Qualitative investigation in explainable artificial intelligence: A bit more insight from social science

Adam J. Johs | Denise E. Agosto | Rosina O. Weber

[1]Department of Information Science,
College of Computing and Informatics,
Drexel University, Philadelphia,
Pennsylvania, USA

Correspondence
Adam J. Johs, Department of Information
Science, College of Computing and
Informatics, Drexel University, Philadelphia,
PA 19104, USA. Email: ajj37@drexel.edu

Abstract

We present a focused analysis of user studies in explainable artificial intelligence (XAI) entailing qualitative investigation. We draw on social science corpora to suggest ways for improving the rigor of studies where XAI researchers use observations, interviews, focus groups, and/or questionnaires to capture qualitative data. We contextualize the presentation of the XAI papers included in our analysis according to the components of rigor described in the qualitative research literature: 1) underlying theories or frameworks, 2) methodological approaches, 3) data collection methods, and 4) data analysis processes. The results of our analysis support calls from others in the XAI community advocating for collaboration with experts from social disciplines to bolster rigor and effectiveness in user studies.

KEYWORDS:
qualitative investigation, explainable artificial intelligence, user studies, rigor

1 | INTRODUCTION

One impetus for advancing the field of explainable artificial intelligence (XAI) has been the incorporation of insights from social science corpora into XAI research [1,2,3,4]. Cognitive scientists with expertise in cognitive psychology have underscored the nuance of how explanations can be used in XAI [5]. Teams of social scientists are serving as pillars to XAI research programs situated in national defense agencies [6,7,8]. Findings synthesized from social science literature now offer a foundation for researchers to inform investigation in XAI [9]; experts have also speculated on XAI systems occupying key social capacities in the future [10].

This paper extends the stream of XAI literature composed of papers from authors who have drawn from the social sciences. Our chief contribution is a focused analysis of the XAI literature on qualitative investigation. In a vein akin to Miller et al. [1], this work is not intended as an exhaustive literature review, but a focused analysis of papers to spotlight the current opportunities in XAI that can be capitalized on to enrich user studies. 'User studies' is a broad term encompassing a gamut of evaluations, like experiments and field studies, where users are involved [11]. We use 'user studies' also in reference to studies conducted for purposes other than evaluation (*e.g.*, exploratory investigation, for discovery, to collect requirements). Coupled with the knowledge drawn from the social science literature, we argue the findings presented in this paper are critical to catalyzing systematic qualitative investigation in XAI; our ultimate aim is to bolster the rigor and effectiveness of user studies in XAI.

Our motivation aligns with arguments raised by others in the XAI community regarding user studies [9,6,12,13]. Of note, Payrovnaziri et al. [12] underline the importance for interdisciplinary cooperation (between AI researchers and medical professionals) toward maximizing the effectiveness of XAI methods in the context of medicine. In their description of a recent multinational, interdisciplinary workshop on explainable machine learning, Bhatt et al. [13] punctuate the need for broad collaboration with experts from social disciplines if the field of XAI is to ensure rigor and effectiveness in user evaluations.

We marry the perspectives of Payrovnaziri et al.[12] and Bhatt et al.[13] with the correlative outlook of researchers not fixed in the traditional bounds of AI. As a human-computer interaction (HCI) expert, Xu[14] emphasizes the importance of spurring interdisciplinary collaboration in XAI to assure rigor in experiments involving users and validity in research protocols. We share the sentiments of Payrovnaziri et al.[12], Bhatt et al.[13], and Xu[14], drawing attention to the qualitative research literature in hopes of galvanizing a systematic approach to qualitative investigation in XAI.

2 | QUALITATIVE RESEARCH: BACKGROUND, DATA, AND METHODS

2.0.1 | Background

There are numerous examples in the AI literature of researchers using 'qualitative' to validly describe analyses that extend beyond quantitative aspects[15,16,17,18,19,20,21,22,23,24,25,26,27]. In qualitative research, however, 'qualitative' refers to a naturalistic orientation to the world used to study the human facets of a topic—namely, the meanings brought by research participants to the phenomena under investigation[28,29,30].

2.0.2 | Qualitative Data

Broadly distinguishing between quantitative data and qualitative data, the former is typically numerical, whereas the latter is typically nonnumerical[31]. Many qualitative researchers also study phenomena quantitatively, for example, using statistical analyses to describe data, and as a means for determining consistency of findings across data collection methods (*e.g.*, interviews vis-a-vis questionnaires)[32]. In qualitative research, many data analysis processes involve some form of 'coding'—a fundamental data analysis technique researchers use to attribute meaning to data to support categorization, pattern identification, and/or theory development[33,34,29]. Coding is a cyclical process encompassing 30+ coding methods researchers can select from to analyze qualitative data[34]; crucially, 'code' in qualitative data analysis is not to be conflated with 'code' as is used in semiotics, nor with 'code' as is used in computer programming.

2.0.3 | Qualitative Methods

As XAI studies may adopt a quantitative, qualitative, or mixed-methods design[35], we limit our focus to XAI papers where *some* degree of qualitative design was adopted by the authors as part of user studies. In line with Mahoney[36], we delimit the scope of our analysis (Section 3) to XAI papers where authors incorporated any form of the three most common methods used for qualitative investigation: 1) observations, 2) interviews, and/or 3) focus groups.

Observations are used in qualitative investigation to inform understanding of research context, affording researchers the opportunity to collect data on the processes, participant behaviors, or programs under study[36]. Observations can be conducted in artificial (*e.g.*, laboratory) settings, but generally occur in natural settings to capture data as data exist in the real-world[37]. Observations afford researchers a means to learn about aspects that may not arise in interviews or focus groups due to the reticence and/or unawareness of participants[36].

Interviews are used to capture insights directly from research participants[36], conducted in-person, via telephone, or on the Web[38]. Interview formats can be structured (rigid), semi-structured (flexible), or unstructured (free-form)[39]. Structured interviews are useful for evaluating systems when a streamlined comparison of responses from multiple users is required; unstructured interviews are useful for obtaining deep insights and affording participants a chance to explore topics they (participants) deem significant; and semi-structured interviews offer utility to researchers who are unfamiliar with a domain or user group and are interested in obtaining deep insights from participants, absent the unbounded nature of unstructured interviews, and absent the limiting nature of structured interviews[39]. Most interviews in qualitative research are semi-structured, as the flexibility of semi-structured interviews afford a versatile happy medium (between the structured and unstructured formats) for data collection[38]. Interviewing offers researchers utility for informing various research activities including initial exploration and requirements gathering[39].

Focus groups, sometimes referred to as 'group interviewing'[40], employ aspects of qualitative interviewing to elicit (otherwise undiscoverable) data from participants by leveraging the nature of group dynamics[36]. The social influences amongst focus group participants can yield insights that would not organically emerge in individual settings[41]. Thoughtful facilitation of focus groups is required not only to capture useful data, but to moderate conflict, stimulate engagement, sustain conversation, manage time,

and navigate through interview guides[39]. Focus groups can be conducted virtually and generally encompass 4–10 participants, lasting no longer than 2 hours[42].

There is some overlap between observations, interviews, focus groups and a fourth research method germane to XAI (as evidenced by Hoffman et al.[6]): questionnaires; our usage of 'questionnaire' is synonymous with the usage of 'survey' in the context of research methods (corresponding to the usage of Lazar et al.[39]). Per Lazar et al.[39], the probabilistic vs. non-probabilistic structure, number of responses obtained, and inclusion of open-ended vs. close-ended questions determine 1) the type of data captured via questionnaires (*i.e.*, quantitative data, qualitative data, or both quantitative data and qualitative data) and 2) how the questionnaire data are analyzed. Open-ended questions yield qualitative data and must be qualitatively coded as part of analysis[39].

Researchers often collect qualitative data within experimental designs toward understanding how participants experienced variables (*e.g.*, interventions)—this is a form of mixed methods study design termed 'embedded mixed methods design'[43]. We refer to the questionnaires disseminated or administered to respondents, and the tasks where participants textually answer open-ended questions pertaining to experimental variables, as 'questionnaire tasks.' Accordingly, as part of our analysis (Section 3), we also include XAI papers where any variant of a questionnaire task was employed during a user study to capture qualitative data.

3 | FOCUSED ANALYSIS OF QUALITATIVE INVESTIGATION IN XAI

Myriad criteria and perspectives have been put forward to guide how quality in qualitative research should be approached[44,29,30,45,46,47,48,49,50,51]. Though a synthesis of the different viewpoints on qualitative research quality is out of scope in this paper, we leverage the discussions of rigor as a fundamental criterion of qualitative research quality to contextualize presentation of the XAI papers included in our analysis.

Rigor as a criterion for qualitative research quality can be characterized as the clear description of steps in one's research, including, appropriateness of the selected data collection method[52]; Tracy[46] cites the proper utilization of theoretical constructs, data, samples, contexts, and data analysis processes as components of rigor. Hence, we conducted our analysis to illustrate the underlying theories or frameworks, qualitative data collection methods, and qualitative data analysis processes used thus far by XAI researchers as part of user studies. And because it is encouraged to use best practices of qualitative methodology as a language to dialogue with quantitatively-grounded research communities[46], we also include qualitative methodological approaches in our analysis.

To begin gathering what has been contributed to the XAI literature on qualitative investigation, we employed a two-pronged approach: 1) examining the findings of recent XAI reviews and 2) exploring the papers published in XAI venues. Myriad reviews were examined[9,7,53,12,54,55,56,57,58,59,60,61,22,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79]. The papers explored were from (including but not limited to) the International Joint Conference on Artificial Intelligence (IJCAI); Association for the Advancement of Artificial Intelligence (AAAI) Conference on AI; ACM Conference on Intelligent User Interfaces (IUI); International Conference on Case-Based Reasoning (ICCBR); International Conference on Autonomous Agents and Multiagent Systems (AAMAS); Conference on Computer Vision and Pattern Recognition (CVPR); Conference on Neural Information Processing Systems (NIPS); and Association for Computing Machinery (ACM)/AAAI conference on Artificial Intelligence, Ethics, and Society (AIES). We prioritized our analysis on AI and computer science venues.

Any reference of qualitative investigation in the XAI papers found via the aforementioned reviews or venues were examined for inclusion in our analysis (as per the contextualization described at the beginning of this section). We also submitted various search queries to Google Scholar (GS), our university libraries, and the DuckDuckGo Internet search engine to expand the scope of our analysis (for example, *qualitative investigation AND explainable AI*, *qualitative investigation AND XAI*, *qualitative explainable artificial intelligence*, *qualitative evaluation AND explainable artificial intelligence*, *qualitative evaluation AND explainable AI*, *user evaluation AND explainable AI*, *explainable AI user evaluation*, *qualitative AND XAI*). Given[29] was referenced to guide our determination of whether a methodological approach, data collection method, or data analysis process used by the authors in any of the retrieved papers warranted inclusion in our analysis. The human-centered XAI papers compiled by Mueller et al.[7] and Vilone and Longo[53] served as valuable bases for our efforts.

Table 1 depicts the results of our analysis. Dashes are used to denote the components of qualitative research rigor not discussed in the reviewed papers. Only underlying theories or frameworks used to inform the design of user studies were included. Duplicate rows were created for papers entailing more than one qualitative data collection method (as described in Section 2).

PAPER	CITED BY	UNDERLYING THEORY OR FRAMEWORK	METHODOLOGICAL APPROACH	DATA COLLECTION METHOD	DATA ANALYSIS PROCESS
Alshehri et al.[80]	*Not in GS*	Folk Theory of Mind and Behavior[81]	—	Questionnaire Task	Thematic Analysis
Conati et al.[82]	*Not in GS*	Framework for User Modeling and Adaptation[83]	—	Questionnaire Task	—
Doyle et al.[84] (*Evaluation Part 1, Evaluation Part 2, and Counter Example Evaluation*)	49	Explanation utility framework[85]	—	Questionnaire Task	—
Dzindolet et al.[86] (*Study 2*)	299	Framework to predict automation use[87]	—	Questionnaire Task	Content Analysis
Dzindolet et al.[88] (*Study 2*)	958	—	—	Questionnaire Task	Content Analysis
Eisenstadt et al.[89]	8	Explanation patterns framework[90]	—	Questionnaire Task	—
Hall et al.[91]	20	Conceptual framework characterizing explainability[92]	—	Semi-Structured Interviews	—
Herlocker et al.[93] (*Experiment 2*)	2088	—	—	Questionnaire Task	—
Huber et al.[94]	*Not in GS*	—	—	Questionnaire Task	Content Analysis (Summative)
Kim et al.[95] (*Pilot Study*)	23	—	—	Questionnaire Task	—
Kim et al.[96] (*Pilot Study*)	92	—	—	Questionnaire Task	—
Krause et al.[97]	277	—	Case Study	Interviews (*format unspecified*)	Exploratory Data Analysis
Kulesza et al.[98]	200	—	Grounded Theory	Questionnaire Task	Coding (*method(s) unspecified*)
Lakkaraju and Bastani[99] (*Human Evaluation of Trust in Black Box*)	62	Model Understanding through Subspace Explanations (MUSE) framework[100]	—	Questionnaire Task	—
Pacer et al.[101] (*Experiment 1*)	35	Various explanation models associated with probability theory, notably,[102]	—	Questionnaire Task	Coding (*method(s) unspecified*)
Putnam and Conati[103]	27	User modeling framework[83]	—	Questionnaire Task	—
Putnam and Conati[103]	27	User modeling framework[83]	—	Interviews (*format unspecified*)	—
Ribeiro et al.[104] (*Husky vs. Wolf task*)	6485	—	—	Questionnaire Task	—
Silva et al.[105]	*Not in GS*	—	—	Questionnaire Task	—
Singh et al.[106]	3	—	—	Questionnaire Task	—
Tonekaboni et al.[107]	87	—	Mixed Methods[108]	Interviews (*format unspecified*)	—
Tullio et al.[109]	153	Theory of developing mental models for expert users[110]	Field Research	Semi-Structured Interviews	Coding (*method(s) unspecified*)
Williams et al.[111]	126	Learnersourcing framework[112]	—	Questionnaire Task	—
Williams et al.[111]	126	Learnersourcing framework[112]	—	Semi-Structured Interview	—

TABLE 1 A depiction of the papers in the XAI literature where some form of qualitative investigation was employed as part of a user study, along with the corresponding components of qualitative research rigor discussed in each paper

And for any papers detailing more than one study, the specific study where qualitative investigation was employed is included next to the citation in the corresponding row. To afford the reader a sense of how much attention has been directed to the papers included in our analysis, we also inserted a column of citation (cited by) counts found in GS; citation counts are current as of Sep. 20, 2021, and not every paper listed had a record in GS at the time of compilation (denoted with *'Not in GS'* in Table 1).

4 | DISCUSSION

4.0.1 | Findings

The most common data collection method used by XAI researchers in the papers reviewed was instantiations of the question-naire task. The underlying theories and frameworks used to inform the design of user studies spanned various domains—*e.g.*, psychology, AI, human-centered computing. Many authors did not specify the format of interviews and/or method(s) of coding used—key details that influence interpretation of results from qualitative investigation. We did not see use of observations or focus groups (group interviewing) for data collection in any of the XAI papers included in our analysis.

As the papers in our analysis were assessed for inclusion based on whether each of the following elements were satisfied:

1. Some degree of *qualitative study design* was adopted by the authors

2. *Studies with users* were conducted

3. *Observations, interviews, focus groups, or questionnaires* were used to capture qualitative data

notable papers were excluded from our analysis and the depiction in Table 1. The work of Madumal et al.[113] was excluded because grounded theory and coding were used to derive a dialogue framework from existing data—not as part of a user study—and the data yielded from the Wizard of Oz experiment[114] were analyzed quantitatively.

Kulesza et al.[115] used a think aloud protocol to conduct a study where participants debugged an intelligent assistant. Bunt et al.[116] used semi-structured interviews and diaries, informed by a contextual inquiry approach, to investigate the explanations of low-cost intelligent interactive systems (IIS). *Diaries*[39] and the *think aloud protocol*[117] are additional methods for capturing qualitative data beyond the scope of this paper, as is the design approach of contextual inquiry[118].

Zhou et al.[119] leveraged the user-centered design process as part of designing an XAI system. Kulesza et al.[120] and Shinsel et al.[121] both used programming-oriented approaches to inform qualitative investigation. The *user-centered design* process[122] and computer programming methodologies also extend beyond the qualitative research corpora we are drawing from in this paper.

4.0.2 | Recommendations for the XAI Community

The results of our analysis underline the necessity for improving the rigor of qualitative investigation employed in XAI as part of user studies. Several of the papers analyzed lacked details required to assess qualitative research rigor. When presenting interview-based studies, investigators must specify the type of interviews conducted (structured, semi-structured, unstructured, or focus group interviewing); an overview of the interview guides used, including the main interview questions asked; and clear representations of the data collected, including data that supported major findings, and data that contradicted major findings.

Many of the papers in our analysis also lacked details describing the methodological approaches and data analysis processes employed. For example, most qualitative researchers across disciplines use thematic analysis and inductive coding as data analysis processes for qualitative investigation. Importantly, 'inductive coding' only describes the process for analyzing qualitative data, and, in and of itself, thematic analysis does not provide a theoretical basis for analysis. 'Thematic analysis' also does not explain how themes are identified, how themes evolve as coding progresses, or connections between the research questions and coding approaches used. Researchers must be explicit with the data analysis processes employed, ensuring analyses are clearly tied to research questions. Per Braun and Clarke[123], the interpretative power afforded in using thematic analysis is limited to description if not applied within a theoretical framework that grounds analytic claims.

Grounded theory is a widely used methodological approach, based on induction and comparison, used in qualitative investigation to construct theory[124,125]. Because the goal of grounded theory is theory development, researchers must ensure to not claim utilization of grounded theory when only using methods of coding regarded as part of—but not exclusive to—the coding canon of grounded theory: *in vivo* coding, process coding, initial (open) coding, focused coding, axial coding, and/or theoretical (selective) coding[34]. In contrast to experimental research, grounded theory begins from empirical observations or data, not preformed theories and hypotheses[39].

Indeed, promising avenues for qualitative research in XAI exist, particularly, with respect to interviewing. The Defense Advanced Research Projects Agency (DARPA) cites interviewing as a valuable method for studying mental models in XAI, namely, by combining tasks where users predict the decisions of an intelligent system with post-experimental interviews[7]. Structured interviewing, in the form of a retrospection task with question-answering, is another example of a method that could prove useful for mental model elicitation in XAI[6]. Others have used interviewing as part of research aimed at improving user

trust in intelligible, context-aware machine learning applications[126]. A lack of user trust is a key impediment to the uptake of AI applications[9]. When studying how satisfied users are with explanations in XAI, 'trustworthiness' is a core attribute used for measurement[6]. Thus, we encourage XAI researchers to consider interviewing as a means to further investigate user trust in XAI.

Rigor is but one dimension of quality in qualitative research. Many other aspects also represent quality in qualitative investigation not illustrated as part of this analysis (*c.f.*, Tracy[46]). As techniques abound for capturing qualitative data, many of which could not be covered in this paper (*e.g.*, the think aloud protocol, diaries, user-centered design process), we contend non-experts in qualitative research should not be encumbered with the additional burden of designing, conducting, and analyzing the results of qualitative investigations in XAI. Ensuring rigor in the user studies conducted in XAI entailing qualitative investigation will promote cross-pollination with social experts outside the XAI community.

Echoing our sentiment from Section 1: we underscore the standpoints of Payrovnaziri et al.[12], Bhatt et al.[13], and Xu[14] and call for the XAI community to collaborate with experts from social disciplines toward bolstering rigor and effectiveness in user studies. The references in this paper afford a foundational basis for XAI researchers to become oriented with qualitative research and facilitate constructive dialogue with experts from social disciplines in support of collaboration. The XAI community is primed for a multidisciplinary convergence on user studies toward the benefit of increasing the trust users place in XAI systems. We hope this analysis will motivate some of the efforts required to converge on such multidisciplinary collaboration.

ACKNOWLEDGMENTS

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Miller T, Howe P, Sonenberg L. Explainable AI: Beware of Inmates Running the Asylum. In: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI 2017); 2017; Melbourne, Australia.

2. Hoffman RR, Klein G. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems* 2017; 32(3): 68-73. doi:10.1109/MIS.2017.54.

3. Hoffman RR, Mueller ST, Klein G. Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intelligent Systems* 2017; 32(4): 78-86. doi: 10.1109/MIS.2017.3121544

4. De Graaf MM, Malle BF. How People Explain Action (And Autonomous Intelligent Systems Should Too). In: 2017 AAAI Fall Symposium Series, Artificial Intelligence for Human-Robot Interaction; 2017. AAAI Technical Report FS-17-01.

5. Byrne RMJ. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence From Human Reasoning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; 2019. doi.org/10.24963/ijcai.2019/876.

6. Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for Explainable AI: Challenges and Prospects. arXiv preprint. arXiv:1812.04608 [cs.AI]. Ithaca, NY: Cornell University Library; 2018.

7. Mueller ST, Hoffman RR, Clancey WJ, Emrey A, Klein G. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. arXiv preprint. arXiv:1902.01876 [cs.AI]. Ithaca, NY: Cornell University Library; 2019.

8. Gunning D, Aha DW. DARPA's Explainable Artificial Intelligence Program. *AI Magazine* 2019; 40(2): 44-58.

9. Miller T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 2019; 267: 1-38. doi.org/10.1016/j.artint.2018.07.007.

10. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI—Explainable Artificial Intelligence. *Science Robotics* 2019; 4(37): eaay7120. doi.org/10.1126/scirobotics.aay7120.

11. Preece J, Sharp H, Rogers Y. *Interaction Design: Beyond Human-Computer Interaction (5th ed.).* Indianapolis, IN: John Wiley and Sons . 2019.

12. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable Artificial Intelligence Models Using Real-World Electronic Health Record Data: A Systematic Scoping Review. *Journal of the American Medical Informatics Association, 0(0):* 2020: 1-13. doi.org/10.1093/jamia/ocaa053.

13. Bhatt U, Andrus M, Weller A, Xiang A. Machine Learning Explainability for External Stakeholders. In: Proceedings of the IJCAI-PRICAI 2020 Workshop on eXplainable Artificial Intelligence (XAI 2019; 2020.

14. Xu W. Toward Human-Centered AI: A Perspective From Human-Computer Interaction. *interactions* 2019; 26(4): 42-46. doi.org/10.1145/3328485.

15. Lapuschkin S, Binder A, Montavon G, Muller KR, Samek W. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016

16. Forbus KD. Qualitative Process Theory. *Artificial intelligence* 1984; 24(1-3): 85-168. doi.org/10.1016/0004-3702(84)90038-9.

17. Druzdzel MJ. Qualitative Verbal Explanations in Bayesian Belief Networks. *Artificial Intelligence and Simulation of Behaviour Quarterly* 1996; 94: 43-54.

18. Howell I, Woodward R, Choueiry BY, Yu H. A Qualitative Analysis of Search Behavior: A Visual Approach. In: Proceedings of the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI); 2018.

19. Clark P, Matwin S. Using Qualitative Models to Guide Inductive Learning. In: Proceedings of the 1993 International Conference on Machine Learning; 1993.

20. Rizzo L, Longo L. A Qualitative Investigation of the Degree of Explainability of Defeasible Argumentation and Non-Monotonic Fuzzy Reasoning. In: Proceedings of the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science; 2018.

21. Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J. Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016); 2016.

22. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. arXiv preprint. arXiv:1806.00069 [cs.Stat]. Ithaca, NY: Cornell University Library; 2018.

23. Park DH, Hendricks LA, Akata Z, et al. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018

24. Wu T, Sun W, Li X, Song X, Li B. Towards Interpretable R-CNN by Unfolding Latent Structures.. arXiv preprint. arXiv:1711.05226 [cs.CV]. Ithaca, NY: Cornell University Library; 2017.

25. Koul A, Greydanus S, Fern A. Toward Learning Finite State Representations of Recurrent Policy Networks. In: Proceedings of the IJCAI/ECAIWorkshop on Explainable Artificial Intelligence (XAI); 2018.

26. Ustun B, Rudin C. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning* 2016; 102(3): 349-391. doi: 10.1007/s10994-015-5528-6

27. Kwon BC, Choi MJ, Kim JT, et al. Retainvis: Visual analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization And Computer Graphics* 2018; 25(1): 299-309. doi: 10.1109/TVCG.2018.2865027

28. Lincoln YS, Guba EG. *Naturalistic Inquiry*. Beverly Hills, CA: SAGE Publications, Inc . 1985.

29. Given LM. *The SAGE Encyclopedia of Qualitative Research Methods*. Thousand Oaks, CA: SAGE Publications, Inc . 2008

30. Atkinson P, Delamont S. *SAGE Qualitative Research Methods*. Thousand Oaks, CA: SAGE Publications, Inc . 2010

31. Schreiber JB. Data. In: Given LM. , ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 186). dx.doi.org/10.4135/9781412963909.

32. Donmoyer R. Quantitative Research. In: Given LM. , ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 714-718). dx.doi.org/10.4135/9781412963909.

33. Miles MB, Huberman AM, Saldana J. *Qualitative Data Analysis: A Methods Sourcebook (4th ed.)*. Thousand Oaks, CA: SAGE Publications, Inc . 2020.

34. Saldana J. *The Coding Manual for Qualitative Researchers (2nd ed.)*. London, UK: SAGE Publications, Ltd . 2013.

35. Chromik M, Schuessler M. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. In: Proceedings of the IUI Workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20); 2020.

36. Mahoney C. Common Qualitative Methods. In: Frechtling J, Sharp L., eds. *The User-Friendly Handbook for Mixed Method Evaluations* 1997.

37. McKechnie LEF. Observational Research. In: Given LM. , ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 574-576). dx.doi.org/10.4135/9781412963909.

38. Brinkmann S. Interviewing. In: Given LM., ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 471-472). dx.doi.org/10.4135/9781412963909.

39. Lazar J, Feng J, Hochheiser H. *Research Methods in Human-Computer Interaction (2nd ed.)*. Cambridge, MA: Morgan Kaufmann, an imprint of Elsevier . 2017.

40. Morgan DL. Focus Groups. In: Given LM., ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 353-354). dx.doi.org/10.4135/9781412963909.

41. Morgan DL. *Qualitative Research Methods: Focus Groups as Qualitative Research*. Thousand Oaks, CA: SAGE Publications, Inc . 1997

42. Liamputtong P. *Focus Group Methodology: Principles and Practice*. London, UK: SAGE Publications, Ltd . 2011

43. Creswell JW. Mixed Methods Research. In: Given LM. , ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 527-529). dx.doi.org/10.4135/9781412963909.

44. Guba E. Criteria for Assessing the Trustworthiness of Naturalistic Inquiries. *Educational Communication and Technology* 1981; 29(2): 75-91. doi.org/10.1007/BF02766777.

45. Flick U. *Qualitative Research Kit: Managing Quality in Qualitative Research*. London, UK: SAGE Publications, Ltd . 2007

46. Tracy SJ. Qualitative Quality: Eight "Big-Tent" Criteria For Excellent Qualitative Research. *Qualitative Inquiry* 2010; 16(10): 837-851. doi.org/10.1177/1077800410383121.

47. Flick U. *The SAGE Handbook of Qualitative Data Collection*. London, UK: SAGE Publications, Ltd . 2018

48. Mays N, Pope C. Qualitative Research in Health Care: Assessing Quality in Qualitative Research. *BMJ* 2000; 320(7226): 50-52. doi.org/10.1136/bmj.320.7226.50.

49. Dingwall R, Murphy E, Watson P, Greatbatch D, Parker S. Catching Goldfish: Quality in Qualitative Research. *Journal of Health Services Research and Policy* 1998; 3(3): 167-172. doi.org/10.1177/135581969800300308.

50. Seale C. Quality in Qualitative Research. *Qualitative Inquiry* 1999; 5(4): 465-478. doi.org/10.1177%2F107780049900500402.

51. Patton MQ. Qualitative Evaluation Checklist. *Evaluation Checklists Project* 2003; 21: 1-13.

52. Saumure K, Given LM. Rigor in Qualitative Research. In: Given LM., ed. *The SAGE Encyclopedia of Qualitative Research Methods*Thousand Oaks, CA: SAGE Publications, Inc. 2008 (pp. 796). dx.doi.org/10.4135/9781412963909.

53. Vilone G, Longo L. Explainable Artificial Intelligence: A Systematic Review. arXiv preprint. arXiv:2006.00093 [cs.AI]. Ithaca, NY: Cornell University Library; 2020.

54. Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion* 2020; 58: 82-115. doi.org/10.1016/j.inffus.2019.12.012.

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

56. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 2018; 51(5): 1-42. doi.org/10.1145/3236009.

57. Lipton ZC. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *ACM Queue* 2018; 16(3): 1-28. doi.org/10.1145/3236386.3241340.

58. Keane MT, Kenny EM. How Case-Based Reasoning Explains Neural Networks: A Theoretical Analysis of XAI Using Post-Hoc Explanation-by-Example from a Survey of ANN-CBR Twin-Systems. In: Proceedings of the International Conference on Case-Based Reasoning (ICCBR 2019); 2019. doi.org/10.1007/978-3-030-29249-2_11.

59. Biran O, Cotton C. Explanation and Justification in Machine Learning: A Survey. In: Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI 2017); 2017.

60. Chakraborty S, Tomsett R, Raghavendra R, et al. Interpretability of Deep Learning Models: A Survey of Results. In: Proceedings of 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation; 2017. doi.org/10.1109/UIC-ATC.2017.8397411.

61. Doshi-Velez F, Kim B. Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint. arXiv:1702.08608 [stat.ML]. Ithaca, NY: Cornell University Library; 2017.

62. Alonso JM, Castiello C, Mencar C. A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field. In: Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems; 2018.

63. Anjomshoae S, Najjar A, Calvaresi D, Framling K. Explainable Agents and Robots: Results From a Systematic Literature Review. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019); 2019.

64. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In: Proceedings of the SIGCHI Conference on Human Factors in Computing System; 2018. doi.org/10.1145/3173574.3174156.

65. Ribera M, Lapedriza A. Can we do Better Explanations? A Proposal of User-centered Explainable AI. In: Joint Proceedings of the ACM IUI 2019 Workshops; 2019.

66. Došilović FK, Brčić M, Hlupić N. Explainable Artificial Intelligence: A Survey. In: Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); 2018. doi.org/10.23919/MIPRO.2018.8400040.

67. Israelsen BW, Ahmed NR. Dave... I can assure you... that it's going to be all right...A Definition, Case For, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. *ACM Computing Surveys* 2019; 51(6): 1-37. doi.org/10.1145/3267338.

68. Sørmo F, Cassens J, Aamodt A. Explanation in Case-Based Reasoning–Perspectives and Goals. *Artificial Intelligence Review* 2005; 24(2): 109-143. doi.org/10.1007/s10462-005-4607-7.

69. Roth-Berghofer TR. Explanations and Case-Based Reasoning: Foundational Issues. In: Proceedings of the European Conference on Case-Based Reasoning (ECCBR 2004); 2004. doi.org/10.1007/978-3-540-28631-8_29.

70. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. arXiv preprint. arXiv:1907.07374 [cs.LG]. Ithaca, NY: Cornell University Library; 2019.

71. Puiutta E, Veith E. Explainable Reinforcement Learning: A Survey. arXiv preprint. arXiv:2005.06247. [cs.LG]. Ithaca, NY: Cornell University Library; 2020.

72. Clinciu MA, Hastie H. A Survey of Explainable AI Terminology. In: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019); 2019. doi.org/10.18653/v1/W19-8403.

73. Longo L, Goebel R, Lecue F, Kieseberg P, Holzinger A. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In: Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction; 2020. doi.org/10.1007/978-3-030-57321-8_1.

74. Mathews SM. Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In: Proceedings of the 2019 Computing Conference; 2019. doi.org/10.1007/978-3-030-22868-2_90.

75. Zhang Q, Zhu S. Visual Interpretability for Deep Learning: A Survey. *Frontiers of Information Technology & Electronic Engineering* 2018; 19: 27-39. doi.org/10.1631/FITEE.1700808.

76. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. arXiv preprint. arXiv:1710.00794 [cs.AI]. Ithaca, NY: Cornell University Library; 2017.

77. Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In: Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing; 2019.

78. Cui X, Lee JM, Hsieh JPA. An Integrative 3C Evaluation Framework for Explainable Artificial Intelligence. In: Proceedings of the Americas Conference on Information Systems (AMCIS); 2019.

79. Tintarev N, Masthoff J. A Survey of Explanations in Recommender Systems. In: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop; 2007

80. Alshehri A, Miller T, Vered M, Alamri H. Human Centered Explanation for Goal Recognition. In: Proceedings of the IJCAI-PRICAI 2020 Workshop on eXplainable Artificial Intelligence (XAI 2019); 2020.

81. Tate C, Malle BF. Explaining the Past, Predicting the Future. In: New York: Oxford University Press. 2006 (pp. 182-209). doi:10.1093/acprof:oso/9780195177664.003.0011.

82. Conati C, Barral O, Putnam V, Rieger L. Impact of Explanations for AI-Driven Hints in an Intelligent Tutoring System. In: Proceedings of the IJCAI-PRICAI 2020 Workshop on eXplainable Artificial Intelligence (XAI 2019; 2020.

83. Kardan S, Conati C. Providing Adaptive Support in an Interactive Simulation for Learning: An Experimental Evaluation. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; 2015. doi.org/10.1145/2702123.2702424.

84. Doyle D, Cunningham P, Walsh P. An Evaluation of the Usefulness of Explanation in a Case-Based Reasoning System for Decision Support in Bronchiolitis Treatment. *Computational Intelligence* 2006; 22(3-4): 269–281. doi.org/10.1111/j.1467-8640.2006.00288.x.

85. Doyle D, Cunningham P, Bridge D, Rahman Y. Explanation Oriented Retrieval. In: Proceedings of the European Conference on Case-Based Reasoning (ECCBR 2004); 2004. doi.org/10.1007/978-3-540-28631-8_13.

86. Dzindolet MT, Pierce LG, Beck HP, Dawe LA. The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors* 2002; 44(1): 79-94. doi.org/10.1518/0018720024494856.

87. Dzindolet M, Beck H, Pierce L, Dawe L. A Framework of Automation Use (Rep. No. ARL-TR-2412). Army Research Laboratory, Aberdeen Proving Ground, MD; 2001.

88. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies* 2003; 58(6): 697-718. doi.org/10.1016/S1071-5819(03)00038-7.

89. Eisenstadt V, Espinoza-Stapelfeld C, Mikyas A, Althoff KD. Explainable Distributed Case-Based Support Systems: Patterns for Enhancement and Validation of Design Recommendations. In: Proceedings of the International Conference on Case-Based Reasoning; 2018. doi.org/10.1007/978-3-030-01081-2_6.

90. Cassens J, Kofod-Petersen A. Designing Explanation Aware Systems: The Quest for Explanation Patterns. In: Proceedings of the 2007 AAAI Workshop on Explanation-Aware Computing; 2007.

91. Hall M, Harborne D, Tomsett R, et al. A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems. In: Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019); 2019.

92. Bohlender D, Köhl MA. Towards a Characterization of Explainable Systems. arXiv preprint. arXiv:1902.03096 [cs.AI]. Ithaca, NY: Cornell University Library; 2019.

93. Herlocker JL, Konstan JA, Riedl J. Explaining Collaborative Filtering Recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work; 2000. doi.org/10.1145/358916.358995.

94. Huber T, Weitz K, Andre E, Amir O. Combining Local Saliency Maps and Global Strategy Summaries for Reinforcement Learning Agents. In: Proceedings of the IJCAI-PRICAI 2020 Workshop on eXplainable Artificial Intelligence (XAI 2019); 2020.

95. Kim B, Glassman E, Johnson B, Shah J. iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction. tech. rep., Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2015-010. Cambridge, MA; : 2015.

96. Kim B, Shah JA, Doshi-Velez F. Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. In: Proceedings of the Advances in Neural Information Processing Systems; 2015.

97. Krause J, Perer A, Ng K. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; 2016. dx.doi.org/10.1145/2858036.2858529.

98. Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I, Wong WK. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing; 2013. dx.doi.org/10.1109/VLHCC.2013.6645235.

99. Lakkaraju H, Bastani O. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020

100. Lakkaraju H, Kamar E, Caruana R, Leskovec J. Faithful and Customizable Explanations of Black Box Models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; 2019

101. Pacer M, Williams J, Chen X, Lombrozo T, Griffiths T. Evaluating Computational Models of Explanation Using Human Judgments. 2013. arXiv preprint. arXiv:1309.6855 [cs.AI]. Ithaca, NY: Cornell University Library.

102. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann Publishers . 1988.

103. Putnam V, Conati C. Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS). In: Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces; 2019.

104. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. doi.org/10.1145/2939672.2939778.

105. Silva CR, Lelis LHS, Bowling M. Teaching Humans with Justifications of Monte Carlo Tree Search Decisions. In: Proceedings of the IJCAI-PRICAI 2020 Workshop on eXplainable Artificial Intelligence (XAI 2019); 2020.

106. Singh A, Balaji JJ, Jayakumar V, Rasheed MA, Raman R, Lakshminarayanan V. Quantitative and Qualitative Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis. arXiv preprint. arXiv:2009.12648 [eess.IV]. Ithaca, NY: Cornell University Library; 2020.

107. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In: Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR 106; 2019.

108. Leech NL, Onwuegbuzie AJ. A Typology of Mixed Methods Research Designs. *Quality & Quantity* 2009; 43(2): 265–275. doi.org/10.1007/s11135-007-9105-3.

109. Tullio J, Dey AK, Chalecki J, Fogarty J. How it Works: A Field Study of Non-Technical Users Interacting with an Intelligent System. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2007. doi.org/10.1145/1240624.1240630.

110. Moray N. Intelligent Aids, Mental Models, and the Theory of Machines. *International Journal of Man-Machine Studies* 1987; 27(5-6): 619-629. doi.org/10.1016/S0020-7373(87)80020-2.

111. Williams JJ, Kim J, Rafferty A, et al. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In: Proceedings of the Third (2016) ACM Conference on Learning @ Scale; 2016. doi.org/10.1145/2876034.2876042.

112. Kim J. *Learnersourcing: Improving Learning with Collective Learner Activity*. PhD thesis. Massachusetts Institute of Technology, ; 2015.

113. Madumal P, Miller T, Sonenberg L, Vetere F. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019); 2019. dl.acm.org/doi/10.5555/3306127.3331801.

114. Dahlbäck N, Jönsson A, Ahrenberg L. Wizard of Oz Studies: Why and How. In: The Proceedings of the 1st International Conference on Intelligent User Interfaces; 1993

115. Kulesza T, Stumpf S, Wong WK, et al. Why-Oriented End-User Debugging of Naive Bayes Text Classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2011; 1(1): 1-31. doi: 10.1145/2030365.2030367

116. Bunt A, Lount M, Lauzon C. Are Explanations Always Important? A Study of Deployed, Low-Cost Intelligent Interactive Systems. In: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces; 2012. doi.org/10.1145/2166966.2166996.

117. Ericsson KA, Simon HA. *Protocol Analysis: Verbal Reports as Data*. The MIT Press . 1984.

118. Holtzblatt K, Beyer H. *Contextual Design: Defining Customer-Centered Systems*. San Francisco, CA: Morgan Kaufmann . 1998.

119. Zhou T, Sheng H, Howley I. Assessing Post-Hoc Explainability of the BKT Algorithm. In: Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES 20); 2020. doi.org/10.1145/3375627.3375856.

120. Kulesza T, Stumpf S, Burnett M, et al. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In: Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing; 2010

121. Shinsel A, Kulesza T, Burnett M, et al. Mini-Crowdsourcing End-User Assessment of Intelligent Assistants: A Cost-Benefit Study. In: Proceedings of the 2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC); 2011

122. Abras C, Maloney-Krichmar D, Preece J. User-Centered Design. In: Bainbridge WS. , ed. *Berkshire Encyclopedia of Human-Computer Interaction*. 2. Great Barrington, Massachusetts: Berkshire Publishing Group LLC. 2004 (pp. 763-768).

123. Braun V, Clarke V. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 2006; 3(2): 77–101. doi: 10.1191/1478088706qp063oa

124. Glaser BG, Strauss AL. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine Publishing . 1967.

125. Bryant A, Charmaz K. *The SAGE Handbook of Grounded Theory*. London, UK: SAGE Publications, Ltd . 2007. dx.doi.org/10.4135/9781848607941.

126. Lim BY. *Improving Understanding and Trust with Intelligibility in Context-Aware Applications*. PhD thesis. , Pittsburgh, PA; 2012. CMU-HCII-12-104.