even if commute patterns are individual and a commuter is the only expert in their patterns. A commuter, using our prototype, can teach the agent selected commutes. Results in this area can give indication and inspiration regarding future studies in shared domains, for example, radiology [14] or personal domains, for example, personalisation of intelligent personal assistants and assistive technology.

We aimed at evaluating interpretability inspired by the considerations that Doshi-Velez and Kim [4] outline. Our results are in line with that work, but our work adds practical knowledge on how the considerations can be applied and used in a real setting. We also unpack local and global interpretability in relation to evaluation. Our focus also add awareness around how explanations and interpretability relates to the end user of the agent. We do this by drawing attention to the term explicability, in this work defined as a term between interpretability and explanations that implies that a domain expert, on their own, can formulate an explanation for an explicable agent.

We proceed as follows, first, we give an overview of related work and introduce our approach. We then outline the study setup that we evaluate in a result and analysis section. Before concluding the results are elaborated in a discussion section.

2 Related work

The report Ethics guidelines for Trustworthy AI [9] highlights three components of AI agents: ethical, robust and lawful. Important for the ethical component is that the agent is explicable, implying a transparency that makes the agent explainable and contestable to those affected [9]. In the report, the terms explicability and explicable are used, according to Merriam-Webster explicable means *capable of being explained*[3]. The report also highlights that demands concerning explicability is highly dependent on the context and severity of the consequences. In this work, we will have a focus on interpretability, explicability, and explainability as an important part of the ethical component and less focus on the components robustness and lawful.

One view on interpretability is the coupling to transparent algorithms, like linear models or decision trees, algorithms that can make the entire model transparent and simultaneity is possible for a human. This coupling is contested [15, 16] in that interpretability challenges has less to do with the choice of model and are more connected to the complexity of the agent and thus, that, for example, linear models are not automatically interpretable and neural networks not automatically black-boxes from an explainability perspective. If the features and labels used in a neural network are understandable, a human can in some cases explain the predictions using the relation between feature values and labels, even if the network as such is a black box.

In an overview of explanations and justification in machine learning Biran and Cotton [1] provides this definition of explainability and interpretability:

[3] Merriam-Webster dictionary, accessed 2020-01-24

> Explanation is closely related to the concept of *interpretability*: systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation.

In this work, we define an end user as anyone that uses the system, a usage that does not demand any ML or domain expertise. Among these end users, there are domain experts/teachers with a teaching goal, this goal is manifested as a knowledge transfer from the teacher to the ML agent. This transfer is performed without any need for ML expertise. We use the terms domain expert and teacher interchangeable.

There exists, in related literature, a limited discussion on requirements on the end user's capability to understand the explanation, although the complexity in providing good explanations are acknowledged [4, 18]. In this work, we will focus on the end user and therefore we distinguish between interpretability, explicability, and explainability in relation to the end user. We define interpretability as a property of the agent in line with the Biran and Cotton [1] definition, which can imply a need for Machine Learning (ML) expertise and/or domain expertise to produce an, for an end user, understandable explanation. For an explicable ML agent, we define that, a domain expert is needed to adapt or rephrase the explanation in understandable terms to the end user. Explainability is then, as we define it, a quality of an ML agent that does not need any human intermediary to adapt or rephrase an explanation to be understood by an end user. An ML agent can then provide, depending on the situation, predictions that are interpretable, explicable or formulate an explanation concerning the prediction. Implementing agents that are capable of providing explanations implies understanding what constitutes a good explanation[18]. The focus in this work is, instead of the problem of formulating good explanations, on interpretability and explicability and ML systems that involves a human domain expert capable of formulating explanations.

Doshi-Velez and Kim [4] emphasize that a need for interpretability builds on downstream goals as fairness, privacy, reliability, robustness, causality, usability and trust. The authors also urge researchers to both define these goals as well as question if interpretability is the right tool to achieve these goals. Interpretability is seen as a quality needed when the problem formulation is incomplete. The authors also state three modes of interpretability evaluation with increasing specificity and cost: Functionally-grounded evaluation with no real humans and proxy tasks, Human-grounded evaluation with real humans and simple tasks and finally Application-grounded evaluation with real humans and real tasks.

2.1 Machine teaching

Simard et al. [21] primarily sees MT as a paradigm shift decoupling the domain experts from the ML experts in order to build systems that can be trained-deplayed and used without the involvement of ML experts. In that work, the importance of a domain-specific language is highlighted, a language that builds

on the taxonomy for the specific domain. MT is then, as we define it, an approach that gives a domain expert control over subjectively selected knowledge to be transferred to a model in a machine learning agent using a domain-specific language. An important part of this language is the user interface that facilitates human-agent interaction.

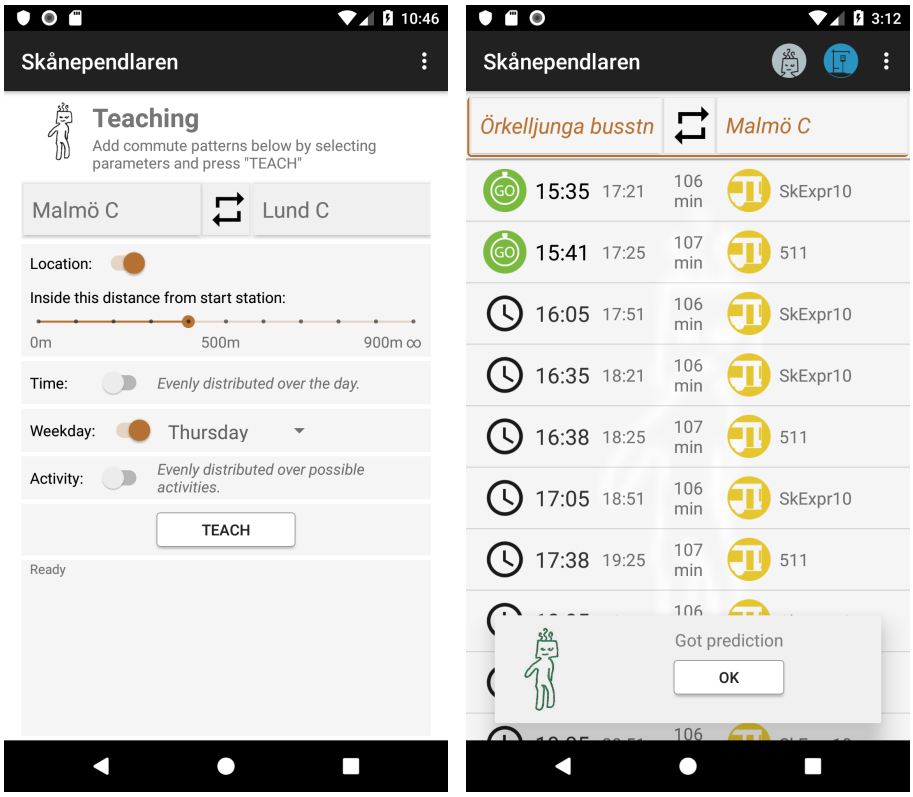

Fig. 1. To the left, the apps teaching interface is shown and to the right, the interface for the predictions is shown.

3 Research setting

The domain commuting has some qualities that makes it interesting as a research domain in the area of interpretable MT, the domain is well-known, accessible but still commute patterns are individual. This implies that a commuter is domain expert in their commute patterns. We built a functioning prototype for the commuting domain that consist of a MT interface and a prediction interface. The final design can be seen in fig. 1.

An important part of any machine learning implementation is the selection of features to use. This is also central in an MT implementation since an available orthogonal feature adds an extra dimension to the feature space and thus gives the teacher further possibilities to separate classes. For commuting the end user's location, day and time seemed to be a natural selection, we also added the end user's activity (still, walking running, in vehicle) as a feature. Since we are interested in commuting we chose the complete journey as our label, consisting of an origin-destination pair.

Neural networks can handle multiple levels of representation and raw input data [13, 10]. By selecting neural networks as our ML method we can simplify interpretation since the learner and the commuter uses the same, for a domain expert, understandable features and labels without any feature engineering. We decided to use a neural network framework from fast.ai[4] since that framework is open source and aims at simplifying experimentation and rapid prototyping. We used the fast.ai tabular application and the automatic learning rate finder implemented in fast.ai.

The black box property of neural networks is a challenge since interpretability only can be evaluated on a agent level and not on a model level. The black box property forces us to aim for model agnostic interpretability which is in line with a growing research interest [20, 22, 17, 12, 8]. Data-hunger is another property of neural networks that has to be mitigated. In a setting like ours, that is cold started, augmented and/or synthetic training data needs to be created during the teaching sessions. For the ML-pipeline in our implementation an online database (Firebase) is used for communicating with the end users Android applications, a Node.js server orchestrates training, prediction and deployment by communicating with a Flask server and the online database. The fast.ai models, one for each commuter, are hosted on the Flask server.

Our initial general target scenario was: a commuter that starts the Android application and expects a journey prediction within a few seconds. Before evaluating interpretability we made sure that training time and prediction time were short enough to be useful in the target scenario. With our ML pipeline and a small neural network we reached prediction times in the area of a few seconds a retraining time around 15 seconds, we assessed these metrics as sufficient for the study.

4 Methodology

To evaluate interpretability from a functional and human perspective we created three personas [19] [5] based on personas created for a local transport provider. For interpretability evaluation on an application level, we conducted an eight-week study. As study participants, we selected eight participants, four males, and four females, with experience from computer science and interaction design. Two were from the industry and six from the University. The participants used

[4] fast.ai

[5] https://github.com/k3larra/commuter#personas

the app daily and we conducted six meetings and workshops. At the meetings, we discussed and compared experiences from last week and discussed the tasks for next week. We analyzed recorded interviews using content analysis [7] focusing on comments on interpretability. The participants were reimbursed by free journeys during the study.

5 Result and Analysis

In this section, we describe the process of evaluating interpretability for the prototype. The steps in this process are presented in a chronological order and for each step we move closer to application level evaluation. The focus is on producing predictions that full-fills the downstream goals of being causal, trustworthy and usable. Causal so that predictions map the commuters intentions, trustworthy and usable so the predictions reflect the context cue of the commuter. Our problem formulation is incomplete and can thus benefit from being interpretable since individual commute patterns cannot be defined in advance and are subject to change [4].

For the prototype, we aimed for an explicable agent, for which, a domain expert/commuter can formulate explanations for performed predictions. For example, based on a given journey prediction, a commuter could formulate the following explanation "Since it is Tuesday evening I am going to town to meet Johan and play boule". When predictions do not match the commuter's expectations there could be three types of reasons. Either, since it is an MT setting, the commute pattern has not been taught to the agent, the commute patterns shifted or the agent is not robust and cannot provide an explicable prediction. The information that is intended to be explicable consists of the predicted journey and the context cue (location, time, day and activity).

5.1 Functionally-grounded Evaluation: no humans, proxy tasks

In this more general and global evaluation, we aimed to make predictions interpretable in relation to the different features. In this MT setting no data exists initially to train from, and the data is created in a teaching session by selecting sub-spaces in the feature space using the interface in figure 1.

We experimented and found that using two hidden layers (200|100 neurons), and creating under 100 examples inside the sub-space made it possible for the training to converge in less than 20 epochs [6]. With this setting, retraining time is still under 30 seconds for training data consisting of up to a few thousand rows on a non GPU cloud server.

We then evaluated the interpretability by keeping all features constant except one so we could investigate how the generalization maps to interpretability. A visualization of one experiment can be seen in figure 2 where all features except location is kept constant and four different journeys are predicted, represented

[6] https://github.com/k3larra/commuter/blob/master/machine_teaching/mt.ipynb

by the different colors. The images shows a situation where the generalization is designed so a journey should be predicted from the closest station.

In the scenarios we created for our personas we defined that a commuter only wants to make predictions centred around an origin station. An alternative edge scenario could be a commuter that wants to teach the prototype to make predictions for someone else, for example, a child, based on a location different from the origin station's location. The result would be, if for example more red teaching data (figure 2) was placed at a location to the left, that classification boundaries would be more complex. To avoid this, since we judged that this would result in predictions that would not be explicable, we only allowed one location, with user-defined size, surrounding the origin station for a journey.

We made similar tests and reasoning using scenarios concerning the other features, but made no restrictions. As an example, if a journey is taught on Monday mornings we deemed that a logical generalization would be to predict the same journey on all weekdays when no other conflicting journeys exists.

The functional test ended with daily usage by the development team to verify the robustness of the implementation.

Fig. 2. Location feature space, markers represent predictions, blobs represent teaching data and areas are classification boundaries.

5.2 Human-grounded evaluation: proxy humans, simplified tasks

Compared to the interpretability evaluation that Doshi-Velez and Kim [4] names as human grounded using real humans and simplified task we saw a need for a more controllable and reproducible evaluation than using real humans could give. Using the design team for this was time-consuming and prone to errors, for example, inaccuracy in GPS position and delays in transport made it difficult to reproduce tests. Here our personas were used as a trade-off. For each of our personas, we initially created three scenarios and trained our agent using

the teaching interface to make predictions matching the scenarios. A scenario formulated for the persona Andrea [7] can be read below:

> It is Monday morning 7:23 and Andrea is as usual late for the bus and runs towards the bus stop "Veberöd förskköningen". She checks the app while running to see if there are any delays.

To get more realistic data we generated three datasets that mimics one-year journeys for the personas. These datasets were used as our test sets to evaluate the models performance. We used our teaching interface to teach the agent and evaluated the result using visual representations as those that can be seen in figure 3. Metrics around how many epochs/time the training needs was balanced towards the amount of teaching data we created (the bar graph). A confusion matrix was used to visualize journeys that are not explicable. Based on these tests the final configurations for the ML pipeline and back-end were decided. (Two hidden layers with (200|100 neurons), 40 synthetic examples created evenly distributed inside the teaching sub-space and training over 7 epochs [8].)

Decisions at this level has consequences for the interpretability in-between global and local since we focus on results based on data for our personas. The diversity and complexity of the commute patterns for the personas has a normative influence for the behavior of the agent.

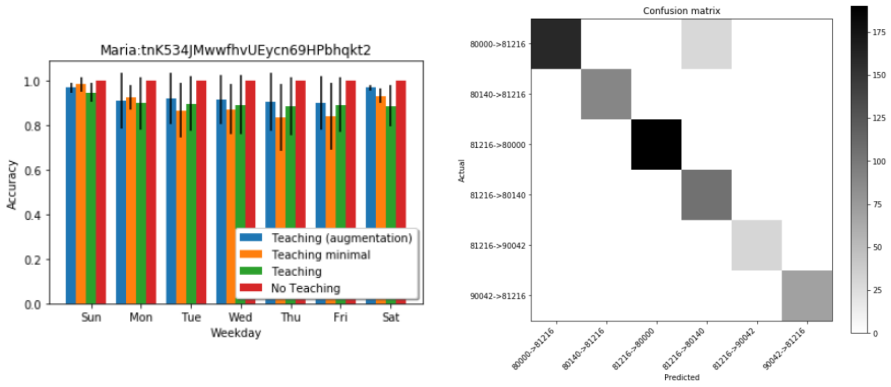

Fig. 3. Tools used during human grounded interpretability evaluation. The test uses synthetic data generated for an envisioned one-year usage for the personas. The data was created using the teaching interface (figure 1). To the left accuracy and variance is aggregated over the week using different amount of generated teaching data (No teaching indicates a baseline), to the right precision is mapped in the form of a confusion matrix.

[7] https://github.com/k3larra/commuter#personas
[8] https://github.com/k3larra/commuter/blob/master/machine_teaching/mt.ipynb

5.3 Application-grounded evaluation: real humans, real tasks

This part of the interpretability evaluation involves real humans and real tasks. A consequence of our MT setting is that it contains a temporal aspect in the form of teaching strategies. The participants explored different strategies, for example, teaching once out of context in the beginning of the week, teaching only when in context, if the predictions were wrong and a combination of these strategies. The focus for this phase was on local explicability, the situation when the end user compares the predicted journey with the intended journey and current context cue.

The participants, in general, found the predictions interpretable in relation to the commute patterns they already taught, or as expressed by one of the participants:

> But it did correct .. around half of the times it was wrong. So it is still logic. Even if i doesn't favor me so to say.

The predictions generalized as intended in the sense that if a journey was taught on one day at a specific location and time it was predicted for all other days at that time. We discussed this inherent quality in machine learning compared to applications based on traditional algorithmic programming. We especially discussed the situation when predictions are made with low class probability and how this should be handled. One participant expressed that predictions should only be given "close" to the teaching data:

> But I would like to say that if you are under some percentage .. and the app has really no idea of where you want to go.. that a dialog comes up.. or that a question-mark is shown to say "I do not know please teach me where you want to go right now"

Over time and with extensive teaching the study participants found that the predictions given deviated from the commute patterns taught and thus the predictions were not explicable. There are different explanations for this. Firstly, our prototype was built and tested using quite simple commute patterns based on scenarios and personas and those do not match more complex commute patterns. Secondly, the deviation can stem from the fact that, in the current implementation, teaching cannot be undone if mistakes in the teaching is made and erroneous journeys are taught.

6 Discussion

In this work we focus on post-hoc agent-level explicability a situation were a human domain expert can explain the predictions given contextual features like time, day and location and previous teaching. In our setting, we have a one to one relationship between a human domain expert and a machine learning agent. The agent is initially untrained and designed to target the commuting domain.

With our work we add to the interpretability terminology by Lipton [15] and Doshi-Velez and Kim [4]. We use the term explicability to indicate a quality of the agent and distinguish between predictions that are interpretable and those that are explicable. Interpretable predictions does, with this distinction, need both ML expertise and domain expertise to be explained whilst explicable predictions can be explained using domain expertise.

When our prototype is used the end user can, for an explicable prediction, formulate an explanation that matches the intended journey. If the prototype is robust, and the wrong journey is predicted, this can be because it has not yet been taught the commute pattern or that the user for one or another reason chose not to teach the pattern. For our prototype, this local in context explicability, works well for simple commute patterns. A possibility to assess the model's knowledge from a global perspective and not only locally and in context would be an important addition from an explicability perspective.

We designed our prototype with the intention of maximizing the part of the predictions that are explicable. By evaluating interpretability during the design phase from a functional-, human- and application level, we found, in line with Doshi-Velez and Kim [4] that the cost and specificity increases with those levels. From a functional level, it was clear that design consideration had a global impact on the final design. By using personas [19] instead of real humans in the human level evaluation we could strike a balance between a human perspective and still get reproducible results. The impact from this evaluation, which sits between a global and local perspective, to a large extent defines the portions of predictions that in the final application are explicable. A larger and more diverse collection of personas with more complex commute patterns would in our case produce a prototype that is explicable for a more diverse user group. The application-level evaluation gave us insights into the temporal aspect of the models taught knowledge and explicability on a local level.

Over time and in more complex teaching environments it will be crucial to assess the learner, in our case a short description, perhaps in the form of a global explanation, using interpretable decision set [11] could be enough like "If weekday=Yes and Location=A and Time $>$ 18:00 and Time $<$ 22:00 then travel from A to B" to make the predictions explicable. If the areas in the feature space could be named during the teaching using concepts and sub-concepts [21] the interpretable decision set could be used in a more easily interpretable fashion for example: "Exercise = Yes, Rugby=Yes then travel from A to B" or rephrased in natural language. Presenting predictions on a global level by selecting features of interest, from a calendar or map perspective, has parallels with work by Lakkaraju et al. [12] in MUSE framework where features of interest can be selected to investigate predictions in sub-spaces of interest.

The TED [8] and MUSE [12] framework could be an interesting addition, implementing TED in our prototype would be beneficial in that meta-information in the form of concepts and sub-concepts can be added to labels corresponding to specific journeys. This would imply that interpretable information can, for example, be added for the concept "Exercise" that includes all journeys the

commuter performs as sub-concepts to "Exercise". The concepts can then be used to inspect sub-spaces of interest in a form similar to MUSE. Our approach has similar goals concerning trust as Teso and Kersting [22] but differs in that the agency in our case is closer to the human teacher that decides the extent of the knowledge that should be transferred.

Our work then suggests that through designing an explicable MT agent for a domain, over time the agent can be trusted. To reach this the agent has to be either, as in our case, matching a significant amount of predictions to the context in an explicable manner and/or support exploration regarding the model so the knowledge transferred can be assessed and understood.

7 Conclusion

This paper set out to explore how interpretability can be evaluated in an MT setting. Addressing evaluation of interpretability is an emerging research field [3, 2], to this, our work adds a case study and evaluates interpretability as a design objective. One contribution is an increased focus on explicability as a relational quality of the ML agent. Explicability indicates that a domain expert can explain a prediction so it can be widely understood. How local and global explicability relates to the evaluation of interpretability are other contributions that can be important in many domains. Our work also shows how initial design goals concerning interpretability and explicability for an ML agent deeply influence the final result.

For future work, we suggest an increased focus on tools that support assessing a model's knowledge from an end user's, domain expert's and ML expert's perspective. Further research in this area can strike a balance between explicable and interpretable predictions and thus MT agents that evolve in incremental short cycles [5] in order to adapt to changes and support personalisation.

References

1. Biran, O., Cotton, C.: Explanation and Justification in Machine Learning: A Survey. IJCAI Workshop on Explainable AI (XAI) **8**(August), 8–14 (2017)
2. Boukhelifa, N., Bezerianos, A., Lutton, E.: Evaluation of Interactive Machine Learning Systems. ArXiv pp. 1–20 (2018)
3. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608 (2 2017)
4. Doshi-Velez, F., Kim, B.: Considerations for Evaluation and Generalization in Interpretable Machine Learning. In: Explainable and Interpretable Models in Computer Vision and Machine Learning., pp. 3–17. Springer (2018). https://doi.org/10.1007/978-3-319-98131-4_1
5. Dudley, J.J., Kristensson, P.O.: A review of user interface design for interactive machine learning (6 2018). https://doi.org/10.1145/3185517
6. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018 (2019). https://doi.org/10.1109/DSAA.2018.00018

7. Graneheim, U., Lundman, B.: Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Education Today **24**(2), 105–112 (2 2004). https://doi.org/10.1016/J.NEDT.2003.10.001
8. Hind, M., Wei, D., Campbell, M., Codella, N.C.F., Dhurandhar, A., Mojsilović, A., Ramamurthy, K.N., Varshney, K.R.: TED: Teaching AI to Explain its Decisions. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society pp. 123–129 (11 2018). https://doi.org/10.1145/3306618.3314273
9. HLEG: Ethics Guidelines for Trustworthy AI (European Commission, 2019). Tech. rep., High-Level Expert Group on Artificial Intelligence (2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
10. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks **2**(5), 359–366 (1 1989). https://doi.org/10.1016/0893-6080(89)90020-8
11. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. vol. 13-17-Augu, pp. 1675–1684 (8 2016). https://doi.org/10.1145/2939672.2939874
12. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and Customizable Explanations of Black Box Models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 131–138. ACM (2019), www.aaai.org
13. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
14. Lindvall, M., Molin, J., Löwgren, J.: From machine learning to machine teaching. Interactions **25**(6), 52–57 (10 2018). https://doi.org/10.1145/3282860
15. Lipton, Z.C.: The Mythos of Model Interpretability. In: ICML Workshop on Human Interpretability in Machine Learning (WHI (2016)
16. Lou, Y., Caruana, R., Gehrke, J.: Intelligible models for classification and regression. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 150–158 (2012). https://doi.org/10.1145/2339530.2339556
17. Lundberg, S., Lee, S.I.: An unexpected unity among methods for interpreting model predictions. arXiv preprint arXiv:1611.07478 (11 2016)
18. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences (2019). https://doi.org/10.1016/j.artint.2018.07.007
19. Nielsen, L.: Personas - User Focused Design. Springer, London (2013). https://doi.org/10.1007/978-1-4471-4084-9
20. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. pp. 1135–1144. ACM Press, New York, New York, USA (2 2016). https://doi.org/10.1145/2939672.2939778
21. Simard, P.Y., Amershi, S., Chickering, D.M., Pelton, A.E., Ghorashi, S., Meek, C., Ramos, G., Suh, J., Verwey, J., Wang, M., Wernsing, J.: Machine Teaching: A New Paradigm for Building Machine Learning Systems. Tech. rep., Microsoft Research (2017), http://arxiv.org/abs/1707.06742
22. Teso, S., Kersting, K.: Explanatory Interactive Machine Learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 239–245. ACM (2019). https://doi.org/10.1145/3306618.3314293