

Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis

JIALUN AARON JIANG, University of Colorado Boulder, USA
KANDREA WADE, University of Colorado Boulder, USA
CASEY FIESLER, University of Colorado Boulder, USA
JED R. BRUBAKER, University of Colorado Boulder, USA

Qualitative inductive methods are widely used in CSCW and HCI research for their ability to generatively discover deep and contextualized insights, but these inherently manual and human-resource-intensive processes are often infeasible for analyzing large corpora. Researchers have been increasingly interested in ways to apply qualitative methods to “big” data problems, hoping to achieve more generalizable results from larger amounts of data while preserving the depth and richness of qualitative methods. In this paper, we describe a study of qualitative researchers’ work practices and their challenges, with an eye towards whether this is an appropriate domain for human-AI collaboration and what successful collaborations might entail. Our findings characterize participants’ diverse methodological practices and nuanced collaboration dynamics, and identify areas where they might benefit from AI-based tools. While participants highlight the messiness and uncertainty of qualitative inductive analysis, they still want full agency over the process and believe that AI should not interfere. Our study provides a deep investigation of task delegability in human-AI collaboration in the context of qualitative analysis, and offers directions for the design of AI assistance that honor serendipity, human agency, and ambiguity.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; **Collaborative and social computing design and evaluation methods**.

Additional Key Words and Phrases: Qualitative research; interview; AI; human-AI collaboration

ACM Reference Format:

Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 94 (April 2021), 23 pages. <https://doi.org/10.1145/3449168>

1 Introduction

“You’ve got 500 codes first and then you’ve got notes everywhere all over them.” (P05)

Hours of interviews and observations, pages of transcripts and field notes, and the large number of codes, labels, and post-it notes that follow are a familiar sight to any qualitative researcher. Qualitative research, known for its ability to deeply understand rich in-situ practices and experiences,

Authors’ addresses: Jialun Aaron Jiang, University of Colorado Boulder, Department of Information Science, 1045 18th St., Boulder, CO, 80309, USA, aaron.jiang@colorado.edu; Kandrea Wade, University of Colorado Boulder, Department of Information Science, 1045 18th St., Boulder, CO, 80309, USA, kandrea.wade@colorado.edu; Casey Fiesler, University of Colorado Boulder, Department of Information Science, 1045 18th St., Boulder, CO, 80309, USA, casey.fiesler@colorado.edu; Jed R. Brubaker, University of Colorado Boulder, Department of Information Science, 1045 18th St., Boulder, CO, 80309, USA, jed.brubaker@colorado.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2021/4-ART94 \$15.00
<https://doi.org/10.1145/3449168>

is widely used in CSCW and HCI for developing human-centered understandings of computing systems. However, qualitative research is also, by its nature, manual and laborious. While the scale of data generated by methods like interviews and ethnography can be viewed as “small” in our current world of “big data,” it nevertheless can be overwhelming to humans, and requires tremendous human effort to analyze. Other qualitative methods like content analysis, on the other hand, often directly tackle the realm of “big data” by using samples like millions of scraped online posts (e.g., [7]), a scale even more clearly impossible for humans to handle. Therefore, qualitative researchers use methods such as subsampling [36] to reduce the sample size to something humans are capable of handling, forgoing the opportunity to take advantage of the richness offered by the full dataset.

The challenge of qualitatively analyzing big data has led some to consider the potential for artificial intelligence (AI) to assist qualitative researchers. While one may argue that qualitative research and AI are incompatible due to their different epistemological traditions, recent work has discussed the conceptual similarities between machine learning (ML), a technique often used in AI, and grounded theory methods (GTM), and proposed hybrid models that take advantage of both types of methodology [33]. Putting this comparison to the test, one study has even shown empirically that GTM and ML can produce similar results on the same data [4]. There also have been various efforts to integrate ML into qualitative research (e.g., [31, 42]), aiming to automate deductive steps of the qualitative research process—applying established codebooks to the dataset—and reduce human effort. However, the inductive step of generating the codebook through discovering patterns and themes cannot be overlooked. Is it possible that AI can also support the inductive step of qualitative research that finds meaning in unstructured data instead of applying already-found meaning on a larger set of data? While human-AI collaboration seems promising for qualitative research, is it something that qualitative researchers themselves desire? And what abilities and qualities do people seek (or not) in AI that would make it a good collaborator? Answers to this question will not only shed light on how we can better design AI assistance for qualitative researchers, but also provide deeper insights into human-AI collaboration in fields typically seen as human-only. In this paper, we do not seek to examine the efficacy or promise of particular models or algorithms, but to uncover the characteristics and needs unique to qualitative research to which AI designers and developers should pay attention. While qualitative research includes many subprocesses such as data collection and result reporting in addition to data analysis, in this paper, we focus on the analysis process in particular because it shows the most promise for potential AI assistance, as both prior research and our participants suggest.

In this work, we examine the practices of qualitative researchers, and the capabilities that they do (or do not) want assistive systems to have, through an analysis of 17 semi-structured interviews with qualitative researchers. While the participants did not participate in the interviews with AI or machine learning in mind, participants imagined systems that were able to automatically offer intelligent inferences and suggestions when they were asked what their imaginary “perfect tool” would be. In this paper, we collectively refer to these systems as “AI”, following the rationale in prior human-AI collaboration literature in CSCW [8]. We first describe the research methods that they use in their work, and the kinds of assistance they would like. Next, we describe the collaboration practices of participants whose methods involve collaboration, highlighting that reaching consensus is a nuanced social practice that extends beyond the actual research. We then discuss participants’ love-hate relationship with qualitative analysis: While the vagueness of qualitative analysis induces confusion and self-doubt, this doubt is essential to their research practice and AI should not remove that doubt in their analysis. Using a framework of task delegability, we close by discussing how qualitative research is a unique case that deviates from the typical tasks humans delegate to AI. While we do not (nor can we) prescribe exactly what algorithms or models should be used, we

argue that honoring uncertainty and serendipity should be a central quality of AI that assists with qualitative research, where data labels are fluid and “negatives” are not well-defined or sometimes nonexistent.

2 Related Work

2.1 Qualitative Research and Methods

Qualitative research accounts for a substantial proportion of CSCW and social computing scholarship—more than half of the CSCW 2019 papers use qualitative methods [20], with new opportunities for qualitative research being a constant topic in CSCW as well [18]. Qualitative researchers use a number of different analytic methods to generate insights, leveraging both inductive logic (to build patterns and themes from the bottom up by organizing data into increasingly abstract units of information) and deductive logic (to check existing patterns and theories against data in a top-down fashion). For example, thematic analysis [6] and constant comparative method [21] both describe processes of identifying, analyzing, and reporting emergent patterns within a set of data. Informed by the perspective of hermeneutics, interpretive textual analysis, a commonly used method in qualitative research, aims to understand cultures and subcultures through situated interpretation of texts [1]. In addition to these methods that focus on analysis, grounded theory [22, 39] alternates collecting and analyzing qualitative data in order to generate theories that are “grounded” in that data [12]. It provides a set of systematic guidelines for analysis while maintaining flexibility to fit the given set of data. Data sources in qualitative research are also diverse and overlapping, and any of the five traditions above could draw from interview data, field notes, and textual or visual content [14]. Many qualitative approaches make use of open qualitative coding [39] to inductively label and categorize emergent concepts while maintaining theoretical freedom, which is particularly useful for sifting through large amounts of unstructured data. Such inductively generated coding schemas can then be deductively applied on the full set of data.

Strengths of qualitative work include the ability to examine a space in great detail and depth, the adaptability of the research framework, collection of data with subtleties and complexities not available in quantitative data, and the situating of data in the context of experience and community [37]. Compared to quantitative research that often relies on positivism that believes that there are observable, measurable, “objective” facts that hold true for everyone, qualitative research often favors interpretivism, which argues that knowledge and truth are subjective and situated, and dependent on people’s experiences and understandings [37]. Instead of goals such as causal determination, prediction, and generalization sought by quantitative researchers, the goals of qualitative inquiries are “illumination, understanding, and extrapolation to similar situations” [23].

However, there are two major challenges of qualitative work. First, qualitative analysis is largely manual and laborious. Qualitative researchers have long used “cut and paste” methods for separating chunks of information, either physically (e.g., index cards) or in databases or spreadsheets [11]. There are now a number of qualitative analysis software tools (e.g., MaxQDA, ATLAS.ti, Dedoose) in which researchers can highlight and assign labels to portions of text, and these tools can produce some lightweight statistics (e.g., count) of these labels, with potentials of having more advanced features like topic modeling. While computational tools can relieve qualitative researchers of the burden of data management, their effects on data analysis are still highly debatable. On the one hand, qualitative researchers have argued that increasing the sample size in qualitative research designs is not an advantage, and researchers should focus on as many distinct cases as possible rather than as many cases as possible [41]. Using software tools that generate coding schemes based on a large number of documents risks opportunities for creativity and serendipity that qualitative research is uniquely able to provide [41]. On the other hand, current qualitative methods such as

content analysis [26] can require researchers to handle such an overwhelming amount of data that researchers may need to rely on subsampling or strategically focusing on a subset of data [36]. However, it can be difficult, if not impossible, to extrapolate the findings from the analysis on a subset of the data to the original dataset.

Second, the rigor of qualitative analysis is more difficult to assess and demonstrate than research rooted in positivist traditions that relies on measurement. There has long been a wealth of literature from multiple disciplines focusing on the assessment of the reliability and validity of qualitative analysis, providing guidance for publication [17], procedures for developing coding schema [10], and design research approaches to ensure rigor in qualitative analysis [30]. CSCW and HCI scholarship have also taken a keen interest in meta-studies of qualitative methods, such as comparing interview methodology through different media [15] and characterizing localized standards of sample size at CHI [9]. McDonald et al. [32] have shown that some CSCW and CHI papers chose to report interrater reliability (IRR) even if IRR might not be compatible with their research design, which highlights a need for demonstrating reliability in qualitative research. These two challenges suggest that qualitative research can benefit from new approaches that aim to help with handling larger amounts of data, and our study explores AI as an option to support qualitative scholars with their research.

2.2 Machine Learning and Qualitative Research

Machine learning (ML) techniques are known for their ability to extract patterns from small samples and generalize them to larger datasets, show promising opportunities to apply traditionally “small” data-focused qualitative methods to “big” data. Muller et al. [33] discussed conceptual similarities between machine learning and qualitative grounded theory methods (GTM)—they both make major claims to be grounded in the data, and they both start with and return to the data. Through a discussion of how Straussian GTM and Glaserian GTM (the two major approaches to GTM) are similar to unsupervised and supervised learning respectively, Muller et al. proposed a number of hybrid iterative approaches that integrate machine learning and GTM. Baumer et al. [4], through an empirical study, showed that ML and GTM could produce strikingly similar output from analyzing the same survey data, and later further showcased the promise of ML through their Topicalizer system [5]. They argue the promise of combining ML and GTM, but also caution that the results generated from machine learning models should be a scaffold for human interpretation instead of definitive answers, a point we echo in our analysis.

Another thread of research has explored the efficacy of using machine learning as a tool to support qualitative research, with an intention to automate parts of the research process. Yan et al. [42] experimented with a human-in-the-loop approach by having a machine learning model code the entire dataset from a pre-defined codebook and having human annotators later correct the labeling, and concluded that creating a one-size-fits-all ML model for all codes in a multi-dimensional coding scheme was impossible. Paredes et al.'s [35] Inquire tool aimed to help qualitative researchers by uncovering semantically-similar content in large-scale social media texts. Marathe and Toyama [31] explored whether qualitative coding could be partially automated through a formative user study and user testing of a prototype. They found that repetitive coding with a well-developed codebook lends itself nicely to automation, but that having good IRR is crucial for automatability. Chen et al. [13] pointed out that the goal of performance optimization in machine learning may be at odds with the goal of discovering patterns in qualitative research where the categories evolve over time, and suggested that a model that identifies points of ambiguity (i.e., disagreements in coding) may be more useful to qualitative researchers. It is important to note that the studies above focused on deductive aspects of qualitative research, typically with well-defined codebooks, and

not inductive practices. Our exploration extends this prior research by providing insights into the inductive step of the initial discovery of patterns and themes.

2.3 Human-AI Collaboration

The research we have surveyed thus far shows the infeasibility of automating qualitative analysis entirely, but also points towards ways that AI could be a collaborator that works alongside humans, rather than a delegate that performs specific tasks. Human-AI collaboration itself is increasingly gaining importance in CSCW scholarship. Emerging work focuses on a variety of topics from information extraction [28] to image understanding [44], consistently highlighting the need for transparency and explainability of AI. Mackeprang et al. specifically pointed out that at high levels of automation, users were unlikely to challenge results provided by AI when unaccompanied by explanations [28]. This situation, to say the least, would be undesirable in the context of qualitative research. Based on interviews with data scientists about their perception of collaborating with AI, Wang et al. reiterated the importance of AI transparency, and further noted that AI should be designed to augment, rather than to automate, a point we echo in our findings [40]. Furthering this concept of augmentation, Yang et al. proposed the idea of “unremarkable AI,” whose interaction with humans should have the right level of unremarkableness but should significantly improve their work [43]. Although Yang et al.’s study is situated in the context of high-stake clinical decision making processes, we see the notion of unremarkable AI as applicable to other contexts where AI should not be obtrusive.

For human-AI collaborations to be successful, the issue of what work the AI should or should not perform needs to be addressed. Lubars and Tan proposed a framework for task delegability to AI [27]. They enumerate four major factors to consider: the human’s motivation in undertaking the task, difficulty of the task, risk associated with incorrectly completing the task, and trust in the AI’s ability to accomplish the human’s goal. These factors, especially risk and trust, resonate with existing concerns around AI fairness, accountability, and transparency that have been increasingly present in human-AI collaboration scholarship. In this study, we use this task delegability framework to look at the case of qualitative analysis, and show how qualitative analysis is a uniquely challenging case for human-AI collaboration.

3 Methods

We conducted interviews to investigate the qualitative research methods used across various fields, with specific focus given to the coding process and tools that assist in data analysis. After acquiring IRB approval from our institution, we focused recruitment efforts on participants who primarily use qualitative research methods in their work. We recruited participants via emails to mailing lists of departments that regularly conduct qualitative research (e.g., communication, anthropology) at our institution, as well as public posts on social media that invited CSCW 2019 attendees to participate. We also encourage participants to share the call for participation with other people, resulting in a snowball sample.

The first two authors conducted 17 semi-structured interviews with qualitative researchers at our institution and among the broader CSCW qualitative research community. Participants were majority women with 2 men, 1 non-binary, and 1 agender individual. Participants ranged in age from 25 to 45, from the fields of Information Science, Anthropology, HCI, Linguistics, and Journalism. A majority of the participants were from the US, in addition to one from the UK, one from Japan, and one from Switzerland. We do not map this specific demographic information to individual participants in order to ensure that they are not individually identifiable. Fourteen interviews were conducted in-person, and three remote. Interviews lasted from 20 to 70 minutes, depending on the depth of the responses given. Table 1 lists our participants along with their field

Table 1. Participant details.

Participant ID	Academic Position	Field of Study
P01	Y1 Ph.D. student	Information Science
P02	Y2 Ph.D. student	Information Science
P03	Y1 Ph.D. student	Information Science
P04	Y3 Ph.D. student	Anthropology
P05	Y1 Ph.D. student	Anthropology
P06	Y4 Ph.D. student	HCI
P07	Y6 Ph.D. student	Information Science
P08	Y3 Ph.D. student	Information Science
P09	Y3 Ph.D. student	HCI
P10	Y4 Ph.D. student	Communication
P11	Postdoc	HCI
P12	Professor	Anthropology
P13	Professor	Communication
P14	Postdoc	Information Science
P15	Y6 Ph.D. student	Linguistics
P16	Y3 Ph.D. student	Linguistics
P17	Y4 Ph.D. student	Journalism

of study and academic position. Participants were compensated with a \$30 Amazon gift card upon completion of the interview.

Our interview protocol focused on how qualitative researchers collect and analyze their data, taking a broad exploratory approach. Interview questions included what methods the participants use, if there are any tools they use to assist in their work, attitudes and opinions about qualitative coding as a process, and relationship dynamics between collaborators. During the interviews, we asked participants to describe a current or most recent qualitative research project they have conducted, including scope, scale, and timing, in as much detail as possible. We then asked participants to describe their collaboration practices if relevant. The next section of questions focused on the software and tools the participants use for their data collection and analysis. We also asked participants for their attitudes and opinions on the coding process, including what they enjoyed about it and where there might be pain points. Finally, we asked participants to talk about what kind of assistance they would like if they had a perfect tool, and areas of their research where they did not want assistance. As mentioned in the introduction, we did not specifically ask participants about “AI assistance” so as to prevent leading them to answers about the term “AI” without a shared, concrete idea of what “AI” means. Instead, we asked what kind of assistance that they envisioned

in a perfect tool, and their responses led us to consider AI as a potential solution, and ultimately human-AI collaboration as the focus of this paper.

We performed a thematic analysis of the interview transcripts [6]. Prior to analysis, all interviews were transcribed, anonymized, and assigned the participant IDs presented here. The first author initially engaged in one round of open coding, using the software MaxQDA. Then the second author, after reading the transcripts, discussed preliminary emerging code groups such as “help with the mess” or “feeling of doing it right” with the first author. Two more rounds of iterative coding helped us combine similar code groups into higher order categories such as “disciplinary differences.” The first author used these categories to produce a set of descriptive theme memos [38] that described each category with grounding in the interview data. All authors then discussed and updated the memos regularly to reveal the relationships between the categories and finally clarified the themes, which resulted in the five main findings we discuss below.

4 Findings

Our findings document qualitative researchers’ current practices and needs that AI needs to support, and discuss their unique, nuanced challenges where AI should not be applied. We begin by describing the methods participants used in their qualitative research, and the kinds of assistance that participants desired from software tools. Next, we discuss the complex collaboration dynamics in collaboration-heavy disciplines by highlighting that reaching consensus is a nuanced social practice that extends beyond the discussion of the research project itself. We then discuss participants’ struggle with confidently conducting qualitative research, highlighting the central role of uncertainty in qualitative analysis. Finally, we discuss participants’ connection with and need for agency over the analysis of their data despite their self-doubt. From here, we conclude by revealing the bottom line that AI assistance should not cross—they should not replace human researchers in doing the analysis.

4.1 Qualitative Research Practices

Our participants used a variety of methods in their qualitative research. Among them, interviews were the most common data collection method, used by 15 participants across 6 fields. Others that were common included participant observation (7 participants), and note taking (4 participants). Participants also described data collection methods specific to their field or discipline, such as Jeffersonian transcribing in linguistics, oral history in journalism, and ethnography in anthropology. Compared to the varied data collection methods, participants’ descriptions of their analysis methods, while having varied names and granular details, broadly converged to a process of inductively identifying interesting pieces of data then finding patterns among them. While we acknowledge that qualitative methods entail a much broader range than what our participants reported, our analysis here only focuses on the methods our participants used. Table 2 shows the complete list of research methods that participants used.

Surveying the practices of qualitative researchers runs into some challenges due to terminology. While many of the participants share similar research practices, they referred to these practices with different terms. These differences not only exist across disciplines, but within disciplines as well. For example, P06 referred to the process of inductively building a codebook then deductively labeling according to the codebook as “thematic analysis,” while P01 and P12 called the same process “content analysis.” Many participants referred to the common grounded theory research method of first identifying data snippets of interest and assigning them descriptive labels, then generating deeper insights from these labels (sometimes with the explicit process of categorizing these labels). However, participants have different terms for both steps. Terms for the first step—identifying interesting units of data—included “qualitative coding,” “open coding,” “inductive coding,” and

Table 2. Qualitative research methods used by participants. Here, we report the methods in participants' own words, so the same term may refer to different methods.

Participant	Field of Study	Type of Data	Data Collection Methods	Data Analysis Methods	Software
P01	Information Science	Interview transcripts, API documentations	Semi-structured interview, content analysis	Open Coding, "seeing themes"	Google Docs
P02	Information Science	Interview transcripts	Interview	Qualitative coding, "forming themes"	MaxQDA
P03	Information Science	Interview transcripts, interview notes	Interview, note taking	Inductive coding, "coming up with categories" Coding, transcribing into International Phonetic Alphabet	None
P04	Anthropology	Field notes, voice recording, video recording	Ethnography, participant observation, interview	"Highlight unusual things", "pulling up bigger themes"	Google Docs
P05	Anthropology	Interview transcripts, voice recording, field notes	Participant observation, ethnographic interview, note taking	Thematic analysis	Yes, but did not remember
P06	HCI	Online comments	Scraping	Transcribe, open coding, axial coding, memoing	None
P07	Information Science	Interview transcripts, field notes	Semi-structured interview, participant observation, expert interview, note taking	Transcribe, code	ATLAS.ti
P08	Information Science	Interview transcripts, field notes	Interview, participant observation	"Classification process to organize data into themes"	Enable
P09	HCI	Interview transcripts, field notes	Interview, participant observation		

P10	Communication	Interviews, field notes, eye tracking data	Interview, experiment, eye tracking	Transcribing, coding, affinity mapping	ATLAS.ti
P11	HCI	Interview notes, interview excerpts	Semi-structured interview	Note taking, coding, “seeing themes”	Excel
P12	Anthropology	Interview transcripts, field notes	Ethnography, participant observation, note taking, interview	Content analysis, open coding, memoing, codebook & closed coding, “seeing themes”	Dedoose
P13	Communication	Interview transcripts	Interview	Transcribing, coding, “coming up with coding schemes”	Word, Excel
P14	Information Science	Interview transcripts	Interview, “kind of ethnographic stuff”, participant observation, participation	Transcribing, micro-coding, “inductively identifying themes”, “grouping into categories using post-it notes”	NVivo
P15	Linguistics	Interview transcripts, clinical texts, online forum posts	Interview	Public discourse analysis, coding, “categorizing codes”	MaxQDA
P16	Linguistics	Video recordings	Recording natural interaction	Transcribing into Jeffersonian system, “making subcollections”	Word
P17	Journalism	Interview transcripts	Interview, oral history	Discourse analysis	Excel, Google Docs

“highlighting unusual things.” Participants’ terms for insight generation are even more varied, from “memoing” “affinity mapping” to descriptive phrases without a proper term—“seeing themes,” “grouping into categories,” and “classification process to organize data into themes.” Furthermore, while P04, a Ph.D. student in linguistic anthropology, also used the term “coding” she was referring

to transcribing the interviews into the International Phonetic Alphabet, instead of identifying interesting data snippets—something we only realized halfway into our interview with her.

Overall, we saw participants use inconsistent vocabularies despite referring to similar methods. While we have not heard stories of participants running into communication problems, it is possible for researchers to accidentally miscommunicate research methods and for reviewers to have confusing expectations. Assistive tools can also create barriers by choosing to use language that favors one particular genre of academic training.

4.2 Assistance From Computational Tools

Qualitative research is a manual, labor-intensive process that would benefit greatly from computational tools, but for many participants, existing tools were not sufficient for their needs. In this section, we first describe the gaps in current tools, and then discuss the assistance that participants desired.

4.2.1 The Inadequacy of Existing Tools. Participants used a variety of software to assist with qualitative research, and only three participants did not use any such tools (see Table 2). However, not only did many tools fail participants in helping them handle large amounts of qualitative data, participants were also unwilling to learn more sophisticated tools. According to P05, the amount of notes and labels quickly became overwhelming, and the software tool that she used was insufficient:

I do everything in Google Docs. You've got 500 codes first and then you've got notes everywhere all over them, which I'm currently trying to decipher and I'm kicking myself for not having a better method. (P05)

Many participants used word processing software like Google Docs to code qualitative data, but these software were not able to handle the overwhelming amount of codes and notes, as was the case for P05. While some participants were aware of other software tools that have more sophisticated functionalities, they did not use or stopped using them for various reasons. For example, P10 stopped using ATLAS.ti because the software broke frequently on her computer, so she returned to her old method of using Excel and handwritten notes. P09 only used software to organize his transcripts, and told us that learning how to use more complicated software was not worth the effort:

Using any tools, I think it gets in the way of the analysis...I think the focus then inevitably becomes on the tool and how I can manipulate and push data in order to make it appropriate for the tool... I need to figure out how I can fluster data in a certain tool, in a certain way so as for it to make it easier for me later on. And I was just thinking it's a waste of time. (P09)

P09's quote is exemplary of participants who chose not to use software tools. Their anticipated cost was not only in learning how to work the software, but also in how to use the software in their own way that is sustainable in the future, and these two layers of barriers precluded these participants from using it. P14, while kept using the qualitative analysis software NVivo, limited herself to the most basic functionalities of coding data and making code categories, then hand-copied these categories onto post-it notes that she later put on the wall. Therefore, participants wanted "more intuitive ways to do qualitative analysis," and something that could help them "sort through this mess":

If I had a [perfect tool] I would like a little bit more input in putting together connections between the data. (P12)

4.2.2 *Desired Assistance.* Participants had different ideas of how a tool could help them make sense of the connections within the data, and cross-sectional analysis and visualization features were common requests:

I really like when I can slice and dice the data. So I want to be able to say, okay, all the people I've talked to who identify as queer, how did they feel about capitalism? I want to be able to do a cross-sectional analysis on multiple codes and domains. (P12)

Just better visualizing the relationships in the data itself. So if I'm missing out codes because I have less time, or if my co-author isn't being participative enough or whatever, then I want to be assured of the fact that I was able to see all kinds of relationships. (P07)

It is important to note that cross-sectional analysis and visualization features, though implemented to various extents, already widely exist in qualitative analysis software such as MaxQDA and NVivo. Prior research has noted that qualitative researchers usually confine themselves to the most basic features of qualitative analysis software [41], but here we see that the participants still want these seemingly basic features, which suggests that the participants never discovered these features in the first place due to the overall difficulty of using the software. Furthermore, P07 specifically pointed out that the tool should not be “like the NVivo type, where I have to really learn a lot of it.” In other words, P07 admitted that an existing tool may be able to achieve what she wanted, but the bar of learning the software was too high.

While participants across disciplines wanted additional help to make sense of the data, some participants would appreciate features that are specific to their fields. For example, P05, an anthropology Ph.D. student, requested features designed for their own research tradition:

Maybe, I don't want to say a translation tool, but kind of almost a sort of node system to be like ... “These were the literal words that were used, but this is kind of the underlying meaning to that.” (P05)

While anthropologists might benefit from features to help them note underlying meanings to spoken words, P15, a Linguistics student, told us that being able to get down to the most granular details of speech was important:

I'm thinking especially in my field where when I'm transcribing interviews, I'm actually not just looking at what is literally being said. There's a process of transcribing for tone of voice, gestures, and things. (P15)

Unlike many of our other participants, for whom the purpose of transcribing is to transform audio data into textual data that is readily analyzable, the detailed transcription—called Jeffersonian transcription—is the core of P15's analysis. Her analysis did not include categorizing snippets of labeled data. Her unique genre of analysis means that a tool designed only for the most common use case will be of no use to her at all.

In addition to features that handle the mechanics and the presentation of qualitative data, participants also wanted features that could take one step further by providing suggestions and inferences, based on the way that the researcher analyzes the data:

If there was some sort of learning algorithm, for example, that would suggest ... other quotes that were similar to that one. (P08)

Or maybe even just the ability to be based on whatever magical software anyone can develop being, “based on the fact that you tagged it this way, we think that these are the more important interviews or not important, but we think that maybe you want to pay more attention to these.” Or be like, “You completely neglected this one. Did you forget about it or is there actually nothing here for you?” (P05)

In sum, participants wanted various types of help from computational tools but often refrained from using sophisticated tools due to their steep learning curve. While their reluctance may indicate usability issues with such tools, their reluctance also suggests opportunities where AI assistance could provide them the desired data representation and visualization, and only offer more sophisticated explanations and customizations if desired. The promise of AI assistance is further highlighted by participants' imagined intelligent suggestions—or, in P05's words, "magical software."

4.3 Collaboration Dynamics

Qualitative research can often be collaborative, involving multiple researchers examining and then reaching consensus on the same data to increase reliability and reduce individual biases. Participants from Information Science and HCI told us that collaborative work with co-authors was common, so it is no surprise that, when asked about features they would like in their imaginary perfect tool, only Information Science and HCI participants mentioned collaboration features:

While I was maintaining this online notebook, I could hypothetically, if my co-author had enough time, the dream scenario would be that he or she would be reading those notes every day and commenting on them ... So the collaborator element's definitely important and could help a lot with the eventual coding and analysis process, even while you collect data. (P07)

With collaboration being the norm, a natural follow-up question is how to reach consensus when there are multiple researchers on a single project. "Consensus," however, takes on many different flavors and nuances, according to the participants.

Inter-rater reliability (IRR) is often used to calculate consensus during a deductive phase of coding, but is often inappropriate for inductive analysis. Consistent with McDonald et al.'s findings [32], reaching consensus through IRR was rare for our participants—only two used IRR to verify consensus, and two others mentioned IRR only to note that they did *not* use it in their research, likely because our participants largely conducted interviews, of which the analysis is often incompatible with IRR.

Our participants most commonly perceived consensus as some flavor of "everyone agrees," though agreement sometimes can be a flexible and loose standard:

Once everyone is pretty much satisfied and doesn't have any complaints or questions or comments anymore that are significant, and by significant, I mean like "we have to change this or it's going to ruin the study." (P02)

Compared to P02, P14's standard for consensus was even more flexible—there is no definition for consensus; it is something that the researcher understands intuitively as they are doing the research:

[Consensus is] just like a kind of tacit knowledge thing where it's you know when you see it, but it's very hard to actually kind of define. (P14)

P14's view of consensus as fluid highlights how it is the outcome of nuanced social practices that is not easily explainable or definable. However, its flexible nature also means that achieving consensus might depend on the researchers' established, implicit norms generated from having worked together for an extended period of time, a facet often overlooked by prior conceptualizations of researcher consensus (e.g., [29]).

These social practices, as we learned from participants, often reduce "consensus" to the decision of a person with the ultimate decision-making power, rather than multiple researchers having

reached some level of agreement. The decision maker, however, varies. Sometimes the decision maker is the project leader:

If I'm leading the paper, to me it's my decision [if we have reached consensus]. (P08)

If the stakes are high, if I am the primary investor in a particular project then yes, I think I should take the decision on this. But if it's maybe my advisor's project I think I would give her the final say in that. (P10)

Sometimes the person with the final say is the most senior person, who may or may not be the project leader:

If [my collaborators] don't [agree], we have a little bit of a debate and I'm assuming the most senior person probably wins. (P03)

While P03 did not personally have experience with the most senior person "winning" the debate, as a junior Ph.D. student this was their default assumption, which suggests a tendency for researchers (especially junior ones) to acquiesce to authoritative figures. P07 elaborated more on this tendency, pointing out that "consensus" is the result of an negotiation of collaboration dynamic and power:

I think it depends a lot on who the collaborators are and what the power relationships with those collaborators are ... These things are not objective ... It's not just about what codes appear or what themes appear the most salient and relevant. So I think it's more about the collaboration dynamics, the power, all of those things, rather than what the codes are seeing ... It's literally about [power]. (P07)

While it might seem ideal for every researcher in a team to have their opinion treated equally, P07 directed refuted this notion, claiming that it would only jeopardize the project:

[Collaborative qualitative analysis] would still definitely take on some hierarchy in what I imagine. There would still be someone who's written more papers. There will still be someone who's written fewer papers. Or still have an expert on the topic, versus someone who's not the expert on the topic. ... I can't imagine there being a flat hierarchy somewhere. And actually, that one time that it happened ... things went really down south. The paper was about everything. It got rejected. Then there was a fight. Then there were all these conversations about who's contributed the most, who's going to be the first author and so on. (P07)

MacQueen et al. [29] argued that in team-based qualitative analysis it is good practice to have one person maintain the team codebook. While MacQueen et al.'s recommendation largely came from a project management standpoint and maintained that the codebook development should still be done collaboratively, here we see a case where one person being in charge of the *direction of analysis* can be more beneficial than a true, equally collaborative fashion. Overall, we observe that in fields where collaboration is common, reaching consensus among researchers can often be a nuanced social practice beyond mere agreement on the research product.

4.4 The Existential Crisis of Qualitative Researchers

Despite the variety in discipline-specific research methods, desired help from software tools, and meanings of consensus, participants across disciplines consistently told us about one feeling: qualitative research is hard and messy.

Sometimes it can be a little bit frustrating when you're sort of in that process where you're like, "Is this a theme? Is that a theme?" Then sometimes when you're going through ... a couple of interviews in and you're like, "Oh there's a theme," and then

you're like, "Oh fuck, that's probably a theme" ... so then you have to go back to see if that [theme] is in [the other interviews]. (P13)

Like P13, many participants described how qualitative analysis could be frustrating, and this frustration often comes from researchers' self-doubt:

I'll just code, and maybe it'll be relevant and maybe it won't be, and I'm not really sure ... If someone just mentions it in passing, do I use that? ... If someone has a few sentences about it then do I code that? Like what is useful to actually code? (P03)

Am I drawing connections between things that aren't necessarily connected? ... which also goes back to not having time to [follow up with] people and be like, "Hey, so you said this thing, I'm just wondering, can you expand that more?" (P05)

P05 doubted herself when drawing what she feared were non-existent connections in the data, and attributed it to having conducted imperfect interviews. The guilt of imperfect research was common among participants, who felt that they could have asked better questions during interviews, could have used more data in the final research product, or could have done a deeper analysis if they had had more time:

I didn't really get to do maybe as much coding as I would've liked, or to look at as many complex issues as I would've liked. ... It was a restriction on time. (P03)

In addition to imperfect research execution being a common source of guilt, others felt apologetic for not following an ideal research methodological paradigm:

I think I would be lying if I said that I completely allow myself to be driven by the data ... especially this study, for example. I did it to inform the design of potential technology that I wanted to build, while I'm doing the questions, I was already looking for particular things. ... So I think that already in the questions you have implicit codes. (P11)

I feel like we're doing it wrong. ... Especially because [we are] saying these are the topics that we're interested in. That's not very grounded theory of me. But, I don't quite know how to do it otherwise, because obviously ... I'm going to look at this interview already from a different perspective than, say, a psychologist. (P15)

Participants thought their own research execution was imperfect—for example, "not very grounded theory." However, in participants' own unique research projects, the "perfect" methodology may not be the most desirable approach. While HCI (as well as CSCW) has distinguished itself as a "discipline that has often proceeded with something of a mix-and-match approach" [16], here we see evidence that such a hybrid paradigm—for example, being driven by the data but also having a topical focus—is also practiced across disciplines.

Overall, self-doubt with all parts of the research process was pervasive among participants. This self-doubt was so strong that P07 told us the moment she had to approach the data she collected was what she "dread[ed] the most." Participants' shared experience of self-doubt highlights the inductive and interpretive nature of qualitative research, and suggests that this vagueness may be inevitable and essential to the qualitative research process itself. While it might be tempting to design AI that can reduce qualitative scholars' anxiety by providing concrete suggestions from their data, we found ample reasons why researchers might resist such assistance, which we detail in the next section.

4.5 Don't Do It for Me

Even though participants consistently told us how qualitative research can be full of doubt and vagueness, they did not want AI assistance to eliminate that vagueness if it meant eliminating the

work to make sense of their messy data. Participants told us that despite the uncertainties, the new discoveries they made through their research process were the most rewarding:

It's really satisfying. ... It's those kinds of exciting Eureka moments that make research kind of worth it. (P14)

My favorite style of coding is more where I'm able to bring out all of these different things. ... So that's what I really enjoy, the process of recollecting everything that happened, and then coming up with something. (P07)

Participants' comments on their unexpected new findings show that serendipity was a major reason why participants enjoyed qualitative research. Integrity was a major factor as well. P12, for example, described a sense of responsibility toward her project:

How do I feel when I'm coding? Overwhelmed. I feel such a great sense of stewardship in the project that we've been discussing. I want to get it right and that makes me feel a lot of pressure. (P12)

The motivations and responsibilities that researchers feel around their data and analysis may present some hurdles for any tool meant to support their work. While some participants mentioned that they would benefit from AI-generated suggestions and inferences, these participants, with their sense of serendipity, responsibility, and intimacy toward their research, explicitly expressed that they needed agency over the analysis of their data, and AI should not do it for them.

I don't want anything to do the analysis for me. That doesn't make sense. This is my interpretation of things ... It's one of the most intimate things. ... I have these little like research crushes with some codes or participants. ... These [data] are my babies. (P11)

P16, while being in Linguistics and doing a completely different kind of analysis than P11's, shared P11's feelings:

As nice as it theoretically sounds to have something do automatic transcription for you, I actually would not want that even if it could do the Jeffersonian level of detail because it's so important for you as a researcher to understand what's happening in your data, doing it yourself. ... Your transcription is part of your analysis as well and that's the only way for you to get your hands dirty with the data is to get in and see what's happening with it. ... I don't know of anybody who uses [professional transcription services] because you take a lot of pride and ownership over your data and it's important to your participants and I would never want to farm that out when it's your research in your research process. (P16)

To P11 and P16, outsourcing analysis for them not only impacts research quality, it also robs them of their emotional connection—the intimacy, pride, and ownership—with the data. Compared to P11 and P16, P06 was slightly more open to the idea of AI-assisted analysis. However, P06 also pointed out that the tool should be limited to the level of suggestions and that making decisions would cross a line:

I don't want it to take out extra information automatically. I don't want it to try to guess or try to ... well, it can try to guess, but I don't want it to decide. ... It can do a suggestion, but I'm not sure if I would trust the suggestion, or actually it can suggest the discussion points that we need, and then we can discuss rather than decide it for us. (P06)

In contrast, P14 felt even suggestions could be harmful to her research process:

Maybe it could make suggestions, but even then I don't know if I want it because it doesn't know what my research questions are. ... I don't want it to pick the good

quotes for me, that's for sure. ... I don't want the computer to do it. I also don't want the computer to influence how I'm thinking about it. ... I don't necessarily want to be primed or how that's going to affect my thinking ... if it's making this suggestion and now I can't unsee it. (P14)

While participants generally resisted complete automation of qualitative analysis, P06 and P14's comments show that even a low-level, suggestion-based automation (according to Parasuraman et al.'s model of autonomy [34]) can also be undesirable. During the interview, P14 asked if there was a machine to interfere, then "why bother?"—why bother having a human researcher at all? According to P14, even if the researcher does not necessarily have to use or trust the suggestions, seeing the suggestions alone can negatively impact the researcher's interpretation and analysis, or even worse, eliminate the meaning of human researchers.

Overall, participants told us about their conflicted feelings about qualitative analysis, one that is marked by anxiety but also intimacy and ownership. Driven by these feelings, they resisted AI to replace them in their analysis, despite seeing its potential to augment their work. These findings suggest that, unlike tasks typical for AI to accomplish, the objective of qualitative research is, often intentionally, ambiguous. Furthermore, qualitative researchers value this ambiguity because it offers them serendipity and opportunities for deep insights. In the discussion section, we detail how qualitative research is a unique case for human-AI collaboration, and provide design implications for AI assistance to honor ambiguity and serendipity.

5 Human-AI Collaboration in Qualitative Research

Our findings reveal complicated feelings participants had toward the practices involved in qualitative analysis. On the one hand, participants struggled with self-doubt in their research analysis and saw the appeal of AI assistance to help them see relationships behind their data. On the other hand, they also liked the analysis process as it is and felt connected with their data, therefore skeptical of AI driving the analysis. This tension between the desire of AI assistance and the need for human agency suggests opportunities for human-AI collaboration in qualitative research, but also draws a line that AI assistance should not cross.

To unpack this tension, we turn to a framework proposed by Lubars and Tan to describe the delegability of a task to AI [27]. While there were likely many small steps in the qualitative research process (e.g., open coding, axial coding, synthesizing), our participants' resistance of complete automation suggested that they viewed qualitative analysis as a singular task in the context of AI assistance, and we can reasonably speculate that AI designers (who are arguably less familiar with qualitative analysis) might do the same. In their framework, they enumerate four factors that must be considered when determining the appropriateness of delegating a human task to AI: **motivation, difficulty, risk, and trust**.

Lubars and Tan conceptualized motivation as having three components: the human's intrinsic motivation of carrying out the task, whether the human's goal itself is mastering the task, and the utility or importance of the task to the human. Our participants' enjoyment and pride in conducting qualitative research, as well as their wish to improve their research skill, show that qualitative research involves high motivation: participants enjoyed and were committed to conducting their analyses. While one may assume that high motivation indicates low delegability, Lubars and Tan did not find statistically significant correlation between the two variables. The lack of correlation coincides with the sentiments of our participants, who were invested in their qualitative work but also acknowledged the appeal of appropriate AI assistance.

Difficulty also consists of three components: a person's perceived ability to complete the task, as well as the required human effort and expertise. Lubars and Tan found a general negative correlation

between difficulty and task delegability: the more able a person perceives themselves to complete the task, the more human effort and expertise required by the task, the less delegable the task is to AI. Qualitative research can be considered high difficulty, as it requires effort and expertise on behalf of the researcher, which suggests its low delegability to AI according to Lubars' and Tan's findings. The difficulty of analysis is evidenced by the low confidence participants had in their analysis. While Lubars and Tan also use required social skills and creativity as dimensions to contextualize the previous three dimensions, they are not central to the conceptualization of difficulty, though one can reasonably argue that qualitative research requires high levels of both social skills and creativity.

Additionally, in the survey that Lubars and Tan used to evaluate their framework, they observed that self-confidence and intrinsic motivations were positively correlated—people enjoy doing tasks that they are good at. However, the case of qualitative research serves as a unique counterexample: not only do people enjoy doing something they are not confident about whether they are doing it correctly, but low-confidence itself is central to the outcome of qualitative research. Therefore, our study on qualitative research challenges the current task delegability framework, and also raises questions around the integration of AI into qualitative research: Should we consider qualitative research as a single, unified “task” for AI to assist with? If qualitative research as a single task should not be delegated, what parts of it are appropriate for AI to assist with?

Qualitative research is harder to characterize in terms of the third factor: risk. Lubars' and Tan's conceptualization of risk is three-fold: personal accountability in case errors happen; the uncertainty, or the probability of errors; and the scope of impact, or cost or magnitude of those errors. These three components of qualitative research all focus on errors, but “error” is inherently difficult to define, if definable at all, in qualitative research to which subjective interpretations are central: What makes a researcher's interpretations *wrong*? Is there such a thing as interpretation *error*? Furthermore, the concept of risk itself may be more nuanced than currently characterized, given that Lubars and Tan did not find significant correlation between risk and delegability, and the complexity of defining “error” in qualitative research may also be present in other kinds of tasks. Designers should carefully consider this flexibility of qualitative research when designing AI assistance, especially supervised models that are entirely dependent on what is “positive” or “negative.”

The final factor, trust, is also nuanced in the context of qualitative research. Trust, which Lubars and Tan also describe as containing three components, includes the perceived ability of AI to complete the task, the interpretability of AI's actions, and the perceived value alignment between AI and human. As our findings have shown, participants had low trust in AI's ability to reliably conduct qualitative analysis. While “reliable” is often defined as being consistently “correct,” our discussion above shows correctness may not be the focus in qualitative research since error itself is hardly definable. Instead, our findings suggest that the reliability lies in the consistent alignment of values—that the value of qualitative research derives from creativity and serendipity [41]—between AI and human researchers, which our participants did not trust AI to have. P14 pointed out that she did not want any assistance because she could not “unsee” the AI suggestions, which indicates an assumption of unwanted, negative influence from AI that values concrete, highest-probability answers, as opposed to generative and nuanced ones sought by human-based qualitative analysis. However, influence is inevitable in collaborative qualitative research—having human collaborators also influences researchers' perspectives, and it is not reasonable to have to self-isolate to achieve the desired research integrity. Therefore, the assumed “badness” of AI influence might be because the suggestions come *from* AI, rather than that the suggestions themselves are of low quality.

The distrust of AI shared across our participants may possibly have to do with the differences between typical behavior of humans and AI. Humans usually make suggestions like “This part of

the data is interesting and maybe you should take a look at it” (and more seasoned collaborators can also explain why it is interesting). However, AI suggestions, generated from classification results, often take the deterministic form of “data X should be assigned code Y” without explanations—even a human collaborator would be difficult to trust if they could only make such suggestions. While our participants did not explicitly mention AI’s interpretability, the well-known opaqueness of the AI’s inner working “black box” could only exacerbate the distrust. P14’s distrust echoes with Gach et al.’s finding that interpersonal trust—the willingness to accept risk based on the expectation of another person’s behavior—often gets mistranslated when it is implemented as impersonal trust that a technical system will behave as expected [19]. This mistranslation of trust resonates with Ackerman’s notion of sociotechnical gap [2]—how can we better technically support the kind of interpersonal trust that we know we must support socially in qualitative research?

Given the challenges related to risk and trust, a key consideration in the design of AI collaborative tools for qualitative research should be: How can we promote more interpersonal trust between human researchers and AI collaborators in ambiguous qualitative research? One way for AI to build trust is to provide humans with the ability to make choices and corrections [25]. If such ability is sufficient, one might think that an ideal collaboration would involve first having the AI automatically label the entire dataset, allowing humans to correct the AI as necessary afterwards. However, not only would such an approach meet the objections of participants (as exemplified by P14), Yan et al. [42] tried such an approach and concluded that creating a one-size-fits-all ML model for all codes in a multi-dimensional coding scheme was impossible. The low precision of their model suggests that humans may have to do *more* work in correcting the AI than if they had coded on their own. While correction is also widely present in human collaboration, it is easier to discuss and resolve the differences with a human than with AI. Furthermore, it is also undesirable to have too much trust—if the data to be coded is ambiguous, our results show that it’s possible that researchers might defer to AI as the *de facto* tie breaker even though the situation deserves human deliberation to uncover the rich qualitative insights underlying that ambiguity. Given the two extremes, it may seem that offering answers that provide precisely the right balance of trust is the natural solution for AI, but “just the right balance” may never be practically achievable—the design of AI assistance for qualitative research may need to forgo the traditional answer-provider paradigm and leave the deliberation to humans.

5.1 From Answers to Ambiguities and Disagreements

How, then, can we preserve human researchers’ agency in ambiguous situations in AI? A good starting point is to consider how it can play out in machine learning algorithms, a core component in many AI systems. Chen et al. [13] point out that the goal of performance optimization in ML may be at odds with the goal of discovering patterns in qualitative research where the categories constantly change: while well-defined data categories are essential anchors for many machine learning algorithms, qualitative analysis does exactly the opposite by iteratively revising, challenging, and sometimes even overhauling existing categories to reveal deeper insights buried in the data; categories are not the end, but only a means to an end. Therefore, a machine learning model that identifies points of *ambiguity* and *disagreement* may have more utility to qualitative researchers. This point is further supported by McDonald et al., who discussed why seeking agreement might not be appropriate when codes are the process instead of the product. This work also highlights the scenario where there is an expert researcher (or a researcher deemed as the “expert”) in a team, in which case the other researchers defer to the expert researcher largely to maintain the social relationship. The expert-takes-all phenomenon was common among our participants, and we can speculate that it also broadly exists given how common advisor-student relationships and mentor-mentee relationships are in research training. McDonald et al. [32] also discussed how

agreement can be harmful in research rooted in critical traditions (e.g., feminist HCI [3]), and the type of “forced” agreement in this work may aggravate this harm.

Taking the above considerations together, we can imagine an example assistive AI system that has the following features. First, to promote ambiguity and preserve human researchers’ agency, the system could have the researcher choose the level of suggestions it provides. For example, the system could default to leaving the researcher on their own (i.e., do nothing) at first, and only later suggest interesting data snippets that the researcher has overlooked based on the researchers’ existing coding/labeling patterns (e.g., through text classification using long short-term memory (LSTM) networks). The system could also suggest upcoming, unexamined data snippets for the researcher to consider, but only if the researcher specifically requests this. As demonstrated in our analysis, unprompted suggestions of new data can lead researchers into uncertain (and possibly unwanted) interpretive directions, a concern expressed by our participants.

While it is possible to provide *code* suggestions (likely based on topic modeling), none of our participants were enthusiastic about such an idea, ranging from hesitant to resistant. Our participants’ reactions show an interesting contrast with the promise of machine learning-based coding shown in prior research [4, 13, 31]. Considering these findings together, we urge AI designers to be careful in making suggestions about the exact codes or labels that a piece of data might take. An AI system capable of making code suggestions should be clear about what this feature is and is not (i.e., that instead of data to be coded, it makes suggestions *of* codes). It should also be available only upon the researcher’s request, and only after it learns enough of the researcher’s coding patterns in a particular project.

Second, given that collaboration was common for many of our participants, the system could have features that further promote ambiguity and serendipity through collaboration. For example, it could suggest data snippets (examined or unexamined) that might be worthy of further discussion among collaborators. The system could also suggest a piece of data that a researcher has already coded to other collaborators who are likely to disagree with that researcher. Both examples aim to encourage discussions among researchers that may spark unexpected insights, and echo the point we made against careless coding suggestions—we envision the system to help set researchers up to do the research, instead of doing it for them.

Our vision above is only an example of what an assistive AI might look like, and the concrete specifications will depend on a number of factors such as the audience for whom it is designed (e.g., AI designed for HCI researchers vs. linguists are likely to be different), and the underlying infrastructure that it needs (e.g. collaborator-based suggestions will need collaboration functionality in the first place). The features that we envision also illustrate that an assistive AI will necessarily be an interactive one with mixed initiatives, some of which are already in line with principles of mixed-initiative designs [24]. While we acknowledge that such a system still does not address the expert-takes-all phenomenon, we are also hesitant to recommend that AI address the nuances of the social relationship between researchers. By revealing the the expert-takes-all phenomenon in the collaboration dynamics between junior and senior researchers, we are not calling for a technical solution to “solve” their relationship; we only hope that qualitative researchers can be aware of this phenomenon, and strive toward more productive collaborations with it in mind.

Uncertainty and ambiguity have long been viewed as a hindrance that costs efficiency and causes mistakes. However, uncertainty is a feature rather than a bug in qualitative research. As our findings indicate, while our participants wanted AI to help them have a better grasp of the uncertainty, they did not want it to be taken away. This very uncertainty allows qualitative researchers to dig deeply into their data, sometimes even to build personal connections with their data, and generate rich interpretations and insights. Our findings not only point to ways that AI can help qualitative researchers, but also reveal ways that qualitative researchers can help each other. For example,

researchers who are not accustomed to collaboration may want to experiment with collaborating as a way to have various perspectives that promote serendipity. Senior researchers may also want to consider teaching junior researchers to embrace uncertainty as part of the qualitative research training, so that they can trust themselves in doing qualitative research on a high level while also maintaining a healthy dose of self-doubt.

6 Limitations and Future Work

Our research has several limitations. First, the majority of our participants were junior scholars, with arguably less experience in qualitative research than more senior ones. While we believe that a student majority is still able to provide valuable insights since they all actively conduct research, it may nevertheless have an impact on their reported experiences and tool use.

Second, we would like to acknowledge that the qualitative research practices documented in this study (with the coding component in particular) are largely based on individuals, despite the ample descriptions of collaboration dynamics that we heard. We suspect the predominance of individual research practices is a product of the common student-advisor collaboration paradigm, and future work may find value in investigating more and other collaborative coding practices. Similarly, the kind of qualitative data that our participants dealt with were predominantly text, but the whole spectrum qualitative data is much broader, such as image, audio, and video, as some of our participants already indicated. We encourage future work to investigate these alternative forms of qualitative data and their analysis.

Finally, as we mentioned in the introduction, our work does not provide guidance for specific models or algorithms to be used. It is a first step toward successful human-AI collaboration in qualitative research by envisioning a world where AI assistance of qualitative research is possible, and laying out the promises and perils of that world. We hope that our findings will inform future in-depth studies of how specific AI systems can be adopted in qualitative research.

7 Conclusion

Qualitative inductive methods are widely used in CSCW and HCI research for their ability to generatively discover deep and contextualized insights, but these inherently manual and laborious processes are infeasible for analyzing large corpora, sacrificing the richness provided by these data. This study explores the potential of AI for assisting humans in conducting qualitative research by investigating current qualitative research practices and revealing where AI can come into play. We describe the methodological practices participants have, the insufficiencies of current tools to support these practices, and desired capabilities of their imaginary perfect tool. We also show that participants, when they collaborate, have complex and nuanced collaboration dynamics, and consensus reaching can be a negotiation of social relationships beyond the research project itself. Furthermore, while participants struggle with the messiness and uncertainty of qualitative analysis, they also want full agency of the process and insist that AI should not take it away from them. We argue that qualitative analysis is a unique case for human-AI collaboration because uncertainty is essential to qualitative analysis itself. The design of AI assistance should embrace this uncertainty and support serendipity by promoting human agency, instead of being the arbiter that aims to reduce uncertainty as AI has been traditionally conceptualized.

Acknowledgments

We thank the reviewers for their significant time and care in improving this paper. We also thank our participants, many of whom sacrificed their conference time to participate in our study. Finally, we thank Brianna Dym, Steven Frost, Katie Gach, Shamika Goddard, Jacob Paul, Anthony Pinter, and Morgan Klaus Scheuerman for their help and feedback on the early drafts of this work.

This work was supported in part by the National Science Foundation Award #1764089.

References

- [1] 2017. Textual Analysis. In *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc, 2455 Teller Road, Thousand Oaks California 91320. <https://doi.org/10.4135/9781483381411.n623>
- [2] Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human-Computer Interaction* 15, 2-3 (Sept. 2000), 179–203. https://doi.org/10.1207/S15327051HCI1523_5 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/S15327051HCI1523_5.
- [3] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, Atlanta, Georgia, USA, 1301–1310. <https://doi.org/10.1145/1753326.1753521>
- [4] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence? *Journal of the Association for Information Science and Technology* 68, 6 (June 2017), 1397–1410. <https://doi.org/10.1002/asi.23786>
- [5] Eric P. S. Baumer, Drew Siedel, Lena McDonnell, Jiayun Zhong, Patricia Sittikul, and Micki McGee. 2020. Topicalizer: reframing core concepts in machine learning visualization by co-designing for interpretivist scholarship. *Human-Computer Interaction* 35, 5-6 (Nov. 2020), 452–480. <https://doi.org/10.1080/07370024.2020.1734460> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/07370024.2020.1734460>.
- [6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [7] Jed R. Brubaker and Gillian R. Hayes. 2011. "We will never forget you [online]": an empirical investigation of post-mortem MySpace comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*. ACM Press, New York, New York, USA, 123–132. <https://doi.org/10.1145/1958824.1958843>
- [8] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 104:1–104:24. <https://doi.org/10.1145/3359206>
- [9] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, San Jose, California, USA, 981–992. <https://doi.org/10.1145/2858036.2858498>
- [10] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods & Research* 42, 3 (Aug. 2013), 294–320. <https://doi.org/10.1177/0049124113500475> Publisher: SAGE Publications Inc.
- [11] Trudy Campbell. 1997. Technology, Multimedia, and Qualitative Research in Education. *Journal of Research on Computing in Education* 30, 2 (Dec. 1997), 122–132. <https://doi.org/10.1080/08886504.1997.10782219> Publisher: Routledge _eprint: <https://doi.org/10.1080/08886504.1997.10782219>.
- [12] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. SAGE.
- [13] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (June 2018), 9:1–9:20. <https://doi.org/10.1145/3185515>
- [14] John W. Creswell. 2013. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. SAGE.
- [15] Jill P. Dimond, Casey Fiesler, Betsy DiSalvo, Jon Pelc, and Amy S. Bruckman. 2012. Qualitative data collection technologies: a comparison of instant messaging, email, and phone. In *Proceedings of the 17th ACM international conference on Supporting group work (GROUP '12)*. Association for Computing Machinery, Sanibel Island, Florida, USA, 277–280. <https://doi.org/10.1145/2389176.2389218>
- [16] Paul Dourish. 2014. Reading and Interpreting Ethnography. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (Eds.). Springer, New York, NY, 1–23. https://doi.org/10.1007/978-1-4939-0378-8_1
- [17] Robert Elliott, Constance T. Fischer, and David L. Rennie. 1999. Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology* 38, 3 (1999), 215–229. <https://doi.org/10.1348/014466599162782> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1348/014466599162782>.
- [18] Casey Fiesler, Jed R. Brubaker, Andrea Forte, Shion Guha, Nora McDonald, and Michael Muller. 2019. Qualitative Methods for CSCW: Challenges and Opportunities. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. Association for Computing Machinery, Austin, TX, USA, 455–460. <https://doi.org/10.1145/3311957.3359428>
- [19] Katie Z. Gach and Jed R. Brubaker. 2020. Experiences of Trust in Postmortem Profile Management. *ACM Transactions on Social Computing* 3, 1 (Feb. 2020), 2:1–2:26. <https://doi.org/10.1145/3365525>
- [20] Eric Gilbert and Karrie Karahalios. 2019. CSCW 2019 Welcome Remarks.

- [21] Barney G. Glaser. 1965. The Constant Comparative Method of Qualitative Analysis. *Social Problems* 12, 4 (1965), 436–445. <https://doi.org/10.2307/798843> Publisher: [Oxford University Press, Society for the Study of Social Problems].
- [22] Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.
- [23] Nahid Golafshani. 2003. Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report* 8, 4 (Dec. 2003), 597–606. <https://nsuworks.nova.edu/tqr/vol8/iss4/6>
- [24] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. <https://doi.org/10.1145/302979.303030>
- [25] Bran Knowles, Mark Rouncefield, Mike Harding, Nigel Davies, Lynne Blair, James Hannon, John Walden, and Ding Wang. 2015. Models and Patterns of Trust. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, Vancouver, BC, Canada, 328–338. <https://doi.org/10.1145/2675133.2675154>
- [26] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. SAGE.
- [27] Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'textquotesingle Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 57–67. <http://papers.nips.cc/paper/8301-ask-not-what-ai-can-do-but-what-ai-should-do-towards-a-framework-of-task-delegability.pdf>
- [28] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 195:1–195:30. <https://doi.org/10.1145/3359297>
- [29] Kathleen M. MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook Development for Team-Based Qualitative Analysis. *CAM Journal* (June 1998). <https://doi.org/10.1177/1525822X980100020301> Publisher: Sage PublicationsSage CA: Los Angeles, CA.
- [30] Carmel Maher, Mark Hadfield, Maggie Hutchings, and Adam de Eyto. 2018. Ensuring Rigor in Qualitative Data Analysis: A Design Research Approach to Coding Combining NVivo With Traditional Material Methods. *International Journal of Qualitative Methods* 17, 1 (Dec. 2018), 1609406918786362. <https://doi.org/10.1177/1609406918786362> Publisher: SAGE Publications Inc.
- [31] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, Montreal QC, Canada, 1–12. <https://doi.org/10.1145/3173574.3173922>
- [32] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 72:1–72:23. <https://doi.org/10.1145/3359174>
- [33] Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. Association for Computing Machinery, Sanibel Island, Florida, USA, 3–8. <https://doi.org/10.1145/2957276.2957280>
- [34] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (May 2000), 286–297. <https://doi.org/10.1109/3468.844354> Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.
- [35] Pablo Paredes, Ana Rufino Ferreira, Cory Schillaci, Gene Yoo, Pierre Karashchuk, Dennis Xing, Coye Cheshire, and John Canny. 2017. Inquire: Large-scale Early Insight Discovery for Qualitative Research. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1562–1575. <https://doi.org/10.1145/2998181.2998363>
- [36] Oliver C. Robinson. 2014. Sampling in Interview-Based Qualitative Research: A Theoretical and Practical Guide. *Qualitative Research in Psychology* 11, 1 (Jan. 2014), 25–41. <https://doi.org/10.1080/14780887.2013.801543> Publisher: Routledge _eprint: <https://doi.org/10.1080/14780887.2013.801543>.
- [37] Gemma Ryan. 2018. Introduction to positivism, interpretivism and critical theory. *Nurse Researcher* 25, 4 (March 2018), 41–49. <https://doi.org/10.7748/nr.2018.e1466>
- [38] Johnny Saldaña. 2009. *The Coding Manual for Qualitative Researchers*. SAGE Publications, Incorporated.
- [39] Anselm Strauss and Juliet M. Corbin. 1997. *Grounded Theory in Practice*. SAGE.
- [40] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 211:1–211:24. <https://doi.org/10.1145/3359313>

- [41] Gregor Wiedemann. 2013. Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 14, 2 (May 2013). <https://doi.org/10.17169/fqs-14.2.1949> Number: 2.
- [42] Jasy Liew Suet Yan, Nancy McCracken, Shichun Zhou, and Kevin Crowston. 2014. Optimizing Features in Active Machine Learning for Complex Qualitative Content Analysis. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics, Baltimore, MD, USA, 44–48. <https://doi.org/10.3115/v1/W14-2513>
- [43] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–11. <https://doi.org/10.1145/3290605.3300468>
- [44] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance Between Human and Machine Understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 56:1–56:23. <https://doi.org/10.1145/3359158>

Received June 2020; revised October 2020; accepted December 2020