Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education

MEI TAN, Stanford University, USA
HANSOL LEE, Stanford University, USA
DAKUO WANG, Northeastern University, USA
HARIHARAN SUBRAMONYAM, Stanford University, USA

Despite the promises of ML in education, its adoption in the classroom has surfaced numerous issues regarding fairness, accountability, and transparency, as well as concerns about data privacy and student consent. A root cause of these issues is the lack of understanding of the complex dynamics of education, including teacher-student interactions, collaborative learning, and classroom environment. To overcome these challenges and fully utilize the potential of ML in education, software practitioners need to work closely with educators and students to fully understand the context of the data (the backbone of ML applications) and collaboratively define the ML data specifications. To gain a deeper understanding of such a collaborative process, we conduct ten co-design sessions with ML software practitioners, educators, and students. In the sessions, teachers and students work with ML engineers, UX designers, and legal practitioners to define dataset characteristics for a given ML application. We find that stakeholders contextualize data based on their domain and procedural knowledge, proactively design data requirements to mitigate downstream harms and data reliability concerns, and exhibit role-based collaborative strategies and contribution patterns. Further, we find that beyond a seat at the table, meaningful stakeholder participation in ML requires structured supports: defined processes for continuous iteration and co-evaluation, shared contextual data quality standards, and information scaffolds for both technical and non-technical stakeholders to traverse expertise boundaries.

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Mei Tan, Hansol Lee, Dakuo Wang, and Hariharan Subramonyam. 2023. Is a Seat at the Table Enough? Engaging Teachers and Students in Dataset Specification for ML in Education. In . ACM, New York, NY, USA, 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Education is a complex and dynamic system [42]. Yet, applications of machine learning (ML) in education rely on generalized approaches with narrow conceptualizations of educational knowledge to analyze learner behavior, interactions, and performance [62]. Consequently, the adoption of ML applications in school administration, instruction, and learning has led to issues of fairness, accountability, transparency, and utility in their implications for practitioners and vulnerable student populations [5, 31, 61, 62, 70]. Harms include systematic inequalities in recommender systems [49] and high-stakes automated decision-making [40], surveillance and civil rights concerns in facial

Preprint, CSCW'24, ,

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

recognition systems [96], data privacy concerns [22, 23, 67], and disparities in student propensities to consent [46].

These issues are rooted in the underlying challenge facing ML design and development, which include undefined policy-level guidelines [79], insufficient teacher education and involvement in ML development [10, 73, 78], the underdevelopment of inclusive and high-quality data systems [24, 56], and the lack of ethical regulation and transparency in data collection, use, and dissemination [61]. Traditional ML development processes undervalue the critical role of trustworthy training data and dataset accountability and largely assume data as given [76].

Further, current engineering processes limit engagement with domain experts and end-users such as educators and students and miss important contextual features of real-world data [85]. When included, domain experts only converge in ML development after crucial data-related decisions have been made [86]. While researchers have created guidelines for downstream data evaluation and documentation (e.g., Datasheets for Datasets [25]), standard practices remain undefined in upstream data specification [30]. Resolving these issues requires addressing the tensions between ML innovations, engineering priorities, and teacher and student needs. Concretely, to create ethical and human-centered ML experiences for education scenarios, we need early collaboration between educators, students, and ML practitioners.

The recent methodological shift in ML practice to re-prioritize the design and quality of training data (i.e., data-centric AI) presents an opportunity to involve teachers and students early in the design of the ML data pipeline [88]. However, mere participation is *not enough* [82]. To truly involve teachers and students, we argue they must be provided with the necessary training and resources to understand and contribute to the design process. This includes ensuring that their input and feedback are considered, working with them to resolve knowledge gaps, contextualizing data needs within domain needs, and negotiating trade-offs around scope and generalizability. Further ML software practitioners should revise their work practices to prioritize domain knowledge and collaboration with domain experts.

In this work, we investigate whether and how teachers and students can work with ML practitioners to define data requirements (the backbone of machine learning models) from the ground up. While prior research has focused on co-designing ML applications with teachers (e.g., [32]), our work looks at the collaborative specification of dataset attributes, labels, and data collection pipeline (i.e., items in the Datasheets for Datasets [25]). We ask the following research questions:

- **RQ1:** What do diverse stakeholders bring to the table when co-designing data specifications?
- **RQ2:** How can we systematically support and amplify diverse stakeholder voices in the ML data specification process?

To investigate these research questions, we conducted a series of co-design sessions engaging experts and stakeholders across domains in collaborative data specification for ML applications in education. Forty participants took part in our study, representing ML engineers, teachers, students, UX designers, and legal experts roles. During these sessions, stakeholders defined dataset characteristics, discussed representativeness and validation criteria, developed labels, and planned ethical data collection strategies for several common application scenarios [55] (e.g., student drop-out risk prediction, automated essay grading, student engagement image classification). We find that teachers and students play a crucial role in contextualizing upstream data-related decisions in downstream use and support the identification of potential biases and reliability threats during data collection and labeling. Further, we identify challenges and needs to deepen stakeholder collaboration to ensure productive participation.

In summary, our work contributes to the emerging practice of data-centric AI, collaborative processes in human-centered AI, and the growing literature on practitioner needs regarding ML applications in education. Through our co-design sessions, we highlight the affordances and limitations of having a seat at the table and discuss directions for future research designing collaborative processes for engaging stakeholders in the education domain. We also discuss the implications of our findings, including developing shared standards, information scaffolds, and supportive tooling to support multi-stakeholder contribution to ML data specification and evaluation.

2 RELATED WORK

The potential for ML systems to create or exacerbate biases, unfairness, and downstream ethical harms has received academic attention across disciplines. The focus of this work investigates the engineering processes in the research and industry environments that build these ML systems. Prior work has highlighted an urgent need for the organizational adoption of tooling and internal processes that support the responsible development and maintenance of fairer systems [44, 80]. These calls to action emphasize two high-level practices: focusing on data work and involving context in the design and development of ML applications [45]. Here we first synthesize existing literature on data practices in ML and then situate our work in current approaches to data documentation and stakeholder collaboration.

2.1 ML Data Pipeline and Current Practices

Compared to software application development, machine learning applications require the complexities of discovering and managing data [2]. The machine learning lifecycle begins with data management, underpinned by a set of system-level requirements, which produces the training dataset used to drive the model learning, model verification, and model deployment stages of the ML workflow [4, 28]. Data management consists of multiple steps, including data acquisition, data annotation, data pre-processing, data augmentation, and data validation [1]. During data acquisition, collecting examples may take the form of searching and indexing existing datasets, distorting and deriving synthetic examples from existing datasets, or creating datasets through data generation techniques [95]. During data annotation, labeling examples may involve the utilization of existing labels, or manually or automatically generating new labels [95]. The data pipeline additionally encompasses the devices and processes involved in storing and moving data [53]. The creation of data used to develop ML systems often requires costly manual work but this work critically affects the trustworthiness of the resulting model [47].

Despite the complexity and significance of data management, current industry practices rely on model-centric development, in which engineering resources are dedicated primarily to iterating the model architecture or training procedure to improve the benchmark performance [47]. Prior work has found 'discretionary' practices in system design [59, 76], ambiguous roles and responsibilities within teams [75], and reliance on individual engineers to identify issues and address ethical concerns [69]. Furthermore, traditional engineering processes limit engagement with domain experts and end-users, separating the work of technical development and understanding end-user requirements [85], and prioritizing technical affordances over the problems of practitioners and real-world contexts [8, 38]. These ad hoc and technology-focused engineering practices have resulted in haphazard data management, in which decisions regarding the definitions of data are forgotten beneath a series of additional decisions, opportunities, improvisations, and assumptions [52]. Practitioners developing ML systems currently face challenges across multiple steps of the data pipeline, including finding, understanding, preparing, and validating data [65, 66]. Audits of dataset development work have found practices that value efficiency over care [60], resulting in an overwhelming majority of datasets that do not meet quality standards [54]. Data-centric practices are undervalued in conventional ML development, resulting in compounding downstream negative effects [72, 76].

2.2 Data-Centric AI and Data Documentation

To address the limitations of model-centric AI practices, recent work has started to focus on data-centric practices, producing supportive tooling for maintaining data repositories and facilitating data annotation and validation [47]. Research in data-centric AI has primarily addressed the downstream harms of low-quality data through the creation of numerous frameworks for facilitating data accountability and transparency through clear documentation practices [3, 7, 15, 18, 25, 68, 71, 93]. The dataset documentation literature introduces standardized processes for datasets to be accompanied by information identifying their motivation, context, composition, features, collection process, biases, recommended uses, and so on (e.g., DataSheets) [25]. Documentation frameworks help engineers understand ethical issues in training data [12] and provide important guidelines supporting accountability in data quality standards.

However, prioritizing data work also necessitates supporting the collection and curation of high-quality data sets in the first place [34] and addressing the upstream work of defining dataset requirements [76]. Ideal data-centric practices begin with specification and defining data requirements according to application needs, but ML systems commonly suffer from incomplete or misinterpreted requirements [1, 14, 20, 36]. Practices that support the specification of dataset requirements early in the data pipeline are understudied in the data-centric ML literature. In education settings, appropriate data specification design in the early stages of ML development is key to mitigating ethical harms in a high-stakes domain [6, 40]. Prior work evaluating AI fairness in education has encouraged research to interrogate the definition of ML problems and data collection procedures and evaluate the quality of training data [40]. Our research investigates the proactive process of data specification, anticipating the evaluative components of documentation frameworks.

2.3 Domain Context and Stakeholder Collaboration

Data is inextricably bound to place and community [89]. The context encoded in data and the context of data production is critical to understanding datasets and their downstream applications [92]. Placing data in their temporal, geographic, and social context, disciplinary norms, and worldly representativeness is a key component of making sense of data [41]. Prior work has called for incorporating more domain knowledge [91], developing domain-specific performance metrics [34, 81], and creating frameworks for documenting context-specific intended use cases [15].

A growing body of research has addressed the elevation of domain context through the study of collaboration and stakeholder participation. AI and HCI communities have increasingly called for more stakeholder participation in the design, development, and maintenance of ML systems [11, 16, 43, 45, 51, 90, 94, 98]. However, meaningful collaborative practice is complicated by the language boundaries of domains and the power dynamics at the intersection of communities of practice.

Firstly, collaboration in social applications of ML involves the complexities of cross-discipline communication. Development practices rooted in silos of expertise limit communication between disciplines. Subramonyam et al. [86] investigated co-creation processes between engineers and user-experience designers and found a separation of concerns between engineers and domain practitioners. Technical experts explore machine learning capabilities independently while making erroneous assumptions about human behavior and contextual needs. Passi and Jackson [58] similarly found a separation of concerns among data science and business analyst experts dividing system accountability tasks. Mao et al. [48] studied the collaborative practices between data scientists and bio-medical scientists and found that these distinct roles often struggle to establish common ground regarding research projects. Work in stakeholder collaboration has additionally emphasized the importance of translation between different forms of knowledge [97]. Hou et al. [35] studied collaborative roles between technical and non-technical workers in a civic data hackathon, noting that the different stakeholders spoke different languages. Collaboration required organizers to understand both data science and context to serve as brokers and translate needs across disciplines. Domain stakeholder involvement requires transparent and interpretable technical explanations [21], but materials for educating stakeholders on ML are scarce [10]. Domain experts face barriers to participation in ML development and decision-making due to persistent knowledge gaps [85].

Secondly, the creative cooperation between stakeholders and technology designers requires the negotiation of values across communities of practice. The involvement of stakeholders in collaborative design efforts evolves from a tradition of participatory design, which highlights an agenda of democratizing innovation by shifting existing power structures and creating a hybrid space between the domains of technology designers and impacted users [77, 83]. Equitable and community-based participatory design emphasize methods that are sensitive to the needs and practices of communities [74]. They aim to foster creativity, learning, and cultural production [19] to design solutions that are considered successful by community metrics [29].

In the education domain, participatory design methods are rarely deployed in the development of AI tools. Though stakeholder involvement is critical for the creation of useful and socially responsible products [13], teachers are often marginalized in technology discussions and engaged only as accessories during the implementation of ML systems [73]. Michos et al. [50] collaborated with educational practitioners to understand practical challenges and iteratively evaluate solutions through workshops and implementation settings, following the structure of design-based research. Holstein et al. [33] involved teachers and students in "participatory speed dating" in order to solicit design feedback regarding AI applications in education. Other studies in education involve end-users through need-finding interviews and product design feedback [99, 100]. While such consultations are valuable, teachers and students are often engaged in a limited capacity as end-users. By participating only in later stages of ML development, long after crucial data-related decisions have been made, opportunities for envisioning equitable design solutions are limited.

In the ML pipeline, data specification is a unique high-leverage stage of involvement for stakeholder participation. End users and domain experts may play a critical role in making transparent what is valued in the data [37]. When engaged early, multi-stakeholder involvement may contribute significant insights to the design of collection and labeling procedures, validation and evaluation measures, modeling choices, and downstream use and maintenance of the dataset and application. By involving diverse stakeholders in education, our work further investigates the expanded role and contribution of teachers, students, engineers, designers and legal professionals in the co-design of data specifications. We position our work at the understudied intersection of stakeholder collaboration in the design of ML data specifications, situated in the unique and high-stakes context of education.

3 METHODOLOGY

To investigate collaborative interactions between stakeholders, we conducted structured co-design workshops with engineers, designers, legal professionals, domain experts, and data subjects (i.e., individuals whose data will be collected). In each workshop, we presented participants with a potential application of ML in the education domain and asked them to collaboratively generate the data specifications for the ML model. The workshop sessions were held virtually via Zoom, with one individual representing each stakeholder's role (a total of 10 workshop sessions). Each workshop session lasted 120 minutes. Our institution's IRB approved the study. Participation was voluntary, and all participants were compensated with $50 for their involvement.

3.1 Participants

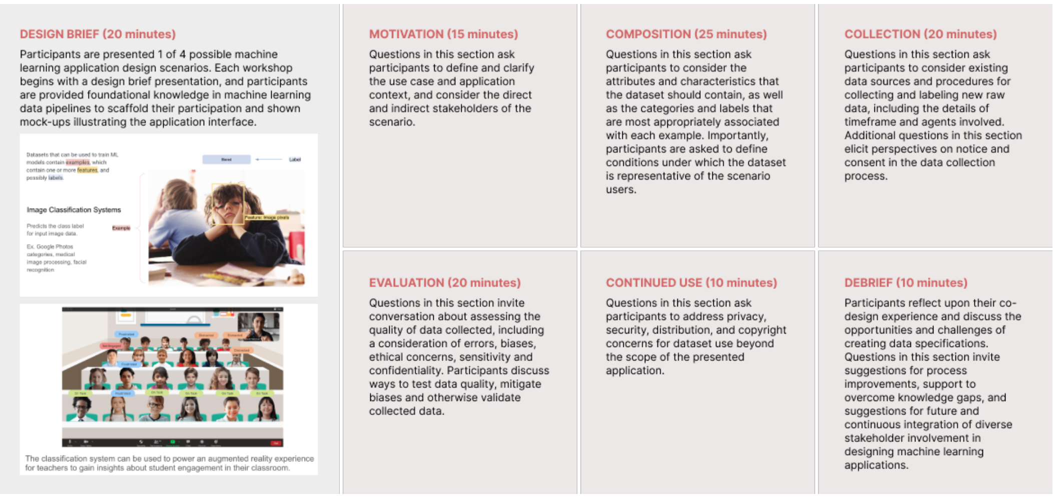
We aimed to recruit one participant from each of the five roles for each session. Because our study is anchored in the education domain, we involved educators in the domain expert role and students in the data subject role. Aside from the student role, all other roles required participants to have at least one year of relevant professional experience. We recruited participants through direct email, mailing lists at university departments and technology companies, and social media posts shared by groups involved in the intersection of AI, ethics, and design. As shown in Table 1, all but one session had participants from four of the five roles. Participants with expertise in the legal and ethical AI domains were challenging to recruit as it is an emerging role in practice. However, in developing the study protocol, we consulted with a legal AI scholar to provide adequate guidance for the group in thinking about legal and ethical requirements and constraints. Further, in cases where two or more scheduled participants were absent, we rescheduled the session and compensated those who were present with an additional $20 for their time (a total of 3 sessions). For session 10, we decided to proceed with the session with three participants as legal professionals were challenging to recruit and schedule. In total, we conducted workshop sessions with 40 participants.

Session	Design Scenario	Participants (Years of Experience)
1	Student Engagement Image Classification	E (25 yrs), T (18 yrs), S, D (2 yrs)
2	Student Engagement Image Classification	E (3 yrs), T (30 yrs), S, D (2 yrs)
3	Student Engagement Image Classification	E (15 yrs), T (2 yrs), S, L (7 yrs)
4	Resume-based Career Recommendation	E (5 yrs), T (9 yrs), S, D (1 yrs)
5	Student Drop-out Risk Prediction	E (3 yrs), T (3 yrs), S, D (5 yrs)
6	Student Drop-out Risk Prediction	E (3 yrs), T (3 yrs), S, D (1 yrs)
7	Student Drop-out Risk Prediction	E (3 yrs), T (8 yrs), S, D (2 yrs), L (5 yrs)
8	Automated Essay Grading	E (2 yrs), T (5 yrs), S, D (1 yrs)
9	Automated Essay Grading	E (7 yrs), T (7 yrs), S, D (2 yrs)
10	Automated Essay Grading	E (1 yrs), T (3 yrs), L (2 yrs)

Table 1. Each workshop session is listed with participants by role (E = machine learning engineer, T = teacher, S = student, D = designer, L = legal/ethics professional) and the associated design scenario. Years of experience in their fields of expertise are indicated parenthetically for each professional stakeholder.

3.2 Workshop Protocol

Motivated by prior research studying collaborative AI design [87], we opted to anchor our workshops on concrete applications of AI in education. Further, we used current *guidelines* on human-centered data specifications and desiderata about *data documentation* as a starting point to develop our workshop protocol. Concretely, the first and second authors analyzed the topics and questions in Datasheets for Datasets [25] to identify those questions that could benefit from multi-stakeholder inputs and can be *proactively* specified before actual data collection. For instance, questions about attributes of each data instance, the meaning of representativeness for the dataset, and collection procedures can all be described upfront. In contrast, questions about sample size and data split between training and test sets are better defined in the later stages of the ML pipeline. As shown in Figure 1, the questions correspond to five main topics for our workshop protocol, including (1) Motivation, (2) Composition, (3) Collection, (4) Evaluation, and (5) Continued Use. Further, to support discussions around each set of questions, we developed guiding prompts and examples based on human-centered data guidelines [27].



DESIGN BRIEF (20 minutes)

Participants are presented 1 of 4 possible machine learning application design scenarios. Each workshop begins with a design brief presentation, and participants are provided foundational knowledge in machine learning data pipelines to scaffold their participation and shown mock-ups illustrating the application interface.

Datasets that can be used to train ML models contain examples, which contain one or more features, and possibly labels.

Image Classification Systems

Predicts the class label for input image data.

Ex. Google Photos categories, medical image processing, facial recognition

The classification system can be used to power an augmented reality experience for teachers to gain insights about student engagement in their classroom.

MOTIVATION (15 minutes)

Questions in this section ask participants to define and clarify the use case and application context, and consider the direct and indirect stakeholders of the scenario.

COMPOSITION (25 minutes)

Questions in this section ask participants to consider the attributes and characteristics that the dataset should contain, as well as the categories and labels that are most appropriately associated with each example. Importantly, participants are asked to define conditions under which the dataset is representative of the scenario users.

COLLECTION (20 minutes)

Questions in this section ask participants to consider existing data sources and procedures for collecting and labeling new raw data, including the details of timeframe and agents involved. Additional questions in this section elicit perspectives on notice and consent in the data collection process.

EVALUATION (20 minutes)

Questions in this section invite conversation about assessing the quality of data collected, including a consideration of errors, biases, ethical concerns, sensitivity and confidentiality. Participants discuss ways to test data quality, mitigate biases and otherwise validate collected data.

CONTINUED USE (10 minutes)

Questions in this section ask participants to address privacy, security, distribution, and copyright concerns for dataset use beyond the scope of the presented application.

DEBRIEF (10 minutes)

Participants reflect upon their co-design experience and discuss the opportunities and challenges of creating data specifications. Questions in this section invite suggestions for process improvements, support to overcome knowledge gaps, and suggestions for future and continuous integration of diverse stakeholder involvement in designing machine learning applications.

Fig. 1. Overview of our study protocol, including a design brief and high-level objectives for each of the data co-design sections.

To establish goals and a common language, workshops began with the presentation of a machine learning design scenario in education and a high-level explanation of the role of data in the intended application. Next, we provided participants with a data specification document with questions detailing considerations in each stage of the data pipeline. While the group collectively brainstormed ideas verbally, a research coordinator facilitated the session and recorded points of consensus in the data specification document screen-shared on Zoom. Our goal was to encourage discussions regarding priorities, trade-offs, and ethical concerns across diverse stakeholders to define data requirements. After each workshop, we asked participants a series of reflection questions to understand challenges in the co-design process and for additional tools or support to improve collaboration. We recorded each workshop session and collected generated data specification artifacts. Below, we detail the main steps in our protocol.

3.2.1 Design Brief. We prepared four ML application design scenarios for anchoring the workshop data specification activity. The scenarios were inspired by recent popular applications of machine learning in education, and our selection favored scenarios using different forms of input data. We aimed to understand whether and how different data types supported and challenged collaboration. The scenarios included a student engagement classification system using *image* data, a student dropout early warning system using *tabular* academic record data, and an automatic essay grading system using *text* input data. Initially, we intended to use resume-based career recommendations as an application of text data in ML. However, based on feedback from session 4, we observed that K-12 educators were less familiar with the application use case. Hence, for the remaining sessions, we opted for AI-based essay grading, which is more familiar to K-12 teachers. In each scenario presentation, workshop participants were shown examples of inputs and outputs to the model, as well as a visual mock-up of the use case and application interface. We broadly described the training data expected through an illustration of the data pipeline with sample features and labels, while noting the many uncertainties in the data specifications left for participants to consider [63]. Workshop participants were then presented with a data specification document and the co-design task. We explained the data specification document as a guidebook for software teams to collect and evaluate the training data used to build the ML application specified in the design scenario. We additionally emphasized to participants that they should work collaboratively, seek the perspectives and expertise of one another, and lean into their unique stakeholder roles to make design decisions. Our study materials are included as a supplement to the paper.

3.2.2 Motivation. In the first stage of the protocol, participants were asked to define the details of the use case and application context for the ML application described in the design brief. Questions in this stage asked stakeholders to identify the people who will directly interact with the system, be directly impacted by the system's operation, or could have a stake in how the system is created, used, or managed. We provided guidance in defining direct and indirect stakeholders and prompted participants to specify the characteristics of the relevant users and environments. Finally, we invited participants to brainstorm and elaborate concrete scenarios to situate the design task and build common ground to support subsequent steps in the protocol.

3.2.3 Composition. In the composition stage of the protocol, participants were first asked to consider the attributes, characteristics, and example instances that the training dataset should contain. We prompted groups that the dataset could contain multiple types or mediums of examples (e.g., documents, photos, people, countries, etc.), and we suggested that groups engage in a generative process with an eye to features that would be predictive of the target ML scenario outcomes (e.g., *"which factors do you think could contribute to whether a student might be at risk of dropping out?"*). Next, participants were tasked with designing the categories and labels that are most appropriately associated with each example. We reminded groups that their chosen labeling schema would define the structure of outputs from the ML model, and we prompted participants to consider the specific context they had chosen in the previous stage. Importantly, participants were asked to define conditions under which the dataset is representative of the scenario users, including the representation and distribution of subgroups. By the conclusion of this stage, stakeholders converge on the specifications for the examples, features, labeling schema, and relative quantities of data representing important traits in the composition of the training dataset.

3.2.4 Collection. In the collection stage of the protocol, participants were asked to design procedures for collecting data based on specifications defined in the previous step. We prompted that approaches could involve the collection of new raw data from relevant communities or accessing and re-purposing existing data sources. We explained that data collection mechanisms could additionally include direct reports by subjects (e.g., survey responses) or inference and derivation from other data (e.g., part-of-speech tags). We additionally asked participants to consider the details of the data collection timeframe and people involved in the collection process, as well as how consent should be requested and provided by individuals represented in the dataset. Participants were then asked to design procedures for labeling the examples in the dataset, specifying the people involved (e.g., domain experts, crowd workers, students), compensation, and timeline. Finally, we invited participants to brainstorm precautions that should be taken to avoid biases introduced by the data collection process and revise the collection procedure to address these risks. At the conclusion of this stage, stakeholders converge on a set of specifications for data collection and labeling and develop an expectation for the shape of the data in preparation for the next step.

3.2.5 Evaluation of Data Quality and Data Cleaning. In the evaluation stage of the protocol, participants were asked to assess the quality of the data collected. To scaffold understanding of data evaluation, we asked participants to first consider how they might measure the quality of a model, noting the desirable and undesirable behaviors in the downstream system. Using the discussion of model quality as common ground, we invited participants to design metrics and validation processes to test for data quality. We prompted groups to consider the potential errors, biases, and ethical concerns in collected data, as well as the characteristics of high-quality data that would inspire confidence in training a high-quality model. Next, we asked participants to design high-level data cleaning procedures, including the removal of low-quality or erroneous examples, the handling of sensitive or confidential data, methods for addressing the under-representation of various subgroups, and the processing of missing data.

3.2.6 Continued Use. In the final specification stage, participants were asked to address privacy, security, distribution, and copyright concerns for dataset use beyond the scope of the presented application. We invited participants to discuss whether the dataset should be distributed for use in future applications and to consider the mechanisms and procedures for data access. We prompted groups to consider the qualities of responsible data stewardship and to brainstorm continued implications of having specified and created the dataset.

3.2.7 Debriefing. At the conclusion of each workshop, participants were invited to reflect upon their co-design experience and discuss the opportunities and challenges of creating data specifications. We asked participants to share the highlights and lowlights of their collaborative co-design experience, recall moments in which they encountered or overcame knowledge gaps, and make suggestions for process improvements in engaging diverse stakeholders in the design of machine learning applications. We explained that the guiding questions used in the data specification design task are an active area of investigation, and we solicited feedback on the order, clarity, and completeness of the protocol.

3.3 Data Analysis

The first author transcribed all workshop sessions first using a Python script with speaker diarization and then, in a second pass, manually verified the transcribed text and speaker roles against the video recordings. Next, we conducted inductive qualitative coding in Atlas.ti [26] using a grounded theory approach [84] beginning with in-vivo analysis. Two authors independently open-coded the same two transcripts and collaboratively developed an initial code book, resolving disagreements by consensus. The resulting codebook consists of 53 codes. The coding scheme included references to procedural data needs (consent, labeling, cleaning, validation, etc.), contextual data needs (representation, bias mitigation, trade-offs, etc.), and collaborative processes (translation, sharing domain expertise, making assumptions, misconceptions, etc.). Using this codebook, we coded the remaining transcripts. The first author applied the code book to analyze the remaining transcripts [17]. Throughout the coding process, the authors wrote reflective memos describing insights and emerging themes and making connections across workshop sessions [9]. Once coding was complete, the research team engaged in multiple discussion sessions. In these sessions, we grouped codes and discussed memos through an iterative sense-making process to identify higher-level themes and synthesize findings across transcripts. Analyses and discussion of themes were informed by the authors' experiences conducting the workshops, as well as by artifacts and notes produced in each session. Our analyses offer insights into the collaborative process for human-centered data specification across stakeholder domains of expertise.

3.4 Positionality

We acknowledge that our research perspectives and approaches are shaped by our own experiences and positionality. Specifically, we are researchers living and working in the U.S., with teaching experience and experiences working with school teachers and district personnel on technology integration, researching the fairness of AI in education, and working with AI practitioners on projects related to human-centered design. In addition, we come from a mix of disciplinary backgrounds, including Computer Science, Learning Sciences and Technology, Education, and HCI, which we have drawn on to conduct prior research into sociotechnical approaches to human-centered AI design practices.

4 FINDINGS

In each workshop session, we provided a diverse group of stakeholders with a specific application of ML in the education domain. We asked them to co-design specifications for each stage of the ML data pipeline. Teams engaged in rich discussions to define dataset composition, collection and labeling procedures, and evaluation metrics. By engaging in generative design thinking, participants shared domain expertise and personal (experiential) perspectives to anticipate challenges and navigate ethical considerations for data subjects and end-users. Across all sessions, knowledge sharing and constant co-evaluation facilitated the conceptualization of a human-centered ML data pipeline from the ground up. We summarize our study findings in terms of (1) contextualizing upstream tasks with downstream use, (2) collaboration strategies across expertise boundaries, and (3) shifting roles, identities, and support needs.

4.1 Contextualizing Upstream ML Tasks within Downstream Use

Typical ML data pipelines are linear and comprised of distinct data and modeling tasks. Our protocol based on current data documentation templates also followed a linear organization. However, participants tended to approach specifications for each component by considering its *interactions* with other stages in the data lifecycle. While current stages in the ML data pipeline are meaningful to engineering tasks, cross-discipline negotiation of concerns transcended discrete steps in the ML data pipeline. As summarized in Table 2, domain experts across all sessions contextualized upstream ML data tasks by considering downstream application context and hypothesized the consequences of collection and modeling decisions in downstream usage. In engaging diverse stakeholders in designing data needs in each stage of the pipeline, we find that collaborative practices can disrupt the backward-looking engineering process of retroactively improving models when performance, utility, or ethical issues surface. Here we present observations about how stakeholders *proactively* anticipate challenges, consider trade-offs, recognize data unknowns, and address bias and reliability threats.

Table 2: Summary of stakeholders' downstream considerations in the education domain associated with upstream data specification tasks and challenges faced by teachers and students in our design workshops.

Upstream Data Task	**Domain Contexts**	**Concerns**	**Unmet Support Needs**
Composition			
Identifying relevant variables	Training data should account for differences across **diverse educational environments** (e.g., public and private institutions, geographic location, grade level, subject of study, and mode of instruction). Representation of subgroups is required along demographic dimensions (e.g., race, gender, socio-economic status) as well as **individual learning needs** (e.g., language proficiencies, neurodiversity, disabilities).	Teachers and students are unsure about the feasibility and ethics of obtaining **sensitive information** (e.g., student perceptions on their relationships with their teachers). Both domain stakeholders express concern about the **fairness and utility** of the model given the numerous critical **factors that data cannot capture** about the student experience (e.g., administrative data does not indicate whether a student is experiencing homelessness or traumas outside of school).	Non-technical stakeholders lack technical knowledge about **data use across the ML pipeline**, including the relationship between training data, application data, and data used for validation (e.g., specifying variables for training data that may be infeasible to collect continuously in application data, hesitating to collect demographic variables under the assumption that they must be model inputs).

Continued on next page

Table 2: Summary of stakeholders’ downstream considerations in the education domain associated with upstream data specification tasks and challenges faced by teachers and students in our design workshops. (Continued)

Upstream Data Task	**Domain Contexts**	**Concerns**	**Unmet Support Needs**
	Nuanced **contextual interpretations** of administrative variables (e.g., separating general absences from excused absences that involve medical leave, student self-perceptions of aptitude detectable from course selection), **student out-of-school factors** (e.g., family and community support, extracurriculars, social network), and **self-reported perceptions** (e.g., writing confidence, classroom trust and safety, boredom) are impactful predictors.	Teachers worry about **misinterpretation of causation** as users attempt to make sense of model inputs and outputs and take subsequent **misinformed action** (e.g., administration blaming student drop-out on teaching quality despite imperfect measures, students learning to insert complex vocabulary rather than improving writing holistically).	Non-technical stakeholders struggle to conceptualize **how variables influence prediction**. This knowledge gap is further complicated by technical handling of different types of data and modeling choices that influence **explainability**.
Developing labeling schema	Labels and attributes should align to **pedagogical goals** (e.g., standards-aligned rubric for essay evaluation along multiple dimensions rather than holistic scoring) and signal actions toward **improving teaching practices** (e.g., identifying lesson activities with low-engagement rather than students who seem bored, identifying specific supports required by students rather than risk of drop-out).	Teachers raise concerns about the complexity of administrative and professional development efforts required to specify **followup action** and **accountability** in response to predictions. Teachers caution against labels that cast assumptions about students and limit **student agency** (e.g., administrative repercussions from classifying students as "drop-outs", behavior management implications from predicting student emotions). Labels impact the design of the final application and how users are trained to interact with it.	Stakeholders lack **common ground**, leaving domain stakeholders to advocate for and explain pedagogical goals, instructional practices, organization of school systems, and sensitive issues in education.
	Labeling schema should account for **multiple standards** across the education system and inherent inconsistencies (e.g., teacher discretion in grading, varying academic standards, varying state requirements for graduation).	Teachers worry about **academic biases** in attributes associated with quality labels in strict evaluative environments (e.g., valuing Standard American English over language familiar to students in their communities)	
Collection			
Identifying data sources	School systems maintain **administrative data** and **historical records** consisting of basic academic and demographic variables. Teachers may also be able to assist with data collection or submit data directly.		Domain-stakeholders struggle with unknown **data ownership** and unknown data management (e.g., deferring to administration without nowing roles responsible for data management or available variables in administrative data, uncertainty about **privacy laws** or what teachers can legally share).
	Educational systems include **third-party partnerships** and interactions with technology, testing, and consulting companies that privately manage data (e.g., College Board, learning management systems, national board for professional teaching).		Stakeholders lack clarity about data collection, management, and privacy **terms from third-party systems**.

Continued on next page

Table 2: Summary of stakeholders' downstream considerations in the education domain associated with upstream data specification tasks and challenges faced by teachers and students in our design workshops. (Continued)

Upstream Data Task	Domain Contexts	Concerns	Unmet Support Needs
Defining collection procedures	Procedures must account for **legal regulations** that govern data collection in protected school-aged populations (e.g., COPPA). Collection may require multiple forms of **data use agreements** (e.g., **informed consent** from parents and legal guardians, informed consent from data subjects, data contracts with administrative data owners and organizations).	Opt-in consent policies may result in **sampling biases** (e.g., overhead of parental consent may deter schools or individual students from participating, volunteered essays may skew toward positive examples).	Teachers and students lack a frame of reference for what they can expect to in terms of **rights and disclosures** detailed in consent forms. Without knowing the highest standards for data privacy and security practices, they cannot evaluate the language of data agreements.
	Consent forms should build trust with **transparency** of purpose and assurances for data management, storage, sharing, and deletion.	Stakeholders worry about the **data privacy** implications of maintaining student identifiers and sensitive information.	
	Procedures should account for contextual factors that may impact data quality such as **temporal variation** (e.g., differing standards and experiences at the start and end of a school year, differing activities at the start and end of a class period), **uncooperative data subjects** (e.g., unreliable or falsified student-submitted data), and **invasive collection methods** (e.g., inauthentic writing tasks, students being aware of being filmed).		
Defining labeling procedures	Labeling should emphasize student agency and fairness through incorporating **student perspectives** (e.g., allowing students to self-identify engagement).		
	Labelers should have **domain expertise** (e.g., experienced teachers, mental health professionals with knowledge of the age group). The **diversity** and representativeness of labelers should match that of the data subjects.	Teachers and students express concern about the **subjective evaluations** and biases that are unavoidable in the education domain (e.g., teacher biases in perceiving student behavior, inconsistent grading standards between teachers).	
Evaluation			
Identifying cleaning and validation requirements	Data may be subject to **missingness** or collection limitations, including biased samples of included schools (e.g., participation from only urban charter schools) and data fields that cannot be collected (e.g., student health details).	Teachers worry about the **transparency of flaws** in the dataset and implications for interpreting model outputs. They note the lack of protocols for documentation and training to name the biases in the data.	

4.1.1 Domain Context Shapes Dataset Specifications. In the composition stage of the protocol, participants initially approached dataset characteristics with broadly defined education contexts. However, in all sessions, teachers and students played a critical role in refining initial specifications in ways that captured the nuances of realistic downstream needs. The first section of Table 2 summarizes how teachers' and students' downstream domain considerations influences their contributions to identifying relevant variables and developing labeling schema. For example, student perspectives provide insights that help to contextualize and re-interpret academic and administrative data. Though models in education commonly predict student outcomes based on an evaluation of academic achievement, student S6 explained that a feeling of academic success and well-being depends on a meaningful combination of variables:

> *S6 (Student Drop-out Risk Prediction): "I think it's not just what their grades are. If somebody is failing out of like AP classes versus like acing like non-AP classes, I feel like, you know, the combination of those things says different stories." (1)*

Teachers similarly identified domain-relevant data features. While normative data practices associate diversity and demographics with a limited set of attributes, teachers highlighted the richness of what diversity means in education. For example, teachers noted the effect **environmental context** such as educational institution type, teacher experience level, urbanicity, teaching quality and subject matter, and community socioeconomic status may have on predicted outcomes. For instance, when specifying student attentional data, teacher T2 explained:

> *T2 (Student Engagement Image Classification): "Diversity can mean so many different things beyond just physical attributes. It's like diversity of environment, what they're working on, because focusing on math might look different than focusing on reading or art. . . or if they're working with other people. . . because those are things that could play into a student's engagement level." (2)*

Further, when defining subgroup representation, teachers advocated for **student-specific context** variables known to significantly differentiate learning experiences and outcomes including student age, language learner status, and first-generation or immigrant status. One contextual variable commonly raised by teachers involves the presence of learning differences or disabilities that would require an individualized education program (IEP). Teacher T1 described:

> *T1 (Student Engagement Image Classification): "The other thing is maybe having data knowing whether a student has an IEP. . . if a student is diagnosed with ADHD they are not going to be as focused as the student without. There are also emotional learning disabilities. . . students who are having particular traumas at home, and these are actually pretty relevant in terms of motivation and engagement." (3)*

Across sessions, the range of encoded information contextualizing diversity stands in contrast to common engineering interpretations of representation standards. By sharing rich anecdotal insights and examples, teachers introduced contextual factors that prompted the group to rethink the **scope and generalizability** of dataset composition. Consequently, participants considered tradeoffs between the size of the dataset (i.e., large-scale collection of contextual variables representing diverse learners and environments) and the scope of the application scenario. Participants expressed concerns about both the amount of data required to ensure equitable distribution of subgroups represented in the dataset and the prediction accuracy for underrepresented subgroups. Subsequently, participants considered whether the model should be designed to apply only to a specific age group, learner status, or type of school. In these discussions, engineers contextualized downstream modeling options and machine learning processes to support data composition and target application decisions. For example, engineer E6 explained the practice of comparing multiple models, in response to a disagreement between non-technical stakeholders regarding semantic differences in public and private school data:

> *E6 (Student Drop-out Risk Prediction): "In machine learning, it's pretty common to have multiple models and it's called ensembling where you just put them all together. You then choose the best results... but then you have different understandings of the same data, and that can be pretty useful, especially if we want to compare..." (4)*

In most sessions, engineers shared expertise highlighting the affordances of machine learning, downstream opportunities to test multiple design choices, and trade-offs in the accuracy and interpretability of technical methods. These technical contributions situate the design of dataset composition in downstream processing and modeling applications, adding methodological context to the real-world factors shared by domain practitioners.

4.1.2 Context Enables Identifying Bias and Reliability Threats in Data Collection. Based on contextual knowledge introduced by teachers, all teams identified threats to data quality in the design of data collection and labeling specifications. The second section of Table 2 summarizes how participants' domain contexts informed their concerns and considerations when identifying data sources, defining collection procedures, and defining labeling procedures. For example, when designing **data collection procedures**, stakeholders leaned on student perspectives as data subjects to anticipate concerns with false, malformed, and missing data. Students shared their personal experiences with survey fatigue, inadequate incentive structures, and creating fake signals. In helping the group to maximize response rates for collecting student resumes, student S4 explained their reactions to various collection strategies:

> *S4 (Resume-based Career Recommendation): "As a student I'm probably not going to bother sending off my resume for no compensation, but compensate me and I might try and game it. However, if you do ask me for the resume with compensation and a survey, I feel like answering the survey questions is maybe going to invest me more than if I was just firing off PDFs." (5)*

Student S2 cautioned that observable signals from students may not align with their needs or experiences. They explained: *"Students always have a way of masking. Even like on Zoom it could look like I'm looking at the screen but I'd be on my phone."* In several sessions, stakeholders grappled with the possible impacts of uncooperative data subjects and adjusted collection procedures to prevent unwanted outcomes. Their hypothesized solutions include information campaigns to improve data subject buy-in, designing non-invasive collection methods to avoid disrupting authentic student behaviors, and enforcing the completion of data fields in the survey instrument to avoid downstream handling of missing data.

Further, data quality threats motivated stakeholders to design **data cleaning procedures** that detect and remove malformed or false data and standardize data fields. Regarding variability in the interpretation of grade labels in collected district writing samples, engineer E10 (Automated Essay Grading) explained: *"Some of the schools are strict on grading, and some of them are not, so maybe we also need to align those... maybe we need to do the data engineering to make all of these standardized."* In another session, a student experience with variable grading encouraged the group to consider labeling schema derived from state standards. Student S8 explained: *"To me, teachers are not always consistent with grading. They all have different perspectives on what A and B are."* Data representativeness and appropriate distributions of diverse subgroups emerged as a recurrent theme regarding data quality in the evaluation stage. Participants designed processes to pursue this standard by returning to the design of collection procedures and augmenting data from underrepresented groups via additional collection processes, extrapolation, or borrowing data from other contexts.

Third, teams brainstormed ways to account for errors in data collection during specification of **data labeling procedures.** Participants committed to hiring multiple labelers and calculating agreement scores, auditing labels in second-hand datasets, and specifying requirements for the qualifications and diversity of labelers as a precaution for obtaining reliable labels. While labeling procedures are discussed in the collection stage of the workshop protocol, participants often discussed data quality standards and data evaluation concepts to anchor their design decisions. Considering the authenticity of labelers, and drawing upon their previous experience as a special-education teacher, designer D2 suggested engaging data subjects in self-identifying labels:

> *D2 (Student Engagement Image Classification): "I think it would be really important actually to have students self identify. The people who are interpreting that facial expression are going to have a different interpretation than the person that made it. But like having the comparison between the students' self reflection and the teachers' perception and measuring that gap in between." (6)*

For sessions in which groups decided to re-purpose an existing dataset rather than collecting raw data, participants expressed skepticism over potentially biased labels and concern for the downstream effects of mislabeled examples.

4.1.3 Validating Specifications by Mapping to Context. Building on the newly acquired contextual knowledge, all teams referenced context as a way of assessing evolving specifications and collection procedures. The third section of Table 2 summarizes how domain context shaped participants' considerations when identifying data cleaning and validation requirements. Few groups noted the importance of transparency of data processes to mitigate applied bias in downstream use cases. As an example, in the automatic essay grading scenario, one of the groups considered reusing second-hand data from a standardized testing service, while observing that work samples from students *financially able* to access the service were over-represented. Specifying the communication of the biases that cannot be mitigated, teacher T8 described:

> *T8 (Automated Essay Grading): "I think also just naming the biases that you cannot reduce, or that you cannot address, so if you're using a certain set of criteria that's constructed by the AP, the emphasis on language conventions is… biased towards standardized academic English. Okay, if we can't eliminate this bias, we can at least name it in our process." (7)*

Further, rather than leaning on technical language and conventional engineering metrics for model performance in the evaluation stage of the protocol, domain experts encouraged groups to prioritize the evaluation of the application as a whole. Across all sessions, groups acknowledged the tradition of standardized quantitative metrics and their inability to capture the real-world effects of a machine learning application. Adding to the conversation about testing procedures validating the accuracy of the model, teacher T9 (Automated Essay Grading) advocated for the most relevant domain-specific performance metric in education: *"The quality should be measured by the learning outcomes of the students… like how they're responding to the feedback that they get from the tool. I know that's really hard."* The challenge of designing evaluation specifications prompted stakeholders to revisit decisions in earlier stages of the data pipeline. This often involved returning to design decisions in the data composition stage and selecting different labels that would better support end-user goals. Despite recognizing the incompleteness of existing metrics, participants faced difficulties creating new ones that better serve contextual needs.

To summarize the section, by moving freely along the data pipeline, participants situated data needs and machine learning processes in domain-aware contexts. They anticipated end-user experiences and proactively mitigated threats to data quality. The separation of concerns between data, modeling, and application work represents an engineering-centric framework. We find that multi-stakeholder groups engage in decision-making holistically, contextualizing data specification with use case elicitation and trade-offs in every stage of application development.

4.2 Collaboration Strategies Across Expertise Boundaries

While role-based knowledge boundaries have traditionally limited opportunities for collaboration between machine learning engineers and domain experts, we observed multi-stakeholder groups engaged in boundary-spanning collaborative practices. Participants employed expertise-specific strategies to overcome knowledge gaps and build cross-discipline understanding. Through practices of translation and advocacy, groups amplified diverse perspectives, built common ground, and navigated ambiguity.

4.2.1 Translation. Non-technical stakeholders often perceive barriers to participation in technical decision-making due to knowledge gaps in machine learning capabilities and processes. Acknowledging the lack of familiarity with ML in the education domain, teacher T9 explained:

> *T9 (Automated Essay Grading): "If you were to go into a classroom, I think the majority of high school teachers in the United States... if you say machine learning and natural language processing and algorithms, they have no idea what you're talking about. That's not because they're stupid, it's just because it's a very niche topic that you don't really hear much about when you're in the classroom. There needs to be some sort of middle ground... some kind of translation to lay folks that don't live in this world of zero and ones." (8)*

While jointly negotiating data needs, engineers across all sessions facilitated collaboration through translation. Concretely, technical experts went beyond merely re-framing practitioner priorities into machine learning terms in data specifications. Engineers in the most successful co-design sessions actively shared technical knowledge to establish common ground and scaffold practical understanding for domain expert participation. One mode of translating contextual data needs into technical specifications involves **evaluating the feasibility** of teacher requests. In the data composition stage, when teacher T5 advocated for augmenting standardized exam scores with local classroom performance metrics, engineer E5 considered the feature in terms of its technical representation:

> *T5 (Student Drop-out Risk Prediction): "Would it be much harder to add in that layer? It's just whether they have passed a class with a certain letter, in this case it's a C or higher."*
> *E5 (Student Drop-out Risk Prediction): "I wouldn't say it's a hard feature to add, it sounds like a binary feature. It's a yes or no, right? You add a column, and the value is yes or no. Yeah, I think that's feasible." (9)*

Engineers across all sessions applied their technical knowledge to support feature requests from teachers whenever feasible. Translation occurred in the encoding of teacher-raised relevant data fields, as well as the planning of technical processes in the data collection and evaluation stages to accommodate use case concerns. Engineers engaged in translation by **clarifying machine learning processes** to support broader practitioner concerns and values. For example, given the publicity surrounding privacy violations and biases along demographic dimensions in machine learning, nontechnical stakeholders displayed a sensitivity to collecting race, age, and gender variables. Addressing confusion and unease about the collection and use of demographic data, engineer E2 explained:

> *E2 (Student Engagement Image Classification): "You can decide to use [demographic variables] to understand your data, and then you already know the data is potentially*

> *biased. So when you build your model, you keep that in mind, and you refine your model to cope with that bias." (10)*

By translating domain priorities into evaluative specifications, the engineer reached across knowledge boundaries and deepened a collective understanding of the use of demographic data in machine learning processes. As a practice, translating allowed engineers to use their technical expertise to amplify the voices of practitioners, enabling non-technical stakeholders to contribute to the construction of human-centered data needs. Using the shared scenario context, engineers **explained trade-offs** in data and modeling choices, building the technical foundation to support domain expert participation. While considering the representation of diverse school settings in the data composition stage, engineer E6 described their considerations in specifying the scope of a model:

> *E6 (Student Drop-out Risk Prediction): "If you train this model for just this one school, then you would be looking at all of the previous data that you have from that school. . . with the downside being that you might not have enough data for the model to learn from, or it might draw the wrong conclusions. One of the benefits of training a model on all the schools in the district is that you have a lot more data points. But the downside of that is like maybe you're at a really poor school, and all the other schools in your district are really rich, so the drop-out patterns might be different."*
> *T6 (Student Drop-out Risk Prediction): "They both seem to have downsides, but maybe per-district is better, because if it's generalized for everyone, then the inaccuracy is higher, but there's a lot more data, so it's better to be more accurate." (11)*

By leaning on the design scenario, the engineer contextualized the effects of technical decisions in domain-relevant terms, enabling the teacher to engage in the evaluation of the trade-off. While technical terms such as "accuracy" and "generalization" had been used previously in the workshop, they had not been taken up by the teacher. By translating the contextual costs and benefits of modeling choices, the engineer empowered the teacher to then take up technical language and contribute to decision-making.

In many sessions, the technical stakeholders took additional care to educate non-technical stakeholders regarding technical details through extended dialogue, actively scaffolding their uptake of technical language in the design process. In a few sessions, engineers employed metaphors and likened ML to familiar analog processes, tailoring technical knowledge in their explanations to serve as a broker between domains. Translation was practiced predominantly by technical experts. Due to the social nature of the education domain, technical experts may more easily make assumptions about educational contexts without requiring translation from domain experts, while the infusion of technology in education is a recent and disruptive change foreign to many domain practitioners.

4.2.2 Advocacy. The high-stakes nature of the educational field necessitates developing machine learning applications prioritizing practitioner experiences. In the collaborative context, domain experts engaged in advocacy, leaning into *extended discourse and emotion-driven language*, urging out-of-domain stakeholders to confront the complexities of education systems and hidden implications of data decisions. As student S4 described, *"you have to try and be an advocate, and if you're going to deploy a system like this you're going to have to come up against people who do advocate for other interests."* Indeed, teachers and students across sessions characterized their collaborative participation as advocacy. They advocated for fairness and utility priorities motivated by their domain contexts and lived experiences while negotiating cross-cutting requirements.

By surfacing critical downstream implications of data labels, feature encoding, and modeling choices, teachers and students voiced values and sensitivities central to the education space. Student

S5 advocated for data subjects by explaining that privacy violations and data misuse put students at risk of negatively impacting future educational opportunities.

> *S5 (Student Drop-Out Risk Prediction): "I would be concerned about teachers or administrators or a committee... overseeing the results... the degree of embarrassment if I did show up as someone likely to drop out... that would imply that you know you're not performing well and something's wrong." (12)*

Teacher T6 similarly expressed concern about downstream harms for students due to the severity of language characterizing labels (e.g., "dropped out" and "did not drop out") in the student drop-out prediction scenario. They warned: *"Then these students will be labeled like dropouts, and then it gives administrators a reason to push students like this out of school."* Teacher T6 instead advocated for student-centric labels and reframed the application scenario to predict whether a student was *"on track to graduate"*. The complex structure of educational systems produces role-based differences in interests, priorities, and interpretations of model results. While school and district administrators value drop-out metrics, teachers prefer a reversed framing featuring student progress towards positive goals, aware of the real-world implications for how students flagged by the system may be treated. By explaining their experience-motivated understanding of mentalities and practices in the downstream application context, teachers in several sessions advocated for data specifications that avoid the perpetuation of system inequalities.

In many cases, groups ultimately adopted the teacher-recommended labels, indicating an openness to identify with practitioner values of supporting student autonomy and avoiding punitive administrative repercussions. However, teachers occasionally received extensive push-back from technical stakeholders. Such back-and-forth patterns between teachers and machine learning engineers are illustrated in Figure 2. In these cases, teachers persisted in their advocacy until groups understood the intent and gravitas behind their concerns. In one session working with the engagement classification scenario, the teacher and engineer engaged in an extended heated exchange regarding the ethics of classifying students with emotion-based labels. Teacher T3 explained the racialized underpinnings of assuming the emotional states of students in classroom practice:

> *T3 (Student Engagement Image Classification): "I would personally feel like that's something I can decide based on myself and my rapport with the students. If a student was frustrated or confused, to use those labels, I would be concerned about stereotyping. It's a really big problem in education, how black students versus white students, how their behaviors read to a lot of white teachers as different, even though it can just be their specific cultural background." (13)*

While advocating against using labels that make assumptions about the emotional states of students, teacher T3 alludes to two other sensitive themes in the education domain: the historical context of racially biased perceptions of student behavior, as well as ML's infringement on the teacher's role of human judgment in the classroom. The exchange emphasized the weight of these critical tensions in education and the prominence of racial and cultural considerations in the domain. A similarly heated discussion in another session involved the collection of teaching quality evaluation data for predicting student drop-out risk. Teacher T5 argued that the use of teacher evaluation data by administrators would impact teacher unions and staffing policies, further complicating an existing district struggle with protections for teachers. By contextualizing the social and political constructs connected to technically objective variables, teacher advocacy enabled groups to collectively situate data decisions in a complex ecosystem and approach designs with respect and sensitivity toward the worldviews encapsulated.

Across many sessions, advocacy operated through the sharing of personal lived experiences. While designing dataset composition in the automated essay grading scenario, teacher T9 argued against relying on quantified rule-based grammatical features to evaluate student writing:

> *T9 (Automated Essay Grading): "When I taught English on the west side of Chicago, ninety-nine percent of my students were African American. I understood what they were saying, but it was not written sort of like traditional academic American English. You don't want to penalize the student for the culture that they live in, and the language that they speak." (14)*

Teacher T9 reflected on their personal struggle to both respect individual student backgrounds and prepare students for future strict evaluative environments, and admitted that they still feel uncertain about the balance. The problem-solving nature of the co-design sessions invited sensitive practitioner stories involving difficulties faced in the classroom. Despite the relative vulnerability required of teachers and students compared to other roles, they met the task and eagerly advocated for the complex realities in the domain.

4.2.3 Ambiguity. During time-limited co-design sessions, participants navigated the balance between big-picture discussions and specifying design details. Engaging in high-level discussions required stakeholders to develop a sense of **comfort with ambiguity**, accepting data unknowns and unfinished design decisions. Though participants expressed uneasiness about ambiguity in the design task, engineer E5 noted how the process stands in contrast to designing with a false sense of certainty:

> *E5 (Student Drop-out Risk Prediction): "When we're talking about designing a system everybody wants to pretend they know more than they think. When we talk about making a decision everybody feels like they already know the answer, like they should know the answer. In this kind of setup…I feel that it's okay for me to not know the answer…and I rely on other roles." (15)*

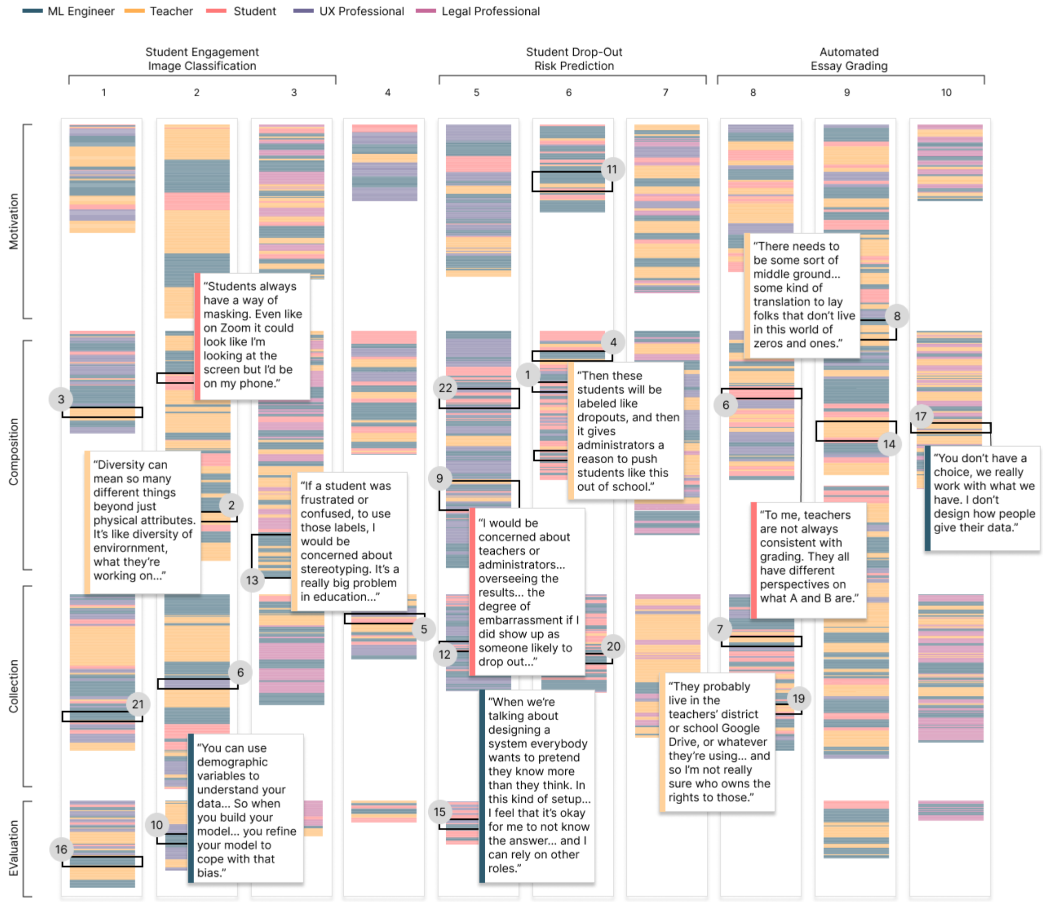
The presence of diverse stakeholders facilitates a collective acceptance of ambiguity while choosing proactive big-picture data planning. Engineers in other sessions echoed their ultimate appreciation of the open-ended nature of design decisions, given the relative infrequency of higher-level conversations in typical machine learning practice.

4.3 Shifting Roles, Identities, and Support Needs

Our study surfaces role-based collaborative dynamics and persistent knowledge gaps and boundaries that *complicate* contribution in multi-stakeholder settings. Although participants engaged productively in the co-design process, we observed groups making assumptions, building on misconceptions, and getting stumped by shared unknowns. Stakeholders struggled with role-based identities and contributions. We identify challenges and support needs for engaging diverse stakeholders in collaborative data specification and summarize these in the final column of Table 2.

4.3.1 Rigid responsibility boundaries. While co-design sessions encouraged many boundary-spanning practices between engineers and domain experts, some role-based boundaries persisted and hindered collaboration. Several engineers maintained a bounded view of responsibilities and liability in data decisions. Especially for evaluations of ethical decisions, fairness for demographic subgroups, and consent practices, engineers were quick to **delegate to specialized entities**. Regarding the representation of subgroups in the composition stage of the protocol, engineer E1 explained:

> *E1 (Student Engagement Image Classification): "This is usually something that you shouldn't be asking just anybody. I'd leave this question up to the ethics review panel professionals." (16)*



"Students always have a way of masking. Even like on Zoom it could look like I'm looking at the screen but I'd be on my phone."

"Diversity can mean so many different things beyond just physical attributes. It's like diversity of environment, what they're working on..."

"If a student was frustrated or confused, to use those labels, I would be concerned about stereotyping. It's a really big problem in education..."

"You can use demographic variables to understand your data... So when you build your model... you refine your model to cope with that bias."

"Then these students will be labeled like dropouts, and then it gives administrators a reason to push students like this out of school."

"I would be concerned about teachers or administrators... overseeing the results... the degree of embarrassment if I did show up as someone likely to drop out..."

"When we're talking about designing a system everybody wants to pretend they know more than they think. In this kind of setup... I feel that it's okay for me to not know the answer... and I can rely on other roles."

"There needs to be some sort of middle ground... some kind of translation to lay folks that don't live in this world of zeros and ones."

"To me, teachers are not always consistent with grading. They all have different perspectives on what A and B are."

"They probably live in the teachers' district or school Google Drive, or whatever they're using... and so I'm not really sure who owns the rights to those."

"You don't have a choice, we really work with what we have. I don't design how people give their data."

Fig. 2. Visualization of role-based contributions in workshop discussions across stages of data specification. Each horizontal line represents one sentence of speech. Selected quotes are marked by number.

In an industry made efficient through role-based specialization, engineering **responsibilities may be narrowly defined**. Ethical standards are often handled by designated professionals in teams separate from engineering processes. Beyond preferring a separation of concerns, engineers indicated being accustomed to a standard industry practice removed from dataset design choices. In the automated essay grading scenario, engineer E10 explained their unease in the data collection stage of the protocol, while deciding between several ideas for labeling schemes:

> *E10 (Automated Essay Grading): "We don't usually have that many options. You don't have a choice, we really work with what we have. I don't design how people give their data." (17)*

While engineers engaged in knowledge-sharing and translating practices, some did so to varying degrees of effectiveness. Despite an engineering effort to break down technical barriers, some teachers continued to feel intimidated by technology blindness. Others further struggled to contribute when additionally perceiving a misalignment between the student age group they had experience teaching and the target student age group for the machine learning application. Teacher T4 explained:

> *T4 (Resume-base Career Recommendation): "I didn't think I had as much to add from an education perspective because my background is with a lot younger students." (18)*

Despite differences in the age group or subject matter in which teachers are more experienced, teachers nonetheless contribute domain expertise. By underselling their understanding of the context from working in the education system as an educator more generally, some teachers viewed their own stakeholder role as a direct end-user rather than co-designer. In several sessions, teachers fell more silent to self-imposed boundaries and misconceptions of qualifications for participation.

4.3.2 Persistent knowledge gaps. Groups face knowledge gaps regarding the identification of data owners. Despite the collection of educational data from academic records, classroom observations, and student work, teachers were unsure of data ownership regulations. Referring to collecting student essay data, teacher T8 explained:

> *T8 (Automated Essay Grading): "They probably live in the teachers' district or school Google Drive, or whatever they're using. . . and so I'm not really sure who owns the rights to those." (19)*

Data ownership is further complicated by the storage of data through ed-tech systems with opaque data privacy terms negotiated with school administration. Even for data assumed to be maintained directly by school administration, teachers could not identify ownership and access processes. With regard to obtaining student academic records, multiple teachers expressed the need for school administrators to be present. Circumstances in education systems introduce variance in roles and responsibilities across districts, complicating the task of bringing stakeholder voices to the table. For domain experts, this knowledge gap may point to an unknown stakeholder and data owner in school administration who should be engaged in designing data specifications. Engineers similarly make assumptions about data availability. Regarding a variety of student academic and financial records, engineer E6 assumed:

> *E6 (Student Drop-out Risk Prediction): "A lot of this data can be directly collected through the College Application, right? Sounds like a safe bet to say the university has all of them, I mean they keep a record of everything. It's probably already marked in their system." (20)*

For engineers, these assumptions may be reflective of their training, in which datasets are given rather than constructed. Despite unknowns regarding data use agreements, shape and composition of data, and the right data owner to contact, engineers maintained confidence that data exists.

Despite involving stakeholders advocating for the legality of data privacy issues, knowledge gaps persist in data security and access. Engineer E1 used an assumption of security to justify haphazard consent practices:

> *E1: "For educational applications where there's going to be very little chance of any kind of deleterious impact or data leak, . . . you can ask the children to check a box saying my parents approve." (21)*

Both technical and non-technical participants expressed relaxed attitudes toward data security, citing a lack of incentive for malicious data breaches. Teacher T2 explained, *"I don't see why anybody would even want to hack into something that the schools were using."*

4.3.3 Shifting stakeholder identities. While the teacher, engineer, and legal professional held clear participatory roles with established expectations for contribution, the collaborative identities of the student and UX professional roles were less defined. The distinct participation patterns across participants is illustrated in Figure 2, indicating the consistent interactions between teachers and engineers in contrast to the variance in the contributions of students and UX professionals. Student S4 explained, *"consistently occupying the student perspective... is difficult because the students are not going to be involved in every stage."* In several sessions, students expressed discomfort being the youngest in a group of professionals and lacking a defined structure of contribution. As a result, they often removed themselves from their primary role of end-user and contributed to the collective task in a general co-designer capacity. Students eagerly participated in technical co-design, offering data collection and modeling ideas unrelated to the student perspective. For example, while the group discussed potential features to include in the dataset composition stage of the protocol for the student drop-out prediction scenario, student S5 explained their ideas for unconventional survey methods:

> *S5 (Student Drop-out Risk Prediction): "it might be interesting to have students survey each other almost, and... this feels kind of creepy, but assigning one person in each of your classes and then say 'does this person seem okay or do they seem like they're gonna fail out of school', I think that that could be interesting."*
> *E5 (Student Drop-out Risk Prediction): "I could see the potential bullying material for that." (22)*

In an effort to contribute to the design task, the student had **lost empathy with data subjects**, requiring the engineer to point out the potential downstream harms. Several students in the student drop-out prediction scenario additionally advocated for collecting mental health and other sensitive information from students. Across all sessions, students often referred to data subjects as "they" and separate from themselves.

We find that, across all sessions, stakeholders lean on prior roles and experiences, often demonstrating multiple competencies, and shifting between these identities throughout the collaborative co-design process. Several legal professionals additionally had technical experience such that they could contribute ML best practices and participate in translation practices. Several designers and engineers had prior experience in various positions in the education sector such that they could contribute ideas motivated by domain expertise to contextualize data. Every stakeholder either remembers the experience of having been a student or is closely connected with a student through their social relationships, such that they could speak on behalf of student interests. Meanwhile, students struggled to always contribute through the role of a data subject and end-user. They instead often opted to participate through the role of a co-designer of data specifications.

UX professionals similarly lacked definition in their participatory roles. In the debriefing of the workshop, engineer E7 reflected, *"this problem didn't have too much to do with the UX perspective, so the collaborators have unequal representation in the discussion."* By organizing co-design workshops that engage diverse stakeholders, the setting of the research study performed some of the traditional roles of designers, confusing the group's understanding of their expected contribution. Across sessions, several designers additionally maintained a role-based separation of concerns, limiting their contribution to technical topics. Removed from technical contribution and lacking the personal domain experiences of teachers and students, designers often faded into the background.

5 DISCUSSION

Our findings demonstrate the vital role that multi-stakeholder collaborations play in the design of dataset specifications. Through conducting workshops anchored in the co-design of ML datasets in the high-stakes field of education, we highlight practitioner efforts to contextualize domain and procedural knowledge, establish common ground, and mitigate downstream harms. Participants engaged in a generative process of negotiating data requirements and quality in each stage of the data pipeline, placing due emphasis on the proactive design of human-centered systems. We emphasize the value of engaging domain experts and discuss the challenges facing the scalable implementation of the collaborative processes explored in this study. We characterize our contributions in terms of implications for the work of data specification and support needs for future integration of multi-stakeholder collaborative processes in responsible data use in education.

5.1 Is a Seat at the Table Enough?

Critical scholarship has explored the tension between developing machine learning systems for scalable production and involving end users in the design of these systems [82]. By establishing a structured co-design environment in which diverse stakeholders were given a seat at the table, many of our participants engaged in organic negotiations to overcome knowledge boundaries to establish common ground through collaborative strategies. We found that participants both made significant contributions to the specification of data requirements and faced challenges in the co-design process. In this section, we discuss the affordances and limitations of stakeholder involvement in our workshop sessions through two participation frameworks and the application context of the education domain.

Delgado et al. describe an analytical framework for the dimensions of participation, designed for practitioners to assess the extent to which a process for participation meaningfully empowers diverse stakeholders in the design of ML applications [16]. Consultation, involvement, collaboration, and empowerment are scaled degrees of participation, assessed at five decision points addressing the motivation, stakes, attendance, form, and power distribution in participation processes. Sloane et al. caution against "participation washing," in which the narrative of participation obscures power dynamics and the extractive nature of collaboration [82]. By critically framing participatory design practices as work, consultation, and justice forms, practitioners may assess the authenticity of collaborative processes. Here we reflect on the nature of participation in our co-design sessions.

5.1.1 Affordances of co-design. Our findings effectively demonstrate the crucial role of collaboration in data specifications, contributing to prior literature by identifying and confirming a critical site of participation in the development of ML applications. Across multiple decision points in Delgado et al.'s framework [16], the participatory structure of our workshops improves upon current practices, which limit teachers to lower degrees of participation, such as engaging in design feedback to improve the user experience of AI systems [33, 73]. In contrast, the co-design of data specification is a participatory structure with stakes that empower stakeholders to contribute to the scope and purpose of ML applications. Data is the backbone of ML models and specifying data requirements is a systematic way to influence system behavior and hold AI accountable to stakeholders. By engaging stakeholders in this high-leverage high-impact stage of the ML pipeline, the design of data attributes and evaluation of data quality can systematically amplify the impact of stakeholder voices.

Teachers shared domain expertise impacting critical features in dataset design and model performance. Out-of-domain experts commonly expressed surprise while recognizing their own knowledge gaps and disrupted assumptions regarding contextual variables relevant to the education domain. When reflecting on their collaborative experiences, engineers across all sessions discussed the narrow technical focus of traditional ML development processes, admitting frequent misplaced efforts and overlooked practitioner priorities. In most sessions, engineers expressed appreciation for the value of collaboration in the early stages of projects and eagerness to incorporate the process into practice. Teachers similarly expressed enthusiasm about contributing to data work. Stakeholders agreed unanimously across sessions that early stakeholder participation in designing data specifications uncovers domain-relevant priorities and potential downstream harms.

Groups additionally realized the value of "big picture" conversations to anticipate future harms by incorporating a contextual understanding of how data choices affect end-users and their environments. Rather than converging on design decisions, groups engaged in a process of ideation and filtration that often resulted in the recognition of multiple possibilities for further exploration. In line with collaborative approaches emphasizing the importance of friction and disagreement [39], diverging perspectives encouraged participants to develop an appreciation for ambiguity. The construction of a jointly negotiated framing, in a participatory process in which diverse stakeholder expertise is valued equally, promotes a plurality of designs in the resulting specifications. In the process, participants admit unknowns and rely on the collective knowledge of those at the table.

5.1.2 Unequal burdens. To engage in effective co-design, participants traversed expertise boundaries by practicing collaborative strategies unique to their roles. We observed domain experts striving to establish common ground by sharing vulnerable personal experiences and advocating for practitioner needs in a complicated historical and socio-cultural context. Out-of-domain stakeholders frequently and confidently made assumptions about the education system. Teachers and students are left with the emotional burden of advocating for their classroom experiences and navigating technology-centric pushback. In contrast, a heavy communicative burden is placed on the role of technical experts as boundary spanners. This finding extends the existing understanding of ML team collaboration where technical members lack domain contexts, in which parties assume a diverse but equal contribution [51, 57, 64]. Because domain practitioners in education feel significant barriers to contributing to the design of highly technical applications, engineers must support the central role of translation in collaborative practice. In the most positive collaborative sessions, engineers exhibited a willingness to teach foundational concepts and processes, patiently explaining technical tradeoffs. By translating domain-specific requirements to data and modeling requirements, engineers were able to address the needs advocated by non-technical domain experts. However, despite the productive ends of these collaborative strategies, they place extensive capacity demands in unique ways on both technical and non-technical participants. Without structured support, multi-stakeholder collaboration is a high-lift endeavor. Across our sessions, groups recurrently fell back on engineer-led linear decision-making when any stakeholder lacked the skill, knowledge, or energy capacity to meet these collaborative requirements.

In Sloane et al.'s framing of the forms of participation, the co-design workshops in this study align most closely with "participation as consultation," in which diverse stakeholders are engaged in various stages of episodic, short-term projects. Consultative forms of participation often take a one-size-fits-all approach, creating a single process and expecting the same form of contribution from all stakeholders. However, diverse stakeholders contribute differently and require different support. Collaborative processes can better engage stakeholder perspectives by designing participation to be context and stakeholder-specific, revisiting processes to ensure the appropriate information is given to and gathered from the appropriate stakeholders.

5.1.3 Unfilled seats. Our co-design workshops represent a lower degree of participation along the dimension of stakeholder selection, as the included community members were chosen by the research team. According to Delgado et al.'s framework, a participation process that truly empowers stakeholders involves engaging community members designated by the community itself. While valuable and necessary, this standard is difficult to achieve in the education context.

The education domain involves complex dynamic systems affected by the attributes of institutions, practitioners, communities, and policy. While teachers contributed eagerly to the co-design of data specifications, participants across sessions grappled with the unknown perspectives of diverse domain-relevant stakeholders not represented in our workshops. Beyond students and teachers, groups named parents, district administrators, school counselors, instructional coaches, support staff and district personnel, community organizations, and policymakers as significant stakeholders in the design scenarios presented. Participants consistently demonstrated knowledge gaps regarding organizational structures in schools and school systems, struggling to identify data owners and match position titles to role-based responsibilities. While participants often referenced "administration" as an agentic entity and critical stakeholder, the specific practitioner role to call upon remains undefined. Further, teachers shared experiences in which they were required to perform the responsibilities of administration and support staff when those roles were unavailable. The complexity of education systems is exacerbated by constant organizational change, situational differences across individual institutions, and overlapping roles due to personnel turnover and resource strain. In the education domain, the non-trivial task of identifying the full set of appropriate stakeholders to bring to the table is a necessary precondition to multi-stakeholder collaboration.

5.2 Supporting Collaborative Data Specification

Prior work assessing engineering processes has noted the lack of defined practices for engaging domain experts and diverse stakeholders, as well as the implications for the fairness and utility of the resulting systems [87]. In response, we formulated our workshop protocol as a preliminary approach to multi-stakeholder co-design of data specifications. Our findings demonstrate that, in order for collaborative data specification to realize its potential of systematically supporting and amplifying diverse stakeholder voices, structural supports are required. We further contribute to the participatory design literature by identifying process needs that facilitate participation by establishing common ground and continuous collaboration practices.

5.2.1 Information Scaffolds. Prior to the co-design workshops, participants were only informed about the research motivations. The design scenario and background about data use in ML development were presented during the session. As a result, participation in our workshops took the form of facilitated group discussions initiated by researchers. While this research design enabled us to observe practitioners navigating knowledge gaps, the introduction of initial groundwork may enable a higher degree of participation. Intentionally designed information scaffolds can establish common ground, overcome technical knowledge gaps, and accelerate proactive contributions across participants.

While the lack of appropriate materials for educating domain experts on foundational ML knowledge is known [10], materials for educating ML practitioners on domain context are equally scarce. Despite a common assumption in collaborative ML work that non-technical experts require scaffolding of technical information, the same assumptions, and requirements and rarely ascribed to technical experts. Pre-reading regarding the social and political context of the design scenario may seed a foundational understanding of domain needs, practices, and motivations. Further, participants reported feeling unsure about the quality of their contributions and uncertain about the expectations of their role. Establishing common ground and defining stakeholder roles prior to engaging in co-design may better prepare participants for richer collaborative discussions.

Information scaffolds may additionally define the collaborative context more effectively. Participants struggled with decision-making due to a lack of clarity regarding the constraints of the design scenario. While the open-ended scenario invited high-level negotiations of priorities and requirements, groups reflected on the potential value of bounding ideation with real-world conditions, factoring financial, labor, and time resources into the initial formulation of the task. Finally, the inclusion of initial groundwork may enhance the generative design process, giving participants time to ideate independently before joint discussion and decision-making.

5.2.2 Shared standards. Across sessions, groups struggled the most with designing specifications for the evaluation of data quality. Participants found this task challenging due to the lack of shared language and metrics for data and system evaluation across disciplines. Non-technical stakeholders were unfamiliar with standard machine learning metrics, such as accuracy, prediction, and recall, and lacked context regarding their applied meanings for the given design scenario. Correspondingly, technical stakeholders were unable to translate practitioner calls for evaluations based on student learning outcomes into actionable data specifications. Despite recognizing the incompleteness of any existing metric, participants faced difficulties creating new ones that better serve contextual needs. We echo the call from prior work that fairness in ML requires the development of domain-specific metrics of quality [81].

5.2.3 Continuous iteration. Applying Delgado et al.'s framework, participation in our workshops is a one-time collaboration needed to better align ML applications with stakeholder needs. While this motivation is representative of a high degree of collaborative participation, the design falls short of empowering stakeholders due to a lack of accountability processes. Without accountability for the quality of implementation of stakeholder contributions, participation in workshops can become performative, failing to actualize the recommendations of diverse stakeholders. According to Sloan et al.'s framework, the most meaningful form of stakeholder involvement (i.e., "participation as justice") requires long-term partnerships with diverse stakeholders, building trust through mutual benefit, reciprocity, equity, and tightly coupled relationships with frequent communication. To establish processes for cross-domain collaboration, prior work has highlighted the importance of design iteration with constant evaluation [87].

Indeed, participants in every workshop session echoed this collaborative requirement, calling for stakeholder involvement at each step in the execution of the data specification. While domain practitioners appreciated the proactive data planning exercise, they expressed concerns about implementation fidelity and the potential for harmful assumptions to re-enter the development process in their absence. Involving multi-stakeholders in continuous collaboration requires the construction of a framework that defines and scaffolds participant roles in iterative data specification in downstream stages of the ML pipeline. Some groups suggested the collaborative creation of a governing set of utility and ethical standards to be used in interval quality evaluations as new data decisions and trade-offs emerge. Support for sustained participation may also involve the development of software platforms to engage stakeholders in the downstream processes of data cleaning and model evaluation. Industry applications of ML development may benefit from the creation of new roles, hiring teachers and boundary-spanners in permanent or semi-permanent positions. The emerging field of education data science may train individuals with expertise at the intersection of technology and education, who are able to translate across domains. Future work should explore these and other processes necessary to sustain iterative and long-term end-user and domain expert participation in data and machine learning development, in each stage of the data lifecycle and beyond.

5.3 Limitations

We present our data specification workshop procedure as a proof of concept with the acknowledgment that our sessions are subject to several limitations. A 2-hour workshop represents an oversimplified data specification process in which time constraints affect the nature of participation. The narrative surrounding the involvement of diverse stakeholders focuses on the prevention of harm by engaging domain context and impacted communities (e.g., [11]). Such framing of priority, combined with time constraints and latent structural inequality between stakeholder roles, limit the contribution of teachers to advocacy around well-recognized challenges in the education domain. When knowledge gaps are large enough, collaborative time is dedicated to establishing baseline domain understanding, leaving the full potential of stakeholder contribution unexplored. We chose to conduct design workshops despite literature acknowledging their limitations in participatory design [29, 74] because they allow us to imagine the research methods that may be adapted to authentic contexts. The demonstration presented here cannot assess the efficiency, feasibility, or affordability of collaborative data specification procedures. However, we identify their promise in addressing downstream issues of model-centric development and invite future work to explore the integration of this practice in industry settings.

Furthermore, our sampling methods may have introduced selection bias, favoring stakeholders who felt fewer barriers to participation and displayed an eagerness to seek collaborative work. Our participants additionally comprised an incomplete representation of stakeholders' communities. Due to convenience and ethical regulations regarding participation in research, the student role was represented by undergraduate students, rather than younger students more accurately impacted by the largely K-12 design scenarios. We recruited ML engineers from a variety of research and industry backgrounds, and teachers specializing in various age groups and subject disciplines, and we do not account for the role of this heterogeneity in the participatory results.

Finally, the four design scenarios produced heterogeneity across sessions and introduced different challenges and design discussions due to the nature of the datasets involved (e.g., tabular, text, and image data). Some scenarios were more emotionally charged, while others were more challenging technically to stakeholders. For example, the undertone of surveillance in scenarios involving student images prompted richer discourse about racial biases, representation, and fairness than in scenarios involving text data. While tabular data was easier to conceptualize, stakeholders were less familiar with image and text processing for data cleaning procedures, and these sessions relied heavily on technical experts to explain processing.

6 CONCLUSION

The emerging fairness, accountability, transparency, and utility concerns surrounding the development of ML applications in education are rooted in the limitations of conventional ML engineering processes. Developing ethical and human-centered ML experiences for education scenarios requires the prioritization of high-quality data contextualized through early collaboration with teachers and students. By engaging diverse stakeholders in a series of co-design sessions, we observed meaningful contributions to dataset specification. Participants shared domain and technical expertise to contextualize data needs, advocate for stakeholder values, anticipate downstream implications, overcome knowledge boundaries, and establish common ground. However, despite the many affordances of our collaborative process, a seat at the table is not enough. Empowering stakeholder perspectives in ML dataset specification requires systematic support, including accountable processes for the continuous involvement of teachers and students in iteration and co-evaluation, shared contextual data quality standards, and information scaffolds for both technical and non-technical stakeholders to traverse expertise boundaries.

ACKNOWLEDGMENTS

We are grateful to the reviewers and our study participants for their time and helpful feedback. This work was funded through a grant from the McCoy Family Center for Ethics in Society at Stanford University. Dakuo Wang was supported by IBM Research as a visiting researcher at Stanford's Human-Centered AI Institute.

REFERENCES

[1] Rama Akkiraju, Vibha Sinha, Anbang Xu, Jalal Mahmud, Pritam Gundecha, Zhe Liu, Xiaotong Liu, and John Schumacher. 2020. Characterizing machine learning processes: A maturity framework. In *International Conference on Business Process Management*. Springer, 17–31.

[2] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[3] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[4] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–39.

[5] Ryan S Baker and Aaron Hawn. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32, 4 (2022), 1052–1092.

[6] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education* 32, 4 (Dec. 2022), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

[7] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[8] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).

[9] Melanie Birks, Ysanne Chapman, and Karen Francis. 2008. Memoing in qualitative research: Probing data and processes. *Journal of research in nursing* 13, 1 (2008), 68–75.

[10] Veronika Bogina, Alan Hartman, Tsvi Kuflik, and Avital Shulner-Tal. 2022. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 808–833.

[11] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint arXiv:2011.13416* (2020).

[12] Karen L Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[13] Amanda Buddemeyer, Erin Walker, and Malihe Alikhani. 2021. Words of Wisdom: Representational Harms in Learning From AI Communication. *arXiv preprint arXiv:2111.08581* (2021).

[14] Harshitha Challa, Nan Niu, and Reese Johnson. 2020. Faulty requirements made valuable: On the role of data quality in deep learning. In *2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*. IEEE, 61–69.

[15] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954* (2022).

[16] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". *arXiv preprint arXiv:2111.01122* (2021).

[17] Norman K Denzin, Yvonna S Lincoln, Michael D Giardina, and Gaile S Cannella. 2023. *The Sage handbook of qualitative research*. Sage publications.

[18] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2342–2351.

[19] Carl DiSalvo, Andrew Clement, and Volkmar Pipek. 2012. Participatory design for, with, and by communities. *International handbook of participatory design* (2012), 182–209.

[20] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* (2020).

[21] Vladimir Estivill-Castro, Eugene Gilmore, and René Hexel. 2022. Constructing Explainable Classifiers from the Start—Enabling Human-in-the Loop Machine Learning. *Information* 13, 10 (2022), 464.

[22] Todd Feathers. Jan 11, 2022. This Private Equity Firm Is Amassing Companies That Collect Data on America's Children. *The Markup* (Jan 11, 2022). https://themarkup.org/machine-learning/2022/01/11/this-private-equity-firm-is-amassing-companies-that-collect-data-on-americas-children.

[23] Todd Feathers. Jan 13, 2022. College Prep Software Naviance Is Selling Advertising Access to Millions of Students. *The Markup* (Jan 13, 2022). https://themarkup.org/machine-learning/2022/01/13/college-prep-software-naviance-is-selling-advertising-access-to-millions-of-students.

[24] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*. 225–234.

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[26] ATLAS.ti Scientific Software Development GmbH. 2023. ATLAS.ti: The Qualitative Data Analysis & Research Software. https://atlasti.com/

[27] Google. 2019. People + AI Guidebook. https://pair.withgoogle.com/

[28] Philip Guo. 2013. Data science workflow: Overview and challenges. *Commun. ACM* (2013).

[29] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[30] Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[31] Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2022. Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 504–526.

[32] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics* 6, 2 (2019).

[33] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2019. Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 157–171.

[34] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[35] Youyang Hou and Dakuo Wang. 2017. Hacking with NPOs: collaborative analytics and broker roles in civic data hackathons. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–16.

[36] Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, Andrew Gelman, and Arvind Narayanan. 2022. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *arXiv preprint arXiv:2203.06498* (2022).

[37] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[38] Hannah Kerner. 2020. Too many AI researchers think real-world problems are not relevant. *Opinion. MIT Technology Review* (2020).

[39] Mahmoud Keshavarz and Ramia Maze. 2013. Design and dissensus: framing and staging participation in design research. *Design Philosophy Papers* 11, 1 (2013), 7–29.

[40] René F. Kizilcec and Hansol Lee. 2022. Algorithmic Fairness in Education. In *The Ethics of Artificial Intelligence in Education*, W. Holmes & K. Porayska-Pomsta (Ed.). Routledge, Chapter 7.

[41] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2021. Talking datasets–understanding data sensemaking behaviours. *International journal of human-computer studies* 146 (2021), 102562.

[42] Matthijs Koopmans. 2020. Education is a complex dynamical system: Challenges for research. *The Journal of Experimental Education* 88, 3 (2020), 358–374.

[43] Sean Kross and Philip Guo. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[44] M Latonero, M Kleinman, and K Hiatt. 2017. Tech Folk:'Move Fast and Break Things' Doesn't Work When Lives Are at Stake. *The Guardian, February* (2017).

[45] David Leslie, Michael Katell, Mhairi Aitken, Jatinder Singh, Morgan Briggs, Rosamund Powell, Cami Rincón, Antonella Perini, Smera Jayadeva, and Christopher Burr. 2022. Data Justice in Practice: A Guide for Developers. *arXiv preprint arXiv:2205.01037* (2022).

[46] Warren Li, Kaiwen Sun, Florian Schaub, and Christopher Brooks. 2022. Disparities in students' propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 564–608.

[47] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (2022), 669–677.

[48] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[49] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. 2022. Equality of learning opportunity via individual fairness in personalized recommendations. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 636–684.

[50] Konstantinos Michos, Charles Lang, Davinia Hernández-Leo, and Detra Price-Dennis. 2020. Involving teachers in learning analytics design: Lessons learned from two case studies. In *Proceedings of the Tenth international conference on learning analytics & knowledge*. 94–99.

[51] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[52] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[53] Aiswarya Raj Munappy, Jan Bosch, and Helena Homström Olsson. 2020. Data pipeline management in practice: Challenges and opportunities. In *International Conference on Product-Focused Software Process Improvement*. Springer, 168–184.

[54] Tadhg Nagle, Thomas C Redman, and David Sammon. 2017. Only 3% of companies' data meets basic quality standards. *Harvard Business Review* 95, 5 (2017), 2–5.

[55] Hannele Niemi, Roy D Pea, and Yu Lu. 2023. AI in Learning: Designing the Future.

[56] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. 2014. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology* 45 (05 2014). https://doi.org/10.1111/bjet.12156

[57] Soya Park, April Yi Wang, Ban Kawas, Q Vera Liao, David Piorkowski, and Marina Danilevsky. 2021. Facilitating knowledge sharing from domain experts to data scientists for building nlp models. In *26th International Conference on Intelligent User Interfaces*. 585–596.

[58] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[59] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (2020), 2053951720939605.

[60] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[61] Francesc Pedro, Miguel Subosa, Axel Rivas, and Paula Valverde. 2019. Artificial intelligence in education: Challenges and opportunities for sustainable development. (2019).

[62] Carlo Perrotta and Neil Selwyn. 2020. Deep learning goes to school: Toward a relational understanding of AI in education. *Learning, Media and Technology* 45, 3 (2020), 251–269.

[63] Nadia Piet. 2019. AI Meets Design. http://aimeets.design/

[64] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[65] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2017. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1723–1726.

[66] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[67] Isak Potgieter. 2020. Privacy concerns in educational data mining and learning analytics. *The International Review of Information Ethics* 28 (2020).

[68] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *arXiv preprint arXiv:2204.01075* (2022).

[69] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[70] Justin Reich and Mizuko Ito. 2017. From good intentions to real outcomes: Equity by design in learning technologies. *Digital Media and Learning Research Hub* (2017).

[71] John T Richards, David Piorkowski, Michael Hind, Stephanie Houde, Aleksandra Mojsilovic, and Kush R Varshney. 2021. A Human-Centered Methodology for Creating AI FactSheets. *IEEE Data Eng. Bull.* 44, 4 (2021), 47–58.

[72] Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* 94 (2019), 15.

[73] Ido Roll and Ruth Wylie. 2016. Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 582–599.

[74] Daniela K Rosner, Saba Kawas, Wenqi Li, Nicole Tilly, and Yi-Chen Sung. 2016. Out of time, out of place: Reflections on design workshops as a research method. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1131–1141.

[75] Jeffrey S Saltz and Nancy W Grady. 2017. The ambiguity of data science team roles and the need for a data science workforce framework. In *2017 IEEE international conference on big data (Big Data)*. IEEE, 2355–2361.

[76] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[77] Elizabeth B.-N. Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *CoDesign* 4, 1 (2008), 5–18. https://doi.org/10.1080/15710880701875068 arXiv:https://doi.org/10.1080/15710880701875068

[78] Daniel Schiff. 2021. Out of the laboratory and into the classroom: the future of artificial intelligence in education. *AI & society* 36, 1 (2021), 331–348.

[79] Daniel Schiff. 2022. Education for AI, not AI for Education: the role of education and ethics in national AI policy strategies. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 527–563.

[80] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).

[81] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818* (2020).

[82] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423* (2020).

[83] Marc Steen. 2013. Co-Design as a Process of Joint Inquiry and Imagination. *Design Issues* 29, 2 (04 2013), 16–28. https://doi.org/10.1162/DESI_a_00207 arXiv:https://direct.mit.edu/desi/article-pdf/29/2/16/1715163/desi_a_00207.pdf

[84] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research.* Sage publications.

[85] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 481, 21 pages. https://doi.org/10.1145/3491102.3517537

[86] Hariharan Subramonyam, Jane Im, Colleen Seifert, and Eytan Adar. 2022. Solving Separation-of-Concerns Problems in Collaborative Design of Human-AI Systems through Leaky Abstractions. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[87] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. 2021. Towards a process model for co-creating AI experiences. In *Designing Interactive Systems Conference 2021*. 1529–1543.

[88] Hariharan Subramonyam, Colleen Seifert, and MI Eytan Adar. 2021. How Can Human-Centered Design Shape Data-Centric AI?. In *NeurIPS Data-Centric AI Workshop. Retrieved from https://haridecoded. com/resources/AIX_NeurIPS_2021. pdf.*

[89] Alex S Taylor, Siân Lindley, Tim Regan, David Sweeney, Vasillis Vlachokyriakos, Lillie Grainger, and Jessica Lingel. 2015. Data-in-place: Thinking through the relations between data and community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2863–2872.

[90] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. 2020. AI for social good: unlocking the opportunity for positive impact. *Nature Communications* 11, 1 (2020), 1–6.

[91] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[92] Janet Vertesi and Paul Dourish. 2011. The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 533–542.

[93] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Michael Muller, Soya Park, Justin D Weisz, Xuye Liu, Lingfei Wu, and Casey Dugan. 2022. Documentation Matters: Human-Centered AI System to Assist Data Science Code Documentation in Computational Notebooks. *ACM Transactions on Computer-Human Interaction* 29, 2 (2022), 1–33.

[94] Michael Weber, Martin Engert, Norman Schaffer, Jörg Weking, and Helmut Krcmar. 2022. Organizational capabilities for ai implementation—coping with inscrutability and data dependency in ai. *Information Systems Frontiers* (2022),

1–21.

[95] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal* (2023), 1–23.

[96] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI now report 2018.* AI Now Institute at New York University New York.

[97] Marian G Williams and Vivienne Begg. 1993. Translation between software designers and users. *Commun. ACM* 36, 6 (1993), 102–103.

[98] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[99] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. StoryBuddy: A Human-AI Collaborative Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement. In *CHI Conference on Human Factors in Computing Systems*. 1–21.

[100] Qi Zhou, Wannapon Suraworachet, Stanislav Pozdniakov, Roberto Martinez-Maldonado, Tom Bartindale, Peter Chen, Dan Richardson, and Mutlu Cukurova. 2021. Investigating students' experiences with collaboration analytics for remote group meetings. In *International Conference on Artificial Intelligence in Education*. Springer, 472–485.