Limits of XAI application-grounded evaluation: an e-sport prediction example

Corentin Boidot, Olivier Augereau, Pierre de Loor, Riwal Lefort

Limits of XAI application-grounded evaluation: an e-sport prediction example

Corentin Boidot[1,2][0000−0001−7857−0714], Olivier Augereau[1][0000−0002−9661−3762], Pierre De Loor[1][0000−0002−5415−5505], and Riwal Lefort[2][0000−0002−5863−8221]

[1] Lab-STICC UMR CNRS 6285, École Nationale d'Ingénieur de Brest, France
{boidot,augereau,deloor}@enib.fr
https://labsticc.fr/fr/poles/interaction
[2] Crédit Mutuel Arkéa, Le Relecq-Kerhuon
riwal.lefort@arkea.com https://www.cm-arkea.com/

Abstract. EXplainable AI (XAI) was created to address the issue of Machine Learning's lack o transparency. Its methods are expanding, as are the ways of evaluating them, including human performance-based evaluations of explanations. These evaluations allow us to quantify the contribution of XAI algorithms to human decision-making. This work performs accuracy and response time measurements to evaluate SHAP explanations on an e-sports prediction task. The results of this pilot experiment contradict our intuitions about the beneficial potential of these explanations and allow us to discuss the difficulties of this evaluation methodology.

1 Introduction

Machine Learning (ML) has made significant progress in the last decade, not just in research, where beating "state-of-the-art algorithms" has become a standard, but also in business. Its use is becoming commonplace, as shown by the growing need for regulation [10]. However, employing these models frequently entails putting a process under the control of a black box: an algorithm whose behavior is unknown to the user [11]. A data scientist can look back on the behavior of his model. However he will rarely be able to specify the role of each parameter in his model, let alone guarantee the calculation's logic. However, not all operations may be left in the hands of a black box for reasons of control, safety, trust, or legal liability: in industries such as medicine, banking, and the military industry, models are needed to provide not only results but also a valid interpretation of those results [24,3]. Explainable Artificial Intelligence (XAI) is a field of study that has grown in popularity in recent years to address these demands.

Depending on the specific ML framework or model, one will find various XAI methods, given that many methods try to be compatible with any model. This variety of methods matches the variety of goals of XAI, and questions about users' needs should be asked before one chooses a method. Do the explanations convey accurate representations, and helpful information? Is the user overloaded

by information or put off by its formulation? Algorithmic properties of the methods can be studied but the effects these explanations will have on different persons remains uncertain. That is partly because the graphical presentation of the explanations and interface ergonomics can interfere with these effects.

XAI therefore requires the implementation of explanations' evaluations in order to know whether the explanations produced the desired effects. One such evaluation technique is human task performance evaluation, also called *application-grounded evaluation*, which is one of the techniques that involve humans in the loop. In this work, we consider the user an expert without ML knowledge: explanations have to support their decision-making process. Evaluating the usefulness of XAI methods by measuring performance requires a given framework: a task, training data, a human operator, and an ML model intended as a decision aid.

The model is trained using the data, and the user makes decisions based on AI guidance and its own understanding of current data. The decision support system can thus be limited to provide the result of the model on the current data, or it can be complemented by additional information-rich interfaces in the XAI framework.

Accuracy is the main metric for binary decision, and decisions are made sequentially so that we can measure speed. Using these performance criteria, we compare the results of decisions made with and without explanations. This performance analysis should provide a quantifiable determination of the quality of human decision-making.

Hoffman et al. proposed different evaluation techniques in their review [14], and measuring performance is the closest to the human agent's actual use. That is why we find it valuable, and we want to apply it to simple data (i.e., tabular) and a simple task format: binary classification.

In this paper, we attempt to evaluate a popular XAI method, SHAP [21], on a simple task: predicting the outcome of an e-sports match. E-sport offers us possibilities because a lot of players could be treated as experts: we can hope to generalize results to other domains' expert explanations. On top of that, we can easily find open data about this game.

Our goal is to predict which of the two teams will win, using event summary data from the first few minutes of a League of Legends match. This task sets the stage for our performance evaluation. Our XAI system measures the accuracy and speed with which human users respond. The difference between human performance from data and an ML analysis explained by SHAP vs. performance without explanation should establish an evaluation of SHAP's influence on decision-making for our task with our interface.

The rest of the article is organized as follows: section 2 presents the state of the art in XAI, focusing on its evaluation techniques, with the work of Jesus et al. [15] that we partially replicate. Section 3 describes our expectations of the results and the methodological framework used. Section 4 reports our results, which are then discussed more broadly in Section 5 before concluding in Section 6.

2 State of the Art

The explainability of ML models can be approached in different ways. It can be seen as the global explanation of a model's functioning, or as a process that seeks to explain the output of a model for a particular input. This study fits into the general "post-hoc" explanation framework [1], which consists in developing an explanatory method for models already designed and trained without concern for interpretability (as opposed to the development of transparent models). We can distinguish the following explanation methods: features importances [25], counterfactuals explanations [27], prototypes [2], model simplification [4], textual explanations [8] and model visualization [18]. These categories are not mutually exclusive: we can find methods at the borders of two categories [16].

One of the most famous XAI methods is SHAP [21], a post-hoc explanation method by feature importance, designed for local analysis (but it can be used for global explanation). It uses calculations from game theory to determine the positive or negative contribution of each input feature to each individual outcome. This method is particularly popular for explaining tree-based models, thanks to an optimized implementation that circumvents the computational costs of the method [20].

2.1 Evaluation of XAI

In addition to creating explanatory methods, one must also be concerned with their evaluation. One of the ongoing problems in XAI field is the lack of a fixed definition of what an explanation should be [9]. This problem is due to the subjective essence of explanations, the diversity of situations in which they are used, and the numerous and potentially conflicting objectives (simplicity, fidelity, completeness) they may pursue from one context to another. In particular, this context includes the nature of the target audience of the explanation (data scientist, layperson, application domain expert, or auditor). The methods created have therefore generally been evaluated more qualitatively than quantitatively [6,22]. However, the thoughts on how to assess XAI methods have flourished so that the main categories can be identified.

On the one hand, we can use purely computer-based evaluations: we test properties of the explanations such as the diversity of the answers or their complexity. On the other hand, we can use human evaluation, either through simple test tasks, or in real conditions. The first ones allow a quantitative evaluation

but can be disconnected from the fundamental objectives of explainability, if the relevance of the tested tasks is not assessed. Moreover, these evaluations are generally adapted to the type of explanation evaluated and do not allow for comparing explanations of a different nature [19].

In the area of human evaluation, a distinction is made between two approaches. Firstly, those which are based on subjective measurements (the user is asked to rate the "comprehensibility" or various criteria related to their feeling) [7]. Then, those that will study the human-AI system from the outside, by measuring the subject response times, or the accuracy of the decisions made, and comparing the use of explanations against different baselines. Such objective measurements can also be done to evaluate more subjective properties: simulatability [12] can be measured to estimate the impact of explanations on user's mental model of the AI.

If we rely only on the former, there is a risk that the research will move towards pleasant but misleading explanations: Ehsan et al. [7] suggest that our positive biases towards AI may prevent us from adequately evaluating its outcome and result and its explanation. However, both are often performed simultaneously in the same experiment.

2.2 Application-grounded evaluation methodology

We want to deploy an XAI system and evaluate it. We restricted this pilot development to a case of binary decision task with tabular data. This case has not been intensively evaluated but the work of Jesus et al. [15] seems particularly significant to us, for the conclusions they draw as for their methodology. Their study is an example of performance-based evaluation that does not rely on hypotheses about the structure of explanation or mental model [17].

Jesus et al. evaluate through practice three types of explanations (SHAP, LIME and TreeExplainer) for financial fraud detection task. Each transaction is scanned out for fraud detection, independently of the others (data is therefore tabular, not sequential). They evaluate the decisions of three expert fraud analysts, through five experimental conditions. Their tests are performed with data sampled around the decision boundary of the model: the three experts are not systematically exposed to the same data but only partially, on a sample used to establish an agreement score. The five experimental conditions are presented successively to each subject, as a long series of decisions to be made, first with the data alone (first condition), then with the data and the ML score (second condition), then with each of the three explanations in the last three conditions.

Their study raises a first half-tone analysis: if the explanations have made it possible to make decisions faster, the accuracy of the experts' judgment has not improved compared to the case where they analyze the raw data and would

even degrade it.

We want to know if this conclusion generalise to other XAI systems designed to support expert decision. In the following, we adapt their methodology to an other application domain: e-sport prediction.

3 Methodology

Hypotheses For our performance measures, our predictions are based on the work of Jesus et al. [15]. Our null hypothesis is that SHAP explanations should have the following impact:

1. The accuracy of user responses should be improved by the explanations
2. The response time should not be affected by the explanations.

These assumptions stem from the fact that we chose a "data + ML score" condition as the baseline. Otherwise, the explanation may represent a gain in time and a loss in accuracy. We measure different indicators of satisfaction, trust, and transparency (described in Appendix), where we expect to have "neutral" indicators with respect to the scales proposed to the user (answers centered on a Likert scale).

User The subjects for these early experiments are students, with a potentially wide spectrum of expertise on the proposed task (some may spend all their free time on the game, and others may meet it for the first time through the experiment). Specifically, we could only to retain data from five research training students with little or no knowledge of machine learning. Of the five, only one knows the game well (user 5 in the results section).

Data Data are aggregate match stats from League of Legends, taken at 10 minutes of play[3]. League of Legends is an competitive online game, known for its high visibility on the e-sports scene. This data thus contains a potentially engaging problem for students, given the popularity of the game; the task makes sense in that there is a market for betting on these matches.
On the presented dataset, 23 columns have been selected to be displayed to the users. Redundant columns have been removed: we preferred using direct statistics applied independently on the two teams instead of differences between both teams. The 39 games displayed were balanced regarding both blue and red teams' victories and error rates in both cases.

Model The model chosen to perform the AI prediction is a Random Forest[4]. We would not use deep learning models but rather tree ensemble methods as

[3] kaggle.com/bobbyscience/league-of-legends-diamond-ranked-games-10-min
[4] sklearn implementation, scikit-learn.org

they constitute state of the art for tabular data [26].

A few remarks about the data: each column corresponds to a performance of one of the two teams, and there is always a symmetrical column representing the result of the other team (except for "FirstBlood" feature). The model does not exploit this property. Moreover, this data is highly aggregated: one may wish to access individual performances for each team's different players, or even to display a video of the match to the users. These data are doubly "incomplete" since the outcome of a match is not defined after 10 minutes: there can be many turnovers so the problem may be considered from a probabilistic angle. Our model achieves a performance of 72% on its test data, which represents 25% of the dataset.

Experimental procedure We use two different experimental conditions: the first starts with explained data, the second starts with just data and row score. In both conditions, the user is exposed successively to "explained" and "non-explained" views in equal proportions. Each participant is first assigned an ID that determines the condition used. The whole experiment is implemented using an interface made with streamlit[5], which presents the context of the experiment, data format, and the explanations format through two example pages before starting the prediction task. Then, the prediction task is done on each match data with the interface in Fig. 1. This graphical block is left empty for data with no explanations. Decisions are made using a cursor set on a scale of seven values, in order to express potential uncertainty on the result. We use sub-series from 4 to 10 matches, after which the interface mode changes (between explained and not explained interface[6]).

For each game, the answer and the response time are recorded. After the predictions, a form is proposed to the user to get feedback and collect information about his profile. The experiment lasted between 30 and 45 minutes for each candidate. In both conditions, users are exposed to the same data, in the same order, only the presence of explanations may vary.

[5] streamlit.io

[6] the users with an odd number have first matches without SHAP explanation, the pairs start with SHAP. From match 19 on, they are exposed to the same interface mode

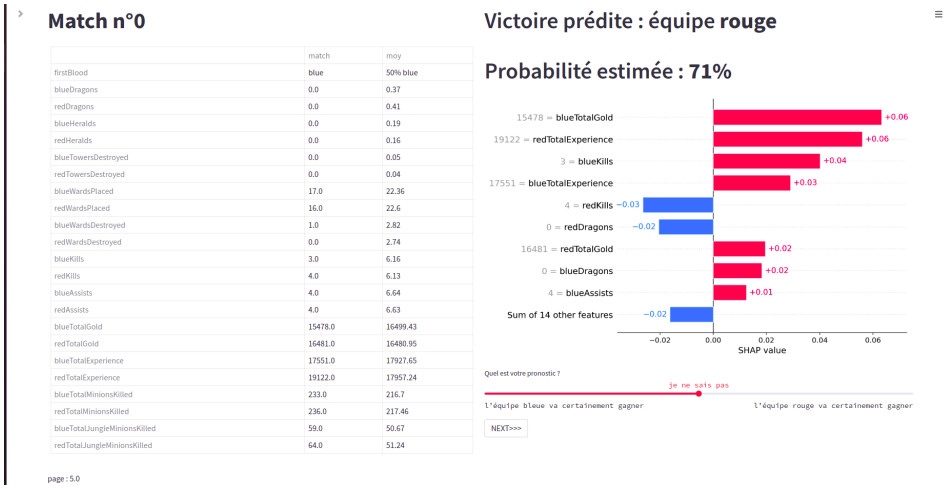

Fig. 1. Decision interface displayed to the user: on the left, the data are presented in three column: feature names, feature value for the current match and averages for reference. On the right at the top, prediction with SHAP in graphical format: the last bar of the graph represents the lesser contributions of SHAP added together. Below, a seven-step cursor that allows the user to express his prediction (translation in Appendix).

4 Results

Response accuracy For the full analysis, we set aside the nuances in the degree of certainty of the responses to keep only three-valued data: they can only be neutral, predictions of the red team victory or the blue team victory. For the accuracy calculations, neutral responses are difficult to interpret. We decided to keep them as they represent about 10% of the responses. As correct answers were counted 1 and incorrect answers were counted 0, we decided to count the neutral as 1/2.
A first observation is that our users have mostly made decisions in accordance with the the AI's suggestions. On the 195 predictions produced, we find 23 neutral predictions (12.3%) that are not or not easily analyzable, 27 predictions going against the AI (13.3%) and 145 predictions that follow the AI (74.4%).
If we consider that AI scores near .5 express uncertainty (we will consider a score of 60% or less as uncertain), we can see that this rate of agreement increases to 86% for cases where the AI looks confident and 64% for cases where the AI looks uncertain. We can guess that the users answered intuitively in this case, while they would rather tend to follow the AI's decision.

In general, the presence of explanations seems to have little influence on users' agreement with the model. At best, we can observe a negative effect on the accuracy of the decisions, as shown in Fig. 2. This result is opposite to the

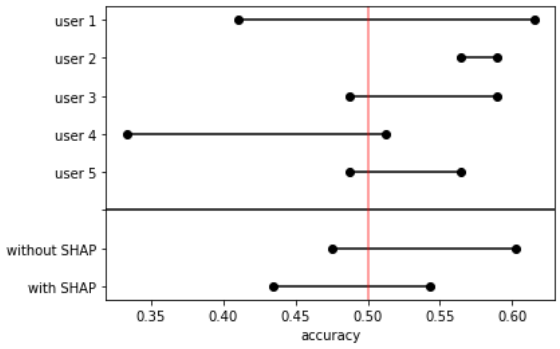

Fig. 2. Accuracy minimal and maximal estimates, considering neutral responses as misses (left points) or as good answers (right points). Our choice to count them as 1/2 equate to using the middle of these intervals as estimates. Above, we consider the global influence of the explanations' exposure on these measures, below we consider the influence of the user.

results of Jesus et al. [15], which suggested that exposure to SHAP explanations should increase accuracy compared to exposure to the ML score alone.

Response time Interestingly, it is difficult to identify any effect on response time: some decisions were made in a matter of seconds, while the longest took about 2 minutes. In Fig. 3, there is no observable effect of the explanations: the variance seems to be dominated by users' internal factors.
Some partial continuity appears in the sequences of response times, with explainable exceptions: peaks at the first decision and at the first change of interface (removal/addition of SHAP to indexes 10 19 25 30 35). We could also see that decision time decreases on average during the experiment, likely because of ha-

user	SHAP	accuracy	time
1	without	0.57	31.6 ± 32.8
1	with	0.5	38.4 ± 19.7
2	without	0.65	18.7 ± 16.8
2	with	0.53	10.8 ± 10.8
3	without	0.48	52.0 ± 27.6
3	with	0.58	49.2 ± 28.7
4	without	0.55	35.6 ± 26.3
4	with	0.39	14.4 ± 17.3
5	without	0.57	24.9 ± 13.3
5	with	0.5	22.3 ± 11.0

Table 1. Mean performances of users, with and without explanations.

bituation, and assume that some data intrinsically require longer analysis.

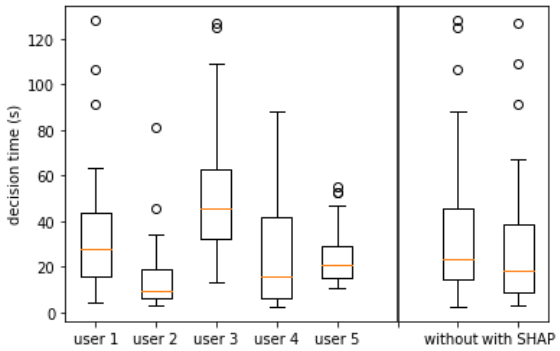

Fig. 3. On the left, box plots of the response time of the different user. On the right, the same data separated given the presence or absence of SHAP explanations.

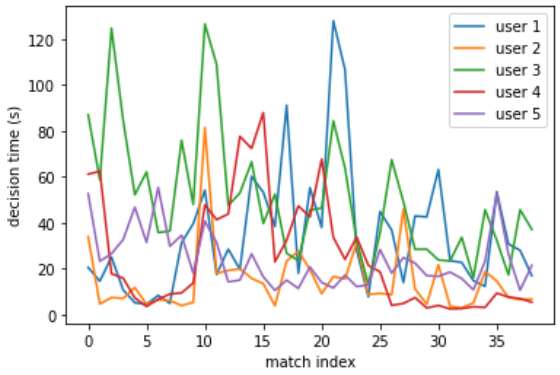

Fig. 4. Decision times graphs of the five users for each match

Although negative, these results indicate the need to solicit a large number of users and not neglect the influence of experiment construction on the measurement. The first trials of the users will necessarily be long: it is necessary to foresee a training phase and avoid changing the users' context inside the experiment too often or too abruptly.

Subjective evaluation This experiment used only one method of explanation, SHAP. The questions were asked in French (an English translation can be found

in the appendix). Likert scales with seven levels were systematically used to collect the answers (adapting the formulation to the question). Two subjects answered the open-ended questions that were offered. Despite the explanation's lack of objective benefit, there was still some positive feedback. For example, the question "The analysis interface[7] was useful to me in making a prediction" was answered with: "somewhat agree", "agree", "agree", "strongly agree", "strongly disagree". Only the last user gave consistently negative responses, commenting in the open field: "I did not use it, it did not always give a true representation of the gap between two teams. A small gap could be represented with a large bar and therefore could be misleading. In addition, not all important values were used" (in response to "How was your experience with the model analysis interface? How did you use it?"). To the question "What would you expect from an AI trying to explain its decision to you?", he answered: "His ability to be very sure of a result but also to say when one cannot conclude anything definite". In addition to user 5, user 3 answered the same questions: "I used it to either help me with aspects I didn't understand/didn't know how to interpret, or to validate what I thought" and "Schematics but maybe also additional captions/additional explanations".

5 Discussion

Our data support our second hypothesis: explanations do not seem to alter the decision time (compared to the case with ML score). However, they do not favor our first hypothesis: the accuracy was not improved but decreased. We must be cautious about drawing any conclusions about these measures. On the one hand, our users do not know much about the game: for example, the most experienced one does not play ranked games, which is however common among frequent players. Moreover, even if a subject would have a deeper experience of the game, understanding the tabular data certainly requires an effort of adaptation, which can reverse our model of what an "expert" would be. Obviously, the small number of users coupled with the small number of predictions made per user is the major limitation of our study: we will scale up the experiment a revised protocol.

One of the troublesome points of this experiment and of Jesus' et al. is the choice to work on a subsample of the data selected by the model. We can consider that the model is relatively inaccurate on these data, and therefore has little useful information to bring through the explanation. This would explain why user accuracies above 50% fall back to the mean once exposed to the explanations but not the cases where it falls below. Of course we do not already know what kind of understanding our model could have, and whether our explanation will indeed convey such an understanding to the user. Nevertheless, that is precisely why we should be cautious that our model has learnt complex patterns before using it on XAI evaluation purpose. Reciprocally, if the human capacities seem wholly exceeded by the model, the information it would provide a sort of popularization or justification (at best teaching [5]) than explanation. Considering this distinction, we see that establishing a framework of measurement to

[7] i.e., the graphical representation of Shapley values

compare humans and AIs before any explanatory experiment is crucial. Only in this way can we have an a priori idea of the explanatory situation in which we are situated and of the benefit that we will draw from a method of XAI. It seems to us that the model's explanations, by analogy with its use between humans, are to be considered in a cooperative framework between entities of comparable expertise. XAI's usefulness would be demonstrated in terms of performance if it allowed a "joint decision-making" strictly more accurate than the one made by the human alone or the model alone. That was the way of proving the existence of the "wisdom of the crowds" for instance [13].
In first approach, we use the accuracy of answers as an indicator of intelligence and expertise in the task. An independent sample must be used to provide an objective comparison of the accuracies of a model and humans. Therefore, the sample should not be artificially balanced against ML's errors. However, this is open to some criticism: it potentially leaves many "simple cases" on which humans and AI would have agreed anyway and where the addition of an explanation does not seem relevant from the point of view of accuracy. It then becomes vital to estimate the intrinsic "difficulty" of each data point, but there is no general approach to this problem. We should not let a model estimate this difficulty for itself, nor humans judge it for themselves. Also, we should distinguish in this difficulty what is an actual complexity, exploitable by a calculation, and what is a simple absence of information.

Beyond this idea of local data difficulty in our experimental sample, the global nature of the task can greatly influence on the human capacity to handle the problem and the possibility for the models to reach high performances: characterizing this nature is not always easy. Two other major factors that are difficult to control for our evaluations are the user profiles, and the influence of the user interface. Beyond the selection criteria for our subjects, which of course introduce biases, humans represent noisy decision systems, which do not necessarily return the same result twice for the same problem [23]. This noise can be ignored when the number of users is high but could present an important limitation to implementing meaningful measures with small populations and should then be estimated.
Finally, there is nothing to tell us that the choice of SHAP is precisely responsible for the user's different decision-making. Maybe the simple provision of a graphical interface with importance values that are consistent with the problem but ultimately independent of the model could have an equal influence on our experience.

5.1 Future works

In general, it seems to us that a map of datasets and corresponding tasks, accompanied by estimates of human performance (expert if possible) and model performance on these tasks, would be very useful for XAI evaluation research. It would help to direct evaluation by performance measures to promising application topics to demonstrate the usefulness of the explanations. Of course, such a

census, represents a considerable amount of work for a moderate epistemic gain. It seems reasonable to put aside possible investigations regarding interface's and local difficulty's influence. We will carry out more precise profiling of our users with respect to the task at hand. This approach must also be accompanied by the development of a training phase in the interface, which gives feedback to the user on these decisions. Otherwise, their predictions cannot fit the problem by themselves: only the response time may decrease according to the user's habituation to the interface. It will be necessary to reduce the number of inopportune changes in the interface that may have affected the experience and to keep only two series (with/without explanation), with an intermediate re-training for the user.
Ideally, our measures should be extended to other tasks and datasets but the need for some form of human expertise, in the face of an ML model, may be limiting. We could also use another explanation format: performance measures' advantage is evaluating wholly different explanations (like counterfactuals, prototypes, or model simplifications) on standard axes.
Finally, our choice for e-sport data as an application domain seems relevant to us because human decisions, although correlated a priori with those of a model, offer a significant margin of variation. For the exploitation of temporal measures, we now have to experiment a baseline without exposure to the ML score, which should allow us to observe an effect. Replicating of the experiment on a larger number of users will allow us to cope with the large variability intrinsic to this measure.

6 Conclusion

This study evaluates SHAP explanations through human performance measures on an e-sport prediction task. This methodological approach is crucial because it allows a firm grounding of XAI evaluations in the human consequences of XAI use, without any assumption about explanation type or mental model. The results of this evaluation indicate that the explanation would have caused our users to lose accuracy. The numerous methodological difficulties of the experiment have been discussed and make us hope for progress in the exploitation of the collected measures, thanks to the development of our methodologies.

Acknowledgements This research has been supported by the group Crédit Mutuel ARKEA. We would like to thank members of the Datalabs service at the Innovation and Operation Pole at Crédit Mutuel ARKEA for their collaboration.

A Translation of the main interface

On the Fig. 1, over the Shap visualisation, user could read:
"Predicted victory: *red* team
Estimated probability: 71%". Under the graph, they could read:

"What is your prediction?"
The likert scale then use the following phrasing:
"the blue team will definitely win, the blue team is likely to win, the blue team has a slight advantage over the red team, I do not know, the red team has a slight advantage over the blue team, the red team is likely to win, the red team will definitely win".

B Translation of the questionnaire

The following question have been asked at the end of the experiment. Questions preceded by an asterisk are open-ended, so people could write whatever they want. Most of questions were asked as affirmative sentence, and the likert scale went from "strongly disagree" to "strongly agree" (centered on "neutral"). Other likert scale went from "absolutely no" to "absolutely yes" (centered on "undecided").

- Do you know anything about the game League of Legends?
- Have you ever played or watched a full game?
- Do you play MOBAs regularly?
- Do you think you can make good predictions about winning after 10 minutes of play?
* (If you play ranked games) what is your rank?

- The analysis interface included all relevant information to help me make a decision.
- The analysis interface allowed me to make a decision more quickly.
- The analysis interface was helpful in making a good prediction.
- The analysis interface was easy to use.
* How was your experience with analysis interface? How did you use it?

- The analysis interface allowed me to understand how the AI worked.
- The AI used is able to make good predictions.
- The model analysis interface explained the model well, in a clear and concise way.
* What would you expect from an AI that tries to explain its decision to you?
- If you were to actually make 10-minute predictions, would you like a model to assist you?

- If you were to actually make 10-minute predictions, would you like to have the mean data?
- If you were to actually make 10-minute predictions, would you like to have the explanations of the model?
- Do you have confidence in the future development of AI?
- Would you be willing to use a similar AI system, with explained results, in another context?

* What are your expectations of using AI in a similar application setting?

– What is your level of education in computer science / engineering sciences?
– Do you have any knowledge of Artificial Intelligence?
* Do you have any other knowledge related to AI or XAI in particular?
– Do you think you are able to estimate the probability of the red team winning?

– Do you feel you made progress during use?
– Were you very focused during the experiment?
– Did the experiment make you tired?

References

1. 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 **58**, 82–115 (Jun 2020). https://doi.org/10.1016/j.inffus.2019.12.012, http://www.sciencedirect.com/science/article/pii/S1566253519308103
2. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
3. Cirqueira, D., Nedbal, D., Helfert, M., Bezbradica, M.: Scenario-based requirements elicitation for user-centric explainable ai. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 321–341. Springer (2020)
4. Craven, M., Shavlik, J.: Extracting tree-structured representations of trained networks. Advances in neural information processing systems **8**, 24–30 (1995)
5. Das, D., Chernova, S.: Leveraging rationales to improve human task performance. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. pp. 510–518 (2020)
6. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
7. Ehsan, U., Passi, S., Liao, Q.V., Chan, L., Lee, I.H., Muller, M., Riedl, M.O.: The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. arXiv preprint arxiv.org/pdf/2107.13509.pdf (Jul 2021), https://arxiv.org/abs/2107.13509v1
8. 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. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 263–274 (2019)
9. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)
10. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)

11. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., Giannotti, F.: A Survey Of Methods For Explaining Black Box Models. arXiv:1802.01933 [cs] (Jun 2018), http://arxiv.org/abs/1802.01933, arXiv: 1802.01933
12. Hase, P., Bansal, M.: Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? arXiv:2005.01831 [cs] (May 2020), http://arxiv.org/abs/2005.01831, arXiv: 2005.01831
13. Herzog, S.M., Hertwig, R.: Harnessing the wisdom of the inner crowd. Trends in cognitive sciences **18**(10), 504–506 (2014), publisher: Elsevier
14. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608 [cs] (Feb 2019), http://arxiv.org/abs/1812.04608, arXiv: 1812.04608
15. Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., Gama, J.: How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 805–815 (2021). https://doi.org/https://doi.org/10.1145/3442188.3445941
16. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: International Conference on Machine Learning. pp. 2668–2677. PMLR (Jul 2018), http://proceedings.mlr.press/v80/kim18d.html, iSSN: 2640-3498
17. Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S.J., Doshi-Velez, F.: Human evaluation of models built for interpretability. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 7, pp. 59–67 (2019), issue: 1
18. Li, J., Chen, X., Hovy, E., Jurafsky, D.: Visualizing and understanding neural models in nlp. arXiv preprint arXiv:1506.01066 (2015)
19. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 31–57 (2018), publisher: ACM New York, NY, USA
20. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: Explainable AI for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610 (2019)
21. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems. pp. 4765–4774 (2017)
22. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019), publisher: Elsevier
23. Mueller, S.T., Weidemann, C.T.: Decision noise: An explanation for observed violations of signal detection theory. Psychonomic bulletin & review **15**(3), 465–494 (2008), publisher: Springer
24. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 629–639. FAT* '20, Association for Computing Machinery, New York, NY, USA (Jan 2020). https://doi.org/10.1145/3351095.3372855, https://doi.org/10.1145/3351095.3372855
25. Ribeiro, M.T., Singh, S., Guestrin, C.: " 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. pp. 1135–1144 (2016)
26. Shwartz-Ziv, R., Armon, A.: Tabular Data: Deep Learning is Not All You Need. arXiv:2106.03253 [cs] (Nov 2021), http://arxiv.org/abs/2106.03253, arXiv: 2106.03253

27. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech. **31**, 841 (2017), publisher: HeinOnline