

https://dspace.org/

Unjustified sample sizes and generalizations in explainable artificial intelligence research Principles for more inclusive user studies

U Peters; Mary Carman

IEEE COMPUTER SOC
https://hdl.handle.net/10539/44401
WOS

Downloaded from a DSpace repository. https://dspace.org/

Unjustified Sample Sizes and Generalizations in Explainable Artificial Intelligence Research: Principles for More Inclusive User Studies

Uwe Peters, *University of Cambridge, CB2 1SB, Cambridge, United Kingdom*

Mary Carman, *University of the Witwatersrand, Johannesburg, 2050, South Africa*

Many ethical frameworks require artificial intelligence (AI) systems to be explainable. Explainable AI (XAI) models are frequently tested for their adequacy in user studies. Since different people may have different explanatory needs, it is important that participant samples in user studies are large enough to represent the target population to enable generalizations. However, it is unclear to what extent XAI researchers reflect on and justify their sample sizes or avoid broad generalizations across people. We analyzed XAI user studies (n = 220) published between 2012 and 2022. Most of the studies did not offer rationales for their sample sizes. Moreover, most of the papers generalized their conclusions beyond their study population, and there was no evidence that broader conclusions in quantitative studies were correlated with larger samples. These methodological problems can impede evaluations of whether XAI systems implement the explainability called for in ethical frameworks. We outline principles for more inclusive XAI user studies.

The artificial intelligence (AI) systems used for decision making in high-stakes contexts (e.g., hiring domains or medical predictions) are often computationally opaque, i.e., they may follow complex algorithms that even AI engineers can no longer fully understand.[1] Since this makes the trustworthiness of these systems questionable, many ethical AI frameworks require the outputs of AI technology to be explainable to people so that they can potentially object to AI-based decision making.[2] One main approach to implementing explainability in practice involves supplementing opaque AI with explainable AI (XAI) models designed to make opaque systems' outputs understandable to humans.

To evaluate whether XAI models' explanations satisfy a principle of explainability, developers often test their systems on human users.[3] Since explanation needs may vary across individuals, the value of such user studies commonly depends on whether these studies' results provide insights into human users' requirements that are generalizable from the particular study participants to broader populations.[1] If findings about XAI systems only hold for a small group of individuals, these systems may not meet the explanatory requirements of other people affected by or using them and hence fail to adequately implement explainability in practice. This can impact trust in AI systems. The generalizability of XAI user study results is thus vital for developers and policy makers to gain insight into whether a given XAI system satisfies an ethical principle of explainability.

One way of achieving generalizability in an XAI user study is to test the whole target population. However, researchers' constraints (e.g., funding) rarely make this feasible. Usually, in human-subject studies in general, a sample is taken from the target population and results are then extrapolated to that population. This can go wrong when the sample is not large enough to reflect the target population.[4] Thinking about and justifying a chosen sample size is thus important for generalizability.

Concerns about underpowered sample sizes, i.e., samples too small to detect effects and generalize

1541-1672 © 2023 IEEE
Digital Object Identifier 10.1109/MIS.2023.3320433
Date of publication 29 September 2023; date of current version 13 November 2023.

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

results, have been raised across the sciences.[5] And many social science journals now require authors to provide sample size justifications, i.e., rationales for selecting a particular sample size (e.g., power analyses).[6] However, although the analysis of these justifications is already a topic of interest in different fields,[6,7,8] no examination of sample size justifications in XAI user research yet exists.

To be sure, if XAI user researchers tailored their generalizations to their samples' size and composition such that the researchers only make broader claims when they have tested larger, more representative samples,[9] then small or unrepresentative samples and omissions of sample size justifications may not be problematic. Similarly, if a given XAI user study was only run to quickly receive feedback and determine whether at least some users understand a particular XAI output, then little generalizability would be sought and no sample size justification may be needed. However, no systematic analysis of how broadly or narrowly XAI researchers do in fact generalize from their studies exists so far.

Yet, investigating concerns about sample sizes, sample size justification, and generalizability in XAI user studies is especially important and intertwined with AI ethics. If XAI user studies have shortcomings in sampling or generalizability, stakeholders may need to recalibrate their trust in the XAI models that these studies often advertise because the studies' findings might then only apply to particular groups of individuals. The implementation of ethical frameworks for XAI systems thus also involves the implementation of certain basic principles of scientific methodology in user studies to ensure that people's trust in XAI systems is aligned with the systems' capability of providing widely acceptable explanations of opaque models.

We therefore systematically reviewed a large number of XAI user study papers ($n = 220$) published between January 2012 and July 2022 to examine sample size justifications and generalizations in them. Most studies did not offer sample size justifications, leaving it unclear whether their samples were large and representative enough for broad generalizations of results beyond the participants. However, most of the studies nonetheless generalized their results far beyond their samples. There was also no evidence that broader conclusions were correlated with larger samples. And such conclusions appeared even when researchers had not checked for relevant background knowledge (e.g., AI expertise) among their study participants that can affect result generalizability. These points suggest that *overgeneralizations*, i.e., conclusions whose scope is broader than warranted by the evidence and justification provided by the researchers, are pervasive in many available XAI user studies.

In summary, this article offers the following novel contributions:

- It is the first systematic investigation of sample size justifications in XAI user studies that provides evidence that such justifications are rare in many available user studies.
- It offers the first quantitative data on overgeneralizations of results in XAI user studies.
- It introduces principles to mitigate these problems and be able to better evaluate whether XAI systems adequately implement explainability for ethical AI in practice.

BACKGROUND AND RELATED WORK

XAI

Although the literature on XAI is extensive, XAI methods can be broadly distinguished into transparent models and posthoc (model-specific or model-agnostic) techniques.[10] Transparent models are, by themselves, understandable (e.g., logistic or linear regression).[11] In contrast, posthoc techniques aim to provide understandable information about how an opaque system produces its outputs for a given input, for instance, by accessing the system's internals (e.g., decision weights), or by analyzing the dependencies between input–output pairs to infer the factors contributing to the system's decisions.[10] Posthoc XAI processing can be local, aiming to explain specific outputs of opaque models, or global, aiming to explain the entire model's behavior, and XAI explanations may be visual (e.g., saliency maps), numerical (e.g., importance scores), or textual (e.g., feature reports).[12]

Sample Size Justification

If XAI user studies find that participants understand a particular XAI system's outputs well, this may be cited to claim that the system satisfies a key component of ethical frameworks for promoting trust in AI. This can facilitate the system's larger-scale deployment. Stakeholders thus need to have confidence that the conclusions drawn from XAI user studies are correct.

Confidence about study results relates to sample size in that narrower confidence intervals (CIs), which indicate more precise results, require larger samples.[9] Larger samples also help detect small group differences for which smaller samples may not be sufficiently powered.

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

That said, even minute statistical effects can be boosted to significance by increasing sample size because this reduces a statistic's standard error, meaning that one can find statistically significant but practically meaningless effects in very large samples. In that sense, samples can also be "too big" for hypothesis testing and potentially magnify the biases associated with mistakes in sampling or study design.[13]

The sample size that a quantitative study needs depends on different factors that researchers may treat differently given their goals. These factors include the α-error level (capturing the risk of false positives, usually set at 5%), the β-error level (capturing the risk of false negatives, commonly set at 20%), and the effect size.[8] Researchers also need to consider sample composition because although randomly chosen larger samples are more representative of a target population, large samples may be homogenous in relevant characteristics and therefore unrepresentative.

Other considerations to factor in are that, unlike large sample studies, small sample studies are cheaper and quicker to perform, which helps prevent the wasting of resources. And in qualitative research, having fewer participants can facilitate closer relationships with them, potentially providing more in-depth insights into personal experiences.[14] Since different decisions may be made based on these and other methodological factors, the provision of sample size justifications in XAI user studies matters because reviewers or other stakeholders cannot assume that the chosen sample size is adequate for a given study.[8]

Common approaches to justifying sample sizes in the social sciences include 1) power analysis, 2) heuristics (e.g., sampling guidelines, consistency with existing research, or rules of thumb), 3) pragmatic considerations (e.g., funding, participant availability, and low response rate), and, for qualitative studies, 4) saturation, i.e., the point at which no new data (ideas, opinions, and so on) can be attained with more participants.[4,7,8] However, it is unclear whether any sample size justifications are reported in XAI user studies and how generalizations of study results are related to sample sizes. A systematic review of XAI user studies is needed to investigate these issues.

A SYSTEMATIC LITERATURE REVIEW

Following a rigorous methodology (as set out by Moher et al.[15]), we reviewed XAI user study papers to answer five research questions (RQs):

1) Do researchers who conduct XAI user studies justify their sample sizes?
2) Do XAI user studies with sample size justification have different (e.g., larger) sample sizes than those without?
3) Do researchers who conduct XAI user studies restrict their conclusions to their participants or study populations or extrapolate beyond them?
4) Are broader conclusions correlated with larger samples?
5) Do researchers who conduct XAI user studies with, for instance, lay users, check whether their participants have a background (e.g., AI expertise) that can influence study results and generalizability?

METHODOLOGY

Literature Search

We used three major databases: Scopus, Web of Science, and arXiv. Scopus and Web of Science index key computer science resources (e.g., the ACM Digital Library and IEEE *Xplore*). ArXiv contains AI papers not yet published in journals, potentially providing insights into recent developments. We also used forward snowballing, i.e., we identified and included relevant papers frequently cited (>10 citations) in the studies selected for full-text analysis. The three databases were searched in July 2022 using search strings containing 15 variants of keywords related to XAI ("XAI," "explainable AI," and so forth) and end users ("user study," "user survey," and so on). More details can be found in the supplementary materials available at http://doi.org/10.1109/MIS.2023.3320433. The results were 2523 papers. After removing duplicates ($n = 535$), the titles and abstracts of the remaining 1988 papers were scanned for studies that met our selection criteria.

Selection Criteria

We included any primary study (article, conference paper, or book chapter) that surveyed people on their perception of AI-based explanations of automated processing and was published between January 2012 and July 2022. We excluded editorials, notes, opinion papers, reviews, or studies that did not contain information about sample size, were not in English, offered only a preliminary data analysis, or had a sample ≤ 5 (an exceedingly small sample to ensure study validity).[7] Two hundred and six articles remained for further screening, during which forward snowballing produced an additional 14 papers, yielding a final $n = 220$ for full-text analysis (see the Preferred Reporting Items for

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

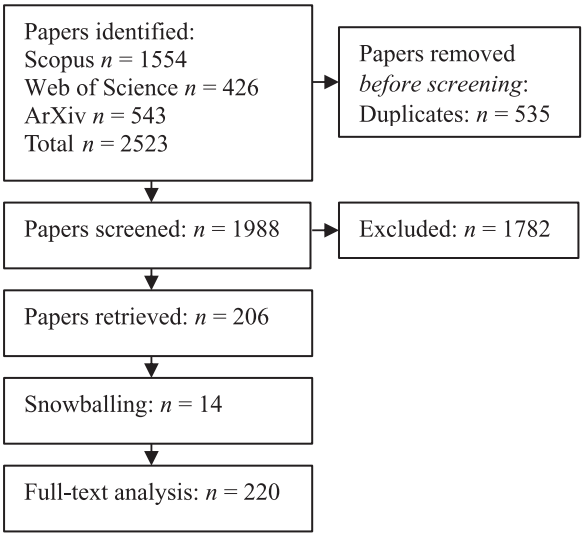


FIGURE 1. PRISMA flow diagram for the systematic review.

Systematic Reviews and Meta-Analysis (PRISMA) flowchart in Figure 1).

During full-text analysis, we (two researchers) separately classified each paper using prespecified criteria for extracting the following five sets of information:

1) *General information*: We extracted publication year, study sample size (final participant number), and study design, categorizing papers as "quantitative" (studies involving statistical analyses), "qualitative" (studies involving qualitative analysis of interviews, free responses, and so on), or "mixed" designs (studies involving both quantitative and qualitative elements).
2) *Scope of conclusions*: In their articles, researchers may limit their study conclusions to their sample, a specific target population, or a minority of individuals by using the relevant qualifiers (e.g., "our participants … ," "U.S. users … ," or "many lay people … "). They may also use past tense to describe findings, limiting their conclusions to their study, sample, or study population. Papers with these features were classified as "restricted." Alternatively, authors may, in their paper's abstract, results, discussion, or conclusion sections, refer to users, experts, people, humans, and so forth in general, not subsets of them (e.g., particular national target populations), or otherwise describe results in ways that suggest they hold across contexts, time, cultures, backgrounds, or groups. Papers with at least one such broad result-related claim in these sections were classified as "unrestricted" (for examples, see Table 1). This label was also applied when an article additionally contained restricted claims, because papers usually undergo many revisions when authors can qualify their broader claims. If that does not happen, this suggests that the authors consider their broader generalizations warranted.
3) *User type*: Our categories were 1) "lay users" (for unspecified crowdsourced participants, unspecified students, or when the study material suggested a lay-user audience), 2) "technical experts" [construed broadly as people with a programming, IT, machine learning (ML), or computer science background], 3) "lay users/technical experts" [for 1) and 2)], 4) "domain experts" (clinicians,

TABLE 1. Examples of unrestricted conclusions.

(1) "Specifically, our results suggest that users prefer more diverse local explanations when they are presented alone, compared to when a global explanation is also available." [12]
(2) "We demonstrated that CX-ToM significantly outperforms baselines in improving human understanding of the underlying classification model." [26]
(3) "Our pilot study revealed that users are more interested in solutions to errors than they are in just why the error happened." [42]
(4) "Explanations lead people to view errors as being 'less incorrect,' but they do not improve trust." [130]
(5) "People prefer item-centric but not user-centric or sociocentric explanations." [142]
(6) "Results indicate that human users tend to favor explanations about policy rather than about single actions." [205]
(7) "Our findings suggest that people do not fully trust algorithms for various reasons, even when they have a better idea of how the algorithm works." [213]

The number in square brackets refers to the identifier of the paper on the data spreadsheet from which the examples are taken (see https://osf.io/vzndw/).

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

lawyers, and so on), and 5) "domain experts/technical experts" [for 4) and 2)].

4) *Relevant background*: Many demographic factors (gender, age, and so forth) may confound user study results. But including or excluding them as control variables requires justification because including control variables decreases available degrees of freedom and power and may in some cases reduce the explainable variance available in outcomes of interest, facilitating false negatives. In contrast, excluding demographic control variables can inflate the explainable variance in the criterion, facilitating false positives.[16] One factor that studies repeatedly found to produce different responses to XAI models is having a background in AI, ML, XAI, IT, or computer science.[1,17,18] People with such backgrounds may display high technical affinity and already be familiar with salience maps, importance scores, and so on. Asking XAI study participants about their potential technical background (i.e., technical affinity, expertise in programming, ML, XAI, IT, or computer science) to control for it during data collection or analysis is thus important if results are to be generalizable to lay users. To record when researchers questioned or categorized users accordingly, we used a simple "yes"/"no" classification per paper.
5) *Sample size justification*: We operationalized sample size justification as any research effort before data collection to obtain adequate sample sizes. We coded papers ("yes"/"no") depending on the presence of at least one of the following: 1) power analysis, 2) heuristics (e.g., consistency with other studies), 3) pragmatic considerations (e.g., resource constraints), or 4) saturation.

Reliability

For each classification, interrater agreement was calculated (Cohen's κ) and was consistently above substantial (between $\kappa = .71$ and .90, $p < .001$). As a further reliability control, for the scope of conclusion variable (our most complex classification), we additionally asked two independent, naïve raters to apply our prespecified criteria to 25% of the data. Interrater agreement between their and our classifications was measured and was substantial ($\kappa = .66$ and .74, respectively). All remaining disagreements were resolved by discussion before the data were statistically analyzed ($\alpha = .05$). Our materials are accessible on an Open Science Framework (OSF) platform (https://osf.io/vzndw/).

RESULTS AND DISCUSSION

Most of the XAI user studies (97.7%, $n = 215$) were published (online or in journals) between 2018 and 2022, with an almost 50% increase from 2020 ($n = 46$) to 2021 ($n = 84$). In the supplementary materials available at http://doi.org/10.1109/MIS.2023.3320433, we present publication numbers by year. Of the reviewed studies, 57.3% ($n = 126$) used quantitative methods, followed by studies with mixed designs (30.4%, $n = 67$), and qualitative research (12.3%, $n = 27$).

RQ1: *Do researchers who conduct XAI user studies justify their sample sizes?* We found that 88.2% ($n = 194$) of the reviewed user study papers did not justify their sample sizes. Looking more specifically at papers by method type, from those with qualitative studies ($n = 27$), none offered a sample size justification. This is problematic because even though qualitative studies often do not aim for generalizability,[14] they still require sample size justification, for instance, to ensure saturation.[7] Focusing on quantitative studies, for these studies, a power analysis is commonly viewed as the gold standard for evaluating study feasibility and for justifying sample size.[4] A power analysis is recommended to prevent undersampling (i.e., too-small, unrepresentative samples that can facilitate false negatives) and oversampling (i.e., too-large samples that can boost practically irrelevant effects to significance). We thus looked more closely at this particular kind of sample size justification in quantitative studies, setting aside qualitative and mixed-study papers.

We found that of the 126 quantitative studies, 88.1% ($n = 111$) did not report any power analysis. Researchers who do not report such an analysis may still have conducted one. However, since conducting a power analysis is a methodological strength, it is hard to see why these researchers did not mention it in their papers if they had performed such an analysis. This suggests that such analyses and sample size justifications were not a part of the study designs. For preliminary, piloting, or explorative studies, asking for a power analysis may be excessive. However, only 5% ($n = 11$) of the reviewed XAI studies indicated that their research was of that kind. Focusing on the more than 88% of the quantitative studies that did not provide any sample size justification, in not offering such a justification, their authors failed to establish that their selected samples were large enough to detect

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

relevant effects and generalize the study results beyond the participants.

RQ2: *Do XAI user studies with sample size justification have different (e.g., larger) sample sizes than those without?* It might be that sample size justification does not affect studies' sample sizes. If so, the sample sizes of studies with such justifications should not significantly differ from those without them. To examine this, we first tested our data for normality, finding that a nonparametric statistic was required (the normality assumption was violated). Treating sample size justification as a binary variable and sample size (final participant n) as a scale variable, a Mann–Whitney U test was run, showing that papers with sample size justification ($n = 26$) had significantly larger sample sizes (*mean rank* $= 141.23$) than papers without it ($n = 194$) (*mean rank* $= 106.38$, $U = 1723$, $z = -2.622$, $p = .009$). A rank–biserial correlation test additionally showed that sample size justification was correlated with larger samples, r_{rb} (218) $= .177$, 95% CI [.042, .306], $p = .008$. To the extent that larger samples are more representative and increase generalizability, this finding suggests a link between sample size justification and generalizability-strengthening sample sizes. But how broadly exactly did XAI user study researchers generalize their results?

RQ3: *Do the researchers who conduct XAI user studies restrict their conclusions to their participants or study populations or generalize beyond them?* Of the 220 papers, 68.2% ($n = 150$) contained unrestricted conclusions, i.e., conclusions that referred very broadly to people, users, or humans as whole categories, or applied across context and time. Table 1 presents seven examples. Importantly, of the 150 "unrestricted" papers, 84.7% ($n = 127$) did not contain any sample size justification. Hence, even though the authors had not reflected on whether their sample sizes would support generalizations beyond the study population, and so did not provide the needed epistemic basis for such generalizations, these papers nonetheless contained conclusions that extended vastly beyond the studied populations. The authors thus overgeneralized their user study results.

RQ4: *Are broader conclusions in XAI user studies correlated with larger samples?* When considering quantitative studies, one may predict a correlation between broader claims and larger samples because randomly selected larger samples tend to be more representative and so can ensure wider generalizability.[9] Finding such a link in *qualitative* studies is less likely because these studies are often not meant to achieve broad result generalizability.[14] Since qualitative and mixed-method studies (due to their qualitative component) may thus skew correlation tests, we excluded them from the analysis for *RQ4* and focused only on quantitative studies ($n = 126$). As data normality was violated ($p < .001$), we first conducted a Mann–Whitney U test to compare sample sizes between "restricted" and "unrestricted" papers. We found that "unrestricted" papers ($n = 99$, *mean rank* $= 66.68$) did not have statistically significantly larger samples than "restricted" ones ($n = 27$, *mean rank* $= 51.83$; $U = 1021.5$, $z = -1.873$, $p = .061$). A subsequent rank–biserial correlation test also did not provide evidence of an association between "unrestricted" papers and sample size, r_{rb} (124) $= .168$, 95% CI [−.013, .337], $p = .061$. Since 78.6% ($n = 99$) of the quantitative study papers were "unrestricted," this suggests that many of the broad extrapolations that these papers contained were insufficiently supported. If they all had been sufficiently supported, we should have found unrestricted conclusions to be correlated with larger samples. However, it might still be that the papers' authors at least took care to control for potential confounding features in their study participants that may affect result generalizability.

RQ5: *Do researchers who conduct XAI user studies with, for instance, lay users, check whether their participants have a background (e.g., AI expertise) that can influence study results and generalizability?* Several studies had different kinds of users as their target audience. Sixty-nine and one tenths percent ($n = 152$) of the studies had only lay users, 10.5% ($n = 23$) only technical experts, 8.6% ($n = 19$) lay users/technical experts, 6.4% ($n = 14$) only domain experts, and 4.1% ($n = 9$) domain experts/technical experts as intended targets. Importantly, focusing on the 152 studies that authors presented as testing only lay users, 70.4% ($n = 107$) did not contain evidence that the participants (typically recruited via crowdsourcing platforms, e.g., MTurk) were questioned about or categorized according to their technical affinity, potential experience with XAI/ML, or computer science background. These participants may thus have had such a background. Since this can significantly influence people's perceptions of XAI outputs, the studies' results cannot be broadly generalized.[1,17,18] Yet, we found that 78.5% ($n = 84$) of the studies did just that (they were "unrestricted" papers).

LIMITATIONS

The studies we included in our systematic review were heterogeneous in user types, XAI outputs, sample sizes, and methods, making extrapolations across papers challenging. Relatedly, we did not record how many of

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

the reviewed XAI papers 1) included user experiments with functional metrics (XAI completeness, faithfulness, robustness, and so forth) tests, 2) introduced new XAI techniques, or 3) offered only comparisons between different techniques in a specific domain. This information could have revealed different correlations regarding the generalizability of study results in subsets of the reviewed papers. Future research on XAI user study generalizations that includes this information would be desirable.

Moreover, we searched only three major databases for XAI research and focused on English publications. We may therefore have missed important XAI user studies.

Additionally, our sample contained 14 not yet peer-reviewed arXiv papers. Focusing only on peer-reviewed articles might change results. However, when we reran the analyses with only peer-reviewed papers, our key findings retained the same trend (more details are provided in the supplementary materials available at http://doi.org/10.1109/MIS.2023.3320433). The only difference was that quantitative papers with broader conclusions now also had larger samples ($p = .027$). However, our focus in this study was on examining XAI researchers' extrapolation tendency in general. Analyzing only peer-reviewed papers is less likely to provide insights into researchers' general disposition to extrapolate rather than into what they do if their inferences are peer reviewed. Since including not yet peer-reviewed papers is more informative on researchers' general disposition and our key trends remained the same, we kept these papers in our analyses.

Finally, we extracted our data from papers manually, not automatically. Human error could have occurred. However, to reduce this risk, we reviewed papers independently, cross-checked each other's classifications, had two author-independent raters classify subsets of the data, calculated interrater agreement, and made our data publicly available for reproduction.

PRINCIPLES FOR MORE INCLUSIVE XAI TESTING

XAI user studies with underpowered samples, missing sample size justifications, and overgeneralizations may lead to an oversight of many people's explanatory needs regarding the AI systems they may encounter. This can hinder the development of ethical AI that uses XAI to implement explainability. We propose three methodological principles to counteract these problems:

1) *Principle of sample size justification*: Studies of new technologies may frequently start small due to funding or feasibility concerns. This may often be acceptable to explore the usability of models and facilitate developmental progress. However, to promote best scientific practice, AI journals should require user studies to include (where feasible) power analyses or other sample size justifications. For XAI researchers who struggle to find method-specific sample size rationales, we recommend three papers with overviews of sample size justifications: Caine,[4] Vasileiou et al.,[7] and Lakens.[8]

2) *Principle of reporting relevant background*: Since having technical affinity, AI expertise, or a computer science background is known to influence XAI user perception,[1] it is important to question participants about it to control for it. This may also apply to other factors (e.g., gender). AI journals should ask XAI study authors to consider, report, and justify the inclusion and exclusion of relevant control variables. Overlooking variables to control can facilitate false positives, whereas including too many can facilitate false negatives. For best practice on control variable usage, we recommend Bernerth and Aguinis.[16]

3) *Principle of generalization checks*: Overgeneralizations may result when researchers need to persuade journals, funding bodies, and policy makers of their studies' importance, or when writing guidelines require more condensed language. We recommend that AI journals reconsider their common emphasis on condensed formats and adopt a principle for authors to provide generality constraint statements in their papers, i.e., statements that articulate generalizability limits and justify the scope of result-related claims. For guidance on what to include in such statements, we recommend Simons et al.[19]

We also suggest that XAI authors use a checklist containing reminders to 1) mention relevant qualifiers (statistical distributions, percentages, "U.S. users," and so on) to tailor their conclusions to the evidence and 2) consider using past tense when reporting results, which restricts results to subsets of individuals.[20] To illustrate this, in the supplementary materials available at http://doi.org/10.1109/MIS.2023.3320433, Table 2 presents restricted reformulations of the unrestricted claims from Table 1.

CONCLUSION

In this systematic analysis of XAI user studies, we found that many XAI researchers did not follow best scientific practice. They did not justify the size of the samples tested, leaving it unclear whether the samples were too small to detect differences and generalize to

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

the target population, or too large, inflating practically irrelevant differences. Yet, most of the studies nonetheless contained broad conclusions extending beyond their samples and study populations to users, people, and so forth in general. Because there was no evidence that these broad conclusions were correlated with larger, more representative samples, and researchers had often not checked for a technical background among their participants that may have undermined generalizability, the conclusions that we found in most of the papers were overgeneralizations. Given the important and valuable role that XAI user research plays for informing stakeholders' assessments of whether a given XAI model implements ethical AI frameworks, future XAI user study methods should be improved such that sample size justifications become routine in XAI research and overgeneralizations are avoided.

ACKNOWLEDGMENT

Uwe Peters is the main author of this article. This article has supplementary materials available at http://doi.org/10.1109/MIS.2023.3320433.

REFERENCES

1. E. Upol et al., "The who in explainable AI: How AI background shapes perceptions of AI explanations," 2021, *arXiv:2107.13509*.
2. J. A. Ferrario and M. Loi, "How explainability contributes to trust in AI," in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, New York, NY, USA: ACM, 2022, pp. 1457–1466, doi: 10.1145/3531146.3533202.
3. X. Wang and M. Yin, "Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making," in *Proc. 26th Int. Conf. Intell. User Interfaces*, New York, NY, USA: ACM, 2021, pp. 318–328, doi: 10.1145/3397481.3450650.
4. K. Caine, "Local standards for sample size at CHI," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2016, pp. 981–992, doi: 10.1145/2858036.2858498.
5. B. Leichtmann, V. Nitsch, and M. Mara, "Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement," *Frontiers Robot. AI*, vol. 9, Mar. 2022, Art. no. 838116, doi: 10.3389/frobt.2022.838116.
6. C. McCrum, J. van Beek, C. Schumacher, S. Janssen, and B. Van Hooren, "Sample size justifications in Gait & Posture," *Gait Posture*, vol. 92, pp. 333–337, Feb. 2022, doi: 10.1016/j.gaitpost.2021.12.010.
7. K. Vasileiou, J. Barnett, S. Thorpe, and T. Young, "Characterising and justifying sample size sufficiency in interview-based studies: Systematic analysis of qualitative health research over a 15-year period," *BMC Med. Res. Methodol.*, vol. 18, no. 1, pp. 1–18, Nov. 2018, doi: 10.1186/s12874-018-0594-7.
8. D. Lakens, "Sample size justification," *Collabra Psychol.*, vol. 8, no. 1, pp. 1–28, Mar. 2022, doi: 10.1525/collabra.33267.
9. N. Asiamah, H. Kofi Mensah, and E. Fosu Oteng-Abayie, "Do larger samples really lead to more precise estimates? A simulation study," *Amer. J. Educ. Res.*, vol. 5, no. 1, pp. 9–17, Jan. 2017, doi: 10.12691/education-5-1-2.
10. T. A. Abdullah, M. S. Zahid, and W. Ali, "A review of interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions," *Symmetry*, vol. 13, no. 12, Dec. 2021, Art. no. 2439, doi: 10.3390/sym13122439.
11. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
13. P. Cohen, *Empirical Methods for Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1995.
14. D. F. Polit and C. T. Beck, "Generalization in quantitative and qualitative research: Myths and strategies," *Int. J. Nurs. Stud.*, vol. 47, no. 11, pp. 1451–1458, Nov. 2010, doi: 10.1016/j.ijnurstu.2010.06.004.
15. D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and Prisma Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, Jul. 2009, Art. no. e1000097, doi: 10.1371/journal.pmed.1000097.
16. J. B. Bernerth and H. Aguinis, "A critical review and best-practice recommendations for control variable usage," *Personnel Psychol.*, vol. 69, no. 1, pp. 229–283, Spring 2016, doi: 10.1111/peps.12103.
17. M. Szymanski, M. Millecamp, and K. Verbert, "Visual, textual or hybrid: The effect of user expertise on different explanations," in *Proc. 26th Int. Conf. Intell. User Interfaces (IUI)*, 2021, pp. 109–119, doi: 10.1145/3397481.3450662.

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.

18. Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, and A. Dhurandhar, "Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable AI," in *Proc. AAAI Conf. Human Comput. Crowdsourcing*, 2022, vol. 10, no. 1, pp. 147–159, doi: 10.1609/hcomp.v10i1.21995.
19. D. J. Simons, D. J. Y. Shoda, and D. Lindsay, "Constraints on generality (COG): A proposed addition to all empirical papers," *Perspectives Psychol. Sci.*, vol. 12, no. 6, pp. 1123–1128, Nov. 2017, doi: 10.1177/1745691617708630.
20. U. Peters, A. Krauss, and O. Braganza, "Generalization bias in science," *Cogn. Sci.*, vol. 46, no. 9, Sep. 2022, Art. no. e13188, doi: 10.1111/cogs.13188.

UWE PETERS is a postdoctoral researcher at the Leverhulme Centre for the Future of Intelligence, University of Cambridge, CB2 1SB, Cambridge, United Kingdom, and the Center for Science and Thought, University of Bonn, Bonn, Germany. His research interests include explainable artificial intelligence, algorithmic bias, and scientific methodology. Peters received his Ph.D. degree in philosophy from King's College London. Contact him at up228@cam.ac.uk.

MARY CARMAN is a lecturer in philosophy at the University of Witwatersrand, Johannesburg, 2050, South Africa. Her research interests include ethical and inclusive artificial intelligence research and application. Carman received her Ph.D. degree in philosophy from King's College London. Contact her at mary.carman@wits.ac.za.



IEEE Security & Privacy magazine provides articles with both a practical and research bent by the top thinkers in the field.

- stay current on the latest security tools and theories and gain invaluable practical and research knowledge,
- learn more about the latest techniques and cutting-edge technology, and
- discover case studies, tutorials, columns, and in-depth interviews and podcasts for the information security industry.

computer.org/security

IEEE COMPUTER SOCIETY

IEEE

Authorized licensed use limited to: University of Witwatersrand. Downloaded on April 12,2024 at 09:07:13 UTC from IEEE Xplore. Restrictions apply.