# Benefits of using multiple post-hoc explanations for Machine Learning

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

# Benefits of using multiple post-hoc explanations for Machine Learning

Corentin Boidot*†, Olivier Augereau*, Pierre De Loor*, Riwal Lefort†
*ENIB / Lab-STICC, Brest, France †Crédit Mutuel Arkéa, Le Relecq-Kerhuon, France
Email: *{boidot,augereau,deloor}@enib.fr, †{corentin.boidot,riwal.lefort}@arkea.com,

***Abstract*—EXplainable AI (XAI) offers a wide range of algorithmic solutions to the problem of AI's opacity, but ensuring of their usefulness remains a challenge. In this study, we propose an multi-explanation XAI system using surrogate rules, LIME and nearest neighbor on a random forest. Through an experiment in an e-sports prediction task, we demonstrate the feasibility and measure the usefulness of working with multiple forms of explanation. Considering users' preferences, we offer new perspectives for XAI design and evaluation, highlighting the concept of data difficulty and of the idea of prior agreement between users and AI.**

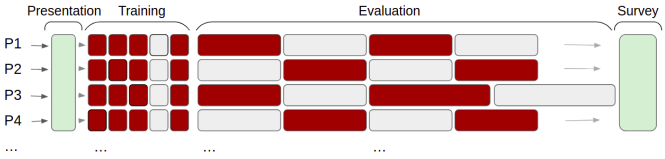


Fig. 1. Experiment's process for different participants. The red blocks represents predictions series with available AI and explanation, while gray blocks represents series under control condition. During the training, the order of games to predict is randomized. During the evaluation, the order of games is fixed, but their number may vary depending on participants' will.

## I. Introduction

### A. Research context

As AI algorithms continue to permeate various aspects of our lives, the need for transparency and comprehensibility in AI decision-making processes has become increasingly imperative. EXplainable AI (XAI) offers multiple algorithms and approaches to compensate for AI's opacity, such as surrogate models, feature importances (FIs), or example-based explanations [1]. With numerous libraries and tutorials, an engineer can now easily make an AI "explainable". However, is this new system useful for its users? A lot of evaluations do not even take users into consideration [2], and most of the time, evaluations are done by comparing only different explanations of the same kind: a FI with other FIs [3], surrogate rules with other rules [4], etc. Nevertheless, application-grounded evaluation can perfectly help to develop benchmarks on diverse forms of explanations [5]. Yet, as the domain is evolving quickly and lacks a unified theoretical framework, the development of benchmarks, evaluations and comparisons of automated explanations seems still hindered.

In this study, explainability is considered in a post-hoc scenario where an AI decision is viewed as a recommendation in a decision task including a human in the loop. Inspired by empirical approaches for XAI evaluations [6]–[8], we designed an experiment to perform simultaneous evaluations of explanations with varied forms at once with few users, from both task-performance and technology acceptance perspectives. We developed a protocol and an interactive environment, meant to evaluate human-AI interaction on a binary classification task, and evaluate it on a winner prediction task on League of Legends (LoL) games data. The effects of four explanations are compared by means of decision time, subjective feedback and use rates. We also take into account the interactions between explanations, and the adaptation to diverse users and data. Our research explores notions of individual preferences and highlights the importance of measuring agreement between experts and AI before any interaction, and the key role of perceived difficulty in explanations' usefulness. In the first part, we will introduce the context of this work, in the second part we will describe the decision system we developed, then we will introduce our experiment's methodology. The fourth part presents our results, and the fifth part will open the discussion about XAI evaluation.

### B. Related works

Among all possible way of evaluating explainability [2], [9], [10], our work is in line with Doshi-Velez and Kim's definition of application-grounded evaluation. We excluded the "proxy-task" approach (ie: human-grouned evaluation) as it uninformative regarding direct performance [11].

In a binary decision problem, task performance is usually measured by accuracy, and the standard research objective is to improve this accuracy through explanations. Some approaches go further and also consider decision time [3], [4], [12]: as a good explanation is asked to be easy to use, the decision should be made faster. Jesus' and Amarasinghe's studies [3], [7] reveal a limit of empirical approach using few experts (three fraud detection experts), as their decision times are very different, on data alone but also when reading different explanations. They used in-subject comparative evaluations of several FIs, and two baselines: decision made without AI, and with confidence score only.

Bansal et al. [6] suggest that it may be too easy to improve on human accuracy by using AI that is significantly better in terms of precision. They show that in the literature, accuracy of the human-AI team has never really progressed by the means of explanations. They carry between-subject studies

on three text-based tasks (two binary sentiment classifications and a question answering task), obtaining cooperative results on accuracy metrics. However, the confidence score alone produced cooperative performance comparable to their best explanations. Baudel et al. [8] also achieved such cooperative performance on a prediction task based on the Titanic dataset, by using only the AI's recommendation. However, when using an AI with higher accuracy, the cooperative performance disappeared, suggesting that there is always a performance threshold where human-in-the-loop decision making becomes detrimental in terms of efficiency. In a context where AI is more powerful than humans, explanation can follow a teaching purpose. For our experiment however, we adopt to a cooperative point of view.

## II. Evaluated system

We designed a task environment for human-AI collaboration that would be suited to compare different explanations methods. To this end, we chose an e-sport prediction as an application domain, with a tabular dataset from League of Legends, on which we trained a Random Forest model. We developed a decision interface enabling users to make predictions on games, with AI assistance (and without it), and implemented four explanations (including the confidence score)[1].

### A. Task, Data and Model

Tasks from medicine or finance are crucial in XAI, but building a pool of participants with the relevant knowledge is indeed a limiting factor. We looked for a task with analogies to fraud detection, so that comparison could be done with these works; we chose e-sport prediction because students can be relevant "experts", thus enabling a better reproducibility. Both are binary classification tasks on tabular data, using statistical data which require specific domain knowledge to be clearly understood. We chose League of Legends (LoL) winner prediction, a popular online game. This game consists of two teams (red and blue) of five players fighting for the control of the opposing camp. The players' characters may gain experience, gold, kill the enemies, destroy enemy facilities, etc., and most players observe these statistics at the end of a game. Using such statistics at 10 minutes of play, we have a binary prediction task on tabular data, with a level of difficulty adapted to observe both human errors and AI errors.

We chose a LoL dataset containing high-level statistics, collected from real high-ranked games, through the game's API. As the dataset contains redundant columns, we kept only 23 pertinent variables. Some of them, like the "gold" accumulated by each team, are known to be good predictors for the winner, but some victories would still surprise most players.
We split our dataset with a 75-25% train-test ratio. Inside this test set, we selected a sample of 80 games to be used for the human experiment. Among those, 30 would be dedicated to human training on the task, and 50 would be used for human evaluation, but because of experimental timing constraints, only 25 were used for the training, and the evaluation could use a variable number of games (see section III).

We trained a Random Forest (RF), using the scikit-learn implementation with 200 trees on normalized data. Our model reaches an accuracy of 72% on the test set. We would not seek a better accuracy as it is nearly about human accuracy on this task (71%).

### B. Decision interface

All the interface was developed in python using mostly streamlit library. A description of the interface is given in Figure 2. The right part of the screen is dedicated to AI (prediction and explanation) and is blank when not accessing the AI. The left part notably contains the button that will "activate" AI. We displayed the data in the middle, in two columns corresponding to the blue and red team, with the same statistics on each line. A third column contains our global means for these statistics, to give reference values, thus helping users to deal with task uncertainty [13]. The slider allows only to select blue or red victory. Time is recorded when the user clicks the validation button (left part of the screen).

When the AI is activated, a new button enables to select any of the four explanation modes, starting on a the default explanation display: confidence score.

Our four explanations were always referenced as 'A' for the confidence score, 'B' for the surrogate rule (skope-rules), 'C' for LIME and 'D' for the nearest neighbors. These explanations have different forms that may complement one another in order to justify AI's prediction, or to calibrate trust. They also have an increasing intended complexity (which matches their display size).

*a) Confidence score:* In this explanation we display the confidence score oriented towards the predicted class, so that the score is always higher than 50%. We did not try to calibrate those scores, because their distribution on our experimental data was already satisfying: few errors are done with extreme confidence, they rather accumulate near the 50% confidence score.

*b) Skope-rules:* We wanted a surrogate model that we could display textually, leading us to choose decision rules. We used skope-rules, an accessible solution that implements this kind of model [14]. We trained two different surrogates in order to justify decisions: one surrogate trained on blue predictions of the RF model, the other on red predictions. This double surrogate solution was motivated by the fact that all learned rules or formulated towards a "positive class" to be detected, which does not fit our task where no team can be said to win a priori. Each of our trained models contained four rules and used only two features, as this configuration gave the best fidelity. Only the first activated rule is displayed to the user, with a conversion to natural language: ie. French. An example

[1] code is available on github.com/CBoidot/benefits_multi_XAI

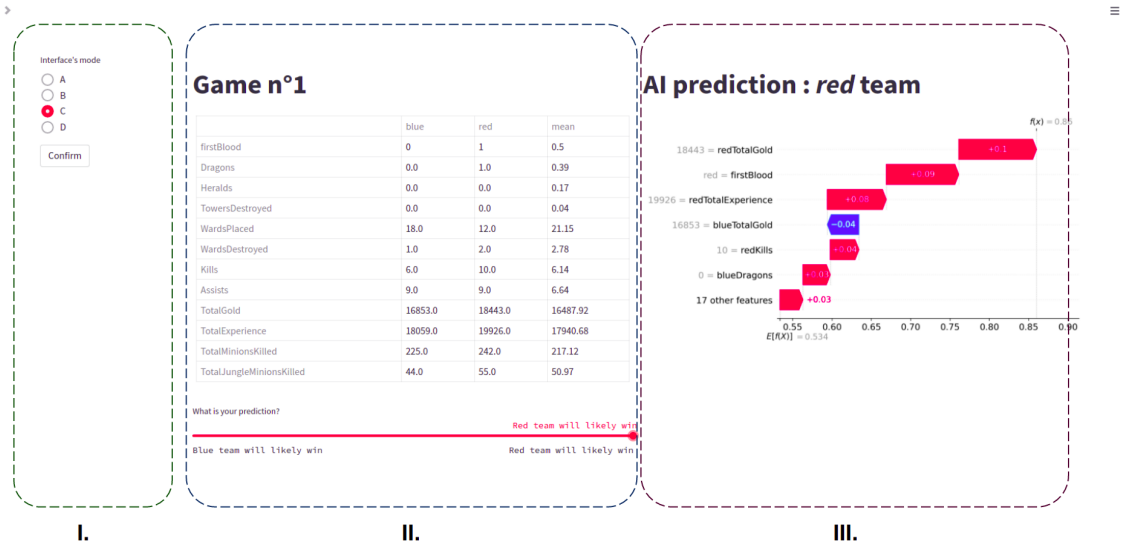


Fig. 2. Example of the interface after accessing AI, with explanation C (LIME). Part I. contains the navigation tools including a validation button, AI recommendation button and explanations' radio button once the previous one is pressed. Part II. contains the current game data, with a common column of reference values, and the slider which enables users to make predictions. When AI is accessed, Part III. display the AI's prediction (here: red team) with the selected explanation underneath. User are free to navigate between the four displays, and can confirm when they want.

of such a rule would be: "As blueTotalGold is lower than 17551 and redTotalExperience is higher than 17394, red team should probably win.". If the surrogate is incorrect regarding AI's prediction, we display: "No rule for this case".

*c) LIME:* Since we used SHAP [15] in an early experiment on the same task (tree-based interventional implementation [16]), and it gave no performance gain, we chose the LIME algorithm [17]. This method has already obtained good task-based evaluation on other tasks [5], [7]. We used the implementation from the original authors without additional parameters, hiding discretizations' thresholds.
We display these FI using a "waterfall plot", that we believe to be more intuitive, using the library shap (an example is visible in Fig. 2). This figure displays the six main FIs plus all the others summed together. Because of their approximations, those explanations could from the the probability displayed in explanation A; for instance if the probability is 55 percent towards the red team, LIME could indicate 45% towards the red team. However, these approximations may in fact help doubt the model when it is useful.

*d) Nearest Neighbor+:* We now require an example-based explanation, coming after the confidence score, surrogate rules, and FIs. We preferred the nearest neighbors (NN) approach rather than prototypes or counterfactuals (CF) because they are simple to describe, compared to the complexity of communicating what a CF really is: data that does not exist [1]. We sampled a thousand examples from our test set in order to build this explanation. We draw the NN from this set, using the L1 distance with normalized features. We display the nearest neighbor in the same manner that we display the match to be analyzed. Because the NN alone would not be helpful enough, we decided to add contextual information in a text under the NN: which team won that match, the associated AI prediction, and how far is a NN with opposite label, measured by the number of neighbors needed to find it.

## III. Methodology

This experiment is designed for a small number of participants, with limited availability (1 hour).With such constraints, each participant has to go through both a control condition and an experimental condition. We interleaved these conditions, so that all the data is processed by a maximum of participants from both conditions. This interleaving should not affect participants overall performance as they all go through a long training phase before (about 30 minutes). Thus our experimental process goes through four stages: presentation of the goals and setting, training on the prediction task, evaluation, and a final survey (see Fig. 1). The explanatory interface is evaluated as a whole, but conclusions can be drawn regarding the explanations used in the interface, thanks to the behavioral and subjective data collected. We assume that by granting the user freedom, we will reach better team performance and system acceptance.

*a) Presentation of goals:* The participant first receives instructions, with information about context, data, and the different explanations, qualified as "displays" for the AI recommendation, that needs justification. The task objective given to them was to take the better decision in the shortest time, so they can define their own balance between these two goals. It was insisted the users were free to use or not the AI, and encouraged to adapt their use to the game.

*b) Training stage:* A training phase is necessary to get stable results during the evaluation, so we designed the following scheme. After each decision, the users are reminded

of their total number of errors, and their decision time is also displayed after the tenth game. Every five games, the users could look back to their decisions, see the ground truth and every displays. We expect that with this feedback, participants could adapt towards both good decision times and accuracy. To prevent that the order of training data creates some bias in participants' perception of the system, we randomized the order of the game for each user. We also disabled the AI recommendations for games 16 to 20, so that users also experienced the control condition, and made decisions with data only. The experimenter had to keep track of time. We chose to keep a fixed number of games to analyze for training (25), and to have a variable number of evaluation games. After 20 evaluation games, the experimenter could change the number of remaining games, depending on time and user's convenience.

*c) Evaluation stage:* Only the predictions done during this evaluation stage are to be analyzed in our results part. As shown in Fig 1, the evaluation goes through two sequences of ten games, and to sequence of variable length, alternating between control and experience conditions. The control condition simply consisted in removing the button that gives access to the AI. The protocol collects behavioral data on decision making, but also multiple subjective feedbacks from the user. After each decision, the participants were asked if they were confident in their decision and if the game was difficult to judge (7-point likert-scales). When the participants had access to the AI interface, they were also asked if they found each of the displays useful (binary answers).

*d) Questionnaire:* We intended to control both perceived use (PU) and perceived ease of use (PEOU) [18] of our system. As the PU of our explanations are already evaluated thoroughly in evaluation phase, we mostly investigate the PEOU by adapting CSUQ questionnaire [19], on both the decision system (considering control condition) and the AI interface. All these questions used 7-point Likert scales, and we take the mean of results (using a scale from -3 to 3). We also asked how interpretable were the explanations.

We recorded information about students' backgrounds at the end of the experiment, to check their degree of expertise. We also recorded information about participants' sociological profile, and especially about their knowledge in AI, as it could be an important factor in AI's perception [20].

### A. Who are our participants?

The experiment took place inside our laboratory, and inside an office associated with our work. We recruited 27 participants, mostly through internal channels (collective mails or chat). We recruited among students and young engineers, some of them had a background in data science: most of them had at least a license degree in informatics. The main prerequisites for participation were knowledge of League of Legends (a minimum of having played a few games). Regarding the expertise on LoL, 12 participants out of 27 had reported that they do not play ranked games, and 9 participants reported they did not care about the meta-game (ie: the strategic discussions surrounding the game and its updates). This "low implication" with the game did not resolve into lower performances, as the level of analysis required here is rather abstract, so we did not add any filters a posteriori to reduce this diversity of profiles. The age of the participants ranges from 19 to 37 y.o.; 96% of participants were males. All of them were able to correctly use the AI interface: we measured a mean PEOU of 1.83 (with Likert scales going from -3 to 3) for the control condition, and 1.0 for the PEOU of the AI itself (worst cases: respectively 0.14 and -0.6). The majority of our participants used the AI assistance during the evaluation: 2 of them did not use it at all, and 2 others would use it only once and twice.

## IV. Results

### A. Explanations' comparison

In order to compare an explanation's effect in an open system, we must observe its regular users. For each of our four explanations, we consider a participant that used it more than once as a "user of the explanation". For each explanation, the use rate could vary a lot: only explanation A (confidence score) was guaranteed to be seen by any participant who wants to access the AI (thus leading to the highest use rate). Comparative results are reported in Table I.

TABLE I
AVERAGE USE AND PERCEPTION OF THE EXPLANATORY DISPLAYS. A: MODEL'S SCORE, B: SURROGATE RULE, C: LIME, D: NN+

| | users | use rate | usefulness | time per use | interp. |
|---|---|---|---|---|---|
| A | 24 | 0.76 | 0.66 | 3.23 | 2.99 |
| B | 10 | 0.19 | 0.44 | 4.56 | -0.97 |
| C | 21 | 0.49 | 0.88 | 6.50 | 1.38 |
| D | 19 | 0.42 | 0.75 | 7.07 | 0.27 |

Usefulness is the mean percentage of positive PU answers to post-decision forms; along with use rate and time per use, it is computed only on the users of the explanation, while interpretability (asked in the exit survey) is a mean from all participants' answers. We see that LIME is perceived as more useful than the confidence score - and NN+ find similar success. It is to be noted that only the confidence is universally perceived as interpretable. The less efficient explanations tested are our surrogate rules (B): only a few participants occasionally use them, and use, with a low perceived utility.

The recorded times per use meet our expectations for the explanations different designs: confidence score (A) is quick and easy to read, followed by skope-rules (B), LIME (C) and finally NN+ (D) as it adds a second table to read. The overall agreement between PU, interpretability and the use rates of B, C and D matches what is expected of rational users.

### B. Behaviors with multiple explanations

For each game, we classify a user's decision into 9 possible strategies: no explanation (referred as 'No'), confidence score alone ('A') and all possible combinations of A and the other displays. Participants had diverse behaviors, and only strategies using B can be seen as marginal (see Fig. 3).

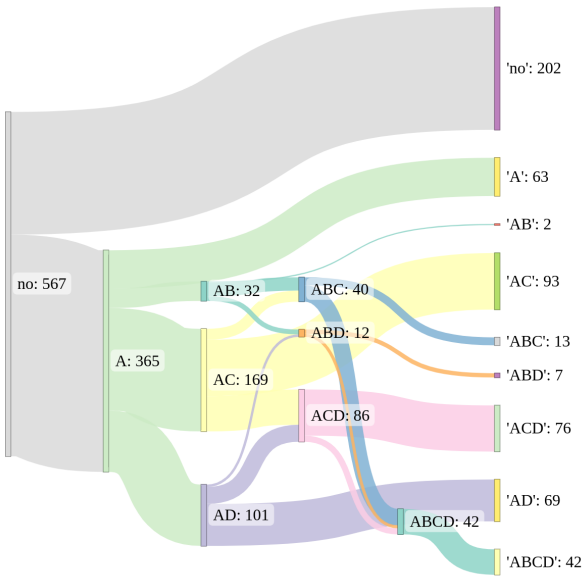


Fig. 3. We recorded the decision path of all predictions under experimental condition. The numbers refers to the numbers of decisions relying on the interface state, all decisions starting with the "no AI" state. The right end of the chart figures the proportions of the nine strategies using explanations or not.

These behaviors can be analyse in the light of participants' characteristics or data characteristics. Links with data will be analyzed in Section IV-C. On average, more than 85% of a participant's usage is accounted for by the "No" strategy (default option) and his two favorite combinations of explanations. This confirms that each participant had preferences rather than a random behavior. Specifically, three users chose a pure strategy (*ie:* using the same strategy on each game), two 'No' and one AC, but the others changed their behavior on different games.

We would not find any link between participants' profiles (expertise with LoL, AI, etc.) and their preference. Preferences could be partially explained by different levels of engagement with the interface rather than rational optimisation, as the more users tend to use AI, the more they tend to also use multi-explanation strategies. Only their natural decision-making speed can explain their engagement, as some participants said they didn't use the explanations because they were too time-consuming.

We grouped our participants by preferred explanation, (the four participants having more than 80% of "No" strategy being another group), leading to groups of 3 to 5 participants. The results of the different groups may be seen in Table II. Because we do not control when each participant will effectively access the AI or not, we do not have one "mean accuracy" of the AI they observe, but it keeps between 70% and 75%. The main effects of the experiment interface on interacting groups is the rise in decision time, and a higher variance in accuracy. Surprisingly, user accuracy remained stable between the control and experimental conditions. This can be explained by a hidden problem of our experimental setting: natural compliance between AI's and humans' decisions. The participants followed the AI 91% of time during the experimental condition, whether using the AI or not, and 93% of time during the control condition. Under such conditions, accuracy gains cannot be a good measure of explanations goodness. To our knowledge, this parameter is rarely controlled in similar user studies. We find no correlation between accuracy and compliance: groups 'ACD' and 'ABCD' both had slightly worse accuracy in experimental conditions while their compliance respectively rise and fall under this condition.

TABLE II
PERFORMANCES BY PREFERENCES: COMPLIANCE (CPL), ACCURACIES (ACC) AND DECISION TIMES (DT) FROM CONTROL CONDITION (CTRL) AND EXPERIMENTAL CONDITION (XP).

| | Acc (ctrl) | Acc (xp) | Cpl (ctrl) | Cpl (xp) | DT (ctrl) | DT (xp) |
|---|---|---|---|---|---|---|
| preference | | | | | | |
| A (n=4) | 0.71 | 0.69 | 0.94 | 0.92 | 13.67 | 17.05 |
| ABCD (n=3) | 0.72 | 0.70 | 0.87 | 0.95 | 17.52 | 23.79 |
| AC (n=4) | 0.73 | 0.72 | 0.98 | 0.89 | 14.64 | 19.53 |
| ACD (n=5) | 0.70 | 0.69 | 0.95 | 0.86 | 11.93 | 19.88 |
| AD (n=5) | 0.70 | 0.76 | 0.93 | 0.91 | 11.27 | 16.33 |
| No (n=4) | 0.72 | 0.72 | 0.88 | 0.96 | 10.24 | 9.58 |

### C. Reactions towards difficulty of prediction

Fig. 4 shows that the more participants find a game difficult to predict, the more likely they would use our explanatory displays. Such trends can also be found with perceived usefulness, and with time spent on each explanation. This trend also matches the following result: on difficult games participants had a tendency to look at more explanations. This results also reminds us that an AI user does not always need an explanation. Perceived difficulty is likely to be affected by the interaction with AI: we used the model score to have a reference for games difficulty, and we found higher correlations with the participants' mean perceptions in experimental condition (respectively Pearson's R=79% and 85% in control and experimental conditions). As we asked our participants their confidence in their own decision, we found a strong negative correlation between perceived difficulty and confidence (Pearson's R=-80%). However, these variables may have causal connections in both directions, so we retained perceived difficulty as the main variable to consider. We are therefore unable to see if explanations really helped participants being more confident in their decisions: this would require a measure of decision change, with intermediate confidence feedback.

## V. DISCUSSION

Even if we still lack capabilities to anticipate people's preferences, this problem can be mitigated easily by letting users change the default explanation. This experiment and analyzes present numerous limits. First of all, refinements are possible on our chosen methods, and especially our surrogate rules: coverage and precision could be added, for instance. However, the main problem remains the strong correlation of results between human and AI: we may find another task to

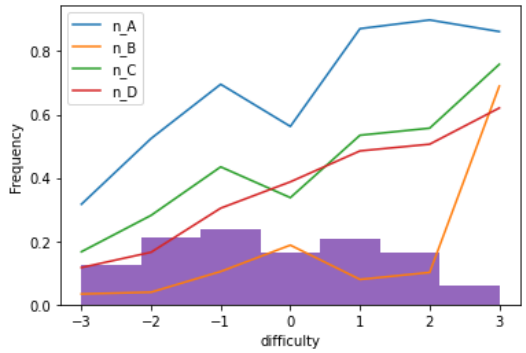


Fig. 4. Use rates of the different explanations, depending on perceived difficulty. The background histogram represents the distribution of perceived difficulty.

evaluate, or conversely use the feedback from this experiment to sort data, and test participants only on "difficult" data.

The existence of individual preferences among explanations open the door to numerous questions about evaluation protocols: do these preferences match their formats? Then we should use several explanations of each form, and subgroups of participants might indicate which is the best among each form. Do these preference match some hidden cognitive strategies, like confirmation or doubt [6], independently of their form? Some participants indeed reported that they use the AI to confirm the a priori they made with data only. At last, we chose not to analyze learning stage results: explanations may find different effects in a didactic perspective. New interactions could easily be developed then, based on users confidence feedback, on their errors, etc.

## VI. Conclusion

In this study, we developed a multi-explanation interface and demonstrated its usefulness through empirical evaluation on an e-sport prediction task. Even if there is no benefit to use multiple explanations on each case, as they bring additive time costs and no accuracy gain, the different formats could be complementary regarding users, uncovering a need for diversity of explanations. Future research should focus on dynamically adapting such interfaces to user preferences, particularly by estimating the current case difficulty.

## Acknowledgment

This research has been supported by the group Crédit Mutuel Arkéa. We would like to thank members of the Data Office of Crédit Mutuel Arkéa for their collaboration.

## References

[1] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh, "EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence," *arXiv preprint arXiv:2102.02437*, 2021.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[3] S. Jesus, C. Belém, V. Balayan, J. Bento, P. Saleiro, P. Bizarro, and J. Gama, "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*, 2021, pp. 805–815.

[4] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez, "Human evaluation of models built for interpretability," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 59–67, issue: 1.

[5] P. Hase and M. Bansal, "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?" *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*, May 2020, arXiv: 2005.01831. [Online]. Available: http://arxiv.org/abs/2005.01831

[6] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[7] K. Amarasinghe, K. T. Rodolfa, S. Jesus, V. Chen, V. Balayan, P. Saleiro, P. Bizarro, A. Talwalkar, and R. Ghani, "On the Importance of Application-Grounded Experimental Design for Evaluating Explainable ML Methods," *arXiv preprint arXiv:2206.13503*, 2022.

[8] T. Baudel, M. Verbockhaven, V. Cousergue, G. Roy, and R. Laarach, "ObjectivAIze: Measuring Performance and Biases in Augmented Business Decision Systems," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*. Springer, 2021, pp. 300–320.

[9] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, pp. 1–45, 2021, publisher: ACM New York, NY.

[10] R. R. Hoffman, S. T. Mueller, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," *arXiv:1812.04608 [cs]*, Feb. 2019, arXiv: 1812.04608. [Online]. Available: http://arxiv.org/abs/1812.04608

[11] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems," in *Proceedings of the 25th international conference on intelligent user interfaces*, 2020, pp. 454–464.

[12] F. Gedikli, D. Jannach, and M. Ge, "How should I explain? A comparison of different explanation types for recommender systems," *International Journal of Human-Computer Studies*, vol. 72, no. 4, pp. 367–382, 2014, publisher: Elsevier.

[13] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[14] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The annals of applied statistics*, pp. 916–954, 2008, publisher: JSTOR.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[16] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee, "True to the Model or True to the Data?" *arXiv:2006.16234 [cs, stat]*, Jun. 2020, arXiv: 2006.16234. [Online]. Available: http://arxiv.org/abs/2006.16234

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[18] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989, publisher: JSTOR.

[19] J. R. Lewis, "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995, publisher: Taylor & Francis.

[20] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I.-H. Lee, M. Muller, and M. O. Riedl, "The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations," *https://arxiv.org/pdf/2107.13509.pdf*, Jul. 2021. [Online]. Available: https://arxiv.org/abs/2107.13509v1