Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach

Henrik Mucha
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB
Karlsruhe, Germany
henrik.mucha@iosb.fraunhofer.de

Sebastian Robert
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB
Karlsruhe, Germany
sebastian.robert@iosb.fraunhofer.de

Rüdiger Breitschwerdt
Wilhelm Büchner Mobile University of Technology
Darmstadt, Germany
ruediger.breitschwerdt@wb-fernstudium.de

Michael Fellmann
University of Rostock
Rostock, Germany
michael.fellmann@uni-rostock.de

ABSTRACT

Explanations in Human-AI Interaction are communicated to human decision makers through interfaces. Yet, it is not clear what consequences the exact representation of such explanations as part of decision support systems (DSS) and working on machine learning (ML) models has on human decision making. We observe a need for research methods that allow for measuring the effect different eXplainable AI (XAI) interface designs have on people's decision making. In this paper, we argue for adopting research approaches from decision theory for HCI research on XAI interface design. We outline how we used estimation tasks in human-grounded design research in order to introduce a method and measurement for collecting evidence on XAI interface effects. To this end, we investigated representations of LIME explanations in an estimation task online study as proof-of-concept for our proposal.

CCS CONCEPTS

• **Human-centered computing → User interface design**; **Visualization design and evaluation methods**; • **Computing methodologies → Machine learning**; *Artificial intelligence.*

KEYWORDS

Human-AI Interaction; Explainable Artificial Intelligence (XAI); Explanatory User Interfaces

ACM Reference Format:
Henrik Mucha, Sebastian Robert, Rüdiger Breitschwerdt, and Michael Fellmann. 2021. Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3411763.3451759

1 INTRODUCTION

Explanations in Human-AI interaction shall support humans in making more informed decisions. The information comes from data and the sense-making inner-workings of decision support systems (DSS) running on machine learning (ML) models. The results and ideally the sense-making process itself – an explanation of the machine behaviour – are presented to decision makers through human-machine interfaces. This is what we will refer to as explanatory interfaces.

The primary function of explanations is to facilitate learning better mental models for how events come about [21]. The concept of explaining machine behaviour is summoned under the term eXplainable Artificial Intelligence (XAI). Successful XAI systems must provide meaningful explanations to people. An explanation is an interaction between two parties, the explainer and the explainee [22] with the goal to align mental models. Human explainers and explainees can manage discrepancies between different mental models by having conversations. Consequently, explanations in Human-AI interaction must enable such conversations, too. Conversations need a common ground and a common language. In Human-AI interaction, such conversations would start with a representation of the explanation.

Designers and developers of XAI systems and their interfaces can draw on lessons from other sciences such as philosophy, cognitive psychology, and social sciences that produced a rich body of knowledge on how explanations work. Miller [22] gives a comprehensive overview. However, these accounts remain rather abstract from a design point of view. In other words, the design space of XAI interfaces is still largely unchartered. At the center is the question how the way we represent machine behaviour and reasoning as interfaces, i.e. specific design elements, affects human decision making and behaviour. The concept of interacting with learning systems is unprecedented and hence necessitates appropriate and most likely novel HCI research methods and tools. XAI research faces many challenges. One is identifying tasks, study designs, and evaluation criteria that allow for human-centered, scalable, and empirical research in the form of experiments to produce evidence for grounding design decisions.

In this paper, we describe estimation tasks and recommend them as a research approach intended to empirically study how different

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451759

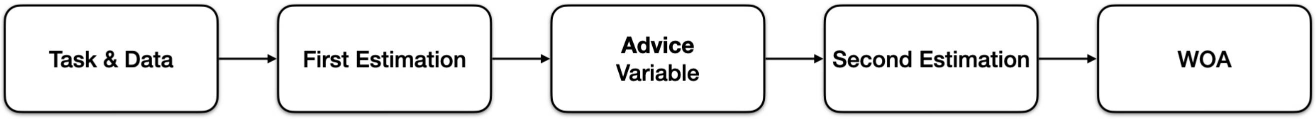

Figure 1: Flowchart of the Elements of an Estimation Task

representations of XAI, i.e. user interfaces, affect human decision making. Estimation tasks are experiments in which participants are asked to make an estimation based on a set of data before being presented advice from a decision support system. Based on this advice and the new information, participants then have the chance to re-assess their first estimation and make a second one. Thus, one can observe if people adhere to the system's advice or not and measure this effect.

In the remainder of the paper, we present related work, an outline of the characteristics of such tasks, and finally an estimation task online study (Mturk) comparing three different explanation designs for a LIME explanation model [29]. Thus, we provide a proof-of-concept example how HCI research can build upon established experiment designs from decision theory research in order to gather data on how different interface designs have an effect on human decision making.

2 RELATED WORK

Despite recent efforts from HCI, e.g., [20] [4] [27], designing human-centered explanatory interfaces for AI/ML remains a challenge. The notion of explaining algorithmic decision making to humans – generally referred to as eXplainable Artificial Intelligence (XAI) [8] [12] [28] [24] – is continuing to move into the scope of recent HCI research efforts. Recent publications cover most aspects of the thinking behind the concepts of XAI and the need for human-centered XAI design, e.g., [15] [27] [13] [2]. We focus on the design of explanatory interfaces and its effects on decision making as our research object. Examples of such interfaces for XAI are [23], [13], [14], [17], [16], [18], and [31]. However, to the best of our knowledge there is little evidence on how representing XAI as interfaces affects human-decision making due to a lack of evaluation in the form of suitable empirical user studies on large enough scales. Yet, such evidence is crucial for exploring the design space of explanatory interfaces. Informed by approaches such as Cheng et al. [4], El Shawi et al. [7] and others [35] [34] [19] [26], we want to add another methodic perspective by proposing to build upon established decision theory experiments from the behavioural sciences and adopt these for human-grounded [6] HCI research on XAI representations and interface design.

Therefore, we build upon the standard setting of a judge-advisor system [3]. In the judge-advisor system (JAS), the judge is the decision maker who receives advice either from an advisor or a computer system (e.g., a DSS), revises her judgment based on the advice, and then makes a (final) estimation. Advice, as is often in estimation tasks, refers to an uncertain or unknown fact that is relevant for the decision and that needs to be estimated as accurately as possible [30] [9]. As is the typical literature studying JAS settings, we are interested in how the judge utilizes the advice, and especially whether there is advice discounting, i.e., under-utilisation of the information. We apply the linear Bayes framework developed by Bates and Granger [1] for measuring the impact of advice. The framework is explicitly or implicitly used in most studies on advice taking (see, [32] [33] for references). In this framework, it is assumed that the judge applies a linear Bayesian inference mechanism to combine his first estimation with the advice to his final estimation. According to Goldstein [10], the linear Bayes approach is particularly relevant in a situation where the judge faces a complex task for which one cannot be expected to make a full prior joint probability estimation, i.e., an estimation of the distributions of both the first estimation and the advice. Instead, the linear Bayes approach only expects the judge to form a prior belief about the precision of his own and the advisor's estimate to form a final estimation. More specifically, the judge will revise her initial estimate according to the linear belief adjustment formula (see Bates and Granger [1], p. 453) as stated in our results section.

3 OUTLINE OF AN ESTIMATION TASK

In order to put this into perspective for HCI research purposes, we include this section as a more general outline of the approach: An estimation task is a task in which participants are asked to make an estimation based on their personal or professional experience and a set of data. Estimations simulate decision making under uncertainty, i.e. participants can most likely not know the true outcome but are required to approximate as closely as possible. As an incentive for making quality estimations, participants are compensated with a basic payoff which is supplemented by a bonus that is higher the better the estimation is. Thus, participants have an incentive to take into account as much additional information as possible.

We present an estimation task experiment which serves as a general example for how estimation tasks can be used in HCI research on explanatory interfaces and their effect on decision making. Such estimation task experiments follow a simple pattern: First, participants are asked to make an estimation based on information (data). Participants are then presented advice from a decision support system (an ML model performing the same task). The advice is constituted by the estimation of the ML model (a number) and an explanation of the estimation in the form of an interface often a diagram, e.g., depicting the influence (weight) certain attributes have on the ML-generated advice. Based on this advice and the information from the explanation, participants then have the chance to re-assess their first estimation and make a second one.

By presenting different interface designs across different treatments as independent variable in a between-subject study, we can

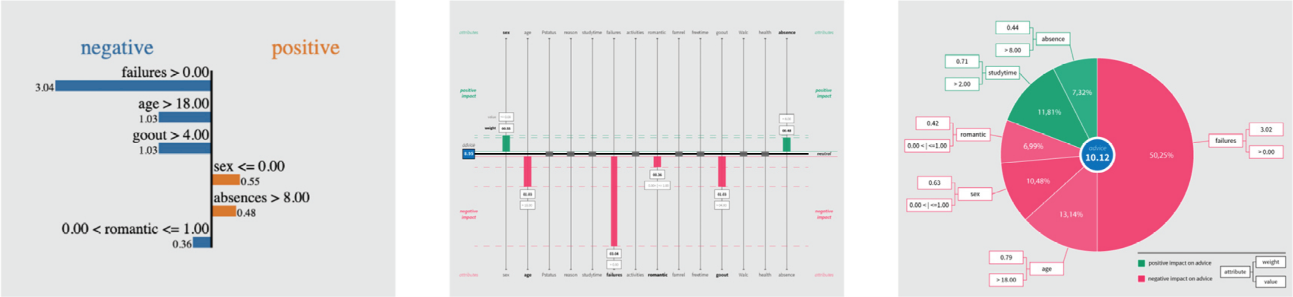

Figure 2: Three different XAI diagram designs used as treatments in the study presented from left to right: Standard LIME graph (BASIC), ABOVE/BELOW(C1), PIE CHART (C2).

compare how different designs have an effect on participants' decision making. In other words, we observe if people adhere to the system's advice or not and record this effect using weight of advice (WOA) as measurement. By comparing the WOA of different designs we can make empirical statements about the quality of each design regarding its influence on participant's decisions which gives also an indication regarding trust. Generally, it is assumed that due to overconfidence a higher WOA, i.e. adhering to the system's advice, leads to better overall decision making performance [25]. We may then build upon this evidence to describe a design space for XAI interfaces. We argue that this research approach and study design can contribute to satisfying the demand for more informed design decisions in XAI interface design in particular and Human-AI interaction in general. The next chapter gives a detailed account serving as a proof-of-concept example.

4 METHOD

We ran a between-subject online study using Amazon's MTurk platform and SoPHIE[1] software for experiment design. The task in this study was to estimate the final grade of a student based on a list of attributes. The procedure followed the process described beforehand (Figure 1). The study had three treatment conditions (Figure 2: The original LIME diagram was the baseline treatment (BASIC); C1 ABOVE/BELOW was the second treatment; C2 PIE CHART was the third treatment (see Figure 2). The different designs serving as treatments were developed in a series of remote design workshops with experts and users. The designs were developed as to emphasise different anticipated effect mechanisms and be sufficiently different. For instance, while BASIC and ABOVE/BELOW visualise absolute relations, PIE CHART visualises relative relations between weighted attributes. BASIC and ABOVE/BELOW are different in so far as the latter visualises all possible attributes and increases informational depth.

For the online study, we recruited participants on MTurk. Participants were presented an introduction to the task before they were asked to estimate the final grade of students based on a list of attributes. We ran three rounds with one estimation task (i.e., one student) in each round. The data we used for this study was from a real secondary school in Portugal [5]. To make an estimation, participants were given information such as academic and personal characteristics also called attributes. This data was collected by using school reports and questionnaires regarding the performance in Mathematics. Participants were explicitly asked to base their estimation on this information. The estimation had to be an integer between 0 (worst score) and 20 (best score). After the first estimation participants were presented the advice of an ML system and an explanation (LIME) in text form and a diagram. The advice in the form of text and the treatment specific graphic representation (diagram) depended on the treatment group participants were randomly assigned to. We asked participants to study these and then make a second estimation in which they could either adjust their first estimation or not. Participants were granted a basic compensation of US $0.50. Additionally, participants could receive a bonus payment of up to US $3.60 to give an incentive for quality estimations. The bonus was calculated as follows: In each round, both the first and second estimation were evaluated. The system compared the predictions with the true value and calculated the deviation. If a participant hit the true value, she received 20 points for this prediction. Each variance reduced the score accordingly (for example, an estimate of 16 with a true value of 13 gave exactly 17 points). After three rounds, all points were added up and converted into US $ with 10 points = US $0.30. Since participants estimated twice in each of the three rounds, they could earn a maximum of 120 additional points = US $3.60. Hence, the exact amount of a bonus depended on the quality of participants estimations. It took 15 minutes on average to complete the study for each participant. We designed the estimation task in this way for various reasons. In our design, we would know that the correct estimate would be an integer between 0 and 20, which considerably reduces noise in the data. Moreover, the task was a true estimation task in the sense that participants were aware that they cannot possibly know the answer but have some information to form an estimate. Finally,

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

Table 1: Values and mean of estimations over all three treatments (between-subject; n=20 per treatment)

Sample	Mean Estimation Round One	Model Prediction	Mean Estimation Round Two	True Value
Student One	10.23	8.95	8.75	8
Student Two	10.60	9.47	9.74	5
Student Three	11.14	10.12	11.16	13

the questions were related to each participant's personal world of experience, which we expected to ensure that the participant could easily understand the task.

5 RESULTS

A total of 60 participants took part in the online experiment. Three sessions were performed with 20 participants in each session. Participants were randomly assigned to one treatment condition, either PIE CHART, ABOVE/BELOW or BASIC. For the analyses, we have a total of 180 observations from 20 participants per treatment. Since we focus on how participants apply advice from a ML system, we use the weight of advice (WOA) as a measure for how much participants adhere to the advice [1].

$$WOA^{pi} = (st_2^{pi} - est_1^{pi})/(advice^i - est_1^{pi}) \qquad (1)$$

with.

$$est_1^{pi} = \text{first estimate of participant } p \text{ in round } i \qquad (2a)$$

$$est_2^{pi} = \text{second estimate of participant } p \text{ in round } i \qquad (2b)$$

$$advice^i = \text{advice from ML system in round } i \qquad (2c)$$

WOA measures the weight the participant assigns to the advice she receives form the ML system, and 1-WOA is accordingly the weight she places on her own initial estimate. Specifically, WOA takes a value of 0 if a participant adheres completely to her first estimate and WOA of 1 if she shifts completely to the advice from the ML system. The measure is not defined in case the advice is exactly equal to the first estimation ($advice^i = est_1^{pi}$). In a post-experimental questionnaire, we collected additional data on participants' motivations to make good estimations, and we asked participants about their perceptions of the graphical representation of the advice.

Figure 3 presents descriptive data on WOA for all three treatments and estimation rounds .[2] It shows that adjustments of participants are on comparable levels for both treatments BASIC and PIE CHART, but lower for the ABOVE/BELOW treatment (0.7737, 0.5461, 0.5959 vs. 0.7107, 0.5557, 0.5715 and 0.5142, 0.4999, 0.4523 respectively). This treatment effect is statistically significant between the PIE CHART and ABOVE/BELOW treatment aggregated over all three rounds of estimation (t-test, t(38) = 2.2018, p = 0.035 two-sided). In contrast, the comparisons of the other two treatments are not significant (all p-levels above 0.05). An analysis of estimations errors (deviation from est_1^{pi} and est_2^{pi} to the true value) revealed no significant results between treatments. Likewise, there are no significant effects in the PEQ Likert Scale ratings concerning usefulness of advice and graphic. However, the free text answers showed predominantly positive feedback, although we should also listen to the seven participants (11%) that stated that they did not take the visualisation into consideration at all or even regarded it as distracting and not useful. In summary, we observed selective and/or adaptive adjustments for treatments BASIC and PIE CHART while adjustment remains constantly low with ABOVE/BELOW treatment. We found there are in fact significant effects between treatments. We conclude that the exact form of visually representing (LIME) explanations is relevant for the design of explanations in Human-AI interactions.

6 DISCUSSION

From the results of the study we conclude that the exact form of visually representing (LIME) explanations has an effect on decision making in Human-AI interactions and thus must be further investigated.

We found a significant effect between treatments, i.e. different designs of the explanation interface. This stands in contrast to relevant other research on the topic, in this case Cheng et al. (2019) who found that "users' trust in algorithmic decisions is not affected by the explanation interface or their level of comprehension of the algorithm" through a similar study design but not using estimation tasks or WOA. Hence, it seems evident that more research is needed on the research object as well as on the methods we use.

Our contribution is to propose an approach for gathering empirical data and evidence that is informed by the behavioural sciences in general and decision theory in particular. In near future work, our goal is to describe a design space for explanatory interfaces grounded in this evidence. It shall address designers and developers and inform design decisions concerning explanatory interfaces in decision support systems.

While conducting this research we became aware of a limitation that needs to be addressed in future studies. While online (MTurk) studies are generally an appropriate research tool [11] they also have weaknesses in the context of design research. While allowing for fast and large-scale data collection as well as – and this is important by the time this paper is written amidst a global pandemic – safe research with human beings, remote online studies lack the possibility to observe and immediately question closely what people are doing and thinking (think aloud methodology). In retrospect, we came to the conclusion that further observational and qualitative data would have been helpful to even better analyse which parts of the explanation design people really focus on. In future work, it will be interesting to compare online studies with lab studies on the matter using the same study design, i.e., estimation tasks.

In conclusion to our research and for the design space to take shape, the following issues need to be addressed in future research:

[2] Since outliers in the data would considerably influence the analysis, outliers below 0 and above 1 were changed to values of 0 and 1 respectively.

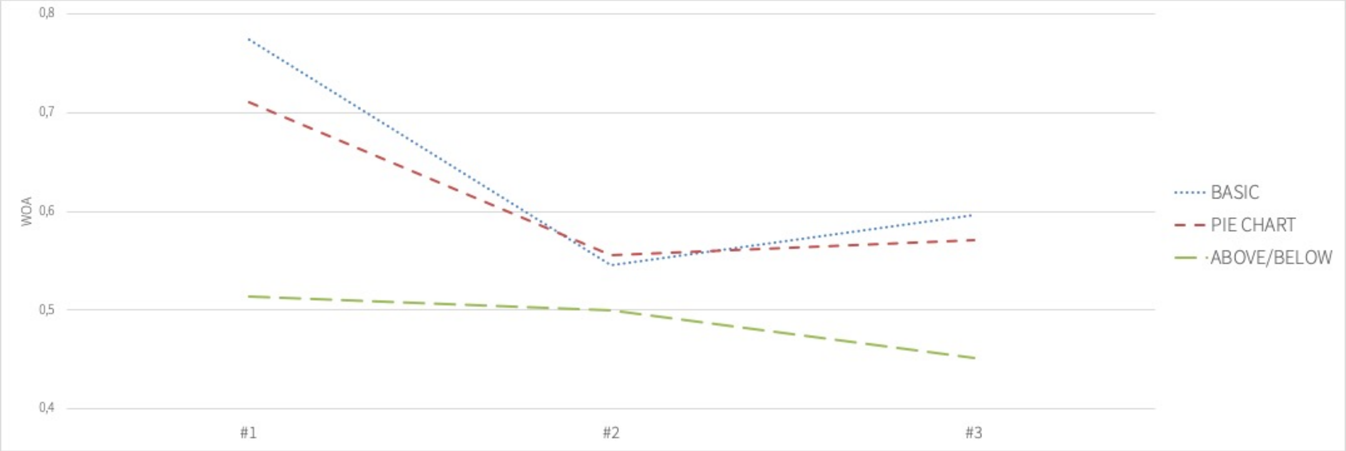

Figure 3: Descriptive data on WOA for all three treatments

- A taxonomy of proxy tasks should be developed based on knowledge from other disciplines, especially behavioural sciences and decision theory, and the learnings of the research presented here as to facilitate empirical human-grounded HCI research in Human-AI interaction that can meet the challenges described previously
- Study designs and experimental set-ups need to be developed further to even better observe effects in the design element (input) and decision making (output) relation
- More thought should be given to qualitative data and user observations (think aloud) to gain a deeper understanding which elements of explanatory interfaces are relevant to user's cognitive processes in decision making which calls for lab studies and/or the comparison of lab and online studies
- The complexity of representations must be gradually increased from static to interactive XAI interfaces as the goal are interactive explanations

7 CONCLUSION

In this late breaking work paper, we made a case for empirical research on the design of explanatory interfaces as part of decision support systems. We introduced estimation tasks for studying the effects of interface design on human decision making and as a proxy for decision making in more specialised domains. We described an online study on the effects of different representations of a LIME explanation model as a proof-of-concept for our proposal. We observed a significant effect between different designs and concluded that XAI representations as interfaces have an effect on decision making in Human-AI interactions and thus must be further investigated as to produce evidence as a basis for future design decisions. We intend to proceed with this research by means of the method we described and communicating the results as a design space for explanatory interfaces.

REFERENCES

[1] John M Bates and Clive WJ Granger. 1969. The combination of forecasts. *Journal of the Operational Research Society* 20, 4 (1969), 451–468.

[2] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[3] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

[4] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[5] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. (2008).

[6] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[7] Radwa Elshawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. 2019. Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 275–280.

[8] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[9] Francesca Gino and Maurice E Schweitzer. 2008. Blinded by anger or feeling the love: how emotions influence advice taking. *Journal of Applied Psychology* 93, 5 (2008), 1165.

[10] Michael Goldstein. 1999. Bayes linear analysis. In *Encyclopaedia of Statistical Sciences*, Kotz et al. (Ed.). Vol. 3. Wiley, 29–34.

[11] Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26, 3 (2013), 213–224.

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[13] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[14] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2017. A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 88–97.

[15] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 162–172.

[17] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the*

2016 CHI Conference on Human Factors in Computing Systems. 5686–5697.

[18] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2018. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 299–309.

[19] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[20] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[21] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[22] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[23] Yao Ming, Huamin Qu, and Enrico Bertini. 2018. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 342–352.

[24] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839* (2018).

[25] Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.

[26] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[27] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[28] Gabrielle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 19–36.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[30] Janet A Sniezek, Gunnar E Schrah, and Reeshad S Dalal. 2004. Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making* 17, 3 (2004), 173–190.

[31] Thilo Spinner, Udo Schlegel, Hanna Sch¨afer, and Mennatallah El-Assady. 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 1064–1074.

[32] Leigh Plunkett Tost, Francesca Gino, and Richard P Larrick. 2012. Power, competitiveness, and advice taking: Why the powerful don't listen. *Organizational behavior and human decision processes* 117, 1 (2012), 53–65.

[33] Lyn M Van Swol, Jihyun Esther Paik, and Andrew Prahl. 2018. Advice recipients: The psychology of advice utilization. (2018).

[34] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[35] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.