Position: Explainable AI Cannot Advance Without Better User Studies

Matej Pičulin [1] **Bernarda Petek** [1] **Irena Ograjenšek** [1] **Erik Štrumbelj** [1]

Abstract

In this position paper, we argue that user studies are key to understanding the value of explainable AI methods, because the end goal of explainable AI is to satisfy societal desiderata. We also argue that the current state of user studies is detrimental to the advancement of the field. We support this argument with a review of general and explainable AI-specific challenges, as well as an analysis of 607 explainable AI papers featuring user studies. We demonstrate how most user studies lack reproducibility, discussion of limitations, comparison with a baseline, or placebo explanations and are of low fidelity to real-world users and application context. This, combined with an overreliance on functional evaluation, results in a lack of understanding of the value explainable AI methods, which hinders the progress of the field. To address this issue, we call for higher methodological standards for user studies, greater appreciation of high-quality user studies in the AI community, and reduced reliance on functional evaluation.

1. Introduction

Explainable artificial intelligence (XAI) plays an important role in artificial intelligence (AI), machine learning (ML), and other areas of quantitative data analysis. As these technologies become more integral to our professional and personal lives and face greater legislative oversight, the importance of XAI continues to grow.

Academic interest in XAI is also growing rapidly. On Dec 29, 2024 there were, according to Scopus, 14077 works with at least one of the terms *explainable AI*, *XAI*, or *explainable artificial intelligence* in the title, abstract, or keywords. More than half of these (8,945) are dated 2024. There are also numerous survey papers that look at XAI from different angles and domains, including a survey of surveys by Schwalbe & Finzel (2023).

We take a closer look at human subject-based evaluation of XAI - evaluating XAI by studying the performance, behavior, and opinions of human subjects (*user studies* for short). In particular, we focus on the quality and scope of user studies and results that have implications for future user studies and the field of XAI as a whole. We do not discuss parts of XAI that are extraneous to user studies and their role in evaluating XAI. For readers interested in these, we recommend Schwalbe & Finzel (2023) as a starting point.

We aim to establish two points:

- **User studies are key to evaluating XAI.** The user is an integral part of XAI and if our goal is to understand the value of XAI methods, there is no alternative to user studies.
- **The current state of user studies in XAI is poor.** There is a clear lack of quality in all aspects of user study design, from defining the purpose of the study and participant selection, to study methodology and task design. In fact, most user studies in XAI are below the threshold of what is acceptable in fields of science with a longer history of human subject-based research. Also, most user studies are conducted with very low fidelity to real-world use.

If we accept both that user studies are key to understanding the value of XAI methods and that the state of user studies is poor, then it follows that our current understanding of the value of XAI methods is poor. And the only way to move forward is to substantially improve at least the quality, but hopefully also the quantity of user studies.

It is our position that we must change our mindset about user studies. Well-designed user studies should be encouraged, because at this point their contribution to XAI is more important than most further theoretical or non user-centric empirical advancement. On the other hand, user studies for the sake of doing a user study or as a methodological afterthought, should be discouraged, as they contribute little to our understanding of XAI.

[1]Faculty of Computer and Information Science, University of Ljubljana, Večna Pot 113, Ljubljana, Slovenia. Correspondence to: Erik Štrumbelj <erik.strumbelj@fri.uni-lj.si>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1.1. Related Work

Evaluating XAI is an active area of research, with several works in part or fully dedicated to user studies. We draw heavily from the seminal work of Doshi-Velez & Kim (2017; 2018), the comprehensive survey by Nauta et al. (2023), and the taxonomies and classifications of evaluation and user studies in XAI (Chromik & Schuessler, 2020; Herm et al., 2022; Lai et al., 2023; Lopes et al., 2022; Rong et al., 2023).

Unlike related work, we focus on the quality of user studies, their fidelity to real-world use, and the implications for our understanding of XAI. Our main contribution is our position, supported by an analysis and synthesis of related work and an analysis of 607 XAI papers that feature a user study.

2. The Role of User Studies in Evaluating XAI

We adopt the perspective of Speith & Langer (2023) that all XAI eventually aims to satisfy societal desiderata, such as trust, fairness, or downstream task performance. Speith & Langer (2023) propose the following model: XAI provides explanatory information, which facilitates understanding, which in turn affects how well the societal desiderata are satisfied. Note that Lopes et al. (2022) also propose a similar but more detailed model of human-centered evaluation.

The satisfaction of societal desiderata provides feedback on the appropriateness of the explainability approach. Speith & Langer (2023) divide evaluation methods into three categories: *explanatory information*, *understanding*, and *desiderata* evaluation methods. Explanatory information methods are concerned with how accurately XAI describe the AI system (for example, fidelity and completeness). Understanding methods are concerned with how well the XAI facilitates understanding of the AI system (for example, subjective understanding or being able to predict model behavior).

Both explanatory information and understanding evaluation methods allow for evaluation without human subjects (or *functional evaluation*, as it is referred to in the popular taxonomy by Doshi-Velez & Kim (2017; 2018)). However, evaluating societal desiderata requires human subjects.

The view that satisfying societal desiderata is the end goal and that explanatory information and understanding are just proxies, is generally accepted: Adadi & Berrada (2018) state four reasons for XAI: to justify, to control, to improve, and to discover. Doshi-Velez & Kim (2017; 2018) state that interpretability is often used as a proxy for other criteria, such as fairness, safety, and trust. They also raise the question of downstream goals of interpretable ML systems and why interpretability is the right tool for achieving those goals. (Vilone & Longo, 2021) state that the construct of explainability is linked with other constructs such as trust, transparency, and privacy. (Lipton, 2018) state for post-hoc interpretability, that work in this field should fix a clear objective and demonstrate evidence that the offered form of interpretation achieves it.

If societal desiderata are the end goal and cannot be evaluated without humans, then user studies are an important part of XAI evaluation. Or, as we and others argue, they are essential (Buçinca et al., 2020; Doshi-Velez & Kim, 2017; 2018; Vilone & Longo, 2021; Zhou et al., 2021).

2.1. Misalignment Between Academia and Practice

Poorly designed user studies, or the absence of user studies altogether, result in a lack of understanding of what works and what does not work in practice. This understanding is also important for guiding theoretical and methodological developments. Without it, there is likely to be a misalignment between academia and practice.

Indeed, this is not only our view, but also an increasingly common theme in related work. Bhatt et al. (2020) emphasize that there is a gap between XAI research and what is needed in practice. Decker et al. (2023) state that the academic XAI toolbox is not fully utilized in practice and practitioners call for tools that do not yet exist. Ghassemi et al. (2021) argue that current XAI methods are unlikely to achieve transparency and mitigate bias in healthcare. Kong et al. (2024) state that human-centered XAI may still lack explicit guidance of methods developing explainability solutions for different stakeholders. Preece et al. (2018) argue that failure to satisfy users of AI technology in the long run will be the most likely cause of another AI Winter. Lai et al. (2023) argue that the focus and design of studies may not align with how AI is or will be used in real-world decision-support applications. And Lopes et al. (2022) state that there is still a clear disconnect between technical XAI approaches and their effectiveness in supporting users' objectives.

3. Challenges in XAI User Studies

In this section we survey and discuss the general sentiment and specific issues that are relevant for designing user studies in XAI.

Several authors point to *a lack of formalism and consensus in terminology* (Jung et al., 2023; Lai et al., 2023; Lopes et al., 2022; Markus et al., 2021; Zhou et al., 2021) and call for *more standardized evaluation and reporting methodologies* (Sperrle et al., 2021). This can be attributed, at least in part, to XAI being a young field. However, standardized user study methodologies can be drawn from fields with longer traditions in such research, such as psychology and HCI. Moreover, there have been developments in formal frameworks for explanation (Adolfi et al., 2025; Bassan et al., 2024; Barceló et al., 2020; Vilas et al., 2024).

Several authors call for *more and better user studies* (Buçinca et al., 2020; Gurrapu et al., 2023; Jacovi et al., 2021; Johs et al., 2022; Keane et al., 2021; Zhou et al., 2021).There is a consensus that *evaluating XAI is an inderdisciplinary effort* (Lopes et al., 2022; Zhou et al., 2021) and that there is a lack of multidisciplinarity in papers (Lopes et al., 2022). Authors argue that the AI/ML community should draw from human-computer interaction (HCI) (Alangari et al., 2023; Lai et al., 2023; Liao & Vaughan, 2024; Williams, 2021), the Human-Human trust community (Vereschak et al., 2021), or social and behavioral sciences (Alangari et al., 2023; Johs et al., 2022; Miller et al., 2017; Miller, 2019). Some authors go further and argue that the HCI community should be the driving force (Chromik & Schuessler, 2020; Vilone & Longo, 2021).

Interactive explanations are also starting to receive more attention. See Bertrand et al. (2023) for a review, Boukhelifa et al. (2018) for evaluation of interactive machine learning systems, and Chromik & Butz (2021) for a review of interactive explanation user interfaces. Nguyen et al. (2024) argue that interactivity is key for real explainability and actionable understanding. Abdul et al. (2018) and Williams (2021) argue that interactive explanations should be explored further. However, most XAI methods and even more user studies, are static (Abdul et al., 2018).

Several authors point to *the importance of taking into account user's mental models and cognitive processes* (Hoffman et al., 2018; Kenny et al., 2021; Lopes et al., 2022; Rong et al., 2023), but there are still few user studies that do so.

3.1. A Diverse Range of Stakeholders

Applications of XAI involve a diverse range of stakeholders, including developers, researchers, end-users, decision-makers, regulators, educators, and policymakers. However, much of academia's focus is on the AI/ML practitioner and iterating between model development and evaluation.

As a result, key stakeholders, particularly end-users, are underrepresented in the literature. The study design choices in current research often fail to align with real-world decision-support applications. For example, tasks based on readily available datasets may not reflect realistic decision-making scenarios. Healthcare might be an exception, as many user studies focus on end-users (Jung et al., 2023).

Ideally, as several authors have also pointed out, proper evaluation of XAI would involve identifying and engaging all stakeholders while understanding the role of XAI in the context of use, domain, and end-users' expertise (Bhatt et al., 2020; Decker et al., 2023; Kong et al., 2024; Lai et al., 2023; Langer et al., 2021; Lopes et al., 2022; Nguyen & Zhu, 2022; Preece et al., 2018).

3.2. Personalized XAI

There is growing evidence that personal characteristics influence how people perceive and interact with XAI. Buçinca et al. (2021) suggest that human cognitive motivation moderates the effectiveness of explainable AI solutions. Reeder et al. (2023) find differences in trust and understanding based on gender and educational background. Millecamp et al. (2019) find that personal characteristics have significant influence in recommender systems, and that this influence is moderated by explanations.

Subsequently, more and more authors advocate for the use of user-centered methods, developing XAI with the end-user in mind, and tailoring XAI to different end-users (Rong et al., 2023; Ribera & Lapedriza, 2019; González-Alday et al., 2023). However, few user studies explore inter-personal differences. Anjomshoae et al. (2019) also find in their survey of explainable agents that only a few works addressed the issues of personalization and context-awareness.

3.3. Links between Functional, Perceived, Proxy, and Real-World Results

Ideally, a user study would evaluate the target XAI method in a real-world context, but such studies are the most challenging to conduct. Evaluation becomes easier with proxy tasks and even easier when relying solely on self-reported percieved quality. Evaluation without users (functional evaluation) is the simplest, but least impactful.

However, validating an easier evaluation method in a more complex context allows us to retain the benefits of simplicity without sacrificing impact. In this section, we summarize empirical findings on such relationships.

Notably, we found no studies exploring the relationship between functional evaluation and user performance. However, we did find several studies that explore other types of relationships:

Hase & Bansal (2020) found that subjective user ratings of explanation quality are not predictive of explanation effectiveness in simulation tests.

Amarasinghe et al. (2024) show the importance of closely reflecting the deployment context, by demonstrating that there is no practical utility of explanations. They also find a mismatch between self-reported metrics and improvement in decision-making.

Buçinca et al. (2020) performed three experiments to compare using proxy tasks and using subjective measures of trust and preference as predictors of actual performance. They found that proxy tasks did not predict the results of the actual decision-making tasks. They also found that subjective measures did not predict objective performance.

Chromik et al. (2021) investigated how non-technical users form their mental models of global AI model behavior from local explanations and found that participants overestimated their understanding.

In summary, we know little about the relationships between these levels of evaluation. What we do know suggests that context plays a crucial role and that designing functional evaluation methods or proxy tasks to predict real-world performance may be challenging.

3.4. The Placebo Effect and Placebic Explanations

The placebo effect is well known in medicine, where the effect of a new treatment is compared to a placebo control, such as a sugar pill or saline injection. The goal is to measure the effect of the treatment itself, beyond the impact of its administration. This principle is also applied in other fields, such as psychology, sports, visualization research (Kosch et al., 2023), and marketing (Vaccaro et al., 2018).

Kosch et al. (2023) identified three types of HCI user studies:

- A *conventional user study* (compare a novel system with a baseline),
- a *placebo-controlled study* (compare a novel system with a system that pretends to have a novel functionality), and
- a *placebo study* (compare a baseline system to a system that pretends to have a novel functionality).

Most XAI user studies are conventional, where the baseline is an existing explanation or no explanation. The positive results of such studies are questionable, as they could stem from a novel explanation or the placebo effect. This is particularly problematic when measuring trust, perceived performance, or other subjective aspects.

Recent research has highlighted this issue and advocates for placebo-controlled studies in XAI (Eiband et al., 2019; Bosch et al., 2024). However, conducting such studies in XAI requires placebic explanations. Unlike in medicine, where the treatment can be easily decoupled from its content, this is a greater challenge in XAI.

A few examples of placebic explanations include Liu (2021), who used tautological statements, which work well with textual explanations, but this cannot be generalized since XAI explanations take many forms. Textual variations were also used by Pias et al. (2024) and Eiband et al. (2019), while Wang & Ding (2024) created placebic explanations for feature importance by randomly shuffling contributions. Similarly, (Kenny et al., 2023) created placebic explanations in reinforcement learning by randomly perturbing prototype images, thus disassociating them with intuitive actions.

Some placebo studies use *no explanation* as a baseline. For example, Eiband et al. (2019) conducted a no explanation/-placebo/real explanation study. They found that placebo explanations invoke levels of perceived trust similar to real explanations. Kosch et al. (2023) and Villa et al. (2023) also show that subjective measurements improve, while objective measurements remain unchanged when the system is described as having AI. Kloft et al. (2024) suggests this effect is not only due to verbal descriptions but also the socio-technical context. Note that these studies used sham systems and did not consider objective metrics like accuracy or time. We did not find an XAI user study presenting a novel explanation with a placebo-control group.

3.5. Concerns with Crowdsourcing Platforms

Crowdsourcing platforms (Amazon Mechanical Turk, Prolific, CloudResearch, etc.) have become very popular in recent years, primarily because of their convenience. XAI is no exception (see Section 4).

As the popularity of crowdsourcing increased, questions about the demographics and data quality of the crowdsourcing samples compared to other samples and the general population have emerged. The demographics and personalities of Amazon Mechanical Turk workers (MTurkers) consistently differ from other samples and the general population (Burnham et al., 2018; Douglas et al., 2023; Goodman & Paolacci, 2017; Paolacci & Chandler, 2014; Weigold & Weigold, 2022). For example, MTurkers are more educated than the general population and older MTurkers may have higher cognitive abilities than the corresponding age group in the general population (Ogletree & Katz, 2021).

MTurkers may be less attentive than students (Aruguete et al., 2019; Barends & De Vries, 2019; Goodman et al., 2013; Tahaei & Vaniea, 2022) when completing tasks. Inattentiveness can also manifest as inconsistencies in answers that don't make sense (Kay, 2024). MTurkers are very much prone to multitasking (Brigden, 2024; Necka et al., 2016). Moreover, recent studies suggest that they often engage in satisficing and low-effort behaviors, potentially compromising previously validated study results (Berry & Burton, 2024; Berry et al., 2024). Additionally, MTurkers can easily misrepresent themselves to qualify for specific studies, posing a significant concern for researchers who require participants with particular traits (Ahler et al., 2021; Dennis et al., 2020; MacInnis et al., 2020; Moss et al., 2021). Besides being dishonest, another issue with screening procedures that include subjective surveys is the potential overconfidence of MTurkers (Tahaei & Vaniea, 2022).

Moreover, a small sample of MTurkers tends to be highly

productive and consequently very familiar with the studies they participate in. This lack of naivety can impact data quality, especially if similar attention checks are frequently used (Chandler et al., 2014; 2015; Necka et al., 2016; Stewart et al., 2017). Another potential issue is the interaction among MTurkers on forums, which can influence the participant pool by preferring some researchers over others (Chandler et al., 2014). Additionally, high attrition rates could be a problem, as MTurk workers can leave studies with just a click (Arechar et al., 2018; Zhou & Fishbach, 2016).

Most importantly, with the growing population of MTurk, the quality of the data seems to have decreased in studies that have been recreated or reviewed over time (Chmielewski & Kucker, 2020; Kennedy et al., 2020; Marshall et al., 2023). Consequently, researchers are starting to focus on alternative crowdsourcing platforms and survey platforms that offer online sample and panel services. It looks like MTurk provides lower quality data compared to Prolific, CloudResearch (Connect), and Qualtrics (Qualtrics Panels) (Albert & Smilek, 2023; Douglas et al., 2023; Eyal et al., 2021; Peer et al., 2023). However, little research has been done yet on how these other platforms compare to other population samples or the general population. There may be some potential data quality issues with these platforms compared to student samples (Novielli et al., 2023; Tahaei & Vaniea, 2022).

These findings at best further emphasize the importance of carefully designing a crowdsourcing user study and at worst put into question results obtained via crowdsourcing.

4. The State of User Studies in XAI

We analyzed user studies from 607 XAI academic papers. Throughout, we grouped the papers into papers published *up to 2020*, papers published *2021 and after*, and papers published in *top 4* conference venues (NeurIPS, ICLR, ICML, AAAI) from 2020 and after. Every paper from the top 4 group is also in one of the former two groups. Note that some results are on all papers, while others use a random subsample from each group. Details of how we collected and analyzed the papers can be found in Appendix A.

4.1. Publication Year

The distribution of the papers over publication year is shown in Figure 1. The rising popularity of XAI in academic research is clear. The lower paper count in 2024 can be explained by the fact that most of the papers were collected in early 2024. However, we have most likely collected proportionately fewer papers in the most recent years, because the sample is biased towards papers published earlier (see Appendix A.3 for a discussion of the limitations).

A more thorough search for relevant papers in the top 4 venues doubled the number of papers for those venues in that period. From this we can estimate that the 607 papers represent at best one half of the total number of XAI papers with a user study in the 2020-2024 period (possibly more prior to 2020).

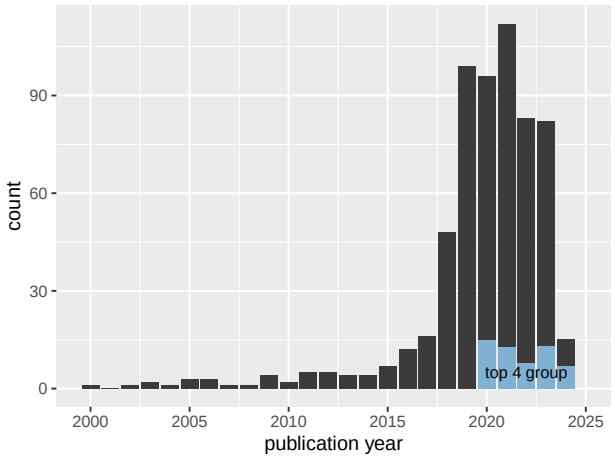

Figure 1. Distribution of papers over publication year. Papers from the top 4 group are highlighted.

4.2. Participant Count and Type

Figure 2 shows the distribution of user study participant count. The distribution is similar for all three groups. Most studies are under 50 participants, with a long tail of studies with more than 50 participants. These roughly correspond to non-crowdsourcing and crowdsourcing studies, respectively.

Figure 3 shows the distribution of user study participant type. The most common approach is to use a crowdsourcing platform, typically Amazon MTurk. The second most common approach is to recruit students, sometimes combined with university researchers and administrative staff. These approaches represent more than two thirds of all studies. Between 20% and 30% of studies are done on domain experts and probably less than 10% on AI experts. Studies from the top 4 group lean more towards crowdsourcing (or not stating particpant type at all) and less frequently include other types of participants, in particular, domain experts. A non-negligible proportion of studies (10% - 20%) do not state participant type.

4.3. Study Design and Quality

The estimates of the variables that measure the methodological quality of user studies are summarized in Figure 4. The standout result is that very few user studies from the top 4 group contain enough information to be considered

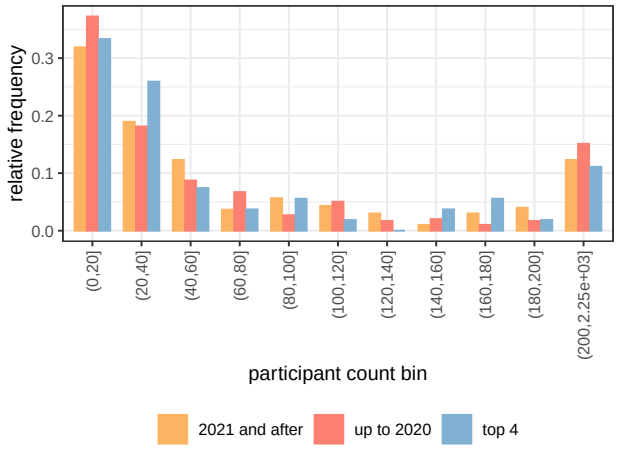

Figure 2. Distribution of user studies over participant count bins. Note that there were a total of 647 user studies in the 607 papers, 48 of which (7.4%) did not report participant count.

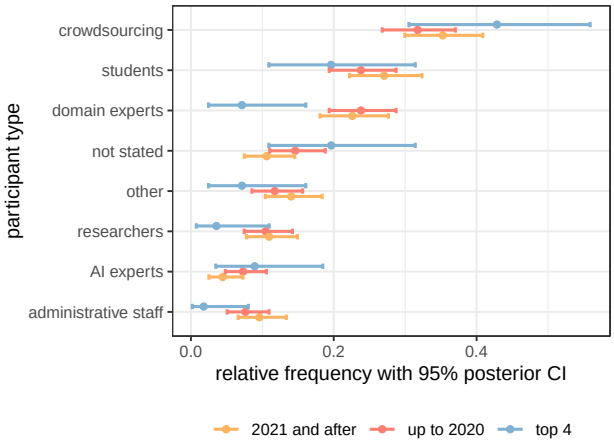

Figure 3. Relative frequencies user study participant type. Note that a user study can have multiple participant types, so a group's relative frequencies might not sum to 1.

reproducible. This is in stark contrast with the rest of the papers, where we estimate between 50% and 90% are reproducible. There also appears to be some improvement for the more recent papers in reproducibility, stating the limitations of the user study, pre-testing, and attention checks in crowdsourcing studies.

Other results are similar for all three groups. Preregistration and pretesting are rare. Validated questionnaires are used in less than half of the studies. Baselines for comparison are included in about one half of the studies, but placebo studies are rare. At least a minimal discussion of the study's limitations is included in about one half of the studies.

4.4. Evaluation Type and Fidelity

The results for the categorization of user studies by evaluation type are shown in Figure 5. Most studies rely on subjective satisfaction or subjective comparison of methods. However, the top 4 group user studies feature more forward simulatability and less subjective satisfaction. The results are similar to the results in Nauta et al. (2023) and the discrepancies can be explained by the difference between the two samples of papers and by our interpretation of subjective comparison (see Appendix A.2.1 for details). The results also align with the results of our additional categorization into objective or subjective evaluation (see Figure 6). While subjective evaluation is more common, both types are common and the difference is less in the top 4 group.

The fidelity and evaluation level results of our additional categorization are shown in Figure 7. Most studies evaluate the user's understanding of the AI system and do so on a toy application or absent an application context. The third most common category are medium fidelity user studies that evaluate downstream performance. There are very few high fidelity user studies. There are no discernible differences between user studies up to 2020 and 2021 and after. However, results suggest that user studies from the top 4 group are more frequently conducted without an application context and focus on understanding.

5. Alternative Views

Our view, that the state of user studies in XAI is relatively poor, is generally accepted and we provide further empirical evidence. Similarly, we do not believe that it is controversial to state that the field of ML should hold itself to high methodological standards when it comes to user studies (see Herrmann et al. (2024) for a similar sentiment regarding empirical research in ML in general). However, there is an alternative view that user studies are not as key for the development of XAI as we claim.

Arguments against user studies are always based on a comparison with the only alternative, which is evaluating XAI without users (or functional evaluation). The two major points of criticism are (a) that user studies are more difficult to do and (b) that they are biased.

It is undisputed that user studies are more difficult to do than functional evaluation. However, if a user study is appropriate for the task at hand and functional evaluation is at most a compromise (and potentially inappropriate), the difficulty of conducting a user study is by itself not a valid argument for using functional evaluation.

When considering this argument and other arguments that

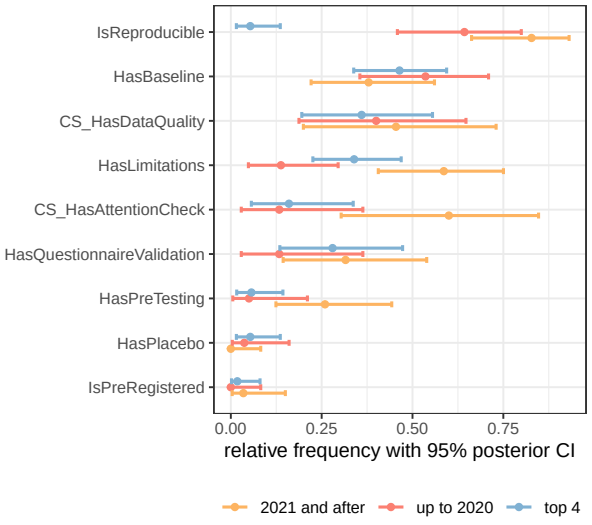

Figure 4. Relative frequencies user study properties. Note that these are based on random subsamples (with replacement) of size 30 for the up to 2020 and 2021 and after groups.

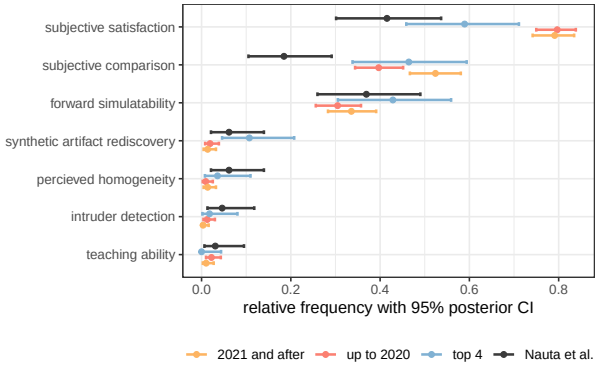

Figure 5. Relative frequencies of quantitative evaluation type categories. We include the results from (Nauta et al., 2023), who introduced the categories. Note that their sample were user studies from 12 top CS/ML/AI venues in the period from 2014 to 2020.

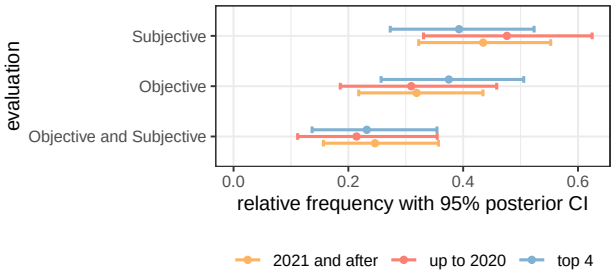

Figure 6. Relative frequencies of objective and subjective evaluations and 95% posterior CI. Note that these are based on random subsamples (with replacement) of size 30 for the up to 2020 and 2021 and after groups.

follow in this section, we should also take into account that there is little to no empirical evidence that validates functional evaluation metrics via downstream task performance. In fact, there is evidence to the contrary and evidence that even in user studies the performance of XAI depends strongly on context. Therefore, proxy tasks or any other deviation from the real-world context can result in misleading results (see Section 3.3 for details).

The other major point of criticism against user studies is bias (Alangari et al., 2023; Kadir et al., 2023). This is most commonly expressed as follows: functional evaluation is objective, while user studies are subjective. With the implied understanding that objective is better than subjective.

For example, Petsiuk et al. (2018) argue that keeping humans out of the evaluation makes it more fair and true to the classifier's own view of the problem, rather than representing a human's view. Rong et al. (2023) state that functional and human-subject based evaluation address two different things. One addresses the general objectivity independent of downstream tasks, while the other contextualize with specific use cases. Markus et al. (2021) write that although quantitative proxy metrics are necessary for an objective assessment of explanation quality, they should be complemented with human evaluation methods before employing AI systems in real-life. Due to bias, Kadir et al. (2023) call for a functional evaluation metric that can be experimentally validated.

Alangari et al. (2023) also argue that, as a consequence of the limitations of user studies, there has been a decline in the use of user studies, with functional evaluation gaining prominence as a more rigorous approach. According to Nauta et al. (2023) the proportion of user studies in all evaluation has remained relatively steady from 2016 to 2020 (around 20%). Our data would support the interpretation that the number of user studies has tapered off since, while the number of papers in XAI keeps growing. However, even if that is the case, we would rather attribute this to the fact that functional evaluation is easier to conduct and not to the limitations of user studies.

We argue that the bias inherent to user studies is by itself not a strong argument against user studies or in favor of functional evaluation. It overlooks the fact that functional evaluation also introduces a bias when used as a proxy for some downstream performance. If we agree that the end goal is to satisfy societal desiderata, any functional evaluation should be validated via user studies (and thus

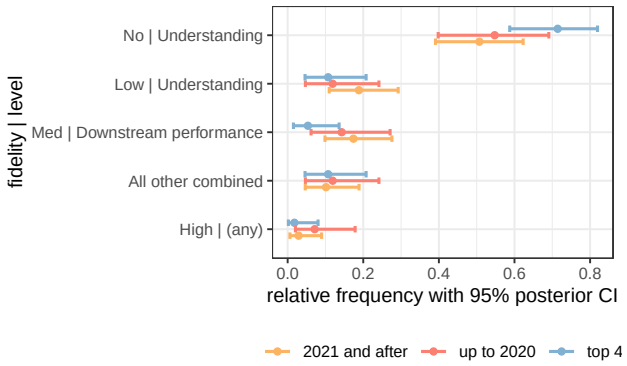

Figure 7. Relative frequencies of the most common combinations of fidelity and level and 95% posterior CI. Note that these are based on random subsamples (with replacement) of size 30 for the up to 2020 and 2021 and after groups.

subjectively). So, user studies, subjective or not, cannot be avoided in a field that is supposed to be focused on humans.

Functional evaluation also has a major limitation that has so far been overlooked. A functional evaluation method cannot be used to compare XAI with different types of outputs. For example, to compare feature importance with counterfactuals or a heatmap with a text-based explanation. Nauta et al. (2023, Table 2) categorize almost 100 example methods, all of which are tailored to specific outputs and cannot be used to compare different types of explanation methods. The underlying issue is that the output (that is, the presentation) of the XAI method is a fundamental part of the method itself, so we cannot standardize outputs without changing the method. Unlike, for example, the task of classification, where prediction outputs (class or probabilities) put all models on the same denominator and simplify comparison. The same issue makes it more difficult to derive placebic explanations.

As an alternative view, we could also adopt the position that XAI is still progressing, thereby circumventing the discussion about potential issues with functional evaluation and the necessity of user studies altogether. That is, if progress continues despite these issues, quality user studies cannot be essential to that progress.

If we measure progress by how much XAI methods are being used, XAI has definitely been progressing. We acknowledge that XAI has produced useful tools for ML practitioners, but would add that in the case of the ML practitioners, the researchers and developers are often the target users, so it can be argued that these methods are being developed close to the target user.

The use of XAI methods has been growing outside of ML as well, due to the popularity of feature contribution methods, in particular SHAP. However, it is not clear if this increase in popularity is due to the progress of XAI or due to the increase in the use of AI in general. And, unlike tabular predictive modeling, natural language processing, or computer vision, where recent developments are already being used in practice, that is not the case with XAI.

Even if we take the position that increased use of XAI methods implies that these methods are useful for users other than ML practitioners, it is not clear how we can reconcile this with growing evidence that the tools being developed are not what target users need (see Section 2.1). And even then we at a minimum have to acknowledge that we have little to no understanding of why, when, and for which type of target user they are useful.

Finally, note that our position does not reject functional evaluation. On the contrary, functional evaluation should play an important role in verifying whether a method meets minimal technical criteria, particularly during rapid prototyping and the early phases of development Miller et al. (2017). It could also serve as a less resource-intensive alternative to user studies, provided that we establish clear links between functional evaluation and the satisfaction of desiderata or performance on downstream tasks.

6. Conclusion

We share the view of Herrmann et al. (2024), who, in their position on empirical research in ML, call for more confirmatory research, comparison studies, replication studies, and meta-studies. And, that the field should *move from being largely driven by mathematical proofs and application improvements to also becoming a full-fledged empirical field driven by multiple types of experimental research*.

In this paper we focused on XAI and on user studies - one type of empirical research that we believe to be key for the advancement of XAI. Historically, the field of ML has relied on theoretical work and improvements over the state-of-the-art with respect to some abstract metric. And in most cases that has led the field very far. However, unlike, for example, supervised learning, where predictive performance is easily measured and plausibly translates to real-world utility, the same cannot be said for XAI. In XAI, the user is an integral component that cannot be easily circumvented, and any attempt to do so risks widening the gap between academic research and practical application.

User studies in XAI are poorly designed (as a whole, with some exceptions) and have (with a few exceptions, such as Evirgen & Chen (2022); Kiani et al. (2020); Kenny et al. (2024); Wong et al. (2024)) low or no fidelity to real-world use. As a result, at best, we know how a ML practitioner's understanding of a model improves with XAI (through self-

reporting and forward simulatability). At worst, we know very little about the practical application and real-world benefits of XAI.

The design and fidelity problem appears to be exacerbated in top ML venues, where it is very difficult to publish a paper that focuses solely on a user study, but at the same time more poorly-designed user studies get published as a part of theoretical or methodological papers. This at least exhibits a consistent view that user studies are not considered that important at these venues. However, we believe that user studies that clearly do not meet even the minimum methodological standards of reproducibility should be rejected, regardless of the focus of the venue. We also believe that methodological developments in XAI that rely exclusively on functional evaluation (as opposed to user studies or theoretical justification) should be subject to greater scrutiny, to deal with the field's overreliance on functional evaluation.

Our results also suggest that the level of relevant knowledge and know-how in user studies is relatively low in the ML community, not only in conducting but also in reviewing them. This is understandable, given the field's history and focus. Some may even argue that other fields, such as psychology and HCI, should take the lead in researching XAI in practice. In particular, researchers from those fields. However, we believe this would be a missed opportunity to elevate the quality of research within the ML community.

How can we encourage better user studies? Recently, venues have emerged to promote empirical research, such as the Journal of Data-centric Machine Learning Research (DMLR), the Datasets and Benchmarks Track at NeurIPS, and the Applied Data Science Track at ECML. A similar venue dedicated specifically to user-centric research in ML would be invaluable. Such a platform could not only foster user-centric research but also enhance the community's expertise in conducting and evaluating user studies. This, in turn, would improve the quality of reviews and elevate the standards of published user studies.

We conclude with a list of XAI topics that we believe are particularly important for the advancement of the field:

- Linking functional evaluation metrics, proxy tasks, and real-world performance. In particular, the development of standardized benchmarks.
- Generalizability of crowdsourced and student-based studies.
- Intra-user and inter-user differences.
- The placebo effect and the development of placebo-controlled studies.
- Neutral method comparison studies of popular XAI methods. We anticipate that the results in the field are overly optimistic, so we share the sentiment of Karl et al. (2024) that we should embrace negative results.

Acknowledgements

We sincerely thank the anonymous ICML reviewers for their insightful feedback, which helped us clarify our contributions and significantly improve the paper. This work was funded by the Slovenian Research Agency (research core funding No. P2-0442).

Impact Statement

This paper presents a position on the importance of user studies in the evaluation of explainable AI. Its goal is to spark discussion and raise the standards of user-centric research in explainable AI and in machine learning in general. There are other potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. 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*, pp. 1–18, 2018.

Adadi, A. and Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

Adolfi, F., Vilas, M. G., and Wareham, T. The Computational Complexity of Circuit Discovery for Inner Interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QogcGNXJVw.

Ahler, D. J., Roush, C. E., and Sood, G. The micro-task market for lemons: Data quality on Amazon's Mechanical Turk. *Political Science Research and Methods*, pp. 1–20, 2021.

Alangari, N., El Bachir Menai, M., Mathkour, H., and Almosallam, I. Exploring evaluation methods for interpretable machine learning: A survey. *Information*, 14(8): 469, 2023.

Albert, D. A. and Smilek, D. Comparing attentional disengagement between Prolific and MTurk samples. *Scientific Reports*, 13(1):20574, 2023.

Amarasinghe, K., Rodolfa, K. T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., Bizarro, P., Talwalkar, A., and Ghani, R. On the importance of application-grounded experimental design for evaluating explainable ML methods. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 20921–20929, 2024.

Anjomshoae, S., Najjar, A., Calvaresi, D., and Främling, K. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Arechar, A. A., Gächter, S., and Molleman, L. Conducting interactive experiments online. *Experimental economics*, 21:99–131, 2018.

Aruguete, M. S., Huynh, H., Browne, B. L., Jurs, B., Flint, E., and McCutcheon, L. E. How serious is the 'carelessness' problem on Mechanical Turk? *International Journal of Social Research Methodology*, 22(5):441–449, 2019.

Barceló, P., Monet, M., Pérez, J., and Subercaseaux, B. Model interpretability through the lens of computational complexity. *Advances in neural information processing systems*, 33:15487–15498, 2020.

Barends, A. J. and De Vries, R. E. Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and individual differences*, 143:84–89, 2019.

Bassan, S., Amir, G., and Katz, G. Local vs. Global Interpretability: A Computational Complexity Perspective. In *International Conference on Machine Learning*, pp. 3133–3167. PMLR, 2024.

Ben David, D., Resheff, Y. S., and Tron, T. Explainable AI and adoption of financial algorithmic advisors: an experimental study. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 390–400, 2021.

Berry, C. and Burton, S. Express: Response satisficing across online data sources: Effects of satisficing on data quality and policy-relevant results. *Journal of Public Policy & Marketing*, pp. 07439156241268707, 2024.

Berry, C., Kees, J., and Burton, S. Response satisficing and data quality in marketing: measurement and effects of satisficing on objective knowledge, experimental results, and replicability of findings. *Journal of Marketing Theory and Practice*, pp. 1–18, 2024.

Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., and Maxwell, W. On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2023.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 648–657, 2020.

Bosch, E., Welsch, R., Ayach, T., Katins, C., and Kosch, T. The illusion of performance: the effect of phantom display refresh rates on user expectations and reaction times. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2024.

Boukhelifa, N., Bezerianos, A., and Lutton, E. Evaluation of interactive machine learning systems. *Human and machine learning: visible, explainable, trustworthy and transparent*, pp. 341–360, 2018.

Brigden, N. Participant multitasking in online studies. *Marketing Letters*, pp. 1–13, 2024.

Browne, S. T., Pike, T. D., and Bailey, M. M. A proposed framework for artificial intelligence safety and technology readiness assessments for national security applications. Technical report, Center for Open Science, 2024.

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

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

Burnham, M. J., Le, Y. K., and Piedmont, R. L. Who is MTurk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion & Culture*, 21(9-10):934–944, 2018.

Chandler, J., Mueller, P., and Paolacci, G. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46:112–130, 2014.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., and Ratliff, K. A. Using nonnaive participants can reduce effect sizes. *Psychological science*, 26(7):1131–1139, 2015.

Chmielewski, M. and Kucker, S. C. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473, 2020.

Chromik, M. and Butz, A. Human-XAI interaction: a review and design principles for explanation user interfaces. In

Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18, pp. 619–640. Springer, 2021.

Chromik, M. and Schuessler, M. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. *Exss-atec@ iui*, 1:1–7, 2020.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., and Butz, A. I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 307–317, 2021.

Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., and Weber, S. H. The thousand faces of explainable AI along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction*, pp. 184–208. Springer, 2023.

Dennis, S. A., Goodson, B. M., and Pearson, C. A. Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, 32(1):119–134, 2020.

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

Doshi-Velez, F. and Kim, B. Considerations for evaluation and generalization in interpretable machine learning. *Explainable and interpretable models in computer vision and machine learning*, pp. 3–17, 2018.

Douglas, B. D., Ewell, P. J., and Brauer, M. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one*, 18(3):e0279720, 2023.

Eiband, M., Buschek, D., Kremer, A., and Hussmann, H. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pp. 1–6, 2019.

Evirgen, N. and Chen, X. Ganzilla: User-driven direction discovery in generative adversarial networks. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–10, 2022.

Eyal, P., David, R., Andrew, G., Zak, E., and Ekaterina, D. Data quality of platforms and panels for online behavioral research. *Behavior research methods*, pp. 1–20, 2021.

Ferreira, J. J. and Monteiro, M. S. 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: 9th International Conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pp. 56–73. Springer, 2020.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3 (11):e745–e750, 2021.

González-Alday, R., García-Cuesta, E., Kulikowski, C. A., and Maojo, V. A scoping review on the progress, applicability, and future of explainable artificial intelligence in medicine. *Applied Sciences*, 13(19):10778, 2023.

Goodman, J. K. and Paolacci, G. Crowdsourcing consumer research. *Journal of Consumer Research*, 44(1):196–210, 2017.

Goodman, J. K., Cryder, C. E., and Cheema, A. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., and Batarseh, F. A. Rationalization for explainable NLP: a survey. *Frontiers in Artificial Intelligence*, 6:1225093, 2023.

Handley, H. A., See, J. E., and Savage-Knepshield, P. A. Human readiness levels and human views as tools for user-centered design. *Systems Engineering*, 27(6):1089–1102, 2024.

Hase, P. and Bansal, M. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, 2020.

Herm, L.-V., Wanner, J., and Janiesch, C. A Taxonomy of User-centered Explainable AI Studies. In *Pacific Asia Conference on Information Systems (PACIS) 2022*, 2022.

Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Rügamer, D., Hüllermeier, E., Boulesteix, A.-L., and Bischl, B. Position: Why We Must Rethink Empirical Research in Machine Learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 18228–18247. PMLR, 2024.

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 624–635, 2021.

Johs, A. J., Agosto, D. E., and Weber, R. O. Explainable artificial intelligence and social science: Further insights for qualitative investigation. *Applied AI Letters*, 3(1):e64, 2022.

Jung, J., Lee, H., Jung, H., and Kim, H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 9(5), 2023.

Kadir, M. A., Mosavi, A., and Sonntag, D. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pp. 000111–000124. IEEE, 2023.

Karl, F., Kemeter, M., Dax, G., and Sierak, P. Position: Embracing Negative Results in Machine Learning. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23256–23265. PMLR, 21–27 Jul 2024.

Kay, C. S. Extraverted introverts, cautious risk-takers, and selfless narcissists: A demonstration of why you can't trust data collected on MTurk. 2024.

Keane, M. T., Kenny, E. M., Delaney, E., and Smyth, B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In Zhou, Z.-H. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4466–4474. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/609. URL https://doi.org/10.24963/ijcai.2021/609. Survey Track.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., and Winter, N. J. The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, 8(4):614–629, 2020.

Kenny, E. M., Ford, C., Quinn, M., and Keane, M. T. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294:103459, 2021.

Kenny, E. M., Tucker, M., and Shah, J. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2023.

Kenny, E. M., Dharmavaram, A., Lee, S. U., Phan-Minh, T., Rajesh, S., Hu, Y., Major, L., Tomov, M. S., and Shah, J. A. Explainable deep learning improves human mental models of self-driving cars. *arXiv preprint arXiv:2411.18714*, 2024.

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J., et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine*, 3(1):23, 2020.

Kloft, A. M., Welsch, R., Kosch, T., and Villa, S. "AI enhances our performance, I have no doubt this one will do the same": The Placebo effect is robust to negative descriptions of AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

Kong, X., Liu, S., and Zhu, L. Toward Human-centered XAI in Practice: A survey. *Machine Intelligence Research*, pp. 1–31, 2024.

Kosch, T., Welsch, R., Chuang, L., and Schmidt, A. The placebo effect of artificial intelligence in human–computer interaction. *ACM Transactions on Computer-Human Interaction*, 29(6):1–32, 2023.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., and Tan, C. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1369–1385, 2023.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473, 2021.

Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., Lange, D., Baydin, A. G., Sharma, A., Gibson, A., et al. Technology readiness levels for machine learning systems. *Nature Communications*, 13 (1):6039, 2022.

Liao, Q. V. and Varshney, K. R. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

Liao, Q. V. and Vaughan, J. W. AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, (Special Issue 5), 2024.

Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Liu, B. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of computer-mediated communication*, 26 (6):384–402, 2021.

Lopes, P., Silva, E., Braga, C., Oliveira, T., and Rosado, L. XAI systems evaluation: A review of human and computer-centred methods. *Applied Sciences*, 12(19): 9423, 2022.

MacInnis, C. C., Boss, H. C., and Bourdage, J. S. More evidence of participant misrepresentation on MTurk and investigating who misrepresents. *Personality and Individual Differences*, 152:109603, 2020.

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113:103655, 2021.

Marshall, C. C., Goguladinne, P. S., Maheshwari, M., Sathe, A., and Shipman, F. M. Who broke Amazon Mechanical Turk? An analysis of crowdsourcing data quality over time. In *Proceedings of the 15th ACM Web Science Conference 2023*, pp. 335–345, 2023.

Millecamp, M., Htun, N. N., Conati, C., and Verbert, K. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 397–407, 2019.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Miller, T., Howe, P., and Sonenberg, L. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

Mohseni, S., Zarei, N., and Ragan, E. D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

Moss, A. J., Rosenzweig, C., Jaffe, S. N., Gautam, R., Robinson, J., and Litman, L. Bots or inattentive humans? Identifying sources of low-quality data in online platforms. *PsyArXiv*, 2021.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Necka, E. A., Cacioppo, S., Norman, G. J., and Cacioppo, J. T. Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PloS one*, 11(6):e0157732, 2016.

Nguyen, T. and Zhu, J. Towards Better User Requirements: How to Involve Human Participants in XAI Research. *arXiv preprint arXiv:2212.03186*, 2022.

Nguyen, T., Canossa, A., and Zhu, J. How Human-Centered Explainable AI Interface Are Designed and Evaluated: A Systematic Survey. *arXiv preprint arXiv:2403.14496*, 2024.

Novielli, J., Kane, L., and Ashbaugh, A. R. Convenience sampling methods in psychology: A comparison between crowdsourced and student samples. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 2023.

Ogletree, A. M. and Katz, B. How do older adults recruited using MTurk differ from those in a national probability sample? *The International Journal of Aging and Human Development*, 93(2):700–721, 2021.

Paleja, R., Silva, A., Chen, L., and Gombolay, M. Interpretable and personalized apprenticeship scheduling: Learning interpretable scheduling policies from heterogeneous user demonstrations. *Advances in neural information processing systems*, 33:6417–6428, 2020.

Paolacci, G. and Chandler, J. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current directions in psychological science*, 23(3):184–188, 2014.

Peer, E., Rothschild, D., and Gordon, A. Platform over procedure: Online platforms that pre-vet participants yield higher data quality without sacrificing diversity. *Working Paper*, 2023.

Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Pias, S. B. H., Freel, A., Trammel, T., Akter, T., Williamson, D., and Kapadia, A. The Drawback of Insight: Detailed Explanations Can Reduce Agreement with XAI. *arXiv preprint arXiv:2404.19629*, 2024.

Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*, 2018.

Qian, J., Li, H., Wang, J., and He, L. Recent advances in explainable artificial intelligence for magnetic resonance imaging. *Diagnostics*, 13(9):1571, 2023.

Reeder, S., Jensen, J., and Ball, R. Evaluating explainable AI (XAI) in terms of user gender and educational background. In *International Conference on Human-Computer Interaction*, pp. 286–304. Springer, 2023.

Ribera, M. and Lapedriza, A. Can we do better explanations? A proposal of user-centered explainable AI. CEUR Workshop Proceedings, 2019.

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., and Kasneci, E. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

Schwalbe, G. and Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pp. 1–59, 2023.

See, J. E. Human readiness levels explained. *ergonomics in design*, 29(4):5–10, 2021.

Speith, T. and Langer, M. A new perspective on evaluation methods for explainable artificial intelligence (XAI). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pp. 325–331. IEEE, 2023.

Sperrle, F., El-Assady, M., Guo, G., Borgo, R., Chau, D. H., Endert, A., and Keim, D. A survey of human-centered evaluations in human-centered machine learning. In *Computer Graphics Forum*, volume 40, pp. 543–568. Wiley Online Library, 2021.

Stewart, N., Chandler, J., and Paolacci, G. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences*, 21(10):736–748, 2017.

Taesiri, M. R., Nguyen, G., and Nguyen, A. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *Advances in Neural Information Processing Systems*, 35:34287–34301, 2022.

Tahaei, M. and Vaniea, K. Recruiting participants with programming skills: A comparison of four crowdsourcing platforms and a CS student mailing list. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2022.

Vaccaro, K., Huang, D., Eslami, M., Sandvig, C., Hamilton, K., and Karahalios, K. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.

Vereschak, O., Bailly, G., and Caramiaux, B. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–39, 2021.

Vilas, M. G., Adolfi, F., Poeppel, D., and Roig, G. Position: An inner interpretability framework for AI inspired by lessons from cognitive neuroscience. *arXiv preprint arXiv:2406.01352*, 2024.

Villa, S., Kosch, T., Grelka, F., Schmidt, A., and Welsch, R. The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior. *Computers in Human Behavior*, 146:107787, 2023.

Vilone, G. and Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

Wang, P. and Ding, H. The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Information Processing & Management*, 61 (4):103732, 2024.

Weigold, A. and Weigold, I. K. Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*, 40(5):1302–1322, 2022.

Williams, O. Towards human-centred explainable AI: A systematic literature review. *Master's Thesis*, 2021.

Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997):177–185, 2024.

Zhou, H. and Fishbach, A. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology*, 111(4):493, 2016.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

A. Empirical Analysis of XAI Papers with a User Study

Here we provide details on how we collected and analyzed the XAI papers that were the basis for the empirical analysis in Section 4. The list of all 607 papers with meta-data is available for download[1].

A.1. Literature Search

We performed three different searches, with Scopus as the starting point in all three. The inclusion criteria were that the paper has a user study (no restrictions on the type or number of participants) and is from the broader field of XAI (explaining AI or ML systems or models; no restrictions on the type of explanation).

A.1.1. Direct Search

We performed a search through Scopus on Apr 7, 2024 with the search string:

```
TITLE-ABS-KEY(("explainable AI" OR "XAI" OR "interpretable AI" OR "interpretable
    machine learning")) AND (participant* OR "user study" OR "user evaluation*" OR "
    user rating*" OR "subjective rating*" OR "human evaluation*" OR "human study" OR
     "human rating*")
```

The query returned 865 candidate papers. We manually checked these papers and found 221 that met the inclusion criteria.

A.1.2. Indirect Search

Next, we searched for XAI papers with a user study indirectly, through XAI papers that were a survey of evaluation in XAI or contained, as part of related work, a collection of XAI papers with a user study.

We performed a search through Scopus on Apr 7, 2024 with the search string:

```
TITLE-ABS-KEY((("explainable AI" OR "XAI" OR "interpretable machine learning")) AND
    ("survey" OR "review") AND "evaluation")
```

The motivation for this search was threefold. First, it doubles as a search for the most related work. Second, it ensures that our collection of papers includes at least all of the papers that are referenced in the most related work. And third, some XAI papers with a user study are difficult to find with a keyword-based search, because they do not contain any of the typical keywords.

The search query returned 192 candidate review papers, 7 of which had a literature review with at least some focus on user studies. We performed forward and backward snowballing from these 7 papers, until we found no new review papers. This resulted in 18 papers: Alangari et al. (2023), Bertrand et al. (2023), Chromik & Schuessler (2020), Ferreira & Monteiro (2020), Herm et al. (2022), Johs et al. (2022), Keane et al. (2021), Lai et al. (2021), Liao & Varshney (2021), Lopes et al. (2022), Mohseni et al. (2021), Nauta et al. (2023), Nguyen et al. (2024), Qian et al. (2023), Rong et al. (2023), Sperrle et al. (2021), Williams (2021), and Zhou et al. (2021). Note that the three papers with the most user studies referenced are Herm et al. (2022) (152, only 25 referenced in the paper), Rong et al. (2023) (97), and Nauta et al. (2023) (65).

We manually checked the papers referenced in the above 18 papers and found 358 that met the inclusion criteria and were not found with direct search.

A.1.3. Top Conferences Search

Finally, we searched for XAI papers with a user study in four of the top venues for ML research (AAAI, ICLR, ICML, and NeurIPS) from 2020 to 2024.

We performed a search through Scopus on Oct 26, 2024 with the search string:

[1] https://github.com/estrumbelj/XAI-user-studies-dataset/blob/main/dataset.csv

```
TITLE-ABS-KEY ( ( "explain*" OR "XAI" OR "interpret*" OR "explan*" ) ) AND ( PUBYEAR
    > 2019 ) AND SRCTITLE ( "ICML" OR "Advances In Neural Information Processing
   Systems" OR "AAAI" OR "ICLR" ) AND ( questionnaire OR crowdsourc* OR "amazon
   mechanical turk" OR "prolific" OR participant* OR "user stud*" OR "user eval*"
   OR "human subject*" OR "user rating*" OR "subjective rating*" OR "human eval*"
   OR "human stud*" OR "human-subject*" OR "subjects" OR "human rating*" )
```

The motivation for this search was twofold. First, these venues are of particular interest, because of their impact on the ML community, both in terms of research directions and standards. And second, a narrower scope of venues allowed us to relax the search string and obtain a larger and more systematic subsample for this subset of venues.

The query returned 359 candidate papers. Note that the query returns papers from conferences other than the four target conferences. We did not include such papers. Also note that NeurIPS 2024 was not indexed by Scopus at the time of the search.

We manually checked the papers and found 47 that met the inclusion criteria. Out of these 47 papers, 28 were newly found and 19 were already found in the previous two searches. Note that nine papers from these four venues were found in the indirect search but not in this search. As expected, all papers found in the direct search were also found in this search, because the direct search had a strictly more restrictive search query.

A.2. Analyzing the Papers

The additional data on the XAI papers with a user study are summarized in Table 1. To make the workload manageable, some of the data that require manual review are included only for a subsample of 116 papers. We sampled (with replacement) 30 papers published up to 2020, 30 papers published 2021 or later, and all 56 papers from the four conferences and published 2020 or later. We used a simple Binomial-Beta Bayesian model with Beta($\frac{1}{2}$, $\frac{1}{2}$) to infer the proportions and we report 95% posterior CI based on 2.5% and 97.5% quantiles.

A.2.1. Categorizing the Type of Quantitative Evaluation

To categorize the type of qualitative evaluation methods used in XAI user studies, we used the 7 evaluation methods identified by Nauta et al. (2023, Table 5).

For *forward simulatability*, *intruder detection*, and *perceived homogeneity* we were able to follow the original definitions.

For *teaching ability* we expanded the original definition, where the participant should be able to predict new instances without explanations, with cases where participants learned a valuable skill (for example, children learned how to better handle their diabetes).

For *synthetic artifact rediscovery* we only included a study if the artifact was added later. For example, if the model was changed to include gender bias, which was not originally there. For example, if the user was to perceive whether or not the model was biased, we counted this as subjective satisfaction.

For *subjective comparison* we included only user studies that compare two or more XAI methods. We included even small changes in presentation. For example, showing feature importance with or without overall model accuracy. We did not include studies that compared different ML models but all with the same type of explanation.

For *subjective satisfaction* we added two groups of user studies. The first are XAI frameworks which allow users to do exploratory analysis, after which the users describe how the framework helped them. The second are user studies that interview participants for their opinion, without an actual application of XAI. Note that fairness is a commonly measured subjective satisfaction item not listed in the original definition.

Our application of the categories resulted in 1.8 categories per paper, compared to 1.2 categories per paper in Nauta et al. (2023). This can be partially explained by our more liberal interpretation of subjective comparison. Furthermore, Nauta et al. (2023) focused on top ML venues, where our results are very similar. That is, the discrepancy can be explained by more frequent use of subjective comparison and subjective satisfaction in venues outside of top ML venues.

Table 1. A summary of the data for the XAI papers with a user study.

	COLUMN	NOTES
PAPER INFO	TITLE	
	DOI	NA if not available.
	VENUE	
	ABSTRACT	
	PUBLICATIONYEAR	
SAMPLING INFO	WHICHSEARCH	Which search found this paper (direct, indirect, top 4).
	SUBSAMPLINGGROUP	NA if the paper was not subsampled for analysis, otherwise its group (up to 2022, 2021 and after, top 4). Comma-separated if more than one applies. See A.2.
USER STUDY (ALL)	PARTICIPANTTYPE	User study participant type (administrative staff, AI expert, crowdsourcing, domain expert, researchers, students, other, not stated). Subtype provided in parentheses (for example, which crowdsourcing platform). Comma-separated if more than one type.
	PARTICIPANTCOUNT	NA if not available. Comma-separated if more than one user study.
	NAUTACLASSIFICATION	Comma-separated if more than one type. See A.2.1.
USER STUDY (SUBSAMPLE)	CATFIDELITY	See A.2.2.
	CATLEVEL	See A.2.2.
	CATOBJSUBJ	See A.2.2.
	HASBASELINE	See 3.4.
	HASPLACEBO	See 3.4.
	ISPREREGISTERED	Was the study pre-registered.
	ISREPRODUCIBLE	Is the study reproducible. True, unless it is missing key information.
	HASPRETESTING	Does the study report any pretesting or pilot study before the main study.
	HASQUESTIONNAIREVALIDATION	Are the user measurement instruments validated in this or a previous study.
	HASLIMITATIONS	Does the study have at least a minimal discussion of limitations.
	CS_HASATTENTIONCHECK	Crowdsourcing user studies only. Does the study include any type of attention check.
	CS_HASDATAQUALITY	Crowdsourcing user studies only. Does the study include any type of post-collection data quality assurance.

A.2.2. CATEGORIZING USER STUDIES ON FIDELITY

User studies vary in how well their findings translate to real-world applications. Therefore, when evaluating the real-world performance of XAI, it is helpful to score or categorize user studies on this dimension. We introduce four categories for the *fidelity* of the user study setting to a real-world application:

- *High:* XAI is embedded in a real-world application and evaluated in a real-world setting. Example: Millecamp et al. (2019) tested their XAI system by directly connecting to the participant's Spotify account to provide them with song recommendations and corresponding explanations.
- *Medium:* Suggests a clear real-world application, but it is not evaluated as such. Example: Paleja et al. (2020) propose a framework for scheduling. The authors do provide an example of a real-world use (Taxi domain), which demonstrates a plausible use for use in real-world situations.
- *Low:* Embedded in a toy or mock application. Example: Ben David et al. (2021) aim to develop a financial algorithmic advisor and evaluate their approach on a simplified lemonade stand game.
- *No:* Not embedded in an application. Example: Taesiri et al. (2022) tested their explanation on a dataset, but there is no sign of how the explanation could be used in real life or what motivates the choice of dataset.

We complement fidelity with the evaluation *levels* inspired by the model by Speith & Langer (2023):

- *Explanatory:* Evaluating how accurately XAI describes the AI system. While a user study can be used for this level, we did not find any in the subsample. That is, functional evaluation is typically used for this level.
- *Understanding:* Evaluating how well the XAI facilitates understanding of the AI system. This can be measured *objectively* (for example, by forward simulation) or *subjectively* (for example, self-reported understanding or trust).
- *Downstream performance:* Evaluating how well XAI contributes to the task the AI system is designed to solve. Can be measured objectively (for example, task performance) or subjectively (self-reported confidence).

Note that we also considered using two existing categorizations: the most popular taxonomy for XAI evaluation methods by Doshi-Velez & Kim (2017; 2018) and the Technology Readiness Levels (TRL) scale, a well-known scale for estimating the maturity of technological development developed by NASA. The taxonomy by Doshi-Velez and Kim subdivides uses studies into application-grounded (real humans, real tasks) and human-grounded (real humans, simple tasks). We found this to be insufficiently granular for our purpose. The TRL scale or its adaptations to AI (Browne et al., 2024; Lavin et al., 2022) or human readiness level (HRL) (See, 2021) are more granular (typically a 9-point scale). However, their purpose is to estimate technology readiness during system development by qualified experts on the design team (Handley et al., 2024). Post-hoc application of this scale to user studies is difficult, because there is in most cases no clear application context, incremental progress, or end goal.

A.3. Limitations

Systematically searching for XAI papers with a user study is difficult. XAI papers might not contain any of our XAI or interpretable ML search terms. Furthermore, they might not even contain any of our user study or participant search terms. The limitations of keyword search are clear from the large number of papers found by indirect search but not direct search. A quick manual inspection of 20 randomly chosen papers found by indirect search showed that 19 of them were indexed by Scopus. Therefore, most papers not found by direct search were not found because they did not contain the search terms (assuming that there are no issues with Scopus data or search).

The chosen search terms were a tradeoff to limit the manual inspection of papers for inclusion to the order of 1000s. The only better alternative would be to manually inspect all papers, which we do not find feasible, unless we severely limit the number of venues and the time period.

As a result, our sample of papers is biased in several ways. It inherits any biases towards venues and publication years of Scopus. However, we argue that Scopus indexes the vast majority of relevant venues, especially in the period of the past 5 years. Indirect search references can only go back in time, so the sample of papers found by indirect search is biased towards older papers. We partially mitigate this by splitting the papers into two groups based on publication year. The indirect sample is also biased towards certain venues. The data we collected does allow us to partition the papers by venue, but we have not done so. Similarly, we could limit our analysis to direct search papers only, which would mitigate at least the biases of indirect search.

In the context of our position, the key question is whether these biases also result in a bias towards lower quality studies. We believe it is very unlikely that a paper that is more easily identified as a XAI paper with a user study, is more likely to have a poorly designed user study. However, it is likely that newer papers have better user studies, which we do investigate by splitting the papers into two groups with respect to time.

Some of the user study variables were not trivial to measure (NautaClassification, IsReproducible, ParticipantType, CatFidelity) and, while we do believe the data are measured relatively consistently (a single rater for each variable), it is possible that there is some between-rater variability. At least for the key claims, this should be mitigated by the fact that when in doubt, we chose to benefit the quality of the user study.

To summarize, this is an exploratory study. A more systematic confirmatory study is required to further validate the findings of this exploratory study. For a more complete view of the quality of user studies in XAI, we also lack a more diverse set of venues, in particular, human-centric research venues (for example, FaaCT) and other top AI/ML venues (for example, IJCAI). However, we argue that the sample is still large enough to draw conclusions and to support our position. For example, if most of the found user studies at top venues are not reproducible, this is reason for concern. Even if we allow for the unlikely scenario that all user studies that we did not find are reproducible.