Making SHAP Rap: Bridging Local and Global Insights through Interaction and Narratives

Michael Chromik

LMU Munich, Munich, Germany
michael.chromik@ifi.lmu.de

Abstract. The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation-generating methods. In practice, Shapley explanations are widely used. However, they are often presented as visualizations and thus leave their interpretation to the user. As such, even ML experts have difficulties interpreting them appropriately. On the other hand, combining visual cues with textual rationales has been shown to facilitate understanding and communicative effectiveness. Further, the social sciences suggest that explanations are a social and iterative process between the explainer and the explainee. Thus, interactivity should be a guiding principle in the design of explanation facilities. Therefore, we (i) briefly review prior research on interactivity and naturalness in XAI, (ii) designed and implemented the interactive explanation interface *SHAPRap* that provides local and global Shapley explanations in an accessible format, and (iii) evaluated our prototype in a formative user study with 16 participants in a loan application scenario. We believe that interactive explanation facilities that provide multiple levels of explanations offer a promising approach for empowering humans to better understand a model's behavior and its limitations on a local as well as global level. With our work, we inform designers of XAI systems about human-centric ways to tailor explanation interfaces to end users.

Keywords: explainable AI · explanation interface · interactivity.

1 Introduction

Many decisions in our lives are influenced or taken by intelligent systems that leverage machine learning (ML). Whenever their predictions may have undesired or consequential impacts, providing only the output of the black box may not be satisfying to their users. Even if the prediction is accurate in regard to the underlying training data, users may distrust the system, have different beliefs regarding the prediction, or want to learn from individual predictions about a given problem domain. Thus, a need for understanding the ML model behavior arises [2]. The field of *explainable artificial intelligence (XAI)* develops novel methods and techniques to make black-box ML models more interpretable. Current XAI research mostly focuses on the *cognitive* process of explanation, i.e.,

identifying likely root causes of a particular event [21]. As a result, some notion of explanation is generated that approximates the model's underlying prediction process. Explanations may be textual, visual, example-based, or obtained by simplifying the underlying prediction model [3]. An approach widely used in practice is *explanation by feature attribution* [3]. Especially local explanations based on *Shapley values* [27] are widespread [4]. Feature attribution frameworks, such as SHAP[1], merely provide visual explanations and leave their interpretation entirely to the user. As such, they are targeting mostly ML experts, such as developers and data scientists. However, Kaur et al. [17] observed in their studies that even experts have an inaccurate understanding of how to interpret the visualizations provided by SHAP. Even if they are correctly interpreted by ML experts, they may still remain opaque to end users of XAI due to their technical illiteracy [6]. This applies especially to end users and subject-matter experts, who often have little technical expertise in ML. Thus, their interpretability needs require even more guidance and attention.

The main idea of this paper is to explore how to improve the accessibility of Shapley explanations to foster a pragmatic understanding [23, 11] for end users in XAI. We believe that an important aspect required to address the call for *"usable, practical and effective transparency that works for and benefits people"* [1] is currently not sufficiently studied: providing end users of XAI with means of interaction that go beyond a single static explanation and that are complemented by explicit interpretations in natural language. As the human use of computing is the subject of inquiry in HCI [22], our discipline *"should take a leading role by providing explainable and comprehensible AI, and useful and usable AI"* [34]. In particular, our community is well suited to *"provide effective design for explanation UIs"* [34]. Our work contributes to the HCI community in two ways: First, we present and describe the interactive explanation interface artifact *SHAPRap* that targets non-technical users of XAI. Second, we report promising results from a formative evaluation that indicates that our approach can foster understanding. With this work, we put our design rationales up for discussion with our fellow researchers.

2 Related Work

We base our work in the interdisciplinary research field of XAI. It aims to make black-box ML models interpretable by generating some notion of explanation that can be used by humans to interpret the behavior of an ML model [31]. An ML model is considered a black-box if humans can observe the inputs and outputs of the model but have difficulties understanding the mapping between them. However, most works focus on computational aspects of generating explanations while limited research is reported concerning the human-centered design of the explanation interface. The social sciences suggest that the explanation process should resemble a social process between the explaining XAI system (sender of an explanation) and the human explainee (receiver of an explanation)

[1] github.com/slundberg/shap

forming a multi-step interaction between both parties, ideally leveraging natural language [21]. Especially, in situations where people may be held accountable for a prediction-informed decision, they may have multiple follow-up questions before feeling comfortable to trust a system prediction. Abdul et al. emphasize that interactivity and learnability are crucial for the effective design of explanations and their visualization [1]. Widely used explainability frameworks, such as SHAP, present their explanations in the form of information-dense visualizations, however, they do not provide any interactivity nor guidance to support users in their interpretation process. As a consequence, even experienced ML engineers struggle to correctly interpret their output and often take them at face value [17]. Humans mostly explain their decisions with words [19]. Thus, it is intuitive to provide end users of XAI with explanations in natural language. We found first work that takes a human-centric perspective on XAI and encompasses interactivity and naturalness. Weld and Bansal [32] propose seven different follow-up and drill-down operations to guide the interaction. Liao et al. [18] compile a catalog of natural language questions that can technically be answered by current XAI methods. Covering multiple of them under a *"holistic approach"* allows users to triangulate insights. Reiter [24] discusses the challenges of natural language generation for XAI. Further, users have been shown to understand technical explanations better if they are complemented by narratives in natural language [9, 10, 13]. For instance, Gkatzia et al. improved users' decision-making by 44% by combining visualizations with statements in natural language [13]. Sokol and Flach [29] present *Glass-Box* an interactive XAI system that provides personalized explanations in natural language. Similarly, Werner [33] presents *ERIC* an interactive system that gives explanations in a conversational manner through a chat-bot like interface. Forrest et al. [12] generate textual explanations from feature contributions based on LIME [25].

3 SHAPRap

3.1 Scenario, ML Model, and XAI Method

Scenario. Our XAI system is centered in a decision-support situation in which the human decision-maker is accompanied by an intelligent and interpretable system. We put our study participants in the shoes of a private lender on a fictional crowd lending platform. We centered our study in a crowd lending domain because we assumed that the participants can relate to decisions about lending or investing personal money. Participants can see demographic information, loan details, and credit history of individuals that request a loan on the platform. Each request is accompanied by an "AI-based intelligent prediction" of the *default risk*, i.e., the probability that the borrower fails to service a loan installment some time during the loan period. The prediction is introduced as an "AI-based" feature that is based on machine learning from historic cases. We build on a tabular data set as many ML models deployed in practice build on this type of data [4, 20]. We used the *Loan Prediction*[2] data set which consists

[2] datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

of 614 loan requests with 13 columns. We relabeled two columns of the data set to be consistent with our scenario[3].

ML Model. We calculated the default risk prediction using a *XGBoost classifier*. Tree-based ensembles, such as XGBoost, are widely used in many real-world contexts because of their practicability [20]. However, they are considered black-box ML models. To limit the cognitive load for participants we chose to train our model on a subset of columns. We used only the seven categorical columns (5 binary, 1 ternary, and 1 with four possible values). We trained a binary XGB classifier with 100 decision trees and class probabilities as outputs. Other than that, we used the default hyperparameters of the *xgboost* package. The accuracy of the predicted default risk on our stratified validation set was 0.83.

XAI Method. In this work, we use the *SHAP (SHapley Additive exPlanations)* [20] framework to compute the model's feature contributions on a local and global level. SHAP belongs to the class of *additive feature attribution methods* where the explanation is represented as a linear function of feature contributions towards an ML prediction. The contributions are approximated by slightly changing the inputs and testing the impact on the model outputs. The framework unifies the ideas of other feature attribution methods (such as LIME [25]) with *Shapley values*, which originate from game theory [27]. Shapley explanations quantify the contribution of individual features values towards a prediction. For a single observation, they uniquely distribute the difference between the average prediction and the actual prediction between its features [20]. For example, if the average prediction over all instances in a dataset is 50% and the actual prediction for a single instance is 75%, SHAP uniquely distributes the difference of 25 percentage points across the features that contributed to the instance's prediction. Despite their vulnerability to adversarial attacks [28] and potential inaccuracies [14], we consider Shapley explanations as relevant to end users for two reasons: (i) they can yield local and global insights because Shapley values are the atomic units of each explanation. As these units are additive, they may be aggregated over multiple predictions or features to learn about the model's global behavior, and (ii) the consistent and model-agnostic nature of Shapley values allows XAI designers to offer a uniform explanation interface to users even if the underlying data or ML model changes.

3.2 Explanation Interface

Local Explanation View. The local explanation view resembles a spreadsheet-like user interface that is overlaid with a heat map of Shapley values for each feature of an instance. We support users' rapid visual estimation of feature contributions through preattentive processing based on a cell's hue [15]. Each cell is shaded depending on their direction and magnitude of contribution towards the prediction (red increases the loan request's risk of defaulting, while green decreases it).

[3] we re-framed the *Loan_Status* column to represent the default risk and the *Credit_History* column to represent a negative item on a credit report.

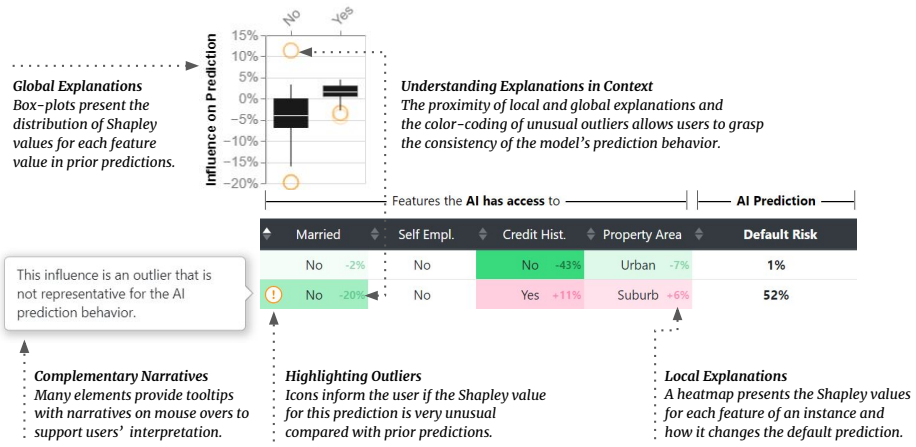

Fig. 1. The components of the *SHAPRap* explanation interface

The local explanation view is *contrastive* [21] as it allows comparing variances *between* feature contributions for individual instances (*horizontal axis*). Further, as we show multiple local explanations next to each other, users can compare variances or regularities *within* feature values across multiple instances (*vertical axis*). To support this, users can sort each column by value to contrast instances with identical feature values.

Global Explanations View. Local explanations yield how an ML model derives its prediction for a single data instance. In contrast, global explanations help users to get an intuition how a model derives its predictions over multiple instances or an entire dataset (*global sample*). For each feature value, we provide a box-plot of how it contributed to the prediction for all instances in the global sample. A narrow box-plot indicates a more consistent prediction behavior, while a wider box-plot indicates that the contributions vary for the same feature value. These variances result from interactions with other features and may require additional judgment (see next paragraph). The distribution of Shapley values in the global view depends on the chosen global sample. If the sample is representative for the population that the ML model will be confronted with in a particular domain, the global view helps users understanding when its predictions are consistent and therefore predictable and when they are not. In practice, the global sample may be the entirety of predictions of an ML model after its deployment across all users, or (if data sparsity requirements apply) a sample of predictions that an individual user has previously been exposed to. Further, it would be possible to let users customize the global sample (e.g., only instances above a certain prediction threshold or instances with a particular feature value). In our prototype, we displayed the distributions of the training and validation sets.

Highlighting Outliers. A post-hoc *explanation by feature attribution* approach, such as SHAP, is always an approximation of the actual prediction behavior of an ML model. Identifying inconsistent contributions and communicating them to the user can improve their interpretation by making it easier to identify ex-

planations that are more representative for the global model behavior. We built around the concept of role-based explanations [5]. We classify each instance's feature value contribution into the roles *normal* (within the *inter quartile range (IQR)* of the global sample), *unusual* (beyond IQR but within whiskers as defined by $\pm 1.5 \times IQR$), and *very unusual* (outliers beyond the whiskers). We highlight *very unusual* contributions in the global and local views as orange warning circles prompting the users to not generalize from these instances to the typical prediction behavior of the model. Further, these outliers may serve as starting points for analyzing feature value interactions. When hovering over an outlier, we highlight features of this instance that are *unusual* and thus provide hints which feature values may be interacting with each other.

Complementing Narratives: It is not easy to understand the concepts of additive Shapley explanations just by looking at plots [17]. It might take some time to interpret a plot, and the user is likely to be overwhelmed at first. Thus, we automatically created textual explanations from Shapley values using a template-based approach and to support their interpretation of the local and global views. We provide users with on-demand textual explanations in form of tooltips on mouseovers for each feature box-plot, instance cell, outlier highlight, and column header. Further, we provided background information about the local and global

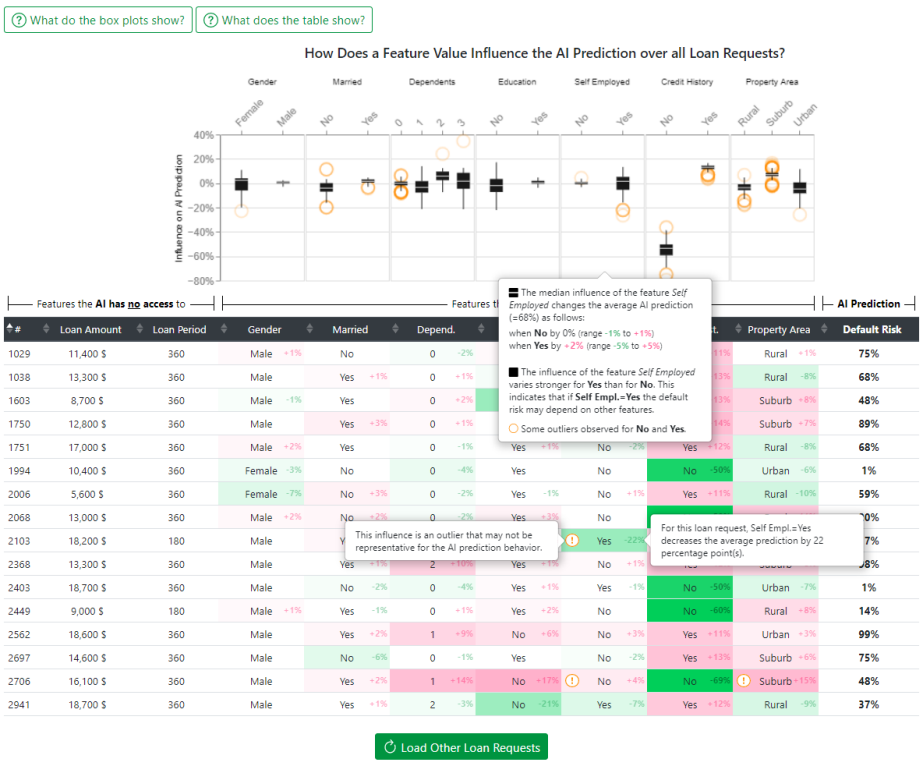

Fig. 2. The explanation interface that participants were exploring.

views during onboarding and accessible through help buttons during interaction. This way, information redundancy can be avoided following the *progressive disclosure* paradigm [30].

4 Formative Evaluation

Method. We conducted a formative evaluation with 16 participants recruited through the online platform *Prolific*. We recruited participants with at least a graduate degree, English fluency, and an approval rate of 100%. 8 participants self-identified as female, 8 as male and were in the age groups 18-24 (3), 24-35 (9), and 35-54 (4). 11 participants agreed to use spreadsheets at least weekly, 6 knew how to read box-plots, and 4 had practical experience with ML. After introducing their role in the crowd lending scenario and the explanation views, users were asked to freely explore *SHAPRap* for 10 to 15 minutes. Then, they rated their level of understanding on a 7-point scale[4] [8]. Afterwards, they completed a *forward prediction* quiz [7]. Participants had to simulate the AI prediction for 6 pre-selected loan requests with the help of the global explanation view. We randomly chose 6 instances with unique feature value combinations and at most two *unusual* contributions to assess participants' understanding of the typical prediction behavior. In the end, they rated the *explanation satisfaction scale* [16] and answered three open questions. On average, participants took 28.1 minutes (SD=10.4 minutes) to complete the study and were compensated £5 per completion (=£10.67/hour).

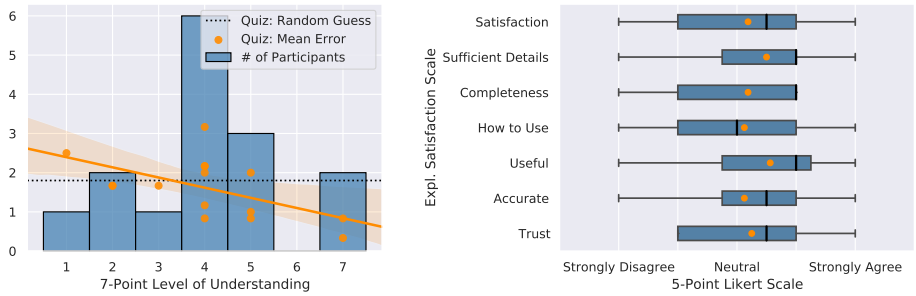

Fig. 3. *(left)* 11 participants perceived they understood at least which features were important for the prediction. 6 of them objectively proved their understanding via a lower than random mean error in a forward prediction quiz. *(right)* Results from the *explanation satisfaction scale*. The orange dots indicate the respective mean.

[4] Level 1: *I understand which features the AI has access to and what the AI predicts as an output.*, Level 4: *I understand which features are more important than others for the AI prediction.*, Level 7: *I understand how much individual feature values influence the AI prediction and which feature values depend on others.*

Results. Overall, our results indicate mixed reactions but show effective gains of pragmatic understanding for some participants. The explanation facility felt overwhelming at first, but the complementary elements of global, local, and textual explanations were considered as somewhat useful and sufficiently detailed to get a general idea about the typical prediction behavior. After exploring *SHAPRap*, participants on average rated their understanding as *"I understand which features are more important than others for the AI prediction"* (mean=4.07, SD=1.67). However, applying this understanding in the quiz turn out to be challenging for 6 participants as they scored worse than random guess (expected error for a random guess was 1.8). For example, P5 *"understood what the box representations meant but found it hard to actually apply this data to the applicants. It might just require practice."* On a positive end, 6 participants rated their gained understanding as at least level 4 and proved this with low mean errors in the quiz (cf. Fig. 3). Participant P6 (no ML experience, mean error of 0.8) *"found the explanations quite complicated to follow but after studying the table and explanations it became clearer as to which factors were being used to measure the likelihood of defaulting on the loan."* P3 (extensive ML experience, mean error of 0.33) found *"the explanations were detailed, and it was interesting to see that credit history was the leading variable for default risk."* Multiple participants appreciated the complementary nature of the natural language explanations. Without them *"the graph was quite difficult to understand on its own"* (P6). P13 liked *"that the [textual] explanations are written simply, everyone would understand it"* and P9 appreciated that the *"language was simple"*. However, it seemed that narratives on a more aggregated or abstract level were missing to understand the bigger picture. P4 found *"this kind of explanations useful just to people who already have studied this but for people with different educational background this kind of explanations are not enough."* P5 suggested adding an executive summary for each loan request and the overall global view. Further, some participants were overwhelmed by the non-linear behavior and interactions of the ML model and seemed to expect to figure them out. P5 found *"the green and red increase/decrease for risk seemed simple and helpful at first, but there seemed to be very random correlations between different aspects."* Similarly, P10 stated: *"I am guessing there are so many intersecting correlations it's hard to read for a non-numbers person."*. This resonates with Rudin [26] that the term *explanation* is misleading as it suggests a full understanding can be reached even if we merely provide pragmatic approximations.

5 Summary

This paper presents the explanation interface *SHAPRap*, which supports end users in interpreting local Shapley explanations in the global context of *normal* and *unusual* model behavior. Further, it provides narratives using a template-based approach. With our work, we contribute to the development of accessible XAI interfaces that enable non-expert users to get an intuition about the probabilistic decision behavior of black-box ML models.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. CHI '18 (2018). https://doi.org/10.1145/3173574.3174156
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access (2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion (2020). https://doi.org/10.1016/j.inffus.2019.12.012
4. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P.: Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020). https://doi.org/10.1145/3351095.3375624
5. Biran, O., McKeown, K.: Human-centric justification of machine learning predictions. IJCAI '17 (2017). https://doi.org/10.24963/ijcai.2017/202
6. Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society (2016). https://doi.org/10.1177/2053951715622512
7. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H.: Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. CHI '19 (2019). https://doi.org/10.1145/3290605.3300789
8. Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A.: I think i get your point, ai! the illusion of explanatory depth in explainable ai. IUI '21 (2021). https://doi.org/10.1145/3397481.3450644
9. Das, D., Chernova, S.: Leveraging Rationales to Improve Human Task Performance. IUI '20 (2020). https://doi.org/10.1145/3290605.3300789
10. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., Riedl, M.O.: Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. IUI '19 (2019). https://doi.org/10.1145/3301275.3302316
11. Eiband, M., Schneider, H., Buschek, D.: Normative vs. pragmatic: Two perspectives on the design of explanations in intelligent systems. In: IUI Workshops (2018)
12. Forrest, J., Sripada, S., Pang, W., Coghill, G.: Towards making nlg a voice for interpretable machine learning. In: INLG (2018). https://doi.org/10.18653/v1/W18-6522
13. Gkatzia, D., Lemon, O., Rieser, V.: Natural language generation enhances human decision-making with uncertain information. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-2043
14. Gosiewska, A., Biecek, P.: Do not trust additive explanations. ArXiv (2020), https://arxiv.org/abs/1903.11420
15. Healey, C.G., Booth, K.S., Enns, J.T.: High-speed visual estimation using preattentive processing. ACM Trans. Comput.-Hum. Interact. (1996). https://doi.org/10.1145/230562.230563
16. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. CoRR (2018), https://arxiv.org/abs/1812.04608

17. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. CHI '20 (2020). https://doi.org/10.1145/3313831.3376219
18. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI '20 (2020). https://doi.org/10.1145/3313831.3376590
19. Lipton, Z.C.: The mythos of model interpretability. ACM Queue (2016). https://doi.org/10.1145/3236386.3241340
20. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A.J., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence (2020). https://doi.org/10.1038/s42256-019-0138-9
21. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2019). https://doi.org/10.1016/j.artint.2018.07.007
22. Oulasvirta, A., Hornbaek, K.: HCI Research as Problem-Solving. CHI '16 (2016). https://doi.org/10.1145/2858036.2858283
23. Páez, A.: The Pragmatic Turn in Explainable Artificial Intelligence (XAI). Minds and Machines (2019). https://doi.org/10.1007/s11023-019-09502-w
24. Reiter, E.: Natural language generation challenges for explainable AI. In: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019) (2019). https://doi.org/10.18653/v1/W19-8402
25. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). https://doi.org/10.1145/2939672.2939778
26. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (2019). https://doi.org/10.1038/S42256-019-0048-X
27. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games (1953)
28. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020). https://doi.org/10.1145/3375627.3375830
29. Sokol, K., Flach, P.A.: One explanation does not fit all. KI - Künstliche Intelligenz (2020). https://doi.org/10.1007/s13218-020-00637-y
30. Springer, A., Whittaker, S.: Progressive Disclosure. ACM Transactions on Interactive Intelligent Systems (2020). https://doi.org/10.1145/3374218
31. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. CHI '19 (2019). https://doi.org/10.1145/3290605.3300831
32. Weld, D.S., Bansal, G.: The challenge of crafting intelligible intelligence. Communications of the ACM (2019). https://doi.org/10.1145/3282486
33. Werner, C.: Explainable ai through rule-based interactive conversation. In: EDBT/ICDT Workshops (2020), http://ceur-ws.org/Vol-2578/ETMLP3.pdf
34. Xu, W.: Toward human-centered AI: A perspective from human-computer interaction. Interactions (2019). https://doi.org/10.1145/3328485