A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Martin Schuessler
Technische Universität Berlin
Berlin, Germany
schuessler@tu-berlin.de

ABSTRACT

The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation methods. However, there is no consensus among the involved disciplines regarding the evaluation of their effectiveness - especially concerning the involvement of human subjects. For our community, such involvement is a prerequisite for rigorous evaluation. To better understand how researchers across the disciplines approach human subject XAI evaluation, we propose developing a taxonomy that is iterated with a systematic literature review. Approaching them from an HCI perspective, we analyze which study designs scholar chose for different explanation goals. Based on our preliminary analysis, we present a taxonomy that provides guidance for researchers and practitioners on the design and execution of XAI evaluations. With this position paper, we put our survey approach and preliminary results up for discussion with our fellow researchers.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

explainable artificial intelligence; explanation; human evaluation; taxonomy.

ACM Reference Format:
Michael Chromik and Martin Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. In *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20) Cagliari, Italy*. 7 pages.

1 INTRODUCTION

We have witnessed the widespread adoption of intelligent systems into many contexts of our lives. Such systems are often built on advanced machine learning (ML) algorithms that enable powerful predictions – often at the expense of interpretability. As these systems are introduced into more sensitive contexts of society, there is a growing acceptance that they need to be capable of explaining their behavior in human-understandable terms. Hence, much research is conducted within the emerging domain of *explainable artificial intelligence* (XAI) and *interpretable machine learning* (IML) on developing models, methods, and interfaces that are interpretable to human users – often through some notion of explanation.

However, most works focus on computational problems while limited research effort is reported concerning their user evaluation. Previous surveys identified the need for more rigid empirical evaluation of explanations [2, 5, 17]. The AI and ML communities often strive for *functional* evaluation of their approaches with benchmark data to demonstrate generalizability. While this is suitable to demonstrate technical feasibility, it is also problematic since often *"there is no formal definition of a correct or best explanation"* [24]. Even if a formal foundation exists, it does not necessarily result in practical utility for humans as the utility of an explanation is highly dependent on the context and capabilities of human users. Without proper human behavior evaluations, it is difficult to assess an explanation method's utility for practical use cases [26]. We argue that functional and behavioral evaluation approaches have their legitimacy. Yet, since there is no consensus on evaluation methods, the comparison and validation of diverse explanation techniques is an open challenge [2, 4].

In this work, we take an HCI perspective and focus on evaluations with human subjects. We believe that the HCI community should be the driving force for establishing rigorous evaluation procedures that investigate how XAI can benefit users. Our work is guided by three research questions:

- **RQ-1**: Which evaluation approaches have been proposed and discussed across disciplines in the field of XAI?
- **RQ-2**: Which study design decisions have researchers made in previous evaluations with human subjects?
- **RQ-3**: How can the proposed approaches and study designs be integrated into a guiding taxonomy for human-centered XAI evaluation?

The contribution of this workshop paper is two-fold: First, we introduce our methodology for taxonomy development and literature review guided by RQ-1 and RQ-2. The review aims to provide an overview of how evaluations are currently conducted and help identify suitable best practices. As a second contribution, we present a preliminary taxonomy of human evaluation approaches in XAI and describe its dimensions. Taxonomies have been used in many disciplines to help researchers and practitioners to understand and analyze complex domains [23]. Our overarching goal is to synthesize a human subject evaluation guideline for researchers and practitioners of different disciplines in the field of XAI. With this work, we put our review methodology and preliminary taxonomy up for discussion with our fellow researchers.

ExSS-ATEC'20, March 2020, Cagliari, Italy

2 FOUNDATIONS AND RELATED WORK

2.1 Evaluating Explanations in Social Sciences

Miller defines explanation as either a process or a product [16]. On the one hand, an explanation describes the cognitive process of identifying the cause(s) of a particular event. At the same time, it is a social process between an *explainer* (sender of an explanation) and an *explainee* (receiver of an explanation) with the goal to transfer knowledge about the cognitive process. Lastly, an explanation can describe the product that results from the cognitive process and aims to answer a why-question. In our paper, we refer to explanations from the product perspective. Psychologists and social scientists investigated how humans evaluate explanations for decades. Within their disciplines, *explanation evaluation* refers to the process applied by an explainee for determining if an explanation is satisfactory [16]. Scholars conducted experiments where they presented participants with different types of explanations as treatments. These experiments indicate that choosing one explanation over another is often an arbitrary choice heavily influenced by cognitive biases and heuristics [12]. The primary criteria of explainees are whether the explanation helps them to understand the underlying cause [16]. For instance, humans are more likely to accept explanations that are consistent with their prior beliefs. Furthermore, they prefer explanations that are simpler (i.e., with fewer causes), and more generalizable (i.e., that apply to more events). Also, the effectiveness of an explanation depends on the current information needs of the explainee. A suitable explanation for one purpose may be irrelevant for another. Thus, for an explanation to be effective, it is essential to know the intended context of use.

2.2 Explainable Artificial Intelligence (XAI)

Interpretability in machine learning is not a monolithic concept [15]. Instead, it is used to indirectly evaluate whether important desiderata, such as fairness, reliability, causality, or trust, are met in a particular context [4]. Some definitions of interpretability are rather *system-centric*. Doshi-Velez and Kim [4] describe it as a model's *"ability to explain or to present in understandable terms to a human."* Miller [16] takes a more *human-centered* perspective calling it *"the degree to which an observer can understand the cause of a decision"*. Human understanding can be fostered either by offering means of introspection or through explanations [3]. A large variety of methods exist for both approaches [9]. The term *interpretable machine learning* (IML) often refers to research on models and algorithms that are considered as inherently interpretable while *explainable AI* (XAI) often refers to the generation of (post-hoc) explanations or means of introspection for black-box models [27, 33]. A model's black-box behavior may manifest itself in two ways: either from complex architectures, as with deep neural networks, or from proprietary models (that may otherwise be interpretable), as with the COMPAS recidivism model [27]. The lines between IML and XAI are often seamless and the terms are often used interchangeably. For instance, DARPA's XAI program subsumes both terms with the objective to *"enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"* [10].

2.3 Evaluating Explanations in XAI

Multiple surveys of the ever-growing field of XAI exist. They formalize and ground the concept of XAI [1, 2], relate it to adjacent concepts and disciplines [1, 16], categorize methods [9], or discuss future research directions [1, 2]. All these surveys report a lack of rigid evaluations. Adadi et al. found that only 5% of surveyed papers evaluate XAI methods and quantify their relevance [2]. Similarly, Nunes and Jannach found that 78% of the analyzed papers on explanations in decision support systems lacked structured evaluations that go beyond anecdotal "toy examples" [24].

Some works have addressed the design and conduction of explanation evaluations in XAI. Gilpin et al. survey explainable methods for deep neural networks and describe a categorization of evaluation approaches at different stages of the ML development process [8]. Yang et al. provide a framework consisting of multiple levels of explanation evaluation [33]. Their definition of *persuasibility* (measuring the degree of human comprehension) focuses on the human and resonates with our notion of human subject evaluation. Our work aims to elaborate on their generic strategy of *"employing users for human studies"*. Nunes and Jannach reviewed 217 publications spanning multiple decades and briefly report findings from applied evaluation approaches [24]. Based on their survey they derive a comprehensive taxonomy that guides the design of explanations. However, their taxonomy omits aspects of evaluation. Mueller identified 39 XAI papers that reported empirical evaluations and qualitatively described chosen evaluation approaches along 9 dimensions [20].

While these works offer valuable ideas, they are limited in their scope and, thus, offer little guidance for XAI user evaluations. Of course, *"there is no standard design for user studies that evaluate forms of explanations"* [24]. However, we believe that a unified taxonomy is needed that integrates the most common ideas related to human subject evaluation and extends them with best practice examples. Such an actionable format can provide great benefit for researchers and practitioners by guiding them through the design and reporting of structured XAI evaluations.

3 METHODOLOGY

In this section, we outline our method of taxonomy development as well as the planned literature review. Our goal is to develop a comprehensive taxonomy for human subject evaluations in XAI. We seek to validate and iterate it through a structured literature review (SLR). Figure 2 illustrates our proposed methodology and the interplay between taxonomy and SLR.

3.1 Taxonomy Development

There are two approaches to constructing a taxonomy. Following the *conceptual-to-empirical* approach, the researcher proposes a classification based on a theory or model (deductive). In contrast, the *empirical-to-conceptual* approach derives the taxonomy from empirical cases (inductive). We follow the iterative process for taxonomy development proposed by Nickerson et al. [23]. Their method unifies both approaches in an iterative process under a shared *meta-characteristic* and defined *ending conditions*.

In line with RQ-3, we defined our meta-characteristic as the development of *a taxonomy for human subject evaluation of black-box*

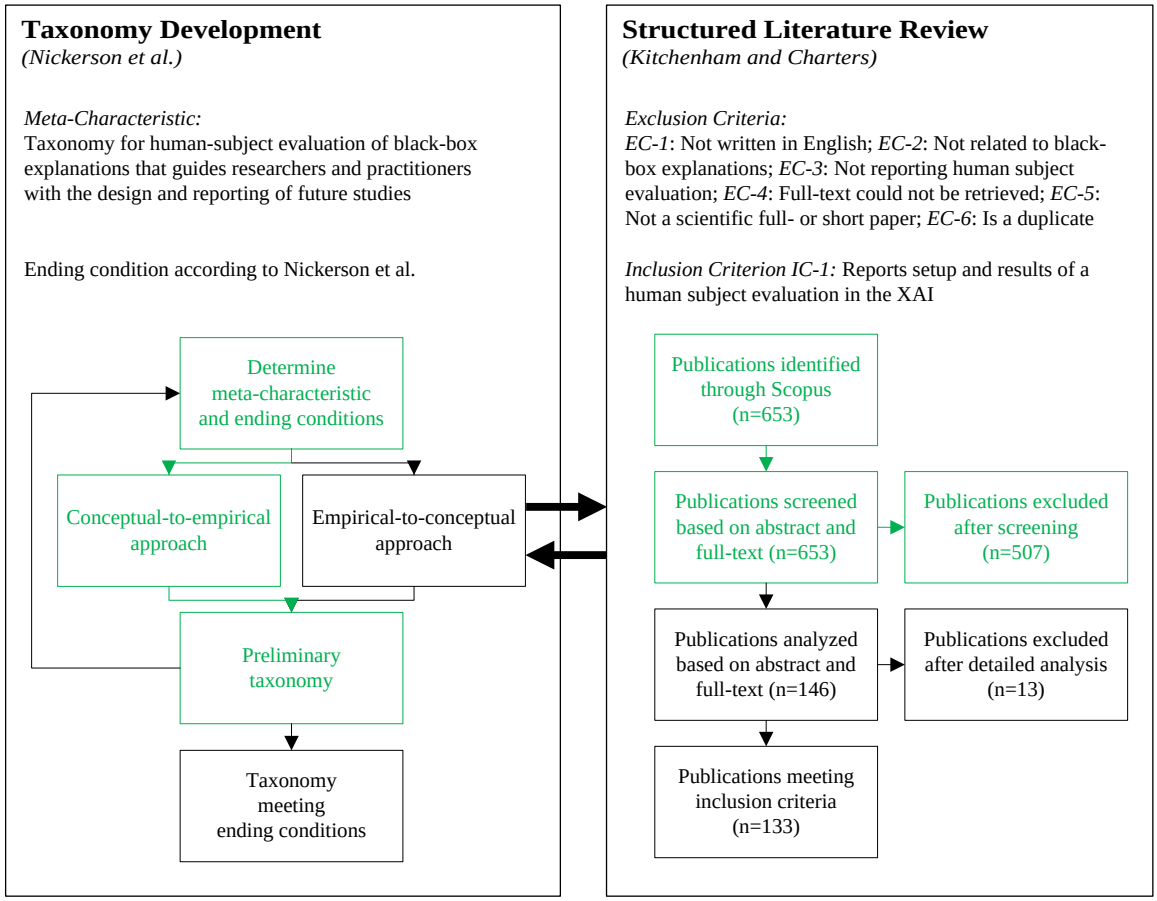

Figure 1: The proposed methodology for taxonomy development with an integrated structured literature review (SLR). Steps highlighted in green describe the preliminary results presented in this workshop paper.

explanations that guides researchers and practitioners with the design and reporting of future studies. We start by applying the *conceptual-to-empirical* approach. To follow this approach, one needs to propose a classification based on a theory or model. We do this by consolidating proposed categories for XAI evaluation in prior work and connecting them with foundational literature on empirical studies. The resulting taxonomy describes an ideal type, which allows us to examine empirically how much current human subject evaluations deviate from an ideal type.

3.2 Structured Literature Review

As part of the *empirical-to-conceptual* iteration, we aim to validate and iterate the taxonomy using a structured literature review (SLR). In line with RQ-2, the review's objective is to capture how researchers currently evaluate XAI methods and systems with human subjects. Through this, we seek to find out how structured and precise we can describe the field using our taxonomy. During this process, we also aim to iterate the taxonomy. The planned SLR follows established approaches proposed by Kitchenham and Charters [13]. In the following, we outline the proposed search strategy.

Source Selection: An exploratory search for XAI on Google Scholar indicated that relevant work is dispersed across multiple publishers, conferences, and journals. Thus, we use the Scopus database as a source as it integrates publications from relevant publishers such as ACM, IEEE, and AAAI.

Search Query: Through our exploratory search, we obtained an initial understanding of relevant keywords, synonyms, and related concepts that helped us to construct a search query. We found that different terms are used between the disciplines to describe the field of XAI and human subject evaluation approaches. Early research does not explicitly state the expressions *XAI* nor *explainable artificial intelligence*. Thus, our search queries are composed of *groups* and *terms*. Groups refer to a specific aspect of the research question and limit the search scope. Terms have a similar semantic meaning within the group domain or are often used interchangeably. We are interested in the intersection of 3 groups that can be phrased using different terms. Table 1 shows our used groups and terms.

Study Selection Criteria: We filtered the search results by six exclusion criteria (EC) and one inclusion criterion (IC). We are interested in primary studies that report the setup and result of human subject

Table 1: Groups and terms used for search query

Group	Terms
1 - Explainable	explainability, explainable, explanation, explanatory, interpretability, interpretable, intelligibility, intelligible, scrutability, scrutable, justification
2 - AI	XAI, AI, artificial intelligence, machine learning, black-box, recommender system, intelligent system, expert system, intelligent agent, decision support system
3 - Human Subject Evaluation	user study, lab study, empirical study, online experiment, human experiment, human evaluation, user evaluation, participant, within-subject, between-subject, probe, crowdsourcing, Mechanical Turk

evaluations in the XAI context (IC-1). We limit the survey to publications addressing the *black-box explanation problem*, according to Guidotti et al. [9] (EC-2). Furthermore, we exclude publications that do not report *human-grounded* or *application-grounded* evaluations according to Doshi-Velez and Kim [4] (EC-3). We applied the exclusion criteria in cascading order, i.e., if we excluded publications due to one EC, we did not assess any following criteria.

Study Analysis: So far, we conducted the search procedure for Scopus in September 2019, which returned a total of 653 potentially relevant publications. Both authors filtered the returned publications by the inclusions and exclusion criteria to control for inter-rater effects. We discussed differing assessments until we reached consensus. We are currently in the process of analyzing the publications that met the inclusion criterion.

4 TAXONOMY OF HUMAN SUBJECT EVALUATION IN XAI

In the following section, we describe relevant dimensions of black-box explanation evaluation with human subjects. We group identified characteristics into *task-related*, *participant-related*, and *study design-related* dimensions. The outlined taxonomy is a preliminary result after the first iterations of the conceptual-to-empirical approach based on propositions in prior work. Furthermore, the taxonomy was validated and refined based on a small subset consisting of 34 publications from the structured literature review following the empirical-to-conceptual approach.

4.1 Task Dimensions

Mohseni and Ragan distinguish two **types of human involvement** in the evaluation of explanations [18]. In the *feedback* setting, participants provide feedback on actual explanations. Experimenters determine the quality of the explanations through this feedback. In contrast, in the *feed-forward* setting no explanations are provided. Instead, humans are generating examples of reasonable explanations serving as a benchmark for algorithmic explanations.

Doshi-Velez and Kim distinguish two types of human subject evaluations that differ in their **level of task abstraction** [4]: *Application-grounded* evaluations conduct experiments within a real application context. Typically, this requires a high level of participant expertise. The quality of the explanation is assessed in measures of the application context, typically with a test of performance. *Human-grounded* evaluations conduct simplified or abstracted experiments that aim to maintain the essence of the target application.

Multiple **types of user tasks** have been proposed to elicit the quality of explanations [4, 18, 33]. We suggest distinguishing them by the information provided to the participant and the information inquired in return. In *verification* tasks, participants are provided with input, explanation, and output and asked for their satisfaction with the explanation. *Forced choice* tasks extend this setting. Here, participants are asked to choose from multiple competing explanations. In the case of *forward simulation* tasks, participants are presented with inputs as well as explanations and need to predict the system's output. *Counterfactual simulation* tasks, present participants with an input, an explanation, an output, and an alternative output (the counterfactual). Based on these, they predict what input changes are necessary to obtain the alternative output. In *"Clever Hans" detection* tasks, participants need to identify and possibly debug flawed models, e.g., a naive or short-sighted predictor [14]. *System usage* tasks are characterized by participants using the system and its explanations for its primary purpose, e.g., a decision-making situation. The quality of the explanation is assessed in terms of decision quality. In *annotation* tasks, participants provide a suitable explanation given input and output of a model.

Explanations are provided to users with very different goals in mind. For their effective evaluation, researchers need to ensure that the **intended explanation goal(s)** are aligned with their intended evaluation goal(s), and vice versa. Also, calibration of the individual goals of participants with the intended explanation goal(s) might be necessary (e.g., through a briefing before the task) [31]. We distinguish 9 common explanation goals, which are derived from [24, 30, 32]: *transparency* aims to explain how the system works, *scrutability* aims to allow users to tell the system it is wrong, *trust* aims to increase the user's confidence in the system, *persuasiveness* aims to convince the user to perform an action, *satisfaction* aim to increase the ease of use or enjoyment, *effectiveness* aims to help users make good decisions, *efficiency* aims to make decisions faster, *education* aims to enable users to generalize and learn, *debugging* aims to enable users to identify defects in the system. In the case of multiple intended explanation goals, their dependencies may be complementary, contradictory, or even unknown (e.g., the impact of transparency on trust).

Hoffman et al. describe multiple **levels of task evaluation** to assess a participant's understanding of and XAI system. Furthermore, they discuss suitable metrics for each level [11]. *Tests of satisfaction* measure participants' self-reported satisfaction with an explanation and their perception of system understanding. On this level, researchers can rarely be sure whether participants understand the system to the degree that participants claim. *Tests of comprehension* assess the participants' mental models of the system and tests their understanding, for example, through prediction tests and generative exercises. *Tests of performance* measure the resulting human-XAI system performance.

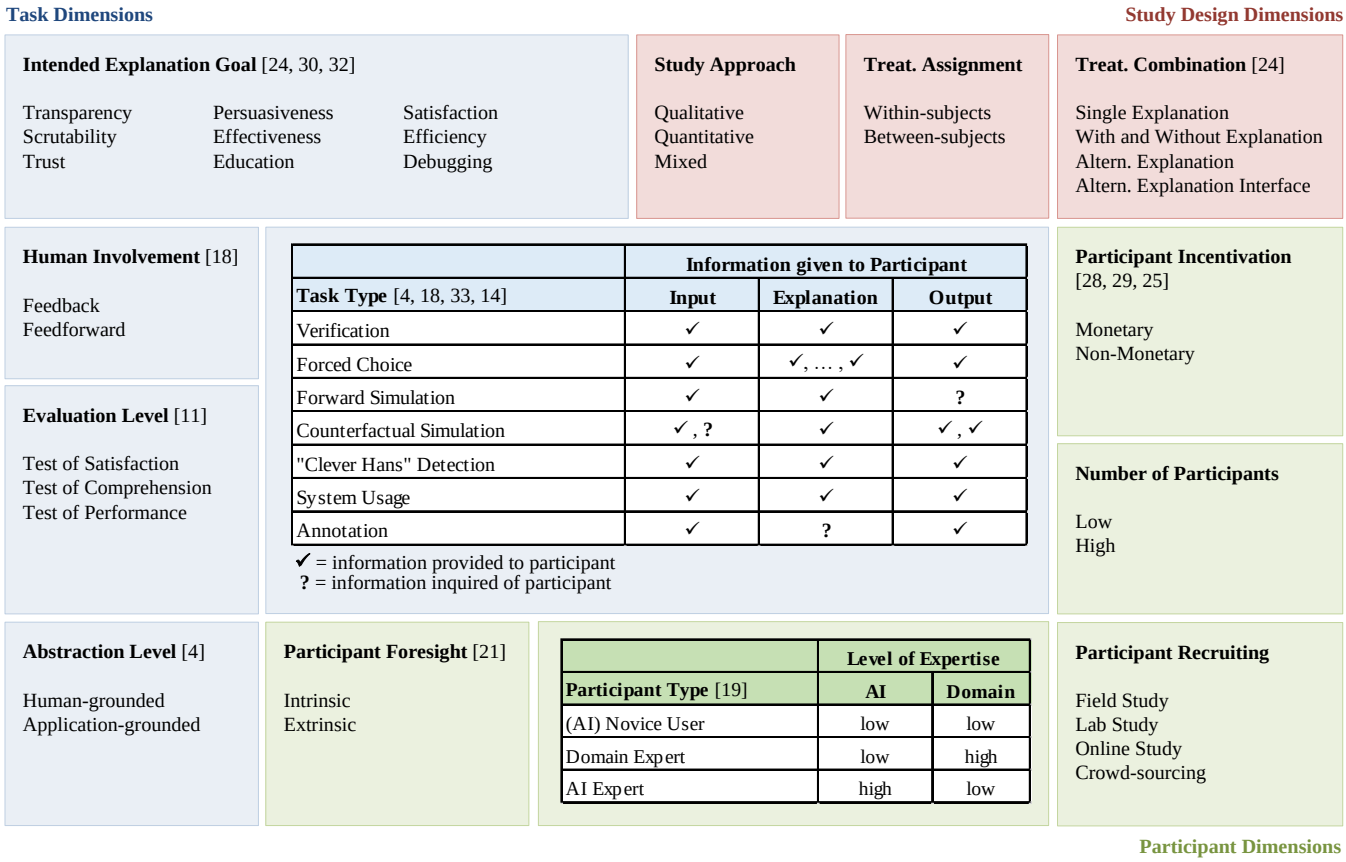

Figure 2: Preliminary taxonomy of human subject evaluation in XAI based on the conceptual-to-empirical approach.

4.2 Participant Dimensions

Mohseni et al. distinguish between several **participant types**: *AI novices* who are usually end-users, data experts (including *domain experts*), and *AI experts* [19]. This distinction is important as user expertise strongly influences other participant-related dimensions. For example, Doshi-Velez and Kim [4], referencing the work of Neath and Surprenant [22], point out that user expertise determines what kind of cognitive chunks participants apply to a situation. The expertise of participants may determine the **recruiting method** and **number of participants**. Recruiting difficulty is likely to increase with the required level of participants' expertise [4]. One can recruit novices in large numbers via *crowd-sourcing*. In contrast, domain or AI experts are usually harder to identify and recruit. They are often invited to a targeted *online study*, a *lab study*, or a *field study*. According to Narayana et al., the user study task may have dependencies with the **level of participant foresight** [21]. In an *intrinsic* setting, the participant's understanding of the context is solely based on the provided information. Thus, all participants are assumed to have equal knowledge about the context. Such types of experiments are usually suitable for novices. In an *extrinsic* setting, participants can additionally draw upon external facts, such as prior experience, that may be relevant for assessing the quality of an explanation, e.g., for spotting model flaws. Such a setting may be more suitable for data experts. However, it also makes controlling for participants' knowledge more difficult.

Incentivization of participants is another relevant dimension. According to Sova and Nielsen, it should be chosen considering study length, task demand, and participant expertise [28]. Stadtmüller and Porst advise us to use a *monetary incentive* for participants [29]. However, several *non-monetary incentives* are known to be effective as well (e.g., gifts for already paid employees) [25, 28]. Prost and Briel found that participants may take part in a study because of study-related incentives (e.g., curiosity, sympathy, or entertainment), personal-incentive (e.g., professional interest or a promise made), or altruistic reasons (e.g., to benefit science, society, or others) [25]. Esser argues that researchers should consider incentives in their combination such that the benefits of participating out-weigh the perceived cost [6].

4.3 Study Design Dimensions

The study design of evaluations may follow a *qualitative*, *quantitative*, or *mixed* **study approach**. In experimental studies, experimenters assign treatments to groups of participants. Applied to the context of explanation evaluations, we can distinguish four common types of **treatments combinations** in line with Nunes and Jannach [24]: *single treatment* (i.e., no alternative treatment), *with*

and without explanation (i.e., no explanation is alternative treatment), *alternative explanation* (i.e., varying information provided in explanations between treatments with other aspects of user interface fixed), *alternative explanation interface* (i.e., varying user interfaces between treatments). Furthermore, we can distinguish study designs by the **treatment assignment**: *Between-subjects designs* study the differences in understanding between groups of participants, each usually assigned to one treatment. In contrast, *within-subject designs* study differences within individual participants who are assigned to multiple treatments.

5 LIMITATION AND FUTURE WORK

Our preliminary taxonomy has limitations. The taxonomy is neither collectively exhaustive nor mutually exclusive. Thus, it does not meet the ending conditions of taxonomy development [23]. We aim to refine and iterate the taxonomy with the results from the proposed structured literature review.

Furthermore, human subject evaluations in XAI are typically embedded in a broader context, which may create dependencies and limit applicable evaluation approaches. Dependencies may arise from the explanation design context, such as the form of an explanation, its contents, or its underlying generation method. Multiple taxonomies have been developed for guiding the design of explanations [7, 24]. Nunes and Jannach proposed an elaborate explanation design taxonomy [24]. However, their taxonomy omits aspects of evaluation. For now, we have abstained from relating our preliminary human subject evaluation taxonomy with this prior work, but plan to integrate them in later iterations.

6 CONCLUSION

In this work, we gave a brief overview of recent efforts on explanation evaluation with human subjects in the growing field of XAI. We proposed a methodology for developing a comprehensive taxonomy for human subject evaluation that integrates the knowledge from multiple disciplines involved in XAI. Based on ideas from prior work, we presented a preliminary taxonomy following the conceptual-to-empirical approach. Despite its limitations, we believe our work is a starting point for rigorously evaluating the utility of explanations for human understanding of XAI systems. Researchers and practitioners developing XAI explanation facilities and systems have been asked to *"respect the time and effort involved to do such evaluations"* [4]. We aim to spark a discussion at the workshop on how to support them along the way.

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* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

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

[3] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[4] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretability. *CoRR* abs/1702.08608 (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[5] F. K. Dosilovic, M. Brcic, and N. Hlupic. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

[6] Hartmut Esser. 1986. Über die Teilnahme an Befragungen. *ZUMA Nachrichten* 10, 18 (1986), 38–47.

[7] Gerhard Friedrich and Markus Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine* 32, 3 (Jun. 2011), 90–98. https://doi.org/10.1609/aimag.v32i3.2365

[8] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (aug 2018). https://doi.org/10.1145/3236009

[10] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (Jun. 2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

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

[12] Frank C. Keil. 2006. Explanation and Understanding. *Annual Review of Psychology* 57, 1 (2006), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100 arXiv:https://doi.org/10.1146/annurev.psych.57.102904.190100 PMID: 16318595.

[13] B. Kitchenham and S Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Keele University and University of Durham, Technical Report EBSE-2007-01* (2007).

[14] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10, 1 (2019), 1096.

[15] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages. https://doi.org/10.1145/3236386.3241340

[16] 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

[17] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*. http://people.eng.unimelb.edu.au/tmiller/pubs/explanation-inmates.pdf

[18] Sina Mohseni and Eric D. Ragan. 2018. A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning. *CoRR* abs/1801.05075 (2018). arXiv:1801.05075 http://arxiv.org/abs/1801.05075

[19] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *CoRR* abs/1811.11839 (2018). arXiv:1811.11839 http://arxiv.org/abs/1811.11839

[20] Shane T. Mueller, Robert R. Hoffman, William J. Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *CoRR* abs/1902.01876 (2019). arXiv:1902.01876 http://arxiv.org/abs/1902.01876

[21] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1802.00682 (2018). arXiv:1802.00682 http://arxiv.org/abs/1802.00682

[22] Ian Neath and Aimee Surprenant. 2002. *Human Memory* (2 edition ed.). Thomson/Wadsworth, Australia ; Belmont, CA.

[23] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3 (2013), 336–359. https://doi.org/10.1057/ejis.2012.26 arXiv:https://doi.org/10.1057/ejis.2012.26

[24] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[25] Rolf Porst and Christa von Briel. 1995. *Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen? Oder: Gründe für die Teilnahme an Panelbefragungen*. Vol. 1995/04.

[26] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and Measuring Model Interpretability. *CoRR* abs/1802.07810 (2018). arXiv:1802.07810 http://arxiv.org/abs/1802.07810

[27] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[28] Deborah Hinderer Sova and Jacob Nielsen. 2003. How to Recruit Participants for Usability Studies. https://www.nngroup.com/reports/how-to-recruit-participants-usability-studies/, accessed December 20th, 2019.

[29] Sven Stadtmüller and Rolf Porst. 2005. *Zum Einsatz von Incentives bei postalischen Befragungen*. Vol. 14.

[30] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW '07)*. IEEE Computer Society, Washington, DC, USA, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[31] Nadya Vasilyeva, Daniel A Wilkenfeld, and Tania Lombrozo. 2015. Goals Affect the Perceived Quality of Explanations.. In *CogSci*.

[32] 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*. ACM, 601.

[33] Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. *CoRR* abs/1907.06831 (2019). arXiv:1907.06831 http://arxiv.org/abs/1907.06831