Adaptation of AI Explanations to Users' Roles

Julien Delaunay, Christine Largouët, Luis Galárraga, Niels Van Berkel

Adaptation of AI Explanations to Users' Roles

Julien Delaunay
Inria/IRISA
Rennes, France
julien.delaunay@inria.fr

Christine Largouet
Institut Agro/IRISA
Rennes, France
christine.largouet@irisa.fr

Luis Galarraga
Inria/IRISA
Rennes, France
luis.galarraga@inria.fr

Niels van Berkel
Aalborg University
Aalborg, Danemark
nielsvanberkel@cs.aau.dk

CHI '23, April 23–28, 2023, Hamburg, Germany
ACM .

Abstract
Surrogate explanations approximate a complex model by training a simpler model over an interpretable space. Among these simpler models, we identify three kinds of surrogate methods: (a) feature-attribution, (b) example-based, and (c) rule-based explanations. Each surrogate approximates the complex model differently, and we hypothesise that this can impact how users interpret the explanation. Despite the numerous calls for introducing explanations for all, no prior work has compared the impact of these surrogates on specific user roles (e.g., domain expert, developer). In this article, we outline a study design to assess the impact of these three surrogate techniques across different user roles.

Author Keywords
Explainability; Interpretability; User Study

CCS Concepts
•Human-centered computing $\rightarrow$ **Human computer interaction (HCI);** *User studies;*

Introduction
Machine Learning (ML) models are increasingly used, spanning from recommendation systems for entertainment applications to decision support for critical tasks such as law [4, 26] and medicine [11, 17]. These algorithms'

Instance x	$fo = 0, a = 18,$ $mc = 1, bm = \text{`}Low', hc = \text{`}No', (o = 30)$
	Explanations
Feature attribution	$(fo = 0) \rightarrow -6,$ $(bm \geq \text{`}Low') \rightarrow -5$
Rule	If $a \leq 20 \wedge mc = 1$ $\Rightarrow$ non-obese
Example	$bm = \text{`}Sometimes', hc = \text{`}Yes', (o = 70)$

Table 1: Explanations for a classifier C computing the risk of obesity $o \in [0, 100]$ with the outcome of 'non-obese' if $o \leq 50$. The attributes consist of the patient's family's obesity antecedents (fo), age (a), monitoring calorie consumption (mc), consumption of food between meals (bm), and high-caloric food (hc).

efficiency has increased at the cost of opaqueness and bias [1, 21, 29]. An increasing focus is placed on transparency and explanations to uncover and mitigate the biases and errors introduced by ML algorithms. Among these explanation methods, surrogate-based-model-explanations (for now on *surrogate explanations*) are the most frequently used [16]. The surrogate methods train a proxy to imitate the classifier's outcomes. This proxy is selected for its simple design, highly transparent, and ease of understanding. In their survey, Bodria et al. [6] grouped the surrogate explanations into three categories: (a) feature-attribution, (b) rules, and (c) example-based explanations. Each of these has a different aim, presented first in this paper, ultimately impacting how the explanation is generated and presented to the user.

While many researchers have pointed out the need for user studies to evaluate novel XAI methods [3, 10], relatively few studies have been conducted. Adadi et al. [2] highlighted that in 2019, from a total of 381 XAI papers, only 5% emphasised users in evaluating XAI methods. Furthermore, although various user roles are involved in the application of ML models (e.g., developers, end-users), evaluations are primarily focused on developers as explanation methods are currently mostly used by developers [5].

Researchers have proposed to create explanations adapted to users' roles, suggesting a total of three different roles: (a) developers, (b) domain experts, and (c) lay users [14, 25]. However, users are more complex entities, and additional criteria may impact their experience with AI systems (e.g. level of trust in AI). We thus propose to conserve the original three roles as one of the multiple dimensions of user roles and present additional criteria. Finally, we introduce a methodology to conduct user studies comparing the impact of the surrogates depending on the context, task, and user role. These studies aim to help select the explanation methods adapted to user roles.

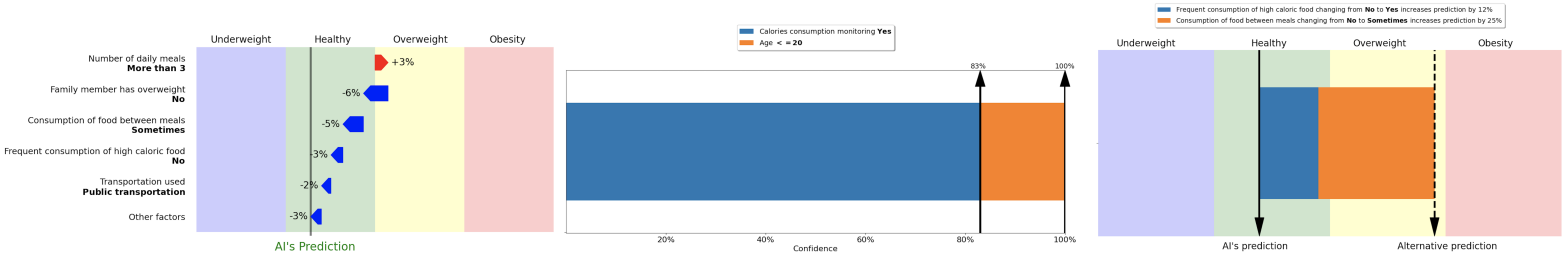
Surrogate Explanations

Each of the three surrogate explanations methods differs in the way they approximate a black-box classifier [6]. Therefore, before elaborating on adapting the surrogate to the user, we first clarify how these methods work and differ. We represent mathematical and graphical explanations for these three surrogates in Table 1 and Figure 1. Each of these explanations shows the main reason for the prediction made by a random forest classifier.

Feature attribution methods associate a weight to the input features to indicate a positive or negative impact on the final prediction. Therefore, Figure 1a shows the explanations in a similar way to what is shown in LIME [23] and SHAP [18], the methods the most commonly used to generate an explanation [16]. Red and blue horizontal bars indicate respectively positive and negative impact. The final score and the vertical bar correspond to the final prediction. Explanations from Figure 1a and Table 1 indicate that the user is less prone to develop obesity due to the absence of obesity antecedents in their family and low consumption of food between meals.

Rule-based surrogates provide the minimum requirements for a given outcome [12, 24]. These requirements take the form of '*if-then*' rules that represent the conditions for a classifier to make a given prediction. These methods are commonly represented as in Table 1, however, Figure 1b depicts similarly to [20], the increasing classifier's confidence in predicting non-obese as the conditions of the rule (i.e., age and monitoring calorie consumption) are met.

Example-based explanations present instances similar to the target with a comparable (prototype [13]) or different



(a) Feature-attribution explanation. **(b)** Rule-based explanation. **(c)** Counterfactual explanation.

Figure 1: Graphical representation of three different explanation proxies for a given instance predicted as non-obese by a classifier.

How to measure the impact of explanations on users?

Intentional measurements

Trust. Cahour and Forzy's scale [7].

Satisfaction. Hoffman et al.'s questionnaire [15].

Understanding. Madsen and Gregor's questionnaire [19].

Behavioural measurements

Trust. The users have the option to modify their prediction after seeing the AI's prediction.

Satisfaction. The time mandatory to solve the simple task or predict.

Understanding. The users have to indicate which factors impact the most toward the prediction

Table 2: Metrics to measure user intent and behaviour.

(counterfactual [9]) classifier's reaction. To the best of our knowledge, no graphic illustration exists for counterfactuals, leading us to develop our own interpretation as shown in Figure 1c. Thus, Figure 1c shows the change in the AI's outcome when modifying a feature value. The counterfactual from Table 1 and Figure 1c shows that increasing both the consumption of high-caloric food and intake of food between meals would have changed the prediction.

Due to a lack of surrogate explanations comparison, XAI users are presently unable to indicate why they might use one type of proxy rather than another. However, the choice of the surrogate and its representation may impact the users (e.g., trust, understanding) [27, 28]. We hence argue that preferring one type of proxy over another should be driven by criteria and situations rather than for functional reasons. We next present different user roles to guide researchers and actors in investigating the impact of selecting a surrogate and representation depending on user roles.

User Roles

Most existing research has focused on three types of roles [14, 25]: (a) developers that create or assess AI systems; (b) domain experts, persons with knowledge or authority in a particular area; and (c) lay users, individuals to whom the AI decision is applied (e.g., bank client). Yet, we argue that users and usage scenarios are more complex than those three well-defined categories. Instead, users of AI systems are multi-dimensional (e.g., roles, goals, trust in AI), and various scenarios affect the suitability of different explanation methods (e.g., data types, explanation representation). We thus propose four additional aspects to consider when selecting explanations adapted to users:

- The motivation to compute the explanation (e.g., increasing performance or trust in the system) is a key criterion for determining the appropriate model.
- The trust in AI systems may differ among the users as not all programmers have blind faith in the systems they code while lay people may place excessive trust in it.
- The challenges of representing data types such as sound or time series is one of the reasons why few explanation methods exist for these data types [6]. As such, the data type influences the choice of the surrogate.

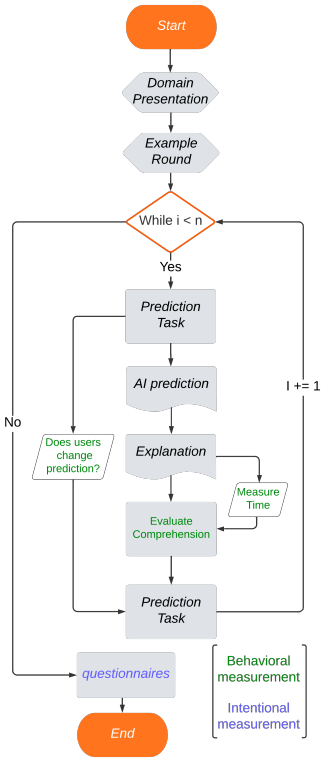


Figure 2: Plan for user study evaluating the impact of a given explanation on users. The tasks are repeated n times, where n is the number of instances predicted by the user. Elements in green are behavioural measurements while in blue are self-reported.

- Selecting one explanation representation over another (e.g., Figure 1 rather than Table 1) is crucial since it has been widely accepted that representations impact how users perceived AI systems [8, 22].

Evaluating how each dimension of the user roles and usage scenarios impacts the perception of surrogate explanation would allow associating surrogate methods adapted to users. We thus elaborate on our evaluation proposal to compare the impact of the three explanations surrogates on various users' roles.

Methodology

Based on various recommendations [15, 28], we outline in this section (i) diverse metrics to measure both user behavioural and perceived impact of an explanation surrogate, and (ii) a roadmap for conducting generic and replicable user studies for a given surrogate and representation.

Perception and Behavioural. Van der Waa et al. highlighted the importance of employing mixed metrics to conduct XAI user studies [28]. We also emphasise differentiating between perceived and behavioural measurements, as the user's perception may differ from their actual behaviour or decision-making. Therefore, we propose combining questionnaires measuring self-reported perception and simple tasks to gauge performance. Table 2 summarises possible metrics and questionnaires to measure both the perceived and behavioural users' (a) understanding, (b) trust, and (c) satisfaction. From these multi-axes measurements, users can envision using one explanation method more than another.

Roadmap. Figure 2 illustrates an experimental protocol to conduct user studies evaluating the impact of a chosen explanation method and representation for one user role.

In the initial steps, we advise introducing the domain and the objective of the experiments. Then, to reduce the possibility of biases led by a short or long training round [28], an example round defines the user's task and the details of the explanation. Following, participants complete the actual study tasks. This can be repeated multiple times to obtain more reliable results. Participants are asked to predict using the same information as the system. Afterwards, participants have access to the AI prediction and its associated explanation. This approach allows for assessing the behavioural understanding, trust, and satisfaction as defined in Table 2. Finally, in the final round, we measure the perceived impact of the explanation through several questionnaires as described in Table 2.

By running this experiment for distinct (a) user profiles, (b) explanation surrogates, and (c) representations, researchers and actors may gain insight into which explanation and representation are the more suitable for a specific user based on various criteria.

Future Work

In this paper, we proposed considering users as more complex than the three original roles but as a multi-axes complexity scale. Comparing the impact of the three surrogate categories over the different aspects of users would benefit the ML sub-community of XAI by allowing them to manage and carefully select the appropriate proxy. Conversely, the HCI sub-community would profit from the roadmap we introduced due to the possibility of conducting generic and replicable user studies. Currently, we launched our experiments on tabular data, with 250 crowdworkers, three surrogates and two representations. Finally, we seek to conduct our investigation with computer scientists from different research laboratories, specialists either in HCI or ML, and domain experts in a relevant domain (e.g., healthcare).

REFERENCES

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proc. AIES*. ACM. https://doi.org/10.1145/3461702.3462624

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. DOI: http://dx.doi.org/10.1109/ACCESS.2018.2870052

[3] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proc. AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=3331806

[4] Michał Araszkiewicz, Trevor Bench-Capon, Enrico Francesconi, Marc Lauritsen, and Antonino Rotolo. 2022. Thirty years of Artificial Intelligence and Law: overviews. *Artificial Intelligence and Law* (06 Aug 2022). https://doi.org/10.1007/s10506-022-09324-9

[5] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proc. on Fairness, Accountability, and Transparency (FAT)*. ACM. DOI: http://dx.doi.org/10.1145/3351095.3375624

[6] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and Survey of Explanation Methods for Black Box Models. *CoRR* (2021). https://arxiv.org/abs/2102.13076

[7] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science* 47, 9 (2009), 1260–1270. https://www.sciencedirect.com/science/article/pii/S0925753509000587

[8] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proc. CHI*. ACM. https://doi.org/10.1145/3290605.3300789

[9] Julien Delaunay, Luis Galárraga, and Christine Largouët. 2022. When Should We Use Linear Explanations?. In *Proc. CIKM*. ACM. https://doi.org/10.1145/3511808.3557489

[10] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* (2017). https://arxiv.org/abs/1702.08608

[11] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *CoRR* (2018). http://arxiv.org/abs/1805.10820

[13] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. 2022. VCNet: A self-explaining model for realistic counterfactual generation. *CoRR* abs/2212.10847 (2022). https://doi.org/10.48550/arXiv.2212.10847

[14] Maryam Hashemi. 2023. Who wants what and how: a Mapping Function for Explainable Artificial Intelligence. *CoRR* abs/2302.03180 (2023). DOI: http://dx.doi.org/10.48550/arXiv.2302.03180

[15] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* abs/1812.04608 (2018). http://arxiv.org/abs/1812.04608

[16] Alon Jacovi. 2023. Trends in Explainable AI (XAI) Literature. *CoRR* abs/2301.05433 (2023). DOI: http://dx.doi.org/10.48550/arXiv.2301.05433

[17] P. Karatza, K. Dalakleidi, M. Athanasiou, and K.S. Nikita. 2021. Interpretability methods of machine learning algorithms with applications in breast cancer diagnosis. In *Proc. Engineering in Medicine & Biology Society (EMBC)*. DOI:http: //dx.doi.org/10.1109/EMBC46164.2021.9630556

[18] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc. NeuIPS*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[19] Maria Madsen and Shirley D Gregor. 2000. Measuring Human-Computer Trust.

[20] Christoph Molnar. 2018. Interprtable machine learning: A guide for making black box models explainable. (2018).

[21] Cathy O'Neil. 2017. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Coll. Res. Libr.* 78, 3 (2017), 403–404. https://doi.org/10.5860/crl.78.3.403

[22] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proc. CHI*. ACM. https://doi.org/10.1145/3411764.3445315

[23] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. SIGKDD*. ACM. https://doi.org/10.1145/2939672.2939778

[24] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proc. AAAI*. AAAI Press. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982

[25] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Proc. IUI Workshops (CEUR Workshop Proceedings)*. CEUR-WS.org. http: //ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

[26] Andrea Tagarelli and Andrea Simeri. 2022. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artificial Intelligence and Law* 30, 3 (01 Sep 2022), 417–473. https://doi.org/10.1007/s10506-021-09301-8

[27] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proc. CHI*. ACM. DOI:http://dx.doi.org/10.1145/3411764.3445365

[28] Jasper van der Waa, Elisabeth Nieuwburg, Anita H. M. Cremers, and Mark A. Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artif. Intell.* 291 (2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

[29] Tomás Zemcík. 2021. Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI Soc.* 36, 1 (2021), 361–367. https://doi.org/10.1007/s00146-020-01053-4