Can we do better explanations? A proposal of User-Centered Explainable AI

Mireia Ribera
ribera@ub.edu
Universitat de Barcelona - Departament de Matemàtiques i Informàtica. Institut de Matemàtica de la Universitat de Barcelona
Barcelona, Spain

Agata Lapedriza
alapedriza@uoc.edu
Universitat Oberta de Catalunya
Barcelona, Spain

ABSTRACT

Artificial Intelligence systems are spreading to multiple applications and they are used by a more diverse audience. With this change of the use scenario, AI users will increasingly require explanations. The first part of this paper makes a review of the state of the art of Explainable AI and highlights how the current research is not paying enough attention to whom the explanations are targeted. In the second part of the paper, it is suggested a new explainability pipeline, where users are classified in three main groups (developers or AI researchers, domain experts and lay users). Inspired by the cooperative principles of conversations, it is discussed how creating different explanations for each of the targeted groups can overcome some of the difficulties related to creating good explanations and evaluating them.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Explainability; XAI; Conversational interfaces; User centered design; HCI

ACM Reference Format:
Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of User-Centered Explainable AI. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 6 pages.

1 INTRODUCTION

Artificial Intelligence (AI) is increasingly being used in more contexts and by a more diverse audience. In the future, AI will be involved in many decision-making processes. For example, in the medical field there will be AI systems that will help physicians to make diagnoses, whereas in companies the support of AI will be used in the interviewing process of recruiting campaigns. In these cases, different types of users, most of them without a deep understanding of how AI is built, will directly interact with AIs and will need to understand, verify and trust their decisions. This change of use scenarios of AI is similar to the one occurred in the '80s with the popularization of computers. When computers started to be produced massively and to be targeted to non-expert users, a need for improving human-computer interaction emerged which would accomplish to make technology accessible to less specialized users. In a similar way, a need for making AI understandable and trustful to general users is now emerging.

In this new broad scenario of AI use contexts, explainability plays a key role for many reasons, since in many cases the user interacting with the AI needs more reasoned information than just the decision made by the system.

Plenty of attention is being paid to the need for explainable AI. In the first part of this paper the current main theoretical concepts involved in explainability are reviewed. In particular, the recent surveys and theoretical frameworks of explainability focus the attention in 5 main aspects: (I) what an explanation is, (II) what the purposes and goals of explanations are, (III) what information do explanations have to contain, (IV) what type of explanations can a system give, and (V) how can we evaluate the quality of explanations. This review reveals, in our opinion, how the current theoretical approach of explainable AI is not paying enough attention to what we believe is a key component: who are the explanations targeted to.

In the second part of this paper, we argue that explanations cannot be monolithic and that each stakeholder looks for explanations with different aims, different expectations, different background, and different needs. By building on the conversational nature of explanations, we will outline how explanations could be created to fulfill the demands set on them.

2 HOW DO WE APPROACH EXPLAINABILITY?

Defining what an explanation is, is the starting point for creating explainable models, and allows to set the three pillars on which explanations are built: goals of an explanation, content of an explanation, and types of explanations. The last key aspect reviewed in this section is how explanations can be evaluated, which is a critical point for the progress of explainable AI.

2.1 Definition of explanation

Explanations are "ill-defined" [17]. In the literature the concept of explainability is related to transparency, interpretability, trust, fairness and accountability, among others [1]. Interpretability, sometimes used as a synonym of explainability, is defined by Doshi and Kim [6] as "the ability to explain or to present in understandable terms to a human". Gilpin et al. [7], on the contrary, consider explainability a broader subject than interpretability; these authors

IUI Workshops'19, March 20, 2019, Los Angeles, USA

state that a model is interpretable if it is "able to summarize the reasons for [system] behavior, gain the trust of users, or produce insights about the causes of decisions". However an explainable AI needs, in addition, "to be complete, with the capacity to defend [its] actions, provide relevant responses to questions, and be audited". Rudin [24] defines Interpretable Machine Learning in a more restricted sense, as "When you use a model that is not a black box", while Explainable Machine Learning is, for this author, "when you use a black box and explain it afterwards".

Miller [19], does an interesting review of social science constructs to find the theoretical roots of the explainability concept. For example, Lewis [15] states that "To explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event –explanatory information – tries to convey it to someone else". Halpern and Pearl [12] define a good explanation as a response to a Why question, that "(a) provides information that goes beyond the knowledge of the individual asking the question and (b) be such that the individual can see that it would, if true, be (or be very likely to be) a cause of". After the review, Miller [19] extracts four characteristics of explanations: "explanations are contrastive" (why this and not that), "explanations are selected in a biased manner (not everything shall be explained)", "probabilities don't matter" and finally "explanations are social".

From these definitions and the recent reviews of explainability [7, 10] we can conclude that there is no agreement on a specific definition for explanation. However, some relevant points are shared in almost every definition. For example, many definitions relate explanations with "why" questions or causality reasonings. Also, and more importantly, there is a key aspect when trying to define what an explanation is: there are two subjects involved in any explanation, the one who provides it (the system), or *explainer*, and the one who receives it (the human), or *explainee*. Thus, when providing AI with explainability capacity, one can not forget about to whom the explanation is targeted.

2.2 Goals of explanations (WHY)

In this section we review, in the context of explainable AI, what the purpose of an explanation is. According to Sameket al. [25] the need of explainable systems is rooted in four points: (a) Verification of the system: Understand the rules governing the decision process in order to detect possible biases; (b) Improvement of the system: Understand the model and the dataset to compare different models and to avoid failures; (c) Learning from the system: "Extract the distilled knowledge from the AI system"; (d) Compliance with legislation (particularly with the "right to explanation" set by European Union): To find answers to legal questions and to inform people affected by AI decisions.

Gilpin et al.[7] mostly agree with these goals, adding specific considerations on two of these points: (a) Verification of the system: explanations help to ensure that algorithms perform as expected, and (b) Improvement of the system: in terms of safety against attacks. Guidotti, et al. [10] enforce for (c) "the sake of openness of scientific discovery and the progress of research" , while Miller [19] directly considers "facilitating learning" the primary function of explanation. Wachter, et al. [26] describe more in detail three aims behind the right to explanation. These three aims are "to inform and help the subject understand why a particular decision was reached, to provide grounds to contest adverse decisions, and to understand what could be changed to receive a desired result in the future, based on the current decision-making model".

Lim et al. [16] add a new goal, relating explainability to: (e) Adoption: Acceptance of the technology. These authors state that "[the] lack of system intelligibility (in particular if a mismatch between user expectation and system behavior occurs) can lead users to mistrust the system, misuse it, or abandon it altogether".

Doshi-Velez and Kim [6] focus on (b) and (d) and see interpretability as a proxy to evaluate safety and nondiscrimination, which can be related to fairness in AI. They also argue that an explanation is only necessary when wrong results may have an important impact or when the problem is incompletely studied. Rudin [24] agrees with that last view, but also mentions troubleshooting (a) as an important goal. On a more theoretical framework, Wilkenfeld and Lombrozo [27], cited in [19], discuss about other functions of explanations such as persuasion or assignment of blame, and they raise attention to the fact that the goals of explainer and explainee may be different.

Interesting discussions can be found in terms of the need and utility of explanations. Abdul et al.[1] see explanations as a way for humans to remain in control. This view is questioned by Lipton [17], who warns against explanations "to simply be a concession to institutional biases against new methods", arising a more deep reflection on how AI fits our society: to empower people or to surpass people. Finally, Rudin [24], in her controversial video seminar, questions the utility of explanations, and states that they only "perpetuate the problem of bad stuff happening", because they act somewhat as a disclaimer. Furthermore, some authors agree that the explainee will only require explanations when the system decision does not match her expectations [8].

Despite the disagreement of some experts on the need of explanations, there are more reasons supporting that AI should be able of explaining its decisions than the opposite. In particular it is very likely that users expect an explanation when the decision of an AI has important economical consequences or it affects their rights. What is clear is that trying to cover all goals with a unique explanation is overwhelming [7]. If we take into account the *explainee* each goal is addressed to, maybe a practical solution could be to create several explanations serving only the specific goals related to a particular audience.

2.3 Content to include in the explanation (WHAT)

There are different opinions about what the content of a good explanation has to include. Lim et al. [16] say that an explanation should answer five questions: "(1) What did the system do?, (2) Why did the system do P?, (3) Why did the system not do X?, (4) What would the system do if Y happens? , (5) How can I get the system to do Z, given the current context?" . These questions are very similar to the explanatory question classes introduced by Miller [19]. Gilpin et al. [7], on the contrary, add a new question related to the data stored by the system: (6) "What information does the system contain?"

Lim et al. [16] relate their five questions to Don Norman gulfs of evaluation and execution, solving questions 1-3 the separation between perceived functionality of the system and the user's intentions and expectations, and questions 4-5 the separation between what can be done with the system and the user's perception of its capacity. These authors tested the questions on an explanatory system with final users and they concluded that "Why questions" (2) were the most important.

Some authors categorize the explanations depending on whether they explain how the model works or the reason of a particular output [7, 10]. Although both aspects are connected, explanations can be more specific when focused on a local result. In the first case, the explanation is more global, and can help users to build a mental model of the system. This global explanation includes also the representation learned by the model (for example, in a Neural Network, what are the roles of layers or units), that allows users to understand the structures of the system. In the latter, the explanation focuses in a specific output and allows users to understand better the reasons why that specific output occurred or the relation between a specific input and its output.

Overall, there are multiple questions that good explanations should provide answers to. We observe, however, a quite consistent agreement on the importance of the "Why" questions. Furthermore, some explanation contents are more interesting or important for some users than others. For example, researchers developing the AI system might be interested in technical explanations on how the system works to improve it, while lay users, with no technical background, would not be interested at all about these type of explanation.

2.4 Types of explanations (HOW)

In this section we review the different ways of classifying explanations according to how they are generated and delivered to the user.

In terms of generation, explanations can be an intrinsic part of the system, which becomes transparent and open to inspection (for some authors this is called interpretability). For example, CART (Classification and regression trees) [2] is a classical decision tree algorithm that functions as a white box AI system. On the contrary, explanations can be post-hoc, built once the decision is already made [17, 20]. For instance, LIME by Ribeiro et al. [23] consists of a local surrogate model that reproduces the system behavior for a set of inputs. Detailed pros and cons of each of these two types are discussed in [20]. In particular, while intrinsic explanations need to impose restrictions on the design of the system, post-hoc explanations are usually unable to give information on the representation learned by the system or on how the system is internally working.

Regarding to the explanation modality, we can find explanations in natural language with "analytic (didactic) statements [...] that describe the elements and context that support a choice", as visualizations, "that directly highlight portions of the raw data that support a choice and allow viewers to form their own perceptual understanding", as cases or "explanations by example", "that invoke specific examples or stories that support the choice", or as rejections of alternative choices or "counterfactuals" "that argue against less preferred answers based on analytics, cases, and data" [11, 17]. Currently visualizations are probably the most common type of explanations (see [28] for a recent review), with a longer tradition of interaction and evaluation methods [13].

We can see there exist many types of explanations and, although visualizations are among the most adopted, it is not clear when or why one type is better than another. In some cases the most suitable modality will depend on the content of the explanation. Furthermore, the user should also play an important role on deciding what type of explanation is the most appropriate according to background, specific expectations or needs.

2.5 Evaluation of explanations

Taking into account all the complexities described above on defining and creating explanations, one can expect that it is not evident how to test the quality of an explanation. Actually, evaluating explanations is maybe the most immature aspect on the research on explainable AI. Lipton [17] and Miller [19] openly question the existing practices for evaluating explanations. Lipton says that "the question of correctness has been dodged, and only subjective views are proposed". Miller [19] argues that most explanations rely on causal relations while people do not find likely causes very useful, and states that simplicity, generality and coherence are "at least as equally important".

In a promising direction, Doshi-Velez and Kim [6] criticize the weakness of current methods for explanation evaluation, and suggest grounding evaluations on more solid principles, based on Human Computer Interaction (HCI) user tests. The authors suggest three possible approaches, from more specific and costly to more general and cheap: (1) application-grounded evaluation with real humans and real tasks; (2) human-grounded evaluation with real humans but simplified tasks; and (3) functionally-grounded evaluation without humans and proxy tasks; all of them always inspired by real tasks and real humans' observations.

The Explainable AI DARPA program (XAI) [11], started on 2017, tries to cover current gaps of this topic and opens many scientific research lines to solve them. The program conceptualizes the goals of explanation as to generate trust and facilitate appropriate use of technology (focusing mainly in adoption, the (e) goal of explanations). The project relates the explanation goals with several elements to evaluate, each one linked to a corresponding indicator.

On the Open Learning Modelling domain, Conati et al [3], based on Mabbot and Bull[18] previous experiments, point out some key considerations on designing explanations such as considering the explainee, as we suggest, and also the reason to build the system, which aspects to made available to the user and the degree it can be manipulated by the user.

On a more technical vein, Gilpin et al.[7], after a review of the literature, cite four evaluation methods. The first two are related to processing (completeness to model, completeness on substitute task), while the last two related to representation (completeness on substitute task, detect biases) and explanation producing (human evaluation, detect biases).

Setting clear evaluation goals and metrics is critical in order to advance the research on explainability and more efforts are needed in this area. How can we say that a system is better than another if we do not know why? More efforts are needed in the area of evaluating explanations. Doshi-Velez and Kim [6], and DARPA [11] proposals have strong points, but they do not cover all the goals set on explainable systems, nor all the modalities and explanation contents.

3 CAN WE DO BETTER?

In this section we critically review the previous sections and give insights on new directions to create better explanations. We build our proposal upon two main axes: (1) to provide more than one explanation, each targeted to a different user group, and (2) making explanations that follow cooperative principles of human conversation.

In order to better contextualize current developments in explainability, we suggest to take into account the communicative nature of the explanations and to categorize explainees in three main groups, based on their goals, background and relationship with the product [4],[5]:

- **Developers and AI researchers**: investigators in AI, software developers, or data analysts who create the AI system.
- **Domain experts**: specialists in the area of expertise where the decisions made by the system belong to. For example: physicists or lawyers.
- **Lay users**: the final recipients of the decisions. For example: a person accepted or rejected on a loan demand, or a patient that has been diagnosed.

Starting with explainability goals, if we take a closer look to the listed goals, we can detect different needs and *explainee* profiles for each of them. (a) verification and (b) improvement goals, clearly appeal to a developer or researcher profile, who wants to improve the algorithm's parameters or optimization. These goals can be attained with the help of domain experts to whom the tool is intended to help: they will be the ones that detect possible failures of the system. However, for the domain experts, the main goal can be to learn from the system (c), to understand the mechanisms of inference or correlation that the system uses in order to improve their decision methods or to hypothesize possible general rules. For domain experts the *explainer* goal providing explanations is to grant the system adoption (e). The last goal mentioned by Samek, the right to an explanation, is clearly targeted to lay users because the system decisions may have economical or personal implications for them, although this goal can be also relevant for domain experts, who might have the legal responsibility of the final decision.

Related to explanation content, Doshi-Velez and Kim [6] argue that different explanations are needed depending on global versus local scope, thematic area, severity of incompleteness, time constraints and nature of user expertise. We can delve a bit more on this idea, particularly in the need to tailor explanations to user expertise, and exemplify it with the following scenario. Let's say we have a system that offers explanations at the representational level, describing data structures; these should clearly not be communicated in the same language for developers as for domain-experts. Even different area domain-experts will require different kind of explanations [22].

In terms of types of explanations, Lipton [17] states that humans do not exhibit transparency, sustaining that human explanations are always post-hoc. On the other side, many authors are concerned about the high complexity of machine learning algorithms and the limits of human reasoning to understand them [26]. This relates to Nielsen heuristic of progressive disclosure or Shneiderman visual information-seeking mantra: "Overview first, zoom and filter, then details-on-demand" as techniques to cope with complex information or tasks. To make explanations more human, Naveed, Donker and Ziegler [21] introduce an interesting framework of explanations based on Toulmin's argumentation model. This model proposal is to communicate decisions giving evidences, like facts or data, that support the decision, and relating both the evidences and the decision with contextual information. Other authors suggest interaction as a way to explore the explanation space: "allowing people to interactively explore explanations for algorithmic decision-making is a promising direction" [1] "By providing interactive partial dependence diagnostics, data scientists can understand how features affect the prediction overall" [14].

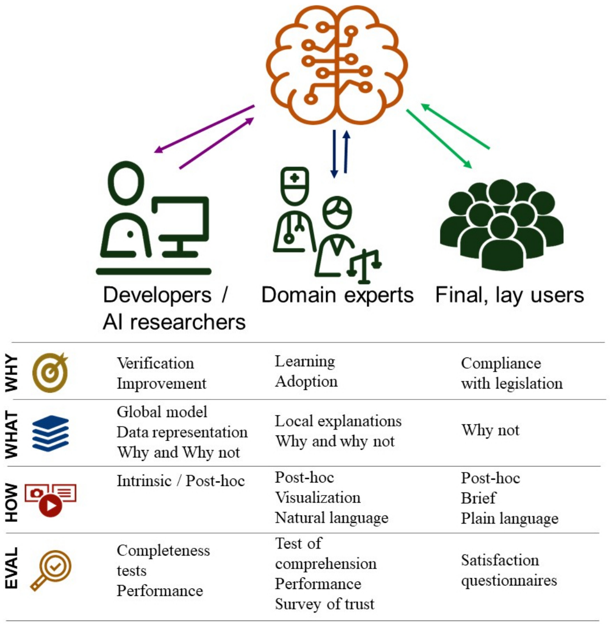
Likewise, Miller [19] criticizes the current proposed explanations as being too static, he describes them ideally as "an interaction between the explainer and explainee". Delving on the fourth feature he identified in social science theoretical constructs: "explanations are social", this author parallels explanations to conversations . Therefore explanations must follow the cooperative principles of Grice [9] and its four maxims: **1. Quality**: Make sure that the information is of high quality: (a) do not say things that you believe to be false; and (b) do not say things for which you do not have sufficient evidence; **2. Quantity**: Provide the right quantity of information. (a) make your contribution as informative as is required; and (b) do not make it more informative than is required; **3. Relation**: Only provide information that is related to the conversation. (a) Be relevant. This maxim can be interpreted as a strategy for achieving the maxim of quantity; **4. Manner**: Relating to how one provides information, rather than what is provided. This consists of the 'supermaxim' of 'Be perspicuous', and according to Grice, is broken into various maxims such as: "(a) avoid obscurity of expression; (b) avoid ambiguity; (c) be brief (avoid unnecessary prolixity); and (d) be orderly".

We observe that (1), (2), and (3) refer to the content of the explanation, while (4) refers to the type of explanation. Notice that these 4 cooperative principles can also be related to other wanted properties of explanations [20], such as fidelity or comprehensibility. Our claim is that Explainable AI for domain-experts and lay users can benefit from the theoretical frameworks developed for human communication.

Finally, considering evaluation, we can also observe that different metrics appeal to different needs and audience. For example, testing completeness or functionally-grounded evaluation are targeted to developers or AI scientists, task performance and mental model appeal to domain experts whereas trust is intended for domain experts and lay users. If we deliver different explanations, targeted to a specific of the above mentioned groups, it will be easier to evaluate them, since we can use the most suitable metric for each case.

4 USER-CENTERED EXPLAINABLE AI

From the literature review and discussions above presented, we conclude that explanations are multifaceted and cannot be attained with one single, static explanation. Since it is very difficult to approach explainable AI in a way that fulfills all the expected requirements at the same time, we suggest creating different explanations for every need and user profile. The rest of this section gives more details on this idea and discusses the different reasons that support our proposal.



	Developers / AI researchers	Domain experts	Final, lay users
WHY	Verification Improvement	Learning Adoption	Compliance with legislation
WHAT	Global model Data representation Why and Why not	Local explanations Why and why not	Why not
HOW	Intrinsic / Post-hoc	Post-hoc Visualization Natural language	Post-hoc Brief Plain language
EVAL	Completeness tests Performance	Test of comprehension Performance Survey of trust	Satisfaction questionnaires

Figure 1: The system targets explanations to different types of user, taking into account their different goals, and providing relevant (Grice 3rd maxim) and customized information to them (Grice 2nd and 4th maxim), as described in section 2. Evaluation methods are also tailored to each explanation

As argued above, we suggest that AI explanations should follow the 4 cooperative principles previously described. In this context, if different explanations are specifically designed for different audiences or users, we can design each one with a particular purpose, content, and present it in a specific way. This procedure makes it easier to follow the principles of *(2) quantity: deliver the right quantity of data and abstraction*, and *(3) relation: be relevant to each stakeholder*. Concretely, taking into account the current research in explainability we suggest these 3 big families of explanations:

- **Developers and AI researchers**: Model inspection and simulation with proxy models. These two types of explanations are very well suited to verify the system, detect failures and give hints to improve it. The mode of communication fits well the audience, who are able to understand code, data representation structures and statistical deviations. Completeness tests covering different scenarios can be set to evaluate the explanation.

- **Domain-experts**: provide explanations through natural language conversations or interactive visualizations, letting the expert decide when and how to question the explanation and led her discovery by herself. Explanations must be customized to the discipline area of the domain experts and to the context of their application, be it legal or medical decisions, or any other, in order to be clear and to use the discipline terminology. Test of comprehension, performance and survey of trust can be set to evaluate the explanation.

- **Lay users**: outcome explanations with several counterfactuals [26] with which users can interact to select the one most interesting to their particular case. This explanation is parallel to human modes and it is very likely to generate trust. Satisfaction questionnaires can be set to evaluate the explanation.

Our proposal is that explanations need to be designed taking into account the type of user they are targeted to, as shown in the pipeline for explanation of Figure 1. That means to approach explainable AI from a *user-centered* perspective, putting the user in a central position. Approaching explainability in that way has two main benefits. First, it makes the design and creation of explainable systems more affordable, because the purpose of the explanation is more concrete and can be more specifically defined than when we try to create an all-sizes all-audiences explanation. Second, it will increase satisfaction among developers or researchers, domain-experts and lay users, since each of them receives a more targeted explanation that is easier to understand than a general explanation. Finally, it will be easier to evaluate which explanation is better because we have metrics that are specific to each case.

Wachter et al. [26] proposal of counterfactual explanations fulfilling the right of explanation is a good example that supports the implementation of these principles. In their paper they abound in the need to make explanations adapted to lay users (user-centered design) "information disclosures need to be tailored to their audience, with envisioned audiences including children and uneducated laypeople" , "the utility of such approaches outside of model debugging by expert programmers is unclear". They also emphasize the need to give a "minimal amount of information" (be relevant), "counterfactual explanations are intentionally restricted". Moreover, when the authors talk about the suitability of offering "multiple diverse counterfactual explanations to data subjects", they could benefit from a conversational approach.

While the proposed scheme of user-centered explainable AI particularly benefits the quantity and relation principles, the manner can also be chosen to be as appropriate as possible to the user. For example, although natural language descriptions can be a suitable modality for any of the three user groups, the specific vocabulary should be adapted to the user background. In particular, technical terms are not a good choice for explanations targeted to a lay user, and explanations for domain-experts should use their respective area terminology. Finally, regarding to the quality principle, we think it has to be always applied in the same way, and it is not necessary to take into account the specific user group.

5 CONCLUSION

While there has been a great progress in some aspects of explainability techniques, we observed that there is a key aspect that is being misrepresented in several of the current approaches: the user to whom the explanation is targeted to. Putting explanations in the user context makes explainability easier to approach than when we try to create explainable systems that fulfill all the requirements of a general explanation. In addition, the user-centered framework gives clues on how to create more understandable and useful explanations for any user, because we can follow the principles of human communication, thoroughly studied.

More generally, the increasing demand of explainable AI systems and the different background of stakeholders of machine learning systems justify, in our view, to revise the concept of explanations as unitary solutions and to propose the creation of different user-centered explainability solutions, simulating human conversations with interactive dialogues or visualizations that can be explored.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya. Icons used in Figure 1, are from Flaticon, made by Freepik and Smashicons. We thank Jordi Vitrià for his review and suggestions on the whole article.

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 *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. Association for Computing Machinery, Montreal, Canada, 1–18. https://doi.org/10.1145/3173574.3174156

[2] L Breiman. 1984. Algorithm CART. In *Classification and Regression Trees*. Chapman and Hall/CRC, Monterey, CA, s.n.

[3] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. (2018). arXiv:arXiv:1807.00154

[4] Alan Cooper et al. 2004. *The inmates are running the asylum:[Why high-tech products drive us crazy and how to restore the sanity]*. Sams Indianapolis, Indianapolis.

[5] Alan Cooper, Robert Reimann, and David Cronin. 2007. *About face 3: the essentials of interaction design*. John Wiley & Sons, United Staes.

[6] Finale Doshi-velez and Been Kim. 2017. A Roadmap for a Rigorous Science of Interpretability. *stat* 1050 (2017), 28.

[7] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. (2018). arXiv:arXiv:1806.00069

[8] Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly* 23, 4 (dec 1999), 497. https://doi.org/10.2307/249487

[9] H.P. Grice. 1975. Logic and Conversation. In *Syntax and semantics 3: Speech arts*. Academic Press, New York, 41–58.

[10] Riccardo Guidotti, Anna Monreale, and Salvatore Ruggieri. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR* 51, 5 (2018), 42 p.

[11] David Gunning. 2017. *Explainable Artificial Intelligence (XAI)*. Report. Defense Advanced Research Projects Agency (DARPA).

[12] Joseph Y Halpern and Judea Pearl. 2005. Causes and Explanations : A Structural-Model Approach . Part II : Explanations. *The British Journal for the Philosophy of Science* 56, 4 (2005), 889–911. https://doi.org/10.1093/bjps/axi148

[13] Jeffrey Heer and Ben Shneiderman. 2012. Interactive dynamics for visual analysis. *Communications of the ACM ACM* 55, 4 (apr 2012), 45–54. https://doi.org/10.1145/2133806.2133821

[14] Josua Krause, Adam Perer, and I B M T J Watson. 2016. Interacting with Predictions : Visual Inspection of Black-box Machine Learning Models. In *CHI'16*. Association for Computing Machinery, New York, 5686–5697. https://doi.org/10.1145/2858036.2858529

[15] David Lewis. 1986. Causal Explanation. In *Philosophical Papers. Vol II*. Oxford University Press, New York, Chapter Twenty two, 214–240.

[16] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, 2119–2128.

[17] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. (2016). arXiv:arXiv:1606.03490

[18] Andrew Mabbott and Susan Bull. 2006. Student preferences for editing, persuading, and negotiating the open learner model. In *International Conference on Intelligent Tutoring Systems*. Springer, Berlin, 481–490.

[19] Tim Miller. 2019. Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[20] Christoph Molnar. 2018. *Interpretable Machine Learning: a guide for making black box models explainable*. Leanpub, British Columbia, Canada. https://christophm.github.io/interpretable-ml-book/

[21] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-Based Explanations in Recommender Systems. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization - UMAP '18*. Association for Computing Machinery, New York, 293–298. https://doi.org/10.1145/3213586.3225240

[22] Forough Poursabzi-Sangdeh. 2018. *Design and Empirical Evaluation of Interactive and Interpretable Machine Learning*. Ph.D. Dissertation. University of Colorado, Boulder. https://scholar.colorado.edu/csci

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. Association for Computing Machinery, New York, 1135–1144.

[24] Cynthia Rudin. 2018. Please stop doing "explainable" ML. (2018). https://bit.ly/2QmYhaV

[25] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. (2017). arXiv:arXiv:1708.08296

[26] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 1–52. https://doi.org/10.2139/ssrn.3063289 arXiv:1711.00399

[27] Daniel A. Wilkenfeld and Tania Lombrozo. 2015. Inference to the Best Explanation (IBE) Versus Explaining for the Best Inference (EBI). *Science and Education* 24, 9-10 (2015), 1059–1077. https://doi.org/10.1007/s11191-015-9784-4

[28] Quanshi Zhang and Song-Chun Zhu. 2018. Visual Interpretability for Deep Learning: a Survey. *Frontiers in Information Technology & Information Engineering* 19, 1423305 (2018), 27–39. https://doi.org/10.1631/fitee.1700808 arXiv:1802.00614