuracy suffers from showing highlights ($\beta_2 = -.321$, p< .05) and the confidence score ($\beta_2 = -.501$, p< .01) for correct AI predictions, while these effects are – by and large – annihilated for incorrect predictions.

4.1. Trust, task ease, and AI err

As a next step, we take a closer look at how specifically Task Ease and AI correctness affected trusting behaviour and accuracy. To do so, Figure 4 displays Trust in AI (left) and Accuracy (right) differentiated by Task Ease (x-axis) and whether the AI was correct (blue) or not (red). As can be seen, both trust and accuracy are markedly higher in case the AI's predictions were correct. In particular, this holds for tasks of equal difficulty.

Figure 4 allows for another interpretation. For the blue dots, Trust in AI *should* be 100% whereas for the red dots, it *should* be 0% (for perfect accuracy). As can be seen, the area/gap between the blue dots and the 100% level is much smaller than the area between the red dots and the 0% level. Hence, in relative terms, trusting (and hence following) wrong

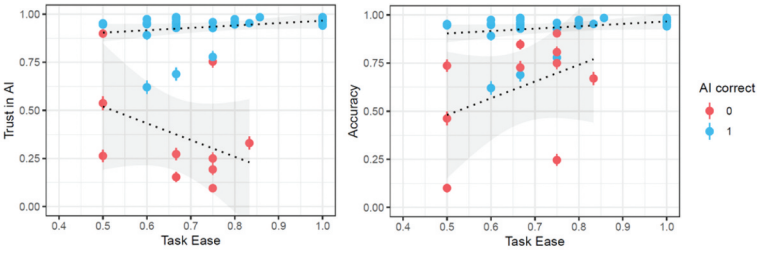

Figure 4. Trust in AI and accuracy by task ease and AI prediction correctness.

Table 2. Confusion matrix and human error types and rates.

		Subject Trusts AI		Human Error	
		No	**Yes**	**Rate**	**Type**
AI Prediction	**incorrect**	1211	725	37.4%	False Positive
	correct	483	7253	6.2%	False Negative

AI predictions is reflected in much larger error rates than deviating from correct predictions. Note, however, that for these cases, the design features for transparency do not seem to contribute to mitigating the problem.

This reasoning is summarised in Table 2. As shown there, the human error rates (i.e. falsely trusting or not trusting the AI) depend on the AI's correctness. Specifically, the error is more than six times higher when the AI's prediction is incorrect (37.4%) versus when it is correct (6.2%). As shown in Figure 4, this observation is not driven by task ease but holds across the scale for different levels thereof.

5. Discussion

Research on new methods of transparent AI is often considered key to building trust into AI systems. As the results of this experiment point out, higher levels of transparency may not necessarily imply higher levels of trust or 'compliance' when it comes to dealing with an AI's predictions, classifications, or recommendations. This finding challenges the common narrative on transparency but also aligns well with recent work (Poursabzi-Sangdeh et al., 2018). To the best of our knowledge, no related work has described the effect of overly trusting wrong predictions. This effect, as pointed out in the previous section, is especially reflected for difficult tasks. An example for too much trust in wrong AI predictions with an explanation is shown in Figure 5: An incorrect AI prediction with high confidence that is followed suit by humans. In this particular example, the accuracy reached by the subjects was 76 ± 5% (mean ± standard deviation) on average when explanations were shown and 85 ± 4% when not. The highlighted words (with positive sentiment) are intuitively understood and appear to lead subjects astray. If no explanations are provided participants are less likely to focus only on these positive words and

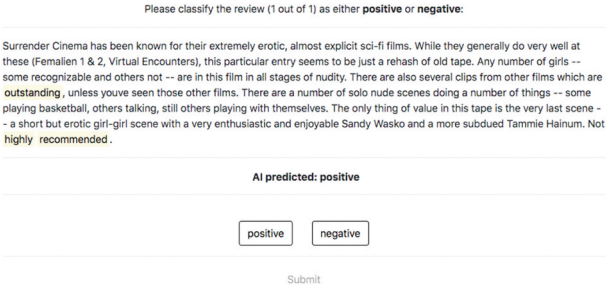

Figure 5. Example for unjustified increase in trust when showing explanations. The AI incorrectly predicted positive sentiment since it ignored the negation and focused on the positive sentiment in some words. As the explanation highlighted these positive words but not the negation, transparency influences subjects to ignore the contextual information important to correctly predict the item as negative.

more likely to also read the negation, which leads to higher accuracy. Note that negation, as many other often subtle linguistic phenomena like sarcasm, can be challenging even for sophisticated state-of-the-art AI systems.

This effect of overly trusting wrong but intuitive predictions is however not what seems to drive the negative effect of transparency on annotators performance. The estimates of Model I(b) in Table 1 demonstrate that the dominant factor are tasks in which the AI's prediction is actually correct and explanations appear to decrease trust. An example of not enough trust in correct predictions due to transparency is shown in Figure 6. Here, the prediction was correct. However, as the explanations are neither related to positive nor negative sentiment intuitively, subjects tended to choose the opposite of what was predicted by the AI, possibly assuming that the prediction was incorrect due to the confusing highlighting. Subjects' average accuracy in this example went down from 96 ± 2% when no explanations were shown to 89 ± 3% when they were.

This example demonstrates that lack of trust can emerge from unintuitive explanations of ML models. Such mistrust is probably justifiable. While we are developing and using ML methods powerful enough to surpass the cognitive abilities of humans in some applications, many of these methods are far from being investigated sufficiently to understand their behaviour (Zhang et al., 2017). Put simply, models with enough parameters can learn classification rules that might just work well on that training data set, but might not be intuitive nor generalise well. An algorithm may in fact learn some flawed sample-specific feature such as a horse website imprint or recognise boats based on the presence of water (Lapuschkin et al., 2019). Latter study demonstrated that explicability can help to debug ML systems. In this case, experts were dealing with the AI's predictions and visualisations. Our results complement these findings with models, predictions, and explanations subject to the evaluation of a general clientele. If those find an AI's explanation unintelligible (accessible), then transparency may lead to distrust (too much trust) into the prediction. Both approaches, debugging ML systems and calibrating human trust in AI, are important for a responsible use of such systems.

Our results indicate that there may exist cases when humans should trust an AI more than they actually do. Indeed, we find that an AI that reports on its confidence may suffer from the 'insecure overachiever' syndrome. As shown in Figure 1, humans are more likely to ignore correct predictions when those are accompanied by an indication of confidence. Examples (such as in Figure 6) indicate that this could be due to unintuitive explanations.

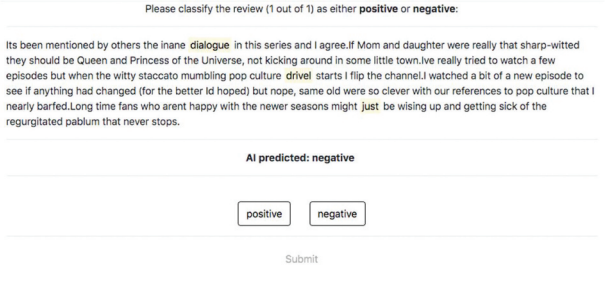

Figure 6. Example for decrease in trust when showing explanations. The AI prediction was correct. But the explanation did not highlight words that could be intuitively related to either positive or negative sentiment from a human 'common sense' perspective.

Such glitches in human-machine interaction could have detrimental consequences on the societal level (Rahwan et al., 2019). Just because the prediction of an AI is not intuitively comprehensible, people should of course not blindly distrust them. In fact, this sort of trust is required for synergetic effects in human-machine interaction: algorithms can identify features and strategies that have thus far been unfamiliar or completely unknown to humans. This has, for instance, reportedly been the case for Alpha Zero's approaches to play the games of Go and Chess.[7] For sentiment analysis based on textual features, humans may intuitively look for signal words such as 'amazing', 'great', or 'awful.' An algorithm, however, may also recognise distinct patterns of punctuation such as an increased use of commas or exclamation marks, or words that are not intuitively related to one of the predicted classes. On the one hand, this may appear uncommon to humans and, when the AI supports its recommendation by such a feature, may lead humans to question its ability and correctness. It can, on the other hand, unlock new insights and improve rules for decision making in the long run.

5.1. Practical implications

Responsible use of AI technology requires the right level of user trust in the system. Our findings demonstrate that humans often fail to trust an AI when they should, but also that humans follow an AI when they should not. In particular, the first effect is exacerbated when explanations are provided along with the AI predictions. These results have important practical implications for the design of AI systems in general and explainable AI in particular. Based on our findings, we propose to not only train AI systems before bringing them to bear – but also to 'calibrate' them to the human users that will ultimately interact with them. Building assistive AI systems and ensuring their responsible use hence requires a multi-disciplinary perspective to fully understand potential trust effects. More concretely, human decision makers, for instance, policy-makers, judges, or medical practitioners that consult AI-based decision support should undergo thorough training of ML basics and diagnostics, learn about accuracy and false positive/negative errors, including illustrative examples for both unfounded trust and unfounded distrust in the provided predictions. In a way, also humans may have to be calibrated to the AI to improve outcomes. In practice this could be implemented through experiments on users of actual assistive AI systems.[8]

Another implication of our results is that the right level of trust depends on the explanations provided to humans about an AI's prediction. Consequently, when designing human-AI interaction, algorithmic transparency and interpretability should be evaluated and implemented carefully to allow for synergies. For instance, displaying explanations that decrease the accuracy of a human-AI team should be avoided. This opens up a research area in and by itself: *Behavioural AI*. Following Gary Kasparov's call for mixed human-AI teams (Quach, 2018), we suggest that research is needed that systematically explores under which transparency regimes, mixed human-AI teams perform best, not only in terms of accuracy but also in terms of time needed for a decision.

Next to these application-specific considerations, on a more general level the responsible usage and right level of trust in an AI also depend on cognitive biases related to technology acceptance of a society. These factors should therefore be considered when designing and deploying assistive AI to the general public. Our findings challenge the common narrative of the positive effects of transparency and suggest that it is important to

not only provide transparency, but also to make sure users also understand the means of transparency. In other words, it is not enough to lighten up the black box – people also need to know whether the flashlight uses regular or black light (to use the same metaphor).

5.2. Limitations and future work

In this study, we deliberately focussed on a task that does not require expert domain knowledge. This approach is motivated by allowing to draw conclusions that are not restricted to a given domain or profession.

As this study was intentionally conducted with a narrow focus for a particular task and model, future work should broaden the scope by including different tasks and feedback mechanisms. The data set used in this study was deliberately chosen such that the model performed comparable to humans (that is, worse than it could have), and the model was not particularly sophisticated either. The question how varying performance would impact the effect of transparency on trust into an AI will definitely deserve closer attention. More sophisticated systems will also pick up more of the language structure needed to correctly interpret the semantics of natural language. Better text classification has the potential to learn representations of text that are possibly more similar to how humans process natural language. On the flipside, these systems can also be more difficult to interpret than the simple linear model used in this study. It is not unlikely that the influence of transparency on human trust will depend both on the model's capability to produce highly accurate predictions as well as its complexity.

Future work should therefore further study AI-based decision support, not only for text but also other input domains such as images. Understanding how trust is driven by different models and their transparency is a valuable contribution to the research community as well as a crucial element for decision makers across many domains.

5.3. Concluding note

We have studied how trust into an AI's predictions is affected by the presence of two auxiliary measures of transparency: relevant feature highlighting and confidence scores. The idea that transparency will foster trust in (assistive) AI technology is one of the main drivers behind XAI research. In light of the public debate around XAI, our results challenge this popular narrative of the desirability of maximal algorithmic transparency. Contrary to that narrative and complementing previous work, we find that transparency can have negative effects on trust and report cases of mistrust when an AI prediction is correct as well as too much trust when an AI prediction is wrong. Hence, responsible usage of AI systems will require both careful calibration of the AI system to perform well in terms of common performance metrics but also a 'calibration' of the level of human trust into its prediction.

Notes

1. https://www.youtube.com/watch?v=DOW_kPzY_JY.
2. Within the scope of this paper, we refer to *Assistive AI, based on ML classifiers for decision support*. For the sake of brevity, we will simply refer to this as 'AI' or 'AI system', knowing that represents a simplification.

3. https://twitter.com/jackyalcine/status/615329515909156865.
4. https://www.nytimes.com/2016/09/15/business/fatal-tesla-crash-in-china-involved-autopilot-government-tv-says.html; https://www.nytimes.com/2016/07/01/business/self-driving-tesla-fatal-crash-investigation.html.
5. https://www.entrepreneur.com/video/287281.
6. http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00000gjdpwwcfcus11t6oo6dw79gw.
7. https://deepmind.com/blog/alphazero-shedding-new-light-grand-games-chess-shogi-and-go.
8. In fact, performance tests of click-workers usually follow that idea.
9. Taken from https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/feature_extraction/_stop_words.py.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Adnan, N., Md Nordin, S., Bin Bahruddin, M.A., & Ali, M. (2018). How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. *Transportation Research Part A: Policy and Practice*, *118*, 819–836. https://doi.org/10.1016/j.tra.2018.10.019

Ajzen, I. (2014). The theory of planned behaviour is alive and well, and not ready to retire: A commentary on Sniehotta, Presseau, and Araújo-Soares. *Health Psychology Review*, *9*(2), *7199*, 1–7. https://doi.org/10.1080/17437199.2014.883474

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., & Kindermans, P.-J. (2018). *iNNvestigate neural networks!* (Working paper).

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Dietvorst, B., Simmons, J.P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dietvorst, B.J., Simmons, J.P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*: *General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., & Beck, H.P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://www.nature.com/articles/nature21056?spm=5176.100239.blogcont100708.20.u9mVh9

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega*, *28*(6), 725–737. https://doi.org/10.1016/S0305-0483(00)00021-9

Gefen, D., Karahanna, E., & Straub, D.W. (2003). Trust and TAM in online shopping: An integrated model. *Information Systems Research*, *25*(1), 1–16. https://dl.acm.org/doi/10.5555/2017181.2017185

Gefen, D., & Straub, D.W. (2003). Managing user trust in B2C e-services. *e-Service Journal*, *2*(2), 7–24. https://doi.org/10.2979/esj.2003.2.2.7

Gino, F., & Moore, D.A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, *20*(1), 21–35. https://doi.org/10.1002/bdm.539

Grzymek, V., & Puntschuh, M. (2019). What Europe knows and thinks about algorithms: Results of a representative survey. *Bertelsmann Stiftung*, 1–38. http://aei.pitt.edu/102582/

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting & Social Change*, *105*, 105–120. https://doi.org/10.1016/j.techfore.2015.12.014

Herlocker, J.L., Konstan, J.A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *CSCW 2000 Proceedings* (pp. 1–10), Philadelphia, PA.

Hind, M., Mehta, S., Mojsilovi, A., Nair, R., Ramamurthy, K.N., Olteanu, A., & Varshney, K.R. (2018). *Increasing trust in AI services through supplier's declarations of conformity* (Working paper).

Horn, F., Arras, L., Montavon, G., Müller, K.-R., & Samek, W. (2017). *Exploring text datasets by visualizing relevant words* (Working paper) (pp. 1–10).

Hutchins, N., & Hook, L. (2017). Technology acceptance model for safety critical autonomous transportation systems. In *AIAA 2017 Proceedings* (pp. 1–5), St. Petersburg, FL, USA.

Koene, A., Clifton, C., Hatada, Y., Webb, H., Patel, M., Machado, C., LaViolette, J., Richardson, R., & Reisman, D. (2019). A governance framework for algorithmic accountability and transparency. *European Parliamentary Research Service*, 1–124. https://doi.org/10.2861/59990

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1–8. https://www.nature.com/articles/s41467-019-08987-4

Lippert, S.K., & Swiercz, P.M. (2005). Human resource information systems (HRIS) and technology trust. *Journal of Information Science*, *31*(5), 340–353. https://doi.org/10.1177/0165551505055399

Lipton, Z.C. (2016). The mythos of model interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning* (pp. 96–100), New York, NY, USA.

Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems* (pp. 1–8), Long Beach, CA, USA.

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., & Potts, C.A. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 142–150), Portland, Oregon.

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Montavon, G., Samek, W., & Müller, K.R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Nickerson, D., & Rogers, T. (2014). Political campaigns and big data. *Journal of Economic Perspectives*, *28*(2), 51–74. https://doi.org/10.1257/jep.28.2.51

Pacaci, G., Johnson, D., McKeever, S., & Hamfelt, A. (2019). 'Why did you do that?' Explaining black box models with inductive synthesis. In J. M. F. Rodrigues et al. (Ed.), *ICCS 2019 Proceedings* (pp. 334–345), Faro, Portugal. https://arxiv.org/abs/1904.09273

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Polonski, V. (2016). *Would you let an algorithm choose the next U.S. president?* Retrieved November 16, 2018, from https://techcrunch.com/2016/11/06/would-you-let-an-algorithm-choose-the-next-u-s-president/?guccounter=2

Polonski, V. (2018). *Humans don't trust AI predictions - Here's how to fix it*. Retrieved May 24, 2019, from https://www.oecd-forum.org/users/80891-dr-vyacheslav-polonski/posts/29988-humans-don-t-trust-artificial-intelligence-predictions-here-s-how-to-fix-it

Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., & Wallach, H. (2018). *Manipulating and measuring model interpretability* (Working paper) (pp. 1–20).

Quach, K. (2018). *Don't try and beat AI, merge with it says chess champ Garry Kasparov*. Retrieved May 5, 2019, from https://www.theregister.co.uk/2018/05/10/heres_what_garry_kasparov_an_old_world_chess_champion_thinks_of_ai/

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., Jennings, N.R., Kamar, E., Kloumann, I.M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D.C., Pentland, A.S., Roberts, M.E., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://www.nature.com/articles/s41586-019-1138-y?_hsenc=p2ANqtz–IRCtofDpjC9xefSHwgX-1wh5xTGoBYy-A7yZ1G2CP25l76yByyqcaOmnwF941clvJTbYVHezHzQXJASWORb6UtLzou7myGSVciGlHtthUAiyPRyw&_hsmi=72127156

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS 2015 Proceedings* (pp. 1–9), Montreal, Canada.

Ribiero, M.T., Singh, S., & Guestrin, C.A. (2016). 'Why should I trust you?' Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining* (pp. 1135–-1144), San Francisco, CA, USA.

Rohde, N. (2018). Quality criteria for algorithmic processes: Analyzing the strengths and weaknesses of selected compendia. *Bertelsmann Stiftung*, 1–24. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Algorithmic_processes_final.pdf

Samek, W. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer Nature.

Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems. In *AAAI 2019 Proceedings* (pp. 1–8), Honolulu, Hawaii, USA.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, *31*(2), 47–53. https://www.cutter.com/article/building-trust-artificial-intelligence-machine-learning-and-robotics-498981

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). *Deep inside convolutional networks: Visualising image classification models and saliency maps* (Working paper) (pp. 1–8).

Söllner, M., Hoffmann, A., Hoffmann, H., Wacker, A., & Leimeister, J.M. (2012). Understanding the formation of trust in IT artifacts. In *ICIS 2012 Proceedings* (pp. 1–18), Orlando, USA.

Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812–820. https://doi.org/10.1109/TII.2014.2349359

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A.A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI 2017 Proceedings* (pp. 4278–4284), San Francisco, California USA.

Venkatesh, V., & Davis, F.D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal studies. *Management Science*, *46*(2), 186–205. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Thong, J.Y.L., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, *17*(5), 328–376. https://doi.org/10.17705/1jais.00428

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yong, E. (2018). *A popular algorithm is no better at predicting crimes than random people*. Retrieved March 2, 2020, from https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/;

Zeiler, M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833), Zurich, Switzerland. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *ICLR 2017 Proceedings* (pp. 1–15), Toulon, France.

Zhao, R., Benbasat, I., & Cavusoglu, H. (2019). Do users always want to know more? Investigating the relationship between system transparency and users' trust in advice-giving systems. In *ECIS 2019 Proceedings* (pp. 1–12), Stockholm & Uppsala, Sweden.

Appendix A. Experimental Instructions

Welcome and thank you very much for participating in this experiment!

In this experiment, you will be asked to classify a series of IMDB movie reviews as either *positive* or *negative*. Your classifications will be compared to the review authors' own classification (i.e. the 'true' review tendency). If you achieve an accuracy at least 85%, you will receive a bonus payment of 0.80 £.

For your decisions, you will be able to draw on the help of an Artificial Intelligence (AI) classifier. This AI was trained on a large amount of movie reviews and has an accuracy somewhat above average human accuracy. Note, however, that the AI is not perfect and that its accuracy is below the bonus payment threshold (see above). Simply adopting the AI's prediction all the time will thus fail to qualify for the bonus payment with certainty.

if (highlights) {

In addition to its prediction (*positive* or *negative*), the AI will highlight several keywords within the movie review that were particularly relevant for its prediction. In this sense, the AI attempts to 'explain' its decision to you.

}

if (confidence) {

if (highlights) {

In addition to its prediction (*positive* or *negative*),

}

else {

Moreover,

}

the AI will indicate its *confidence* with regard to prediction. This score ranges from 0 to 1, where lower scores mean that the AI is uncertain about its prediction (i.e. it thinks that it is close to guessing) while higher scores mean that the AI is confident about its prediction.

}

After you have submitted your classification for one movie review, the next review appears on the screen. Overall, there will be 50 movie review to classify as either positive or negative.

After you have completed all classification tasks, we will ask you to fill out a short questionnaire. After that, a completion link will lead you back to Prolific. Please note that it is important to conduct the entire experiment in order to qualify for any payouts.

Once again, thank you very much for taking part in this experiment!

Let's start!

Appendix B. Machine Learning Model

The support tool used employed a classifier that operated on unigram bag-of-words features (word frequency counts in each review), normalised by term frequency and inverse document frequency. English stopwords[9] were removed prior to feature extraction. Bag-of-words feature vectors were used to train a regularised multinomial logistic regression model with stochastic gradient descent (SGD), using the python library sklearn and a regularisation parameter of 10^{-4} (Schmidt & Biessmann, 2019). Training data were the 25,000 movie reviews in the training data set of Maas et al. (2011), these data points were not used for anything but the training phase; model evaluation data as well as the data for the main experiment of this study were taken from the other 25,000 movie reviews in the referred study's test set.

Appendix C. Explanation Generation

We establish algorithmic transparency by 'explaining' the AI's predictions with *word highlighting*. More specifically we highlighted the top three words most relevant to the ML model prediction. This requires to compute feature importance scores for each word in a review. For computing such importance scores, one needs to choose from a plethora of methods (for more comprehensive lists, we kindly refer to Guidotti et al. (2018) and Samek (2019)). For this study, we employ the method proposed by Haufe et al. (2014), based on two reasons. For one, it is simple to implement and second it can be shown that this method compares favourably to other popular methods in empirical studies on interpretability quality (Schmidt & Biessmann, 2019). The generation of the importance scores for each word (or unigram feature; following Equation 6 from Haufe et al., 2014) yields class-specific feature scores of $a_k = X^\top y_k$where a_k is a d-dimensional vector of importance scores for each of the d words in the entire training corpus, X is a n-by-d matrix of bag-of-words features with n rows corresponding to the n reviews in the test data set and y_k is a n-dimensional vector with the predictions of the ML model for class k. For simplicity of notation we here assume that both the data as well as the predictions are centred and normalised to unit variance. With this assumption the importance scores are simply the covariance between the predictions of each class and each bag-of-word feature. To generate the explanations (presented to participants as highlighted words for each review text), we selected the feature importance scores a_k associated with the predicted class k, and ranked the words in a text according to the element-wise product of features x_k and feature/prediction co-variances 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. In our study we chose to highlight the three most highly ranked words by the aforementioned score. A python notebook with the entire data preprocessing, training and generation of explanations is provided in the accompanying GitHub repository.