Bridging the Gap between Human-Computer Interaction and Machine-Learning on Explainable AI: Initial Observations and Lessons Learned

Julien Albert, Adrien Bibal, Benoît Frénay, Bruno Dumas

Bridging the Gap between Human-Computer Interaction and Machine-Learning on Explainable AI: Initial Observations and Lessons Learned

Combler la distance entre l'interaction humain-machine et le machine learning sur l'IA explicable : premières observations et leçons apprises

JULIEN ALBERT, NaDI/PReCISE, Faculty of Computer Science, University of Namur, Belgium
ADRIEN BIBAL, University of Colorado Anschutz Medical Campus, USA
BENOÎT FRENAY, NaDI/PReCISE, Faculty of Computer Science, University of Namur, Belgium
BRUNO DUMAS, NaDI/PReCISE, Faculty of Computer Science, University of Namur, Belgium

Explainability in artificial intelligence (XAI) is a rapidly growing research field nowadays, in particular in machine learning (ML). XAI concerns both the technical capacity to understand the functioning of ML models and the adequacy of the explanations with the targeted users and the contexts of use. It thus joins both the concerns of ML and of human-computer interaction (HCI) researchers. We therefore organized a workshop on XAI during the IHM'22 conference and gathered about thirty researchers from the HCI and ML communities. The contribution of this paper sums up the main teachings and the most promising avenues for collaboration that emerged during the discussions.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: explainable artificial intelligence, machine learning, human-computer interaction
L'explicabilité en intelligence artificielle (XAI) est un domaine de recherche en plein essor aujourd'hui, en particulier en machine learning (ML). Ce domaine concerne à la fois la capacité technique à comprendre le fonctionnement des modèles de ML et l'adéquation des explications avec les utilisateurs et les contextes d'utilisation ciblés. Il rejoint ainsi les préoccupations des chercheurs à la fois en ML et en interaction humain-machine (IHM). Nous avons donc organisé un atelier sur l'XAI pendant la conférence IHM'22 et avons réuni une trentaine de chercheurs issus des communautés IHM et ML. La contribution de ce présent article résume les principaux enseignements et les pistes de collaboration les plus prometteuses qui ont émergé lors des discussions.

Mots-clés additionnels : intelligence artificielle explicable, machine learning, interaction humain-machine

Reference:
Julien Albert, Adrien Bibal, Benoît Frenay, and Bruno Dumas. 2023. Bridging the Gap between Human-Computer Interaction and Machine-Learning on Explainable AI: Initial Observations and Lessons Learned. IHM'23 : Actes étendus de la 34ème conférence Francophone sur l'Interaction Humain-Machine, April 03–06, 2023, Troyes, France.

1 INTRODUCTION

Explainability in artificial intelligence (XAI) regroups "movements, initiatives and efforts made in response to AI transparency and trust concerns" [2]. Its goal is to develop methods and tools "to explain or to present in understandable terms to a human" the working of AI systems [7].

XAI is a rapidly growing research field nowadays, in particular in machine learning (ML) [2]. This is due to multiple factors stemming from the needs of different stakeholders in the development and use of ML techniques. These include developers (from research and industry), service providers (private and public companies), end-users and third parties

(authorities, audit agencies, people in charge of GDPR compliance, etc.). In particular, the increasingly widespread use of deep learning (DL) techniques, due to their ground-breaking results in many fields and application domains, is one of the most important factors. Indeed, the inner working of DL models is particularly hard to understand and has therefore strongly concentrated research activities to address this so-called "black-box problem" [12].

XAI is concerned with the technical capacity to understand the functioning of the different types of ML models, but it is also concerned with the suitability of the explanations according to the users and the targeted contexts of use. It thus joins the concerns of both researchers in ML and in human-computer interaction (HCI). Indeed, the explainability of ML models has the user at its center and therefore cannot really improve without advances in HCI.

Recently, HCI community has actually begun to take interest in XAI topics. This is notably evidenced by the increasing presence of contributions and even tracks dedicated to the subject at recent editions of *CHI* and *IUI* conferences. Works like Abdul et al. [1], which tries to link multi-disciplinary state-of-the-art works in XAI with concerns specific to the HCI community, illustrate this new interest. DARPA's XAI research program [13] also illustrates the growing awareness that XAI research needs to involve other disciplines than ML, and in particular HCI. Moreover, a new trend in the field, called *human-centered XAI*, is beginning to emerge. Its aim is to put back the user at the center of concerns and take a more holistic view of XAI issues by taking into account insights from HCI and social sciences. This trend was initiated by the former work of Ehsan et al. [9], which focuses on the importance of involving the user in XAI research. More recently, Liao et al. [15] propose, in a position paper, to define the main principles of this trend and review the state-of-the-art.

As stated above, XAI research would benefit greatly from collaborations between ML and HCI researchers and we believe that it is essential to foster them. We therefore decided to organize a workshop on XAI during the *33rd International Francophone Conference on Human-Computer Interaction*[1] (IHM'22). This workshop, entitled *Human-Computer Interaction and Explainability in Artificial Intelligence*[2], aimed at gathering researchers from both communities so that they can present their works and exchange ideas. By this mean, we hoped to encourage further and more concrete collaborations between HCI and ML researchers on XAI topics.

The contribution of this paper will therefore review this workshop. First, it will briefly introduce the set up activities and the general course of events. Then, it will present the main ideas that emerged during the discussions between participants. Finally, it will conclude with the main teachings and the most promising directions for further collaborations.

2 A WORKSHOP BRIDGING HCI AND XAI

The general objective of the *HCI and XAI* workshop was to bring together researchers in HCI and ML in order to provide an overview of research activities on XAI involving both communities. The planned activities were research presentations in the morning and a world café in the afternoon. The aim of the morning presentations was to propose a panorama of research activities around XAI carried out by researchers in the HCI and ML communities. The afternoon world café was designed as a moment of meeting and discussion between participants on themes selected to encourage points of convergence between researchers with potentially very different backgrounds. The workshop took place during the *33rd International Francophone Conference on Human-Computer Interaction*, which was held from April 05 to 08, 2022 in Namur, Belgium. It gathered about thirty participants from various French and Belgian universities and research centers, working in HCI and ML domains.

[1]Conference Website: https://ihm2022.afihm.org/en/
[2]Workshop Website: https://projects.info.unamur.be/ihm-xai/index-en.html

3 THE WORLD CAFÉ

In order to encourage exchanges on research ideas, activities, events, initiatives, etc. in the field, a world café [19] was set up in the afternoon of the workshop. This conversational setup was chosen to encourage everyone's participation (through the formation of small discussion groups), to confront the sometimes very different visions, and to bring out original ideas based on everyone's input (which is enabled by the rotation process explained below).

In practice, the participants were divided into different thematic discussion tables. Periodically, they were invited to change tables and therefore discussion themes. During the discussions, they were also asked to write their ideas on sticky notes to keep track for the following groups. Finally, once the rotation cycle was over, the participants returned to their starting tables in order to structure the ideas collected during the discussions via the sticky notes and to present a synthesis to all other participants (see Fig. 1).

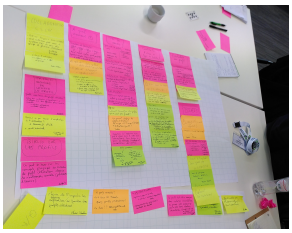

Fig. 1. Example of a panel with sticky notes generated during the first four rotations and being rearranged for presentation.

The four discussion themes for the world café were chosen based on their importance in the XAI field and their potential to be fruitful meeting points between ML and HCI researchers. The first theme, **User Profiles**, was chosen because of the crucial importance of understanding users' needs and context to design relevant XAI systems, and more generally to conduct meaningful XAI research [15]. The second theme, **Model-Representation-Presentation**, questioned whether the explanations should focus on the abstract mathematical entity that is the model, its representation (e.g., as a tree, as rules, as a text, etc. for a decision tree) and/or its visual presentation to the user (e.g., the choices of colors, shapes, etc. in a decision tree). As pointed out by Bibal and Frénay [5], it is important to distinguish these three stages in the interaction between the user and the model, whether for design or evaluation. The third theme, **Interaction and Actionability**, was motivated by the will to put back the explainability in an interaction perspective [13]. The focus is therefore on the actionability of the explanations, i.e., their efficiency and the effectiveness w.r.t. to the addressed use case, in particular the user's task. The fourth theme, **Evaluation**, concerns a mandatory step to validate and assess research outputs w.r.t. the state-of-the-art. However, current XAI approaches and methods are subject to much discussions given the conceptual difficulties of evaluating explanations [22].

4 RESULTS

This section sums up the results obtained during the world café. The goal is to present the ideas generated during the workshop, to discuss them in a prospective way and to map them with references from the XAI literature.

4.1 Theme 1: User Profiles

All participants agreed to consider that the user is of central importance in XAI. Indeed, more and more researchers from the field consider that one-size-fits-all XAI methods that can adapt to every user is an illusion [15]. Therefore, an

understanding of user characteristics and needs is essential to conduct meaningful XAI research. All HCI researchers pointed out that the HCI literature is quite abundant on this topic. However, as all users are different, it is sometimes hard to define relevant, yet generalizable, research contexts. Therefore, the most often used compromise between addressing particularities of each user and preserving the generalization potential of the results is to define user profiles.

Different types of profiles have been identified by the participants: (i) lay users without particular expertise in the domain or in AI, (ii) domain experts with strong knowledge of the application domain but no specific expertise with AI techniques, (iii) AI experts (e.g., data scientists and ML researchers), and (iv) regulators and auditors with strong legal and/or ethical concerns (e.g., GDPR). These profiles make it easier to design meaningful explanations. For instance, one can present explanations by using terms and concepts from the domain when targeting expert from this very domain. Another example is to adapt the explanation level of details to the profile. For instance, in the case of lay users, the accessibility is particularly important and the use of metaphors can help. As a final example, regulator and auditor profiles have very specific requirements about what should be a good explanation for a ML model [6].

From the perspective of HCI researchers, it is really important to adapt explanations to the targeted profile(s). However, spontaneous questions emerged, especially from ML researchers, during the world café, such as: Who can/should define the user profiles? (the expert, the developer, the user, the jurist?); How to discover relevant profile(s)?; How to technically take the profile into account in explanation process? These questions raise the need to have user research and modeling methods, which is a subject where the HCI literature can help. For instance, an interesting framework is proposed by Ribera and Garcia [18] to help define meaningful explanations based on some user profiles and corresponding archetypal tasks and goals.

Beyond the profile, some participants pointed out that it would be necessary to define the broader usage context, by including also the user task, the used device and the environment in which the interaction takes place. This is partly explored by Ferreira and Monteiro [10], who provide an overview of the different contexts in which explanations can be provided. About the user task, one of the main questions is: How is the user going to use the explanations? (e.g., to have some trust in the model, to modify its behavior, etc.). It is also important to understand the needs related to the usage context (e.g., regulatory context, business goals, trust or safety aspects, risk of the task, responsibility of the user, etc.). It is also especially important to define the level of explanation required (from very detailed, to no need at all). Furthermore, some users can also need different levels of explanation based on contextual elements (e.g., if an explanation does not match his expectations, the user may want to further investigate the model). Therefore, in those cases, the system must be able to adapt the level of explanation accordingly to the current context. Finally, the timing of the explanations is also important. This aspect can be considered from the perspective of the life cycle of a ML model: data collection, model training, model testing, deployment and usage, requires adapted explanations. The explanation can also be considered from the perspective of an AI-assisted decision process. The question can therefore be "when the explanation moment happens?" During the decision process? At the end of the process? Should the explanation be presented before or after the result? Etc.

Cognitive aspects, in particular biases, are also something that must be taken into account. Examples of perception biases related to XAI are: wrong/over interpretation, over/under reliance, narrative bias, confirmation bias, mere exposure effect, anchoring effect, beauty/soundness of the explanation (e.g., aesthetic qualities of dimensionality reduction methods like t-SNE and PCA when used for visualization, and other presentation elements like colors, distances, forms, etc.). These aspects are beginning to be explored in XAI. For instance, the conceptual framework proposed by Wang et al. [23] aim at helping explanation design choices based on potential users' cognitive processes.

The main issue when taking biases into account is to relate those to a specific profile (with questions like: Are there specific biases by profile? What is the prominence of some biases for a specific profile? Etc.). While it can seem at first sight that biases can be related to profiles based on some characteristics like data literacy or expertise/knowledge, there may be no direct link between biases and profiles, as no one is immune to biases. For instance, Ehsan et al. [8] show, among other biases, an "unwarranted faith in numbers" by the users when interacting with explanations with numerical values, independently of their expertise. Finally, explanations can impact the user's perception of the system, but also other parts of the development process, like the data annotation process (variability of the annotations w.r.t. the annotator, or even conflicting ground-truth).

4.2 Theme 2: Model-Representation-Presentation

This theme focused on what the explanation should apply (the abstract mathematics behind the models? The representation of it? And/or the final visual presentation of the chosen representation?). For the workshop's participants, fundamental questions on this topic were: What are legitimate representations for the inner working of a particular model?; Via what visualization tools/techniques do we deliver the chosen representation?; And what are we losing in the process? Those questions are of course strongly dependent of the addressed context, as described above. What is important to note in this three-stages perspective (Model-Representation-Presentation), is that while the explanation addresses the model, users do not interact directly with the model but rather with the visual presentation of the chosen representation of the model, as shown by Bibal and Frénay [5].

In particular, participants put forward that representation, and therefore presentation, depend on the addressed profile. It is important to rightly choose the representation and the presentation based on the addressed profile(s), by considering, among other things, the comprehension level of the user, his experience, his data literacy, etc.

While participants agreed on the importance of user profiles, a question that was still open was who is responsible for the choices of the representation and the presentation. Although the designer of the system may seem to be the best able to make an informed choice, it may sometimes be better to leave the choice to the user.

In particular for representations, explanation modalities are quite diverse and can be regrouped under the following categories [21]: numerical, rules, textual, visual and mixed (i.e., when modalities are combined). The identification of the strengths and weaknesses of each modality is an open question, but it seems that it heavily depends on the addressed context [18]. User characteristics are also important here, in particular the data literacy, to choose adapted modalities in a very broad design space. But here again, the choice of modality could be (partly) left to the user.

Another important aspect for the workshop participants was the loss of information on the path from the abstract mathematical model to its visual presentation (for instance, the latent space of the model may not always be directly interpretable without simplifications). This loss of information is of course inevitable but it must be managed carefully. The choice of a complete, or reduced, visual presentation of the model is based on the target user (e.g., an AI expert may need all the mathematical details, but a lay user only needs the main features used by the model) and the user task (e.g., must the model be explained globally or locally?). Once again, a more open solution could be to let the user have the possibility to choose the level of details.

As a final word on this theme, the participants emphasized that the model-representation-presentation path does not imply that the same choices should be made through all the stages. However, there exists some dependency between those. As an obvious case, a chosen model constraints the representation and presentation choices. For instance, choosing to use a decision tree will impact the possible representation (tree, rules, text, etc.) and the underlying visual presentation.

4.3 Theme 3: Interaction & Actionability

Explanations are not an end in themselves, but must allow the user to achieve his goals or tasks. To do so, it is important that the user can adequately interact with the system by means of explanations, possibly in an iterative way with multiple exchanges between the user and the explanation system. Actionability of the explanation is therefore an essential focus in order to design suitable systems. Among goals and tasks cited by the participants were the exploration (and possibly the modification) of black-box AI models, understanding and learning, the improvement of the trust towards AI models and systems and the will to get more control. In this context, an insightful tool is the guidelines about human-AI interaction proposed by Amershi et al. [3] based on existing literature and interviews with design practitioners. Those can help to meaningfully integrate explanations in some interaction scenarios.

Mentioned by the participants, human-in-the-loop ML regroups approaches that try to involve people in the development process of ML models [17]. For instance, in active learning, the user is asked by the model to label a subgroup of data for which it performs poorly. In XAI research, some active learning scenarios could also involve interactions where the user is asked by the model to correct some generated explanations. In the same vein, one participant suggested an interactive process to determine the amount of information to display, with the aim of converging towards a more suited visual presentation.

Taking into account the context, and especially target profile characteristics, is still essential here. More specifically, the complexity level of the representation impacts the user's ability to interact diligently with the underlying model. Another issue raised during discussions on integrating users in the loop is how the system can take into account the potential evolution of the context or the user. Indeed, tastes, opinions and habits are changing over time.

Finally, participants insisted on the fact that interactivity is a really desirable feature for XAI systems. This is in line with some works in literature, like the one of Sokol and Flach [20], which developed an interactive system that allows the user to personalize the explanations. This work highlights the iterative nature of such interaction where the user wishes to improve the system by interacting with it (in a kind of "what if..." process). According to the participants, allowing multiple interactions is also a good mean for the user to experience desirable feelings like trust, control, freedom and avoid unwanted ones like frustration. However, to do so, the interaction flow should be designed and validated with care, by taking into account these feelings. In order to achieve that, interaction strategies like gamification seem to offer promising perspectives [11].

4.4 Theme 4: Evaluation

Although all participants agreed that it is essential to evaluate XAI methods and systems, it is on this theme that the difference of research practices between HCI and ML participants appeared the clearer, especially with regard to the two categories of evaluation method: heuristic-based and user-based [5].

The principle of heuristic-based evaluation is to define an evaluation method based on one or more scoring functions. The purpose of these crafted scoring functions is to capture one or more properties that the researcher wishes to assess. It is therefore easier to rank XAI methods by their score and to identify the most effective ones based on the score they obtained (see, e.g., Li et al. [14]). Evaluations of this kind are also called, in the literature, *functionally-grounded* [7]. Those have the favor of ML researchers mainly for their ease of use and their objectivity.

The main advantage of heuristic-based evaluation, and its main limitation, is that no real user is involved, which greatly facilitates its use. However, its objective nature, as opposed to the subjective nature of user-based evaluation, should be more challenged and more subject to discussion, according to the participants. For instance, the formulation of

the scoring function is occasionally arbitrary and the link with the property being evaluated is sometimes questionable. It would therefore be interesting to have additional validations of these heuristic-based metrics, e.g., by validating a heuristic-based method by a user-based one. Another important issue with the heuristic-based evaluation of explanations is the need of ground-truth explanations to compare to the generated ones. Collecting those is not trivial for many reasons: they require human annotations which is a lengthy process, those annotations can be conflicting, and features used by the model and the human can be different without one of the two being wrong.

On the other hand, for HCI researchers, the importance of having real users is fundamental and used-based evaluation is therefore essential. Having real users makes it possible to confront the system with the outside world and the potential variety of usages. Even a seemingly relatively well-defined profile may in fact contain a wide variety of users. It is therefore the occasion to collect rich insights about the evaluated explanations. However, this variety of users can be challenging when assessing an XAI system. Indeed, the concept of explanation includes by definition a strong subjective dimension, which implies a broad range of different and plausible explanations between humans and models [16].

A quality of user-based evaluation is its capacity to adapt to a wide range of research goals and contexts. As pointed out by HCI researchers, user-based evaluation makes it possible to assess a bigger range of properties from XAI methods and systems than heuristics-based evaluation, i.e., the whole user experience. Furthermore, available methods and tools are also more diverse: questionnaires, interviews, eye-tracking, physiologic measures, user testing, etc. Given all these methods and tools, it is therefore easier to build some suitable evaluation protocol for a specific research context. However, external constraints can limit the available choices, especially the time and budget available.

Depending of the context and the addressed profile, having access to a sufficient pool of relevant users is not always easy. It is occasionally needed to find some proxy users and sometimes proxy tasks. This distinction is made in XAI evaluation literature with the terms *application-grounded* (real task and user) and *human-grounded* (proxy task and user) [7]. An aspect of user selection that is sometimes underestimated is the risk of bias due to the way users are recruited. For instance, using Amazon Mechanical Turk entails a risk of having a disproportionate number of people familiar with IT. Other biases were also evoked by the workshop's participants, such as the experimenter's bias when using physiological measures or the various user biases when using subjective methods.

As reminded by the participants, the first step when evaluating XAI methods or systems is to define an hypothesis. This hypothesis must include the contextual aspects (like the users' profile, but also the tasks, goals, constraints, etc. as discussed above), the ensuing needs, and the relevant properties to evaluate. Evaluation properties are also deeply associated to the notions related to explainability, as outlined by Vilone and Longo [22] who collected numerous notions that occurred in the existing XAI literature (e.g., faithfulness, robustness or sensitivity). It is also important to distinguish and prioritize desirable properties. For example, having explanations that are very faithful to the model and very accessible to lay users at the same time is quite complicated and some trade-off will generally be necessary. Finally, the discussion over evaluated properties highlighted a cultural difference between ML and HCI researchers. The former tend to consider properties in such a way that they can be objectified, and ultimately operationalized (i.e., formulated as a computable metric), While the latter tend to view properties through the prism of the user.

As the discussion between ML and HCI researchers progressed during the world café, a consensus seemed to emerged about the interest of combining heuristic-based and user-based evaluations. For instance, the first could be used to select the most promising approaches and the second to confirm and refine. Another way is to proceed as a block of layered evaluations by going in the direction of the evaluation of interactive and adaptive systems and use the most suitable method at each step. The need for robust guidelines to conduct evaluation in XAI could be a nice way for fruitful collaboration between ML and HCI researchers [4]. For instance, some experimented HCI participants insisted that it

could be easy to derive or adapt several evaluation methods from interactive systems to propose methods dedicated to XAI research, and with several possible categories (with or without users, with or without systems).

5 CONCLUSION AND FUTURE WORKS

The goal of this workshop was to foster exchanges between ML and HCI researchers about XAI, and to encourage future collaborations. About **User profiles**, participants pointed out that user research and modeling are essential steps in any XAI research project. However, although some dedicated methods and frameworks are beginning to appear, it seems that the question of adapting explanations to the profile and the context of use remains largely unexplored.

The **Model-Representation-Presentation** path from the working of the model towards its visual presentation to the user is very useful to understand the implications of design choices in the development of XAI systems. However, the importance of seeing the problem of explanation through different stages is often underestimated, especially the representation stage. A more dedicated study of this stage is lacking, questioning what could be done based on the studied model (i.e., what are the possible representations), and what should be done based on the addressed context.

About **Interaction & Actionability**, the participants agreed on the fact that displaying explanations is not an end in itself and that those explanations must be actionable by the user to fulfil his goals. A crucial point here is therefore to improve the user experience by designing suitable interaction flows and improving pragmatic and hedonic qualities. More specifically, the design of interaction flows allowing multiple round-trips between the system and the user seems to be a promising path to explore, especially because it is more in line with the human way of exchanging explanations.

On the **Evaluation**, there was an agreement on the fact that heuristic-based and user-based evaluations are more complementary than opposed. However, there is an urge to provide HCI-based guidelines, and more concrete collaborations to conduct relevant evaluations, as current practice remains quite different between ML and HCI researchers. In particular, robust methods combining heuristic-based and user-based evaluations seem to be a promising direction.

Finally, the main lesson of this workshop is the confirmation that more collaborations between researchers from ML and HCI communities are essential to address the main issues and challenges in XAI. Indeed, for ML researchers, there is a lot of knowledge and experience to take into consideration from the HCI community. For HCI researchers, XAI topics offer many research opportunities, and even more so given the ever-widening range of AI application domains. Therefore, we hope that this paper and the results of our workshop will be perceived as an invitation for both communities to collaborate more in the future.

ACKNOWLEDGMENTS

Julien Albert is supported by Service Public de Wallonie Recherche under grant n° 2010235-ARIAC by DIGITALWALLONIA4.AI. Adrien Bibal is supported by a Belgian American Educational Foundation (BAEF) grant. The workshop organizers would also like to warmly thank all the participants of the workshop for the quality of their involvement.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Conference on Human Factors in Computing Systems*. 1–18.

[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.

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Adrien Bibal, Bruno Dumas, and Benoît Frénay. 2019. User-Based Experiment Guidelines for Measuring Interpretability in Machine Learning. In *EGC Workshop on Advances in Interpretable Machine Learning and Artificial Intelligence*, Vol. 177.

[5] Adrien Bibal and Benoît Frénay. 2016. Interpretability of Machine Learning Models and Representations: An Introduction. In *The European Symposium on Artificial Neural Networks*. 77–82.

[6] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29, 2 (2021), 149–169.

[7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[8] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *arXiv preprint arXiv:2107.13509* (2021).

[9] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). 449–466.

[10] Juliana J. Ferreira and Mateus S. Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments (Lecture Notes in Computer Science)*, Aaron Marcus and Elizabeth Rosenzweig (Eds.). 56–73.

[11] Laura Beth Fulton, Ja Young Lee, Qian Wang, Zhendong Yuan, Jessica Hammer, and Adam Perer. 2020. Getting playful with explainable AI: games with a purpose to improve human understanding of AI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2019), 1–42.

[13] David Gunning and David W. Aha. 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine* 40, 2 (2019), 44–58.

[14] Xiao-Hui Li, Yuhan Shi, Haoyang Li, Wei Bai, Caleb Chen Cao, and Lei Chen. 2021. An Experimental Study of Quantitative Evaluations on Saliency Methods. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3200–3208.

[15] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[16] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38.

[17] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2022. Human-in-the-Loop Machine Learning: A State of the Art. *Artificial Intelligence Review* (2022), 1–50.

[18] Mireia Ribera and Àgata Lapedriza García. 2019. Can We Do Better Explanations? A Proposal of User-Centered Explainable AI. In *IUI Workshops*. 2327.

[19] Nikki Slocum. 2006. Le World Café. In *Méthodes participatives : Un guide pour l'utilisateur*. 173–183.

[20] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Künstliche Intelligenz* 34, 2 (2020), 235–250.

[21] Giulia Vilone and Luca Longo. 2020. Explainable Artificial Intelligence: A Systematic Review. *arXiv preprint arXiv:2006.00093* (2020).

[22] Giulia Vilone and Luca Longo. 2021. Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence. *Information Fusion* 76 (2021), 89–106.

[23] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.