Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development

Aspen Hopkins*
Massachusetts Institute of Technology
Cambridge, USA
dataspen@mit.edu

Serena Booth*
Massachusetts Institute of Technology
Cambridge, USA
sbooth@mit.edu

ABSTRACT

Practitioners from diverse occupations and backgrounds are increasingly using machine learning (ML) methods. Nonetheless, studies on ML Practitioners typically draw populations from Big Tech and academia, as researchers have easier access to these communities. Through this selection bias, past research often excludes the broader, lesser-resourced ML community—for example, practitioners working at startups, at non-tech companies, and in the public sector. These practitioners share many of the same ML development difficulties and ethical conundrums as their Big Tech counterparts; however, their experiences are subject to additional under-studied challenges stemming from deploying ML with limited resources, increased existential risk, and absent access to in-house research teams. We contribute a qualitative analysis of 17 interviews with stakeholders from organizations which are less represented in prior studies. We uncover a number of tensions which are introduced or exacerbated by these organizations' resource constraints—tensions between privacy and ubiquity, resource management and performance optimization, and access and monopolization. Increased academic focus on these practitioners can facilitate a more holistic understanding of ML limitations, and so is useful for prescribing a research agenda to facilitate responsible ML development for all.

CCS CONCEPTS

• **Social and professional topics** → **Socio-technical systems**; **Computing organizations**; **Codes of ethics**.

KEYWORDS

Machine Learning Practice, Contextual Inquiry, Responsible AI, Big Tech

ACM Reference Format:
Aspen Hopkins and Serena Booth. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3461702.3462527

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution International 4.0 License.

AIES '21, May 19–21, 2021, Virtual Event, USA.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462527

1 INTRODUCTION

In past research analyzing ML practice, the vast majority of studies draw participants from Big Tech companies or academia [1, 5, 22, 24, 28, 29, 29, 30, 36, 37, 41, 45, 58, 60], with few exceptions [9, 25, 42]. However, wealthy Big Tech and academic communities offer privileges and perspectives that are *not* universally representative. For example, Simonite [50] chronicled how a collaboration between Google and Carnegie Mellon University collected 300 million labeled images and used fifty GPUs for two months—a scale of development which is increasingly the norm, yet is untenable for less resourced or less experienced organizations. This leads to the question: how well do past studies of Big Tech and academic practitioners encompass the needs of other data and ML workers?

Pereira et al. [42] observed that the diversity of data science teams' compositions, goals, and processes remains understudied—particularly for practitioners outside of Big Tech. This is certainly not the only understudied component of data and ML work outside of Big Tech and academia. We ask: what problems do smaller companies, organizations, and agencies face? What are their practices? How can the AI research community ensure our work is targeted not just at well-resourced organizations but also those with limited fiscal resources and increased *existential risk* [51]? These questions are particularly consequential to future work on ethical and fair practices [12], as these organizations often find applying current best practices in responsible AI development to be too costly.

To answer these questions, we conducted 17 interviews with practitioners working outside of Big Tech and academia in which we discussed current practices, fairness, and risk mitigation in ML. We used thematic analysis to assess these semi-structured interviews and uncovered six broad themes. We explore tensions between privacy and ubiquity, resource management and performance optimization, and access and monopolization. We focus on the subdued impacts of GDPR and privacy legislation, the limited usefulness of model explanations, the trend of deferring responsibility to downstream users and domain experts, and Big Tech's monopolization of access. These tensions reflect underlying and competing concerns of growth and cost, with frequent and complex trade-offs.

While our findings often overlap with those of past practitioner studies, we find that resource constraints introduce additional challenges to developing and testing fair and robust ML models. Further, even universal challenges of responsible development are exacerbated by an organization's resource constraints—a particularly concerning trend considering ML's growing ubiquity and the rapidly developing support for its democratization. Finally, we discuss how the research community can direct future efforts to assist in managing these trade-offs and advocate for giving more research attention and oversight to practitioners in the long-tail of ML development.

2 RELATED WORK

Efforts to understand ML practitioners' challenges are commonly assessed through contextual inquiry, surveys, and interviews. While this paper explores the holistic experience of developing, deploying, and monitoring ML systems, past efforts have typically focused on one of explainable AI, responsible AI, or end user requirements.

2.0.1 Interpretability. Much of the literature on ML practice focuses on interpretability. Kaur et al. [28] conducted a contextual inquiry to explore data scientist use of interpretability tools. Drawing their population from a "large technology company" (presumably Microsoft), they found data scientists over-trusted and misused interpretability tools, and could not generally explain interpretability visualizations. Brennen [9] conducted interviews with ML practitioners working in a wider variety of contexts. Through their interviews, they observed practitioners in academia and research labs want explanations to provide insight into model mechanics, while other practitioners want explanations that use model outputs "more effectively and more responsibly"—an understudied use case for explanations. Our study of practitioners at startups, non-tech, and public service organizations reinforced this need. Hong et al. [25] interviewed ML practitioners who use interpretability methods and found that a core use case for explanations is in building trust both between people and models, and between people within an organization. While our interviews exposed these same desires, few organizations had invested in building interpretability tools, and those that had found only limited success incorporating them.

2.0.2 Responsible AI. Responsible AI—particularly focusing on bias and fairness—is another area of ML which has received extensive research attention and contextual inquiry. Holstein et al. [24] conducted semi-structured interviews and surveys with ML practitioners at "10 major companies" to gain insight into existing practices and challenges in ML fairness. Holstein et al. [24] uncovered that—while the ML fairness research community is largely focused on de-biasing models—these stakeholders instead focus on the problems of data collection and data diversity. We found this approach of focusing on data diversity to be similarly common with practitioners from smaller or less visible organizations.

Rakova et al. [45] introduced a framework for analyzing how company culture and structure impact the viability of responsible AI development. They noted a lack of clarity in how the multitude of proposed frameworks and metrics for responsible AI are translated into practice. When companies introduced such frameworks, practitioners were concerned by the risks of inappropriate or misleading metrics. Further, they found practitioners at Big Tech companies commonly contend with deficient accountability and decision-making structures that only react to external pressures. We found these experiences and structures to be less applicable for smaller and less visible organizations, where individual contributors generally have more decision-making power, but fewer resources and less developed processes for assessing bias and fairness.

2.0.3 User Requirements and Expectations. Another common form of contextual inquiry for ML evaluates how stakeholders will interact with ML systems. To this end, Cai et al. [10] interviewed pathologists before, during, and after presenting neural network predictions of prostate cancer, with a focus on their needs for onboarding this technology. They discovered that these stakeholders wanted more insight into the expected model performance: they wanted insight into the model's strengths and weaknesses, as well as the design objective. Relatedly, Amershi et al. [1] proposed 18 design guidelines for human-AI interaction, and validated these guidelines with design practitioners recruited from a "large software company" (presumably Microsoft). While these design insights are useful for all practitioners, this line of contextual inquiry focuses only on the end human-AI interaction portion of ML development.

2.0.4 Characterizing Practitioner Needs. ML practitioners span diverse fields, industries, and roles. Data scientists are a unique example. This field has been studied separately from ML due to its emphasis on business-adjacent analytics [41], but data scientists are increasingly adopting ML methods. Researchers now study "Human Centered Data Science," which encompasses interest in the shared practices of data scientists and ML practitioners [2, 35, 37]. We adopt this broad view of *who* constitutes an ML practitioner, and we do not differentiate data scientists who use ML methods.

Past research on both ML and data science practitioners categorized various approaches and motivations [22, 36, 41] and assessed stakeholders' communication needs [41]. These past works highlight two key challenges: the need for continuous refinement, and a desire for clarity in communication and objectives. Data work is inherently "messy" [41], and practitioners rarely have domain expertise for any given application. Instead, Viaene [56] found that practitioners learn as they collaborate with new domain stakeholders and grow familiar with the data. The practitioners we interviewed similarly collaborated with domain experts. By and large, interviewees found this process challenging, with communication greatly affected by stakeholders' varied levels of data experience.

Passi and Jackson [41] highlighted how teams iteratively negotiated and justified the worth of data science solutions; similarly, Hohman et al. [22] described the *data iteration* processes involved in ML development at Apple. In contrast to Passi and Jackson [41], Hohman et al. [22] focused on functional iteration, updating datasets to improve models and reduce bias, and not iterating on the communication needs of data science and ML teams. Our interviews with practitioners at smaller or less visible organizations exposed similar trends of constant iteration and reinforced the need for communication as an irreplaceable mechanism for addressing the "messiness" of data work and for building trust.

3 METHODS

We used a combination of targeted, convenience, and snowball sampling to invite participants [16, 17]: we invited participants by cold-emailing, leveraging our professional networks and the networks of participants. We chose interviewees to represent a variety of contexts, from municipal analytics teams to small startups to publicly traded non-tech corporations. We spoke to a combination of CTOs, directors, investors, engineers, and analysts. All practitioners used or directed the use of ML in their respective work. We categorized interviewees based on available resources; this categorization was inferred through interviews and supported by Public Financial Planning and Budgeting Reports and Crunchbase. We present an overview of these interview contexts in Table 1.

ID	Type	Company Description	Interviewee Title	Resources
R1	Publicly Listed	Shopping/recommendations	Data Engineer	$$$$
R2	Startup	Shopping/recommendations	VP of Product	$$
R3	Startup	Shopping/recommendations	VP of Strategy	$$
R4	Publicly Listed	Pet care (diagnostics)	Senior Data Scientist	$$$$
R5	Startup	Healthcare (diagnostics)	Chief Operating Officer	$
R6	Startup	Fitness	Chief Technology Officer	$$$
R7	Startup	Real estate	Chief Technology Officer	$$
R8	Small Company	Real estate	Head Of Analytics	$$
R9	Startup	Real estate	Senior Product Manager	$$
R10	Startup	ML consulting and tools	Chief Technology Officer	$$
R11	Startup	ML consulting and tools	Chief Executive Officer	$
R12	Startup	Data automation	Board Member/Investor	$
R13	Startup	Pet care	Director of Engineering	$
R14	Public Sector	Municipality	Asst. Director of Data Analytics	$
R15	Venture Capital	Investment	Startup/ML Investor	-
R16	Startup	Language learning	Chief Technology Officer	$
R17	Startup	Language learning	Chief Technology Officer	$

Table 1: Overview of interviews, including the type of organization, an overview of the company's main product, the interviewee's title, an interview ID, and an estimate of the companies' overall available resources based on public records and Crunchbase financing reports. One participant's distinctive title has been changed to a functional equivalent to preserve anonymity. In general, $ companies had less than 15 total employees; $$$$ had on the order of 100 engineers.

3.1 Interviewees

All interviewees did some form of advanced ML development, though these efforts were often pursued for internal tools or as unreleased, experimental products in development. Still, several organizations had released products which used ML methods extensively (R1, R4, R5, R10, R11, R13, R14, R17, R16). Every interviewee we spoke with incorporated—at minimum—linear regression methods as a core part of their work, and many used significantly more complex ML techniques. Despite this, we found these organizations often expressed that "we don't do ML," even when ML was advertised in the company's description, marketing, and website. Several companies rejected our requests for interviews on these grounds.

Across our interviews, R1 works at a company most similar to Big Tech: they have extensive resources and advanced engineering practices, and they use ML throughout their business. R2 and R3 work at an early stage shopping and recommendations company, which is actively migrating to a learned recommendation engine. R4 works at a publicly listed, well resourced company that has substantial ML integrations despite not being tech-first. R5 works at an early stage healthcare diagnostics company, and ML is core to their business. R6 is a well-funded, late-stage fitness startup with an experienced ML team, but they use ML in a limited capacity to drive investment. R7, R8, and R9 work in real estate domains, and all three are relatively early in their ML integrations. Of all our interviews, R9 uses the least ML. R10 and R11 work at ML consulting and tooling companies; both have strong expertise and advanced practices. R12 works at a data management company. R13 works at an early ML-focused startup with an emphasis on computer vision. R14 works in a municipal analytics office; their ML models are typically rapid prototypes. R15 supports and invests in ML-focused startups. R16 and R17 work at early stage language learning companies; both are actively developing core ML features.

3.2 Interview Process

We followed a semi-structured interview protocol. While we adapted our questions for each interviewee, we derived these questions from a set of common themes. We sought to survey broader ML and data processes within these organizations—along with the specific challenges these practitioners faced—by eliciting examples, descriptions of existing processes, and anecdotes related to deploying models. We provide the core interview protocols in the supplementary materials, but include samples of questions here:

- What is your process for launching a data-related project?
- How do you evaluate your models or data?
- How do you track projects over time?
- What do updates and changes look like?
- What setbacks have you experienced after launching?
- How do you think about representative data and/or testing?
- How do you think about bias?
- How has GDPR or other data legislation affected you?

Before starting each interview, we obtained permission to transcribe the interview. We intentionally chose not to audio record for interviewee comfort, instead taking notes that included relevant quotes and building transcripts post-interview as described by Rutakumwa et al. [48]. We conducted these interviews by video chat; these lasted an hour on average with the longest lasting two hours and the shortest forty-five minutes. This work is IRB-approved.

3.3 Analysis

We systematically coded and analysed transcriptions of these interviews using *thematic analysis* [8]. This process involves codifying statements and observations from each interview, grouping statements by initial themes, refining and evaluating themes, and then

extracting final themes. The first summary step converted the transcriptions to 945 individual codes. The second step constructed 101 detailed themes. The final step extracted 6 overarching themes.

4 THEMES

In reviewing these interviews, we characterize tensions between development and production, privacy and ubiquity, resource management and performance, and access and monopolization. In presenting these themes, we focus on both those sentiments which support prior research analyzing subsets of ML practitioners and those sentiments which are underrepresented in existing literature.

4.1 "It's Tough." Tensions Between Expectations & Feasibility

Human expectations typically manifest as projections of how the model should behave (R2, R3, R5, R7, R8, R12, R13, R14, R17), and balancing human expectations and feasibility was a recurring theme. Users, leadership, and even engineers often had unrealistic expectations, strong beliefs on the state of ML capabilities, and exceedingly high standards based on prior experiences with other ML products. Practitioners struggled to realize these expectations.

4.1.1 Users' Expectations. Concerns of *user expectations* troubled interviewees. R16 and R17 quickly learned that users wanted products that performed as well as Google Translate *at a minimum*, but that achieving that level of translational proficiency is infeasible for such early language-learning startups. Failing to meet user expectations would lead to a loss of trust (and ultimately users), but the data, compute, and expertise required to reproduce this proficiency were inaccessible. Both companies had contemplated paying for access to Google Translate's API, but this cost was equally untenable and its use would introduce questions of flexibility and transparency for future development. Instead of using proprietary systems, R16 and R17 were building their own models and datasets through in-house labeling and user-provided data; both companies ultimately limited the scope of their respective products to reduce risk. For both companies, this decision potentially cost user engagement.

R8 similarly described experiencing "the big pushback" when attempting to balance expected user behavior, convenience, and the inherent opaqueness of external, proprietary models and APIs. They considered proprietary models like Amazon's AWS tooling to be "good products," if "a lot of black box work." R8 struggled to balance meeting the users' expectations—for which these proprietary tools are often helpful—with the desire to audit, debug, and maintain control of their software. This fundamental tension between meeting expectations—informed by exposure to Big Tech products and models—and managing resource constraints was a concerning and recurring message. These companies found this balancing act to be exceptionally stressful, as they felt user expectations set an untenably high bar given available funding, time, and experience.

4.1.2 Management's Expectations. Interviewees found communicating the limitations of ML to be challenging—especially with non-technical management (R2, R3, R8, R10, R12, R15). For example, R3 described a situation in which management identified a seemingly common pattern, but their ML team found themselves unable to extract this "data story" given their available resources. Their leadership team identified a common pattern of people first exploring wedding-related content such as dresses, then travel and honeymoon ideas, and finally topics related to child rearing. The team hoped to predict when users began this journey for content curation. In practice, R3's team was unable to extract this distinctive trajectory from their data, disappointing leadership. R3 speculated on the causes underlying this failure: they lacked access to labeled groundtruth to assess predictions and needed additional context *outside* the scope of their data to differentiate signal from noise. They ultimately suggested this normative story of marriage, then honeymoon, then child rearing might be less common in practice.

Some interviewees (R3, R10, R12, R8) regarded strong leadership or stakeholder intuition as a warning sign when employing ML tools and techniques. These interviewees found the combination of low data literacy and strong intuition to be most concerning, and remarkably common in both startups and non-tech companies. As R12 explained, "CEOs and executives don't really understand what it takes [to develop and deploy ML]," particularly outside tech-first organizations. One interviewee (R10) declared that in cases of low stakeholder technical and data literacy, they opted not to pursue contracts to avoid wasting time and resources.

4.1.3 Predicting Performance and Cost. The ML community is increasingly aware of *model under-specification*, one symptom of which is that training the same model multiple times on the same data results in different defects [13]. Practitioners—particularly in under-resourced environments—are patently aware of this challenge (R6, R13). At Big Tech companies, a common mitigation strategy is to train a model multiple times on the same data, assess the resulting models, and deploy the best of the bunch. In our interviews, only the most technology-first company (R1) adopted this approach. In line with other less-resourced organizations, R13 discussed how finances expressly prevented them from doing so.

R13 shared that they did not have the resources to train multiple models for the same set of data. As they scaled their business, they collected more data—in their case, data which mostly consists of videos of dogs. With each new data collection iteration, they retrained their model using all available training data, resulting in unpredictable performance drops. They had been debating decomposing their model into simpler models corresponding to specific characteristics, such as dog breed. In this manner, they could scale their product by introducing new models without the risk of compromising past performance on other breeds. Of course, the downside is substantial: instead of testing a single model, they would need to test many. Further, they believed the larger model would be more robust as it should better generalize to underrepresented data—such as mixed breeds, or atypical characteristics—whereas breed-specific models would be substantially worse.

In response to these challenges in developing models and predicting ML development costs, many interviewees (R6, R7, R9, R15, R12, R14) simply wondered whether deploying ML models was worth the trouble at all. One interviewee (R7) had concluded it was not. In reference to to their work in real estate they stated, "I don't have any worries about automating the quality checks," but "we'll never have the density of data to really automate some of these things... the ROI [Return On Investment] on automation might be low." After extensive preliminary development, R7 came to believe that human

intuition and domain knowledge was irreplaceable—models make predictions based on known prior behavior, but domains like urban property appraisals are inconsistent, constantly fluctuating based on ever-changing human priorities.

4.1.4 Discussion: Expectations and Feasibility. Practitioners struggled to meet human expectations—whether from users, from leadership and management teams, or even from themselves. User expectations are dynamic, but are largely informed by the practices of Big Tech companies (R16, R17). When extremely well-resourced organizations release an ML product into the world, it raises the bar of expected quality and effectively *monopolizes access*. To participate, organizations must opt in to use of released, opaque models and APIs or invest beyond their means in collecting data and modeling.

The lore of ML has resulted in non-technical management believing that pattern recognition should be straightforward, but practitioners (R10, R3) often find themselves unable to meet these unrealistic expectations. When we researchers talk about AI and ML, and especially when we engage popular press in these conversations, we establish expectations about the capabilities of ML. As we release new models, tools, and techniques, we set user expectations for feasibility and quality standards. Small and less-resourced organizations struggle to meet these expectations. With this monopolization of access, many questions arise: if organizations use proprietary models downstream, how do they introduce transparency or maintain autonomy? Who is responsible? And what are the implications of these participation monopolies?

When developing models internally, interviewees cited several potentially viable approaches to meeting these expectations, such as replacing large, complex models with multiple, specialized models targeting specific tasks (R13). But this is a relatively new topic, and practitioners would benefit from guidance on *when*, *how*, and *whether* to transition from a single model to multiple specialized models. Understanding the costs and benefits for such transitions is crucial as well—what are the monetary and environmental expenses involved in training and evaluation for each choice [53]?

Further, while we often think of ML as a useful and perhaps inevitable tool, these organizations question that narrative (R6, R7, R9, R15, R12, R14). The decision to implement ML typically falls to management and team intuition, and is sometimes merely employed as a mechanism to drive investment (R6, R12). This community of long-tail, less visible ML practitioners would benefit from standardized recommendations for assessing when ML is most appropriate.

4.2 "A Hotbed of Bias." Efforts to Assess, Prevent, & Mitigate Bias

Nearly every organization we interviewed expressed substantive concern over inadvertently deploying biased models (R1, R3, R4, R5, R7, R10, R11, R13, R14, R16, R17). R11 stated this to be the "biggest business concern" for companies incorporating ML, and other interviewees shared similar sentiments (R16, R17). Yet strategies for uncovering and mitigating bias and its impacts were consistently underdeveloped, pointing to the potency of this problem.

4.2.1 Bias Mitigation Through Diversity or Personalization. A remarkably common mitigation attempt to alleviate bias was through the acquisition of sufficiently diverse data for model training and evaluation (R16, R5, R8, R10, R3, R13, and R14). This strategy contrasts with academic approaches which aspire to debias models assuming fixed data, but mimics the broader practices of Big Tech [24]. Mechanisms for acquiring this diverse data include attempting to develop a sufficiently diverse userbase (R16), ingesting data from varied sources (R8, R10), augmenting available data (R11), collecting diverse data in-house by assessing *axes of diversity* (R13), and incorporating participatory design principles directly into data planning and collection mechanisms to ensure representational data (R14). Another commonly proposed bias mitigation strategy considered model personalization (R17, R16, R2, R13). By giving users increased control over their models, these companies argued that their users could tailor outcomes to their personal needs—circumventing the problems of biased models, or so the logic goes.

Still, data-focused mitigation strategies suffered from drawbacks: practitioners developed ideas of how models would behave on a given dataset, but after each modification or retraining found the resulting performance to be unpredictable. As such, R13 adopted a slow and cautious update protocol: they avoided introducing new, diverse data until user requirements deemed it strictly necessary. By doing so, they minimized how frequently they updated—and potentially broke—their models. R13 pursued this strategy as their product was in Beta, and they found it offered greater consistency. For a deployed model, though, this slow and cautious approach to adding new and diverse data can exacerbate or prolong model biases. Despite the increased risks, this slow and cautious strategy to data changes is common in production: teams frequently use model accuracy to appease investors (R5, R15), which may lead to them opting to deprioritize data diversity to reach investor expectations.

In spite of their varied bias mitigation strategies, interviewees nonetheless remained concerned about releasing biased models. They lamented they had no systematic way of acquiring diverse data (R13), of assessing the axes of diversity (R10, R13), or of assessing and evaluating the performance of personalized models (R17, R16, R2, R13). For example, R17 explained they were considering investing in federated ML for personalization—but assessing the quality of each personalized model is hard [55], as is the engineering challenge behind such a strategy. In two separate vision applications, R13 and R6 further lamented that the scope of diversity needs was *far greater* than they initially anticipated: they must consider not only diversity of image subjects, but also the diversity and quality of seemingly unimportant image backgrounds.

4.2.2 Assessing Blind Spots. A related concern is creating models which suffer from *blind spots*: undetected failures due to missing or biased data and inadequate test coverage. Many interviewees identified human subjectivity in data collection and labeling consistency as the root cause of these failures (R8, R15, R7, R3, R11), contrasting in part with the research community's broader concerns that models do not learn the causal relationships within an available, finite dataset [13]. Human subjectivity affects the full development pipeline from deciding what data to collect to data labeling to model assessment and evaluation. Some interviewees expressed optimism about efforts to use internal quality metrics as mitigation mechanisms to find and ultimately remove these blind

spots (R15, R7, R11, R3). R11 was similarly optimistic that by holistically considering their data and performance, they would be able to isolate data factors related to computational *underperformance*.

Still, interviewees were optimistic that ML could be deployed responsibly: they believe model biases to be more scrutable than human decisions (R7, R14, R5). One interviewee (R5) explained that because they drew training and test distributions from the same imaging devices, they believed they strongly upheld the assumption that training and test data are independent and identically distributed. Further, R5 had collaborations with insurance companies; they argued this resulted in increasingly representative data. As a consequence, they asserted that they were insulated from blind spots. Interviewees expressed desires for better assessments of blind spots and model fairness (R10, R16, R13, R14).

4.2.3 Deferred Responsibility. Several interviewees expressed that biased models were a possibility, but that these models could still be deployed safely by leveraging a final barrier of human judgment (R8, R5, R13, R11, R14): when faced with an incorrect assessment, a human would be capable of overriding the model's decision (R5, R8, R11) or disusing the model (R14). We find this trend of deferred responsibility to be common among these practitioners and troubling. Past research has demonstrated that even when a human is capable of outperforming a model when acting independently, they tend to defer to model predictions when available [15, 54]. One interviewee (R4) expressed concern about this possibility of unintentionally undermining human expertise. To safely deploy their models, R4 was actively collaborating with a design team to emphasize decision uncertainty. However, several other practitioners adopted a notion that building tooling to support downstream users and organizations was unnecessary, believing that users were sufficiently capable without this hand-holding (R11, R5)—a risky position. This highlights the murky notion of responsible parties: are developers responsible for the consequences of introducing ML, or are end users and domain experts who use these models?

4.2.4 Discussion: Assess, Prevent, & Mitigate Bias. Risk and harm assessments were recurring themes. When interviewees deferred responsibility to others, they assumed their products could cause minimal harm. When practitioners considered risks to be sufficiently low, they felt little responsibility to consider the potential for harm. These risks are not distributed equally: prescribing home valuations (R7) is demonstrably more risky to human wellbeing than ranking "cool" shirts (R3). Still, this is a slippery slope of complacency. We assert assumptions of minimal risk are inherently biased and not appropriately calibrated against feedback from users and stakeholders. Even a shirt recommendation can cause harm and perpetuate inequality—e.g., if the model consistently demotes Minority-owned brands. While many interviewees expressed concern for the broader implications of their work (R1, R2, R3, R10, R11, R13, R14, R17), several did not (R9, R6, R8, R12, R5). This apathy was exacerbated when practitioners relied on humans to arbitrate decisions. R5 explained that blind spots were "low risk," with bad outcomes resulting in a "$300 procedure instead of a $50-60 procedure," but the person receiving this treatment might not agree.

Trade-off considerations between robustness and accuracy were a common theme. Instead of the intuitive or ad-hoc methods that our interviewed organizations currently use to assess risk and mitigate bias, we suggest developing methods to recognize risks and harm through atypical perspectives—such as by characterizing expected users and then assessing any embedded biases within these expectations for rectification. Some companies bring in consultancies to do this, but this is untenable with resource constraints. We assert that the research community can help by designing tools which actively encourage critical thinking, monitoring, and assessment for data planning and modeling tasks [33, 34]. Such efforts can better support these less-resourced organizations.

4.3 "You can poke and prod black box models, right?" Black Boxes & Overconfidence

Concerns of bias are often associated with ML techniques operating as "black boxes" [31]. Practitioners had a broad swathe of opinions and strategies for managing, using, evaluating, and avoiding black box models. Two comments were particularly surprising. First, practitioners explained that ML models are not the only black boxes they engage with—so this is familiar territory. Second, practitioners were more optimistic about engaging simple prodding-and-probing methods to understand black boxes than previously reported [25].

4.3.1 Black Boxes, Explanations, & Transparency. Prior work has extensively engaged with concerns of learned black box models. These works point to widespread desires for explanations [9, 25, 28]. We instead found practitioner opinions on ML explanations vary widely. We found *ambivalence* to be a common sentiment toward using black box models, particularly when the perceived potential for harm was low (R2, R3, R5, R1, R13, R6). For example, R2 stated they "don't mind the *unknown* of black box," though they "do mind if it causes harm." R5 stated explanations were not necessary as their product only *supports* a human in diagnosis—so, again, the black box is low risk. Similarly, R1 felt little concern using black box models, with the caveat that "From an ethical perspective, I don't think one should write a model if they have no idea how to tell when something goes wrong."

Many interviewees (R2, R1, R9, R4, R3) pointed to the usefulness of *transparency through example*, where examining sets of inputs and outputs provided sufficient insight into the model behaviors [7]. R3 wanted awareness for how a model reached its goal through increased cognisance of the impacts their objective functions. They detailed how, after adapting their recommendation model to optimize for click-through rate, their model began recommending fuzzy pictures and "weird" content. Despite fulfilling their objective, the content was not useful to their users. To address this, they introduced more stringent prodding-and-probing processes. While Hong et al. [25] suggested this style of proding-and-probing experimentation did not meet practitioners' needs for transparency, we instead discovered that interviewees found this process of experimentation and rigorous documentation and versioning mitigated their concerns about using black box models.

In addition, R4 and R10 desired classic ML explanations. R4 wanted to provide explanations to help users assess predictions, believing this especially necessary for calibrating trust. For internal interpretability use, R4 had implemented LIME [46] and feature importance methods, but found these explanation methods unhelpful for their end users. R10 also desired explanations, stating "nobody

wants a black box model." As an intermediate solution, R10—a consulting company—relied on Shapley Values [49] to increase customer trust, though they noted it was not an ideal solution. They cited two challenges: first, for each data type, they needed to design a custom interface for providing these explanations. Second, they found their users did not know how to interpret these values.

Finally, R8 presented a unique perspective on black boxes, stating that "black box" could equally describe teams, processes, decisions, proprietary models, and obfuscated APIs. R8 found these non-ML black boxes to be equal sources of frustration. They opted to stop using proprietary models as they found these hard to inspect when problems arose. Similarly, undocumented API changes often introduced data errors. Even people could act as "black boxes" when not effectively communicating, and described letting go of an engineering team that wouldn't explain their work. Ultimately, R8 saw black boxes—whether models, processes, or people—as a tool to avoid.

4.3.2 Efforts to Mitigate Overconfidence. Overconfidence is a common ML problem wherein a model's average confidence greatly exceeds its average accuracy [21]. This is related, in part, to the question of deferred responsibility: companies often sought to catch errors by having humans assess low confidence predictions (R13, R5, R4). In our interviews, discussions of overconfidence did not focus on strategies for calibrating the confidence of models, as is trendy in research [21]. Instead, interviewees typically described user-facing mitigation strategies such as onboarding or uncertainty communication. For example, R5 relied on onboarding users for their predictive healthcare models and explained how healthcare practitioners should interpret their confidence metrics. They stressed these confidence metrics are *not* probabilities—100% confidence does not indicate certainty, nor did 50% randomness. These proprietary scores were designed with user feedback and practitioner experience. The team also developed sensitivity thresholds for categories of healthcare practitioners, noting these practitioners tended to have higher or lower tolerances for false positives. During their onboarding process, they explained these sensitivity thresholds and when to anticipate false positives or negatives.

In contrast to R5's approach of relying on onboarding to mitigate overconfidence, R4 sought methods to continuously calibrate user trust and understanding. To do so, R4 collaborated with designers to explore better presentation of uncertainty and curated model explanations. In working with these designers, they expressed a desire to present information in a "non-definitive way." We argue that the former approach of using onboarding and education to mitigate overconfidence is unlikely to be effective: prior research in the explainable AI and human factors communities has shown that managing human responses to automated decisions is a major challenge [15, 54]. Presenting information in a non-definitive way has more promise, but uncertainty communication remains a well studied yet unsolved problem [14, 39].

4.3.3 Discussion: Black Boxes & Overconfidence. Some interviewees described black boxes as part of life—unavoidable, perhaps frustrating, but certainly normal. Despite the ambivalence some interviewees (R2, R3, R5, R15, R9, R13, R6, R1) displayed when discussing automated explanations for black box systems, others (R4, R8, R10, R11) were deeply worried about their implications. While these companies were willing to use these black box models as a supporting measure—particularly for things humans are bad at, like repetitive tasks (R13)—they were not willing to substitute ML in place of human intuition. R7 explained, "I don't have any worries about automating the quality checks." However, they also said this of their higher risk property assessment use case: "the value of something is not objective. If you had a perfect algorithm for what something is worth, but then someone buys it for three times as much, then that's what it's worth." According to R7, no amount of transparency could make these techniques sufficiently safe in this context—they would always be conducted by a human.

In the explainable AI community, there is a latent assertion that poking and prodding a model can never be sufficient for understanding a model; instead, we need a mathematically-meaningful yet to-date-indeterminate mechanism to provide actionable insights into the decision-making mechanics of ML models. Past research with practitioners from Big Tech has confirmed a desire for this form of meaningful explanation method [25]. However, many of the practitioners we interviewed (R1, R2, R9, R4) expressed more optimism about the idea of transparency by example, where examining sets of inputs and outputs allows developers or users to build accurate mental models of the ML model. In contrast with past research, these practitioners believed that adopting this testing procedure could suffice for model understanding.

Still, some interviewees did desire explanations (R4, R10); their desires reflected insights from Hong et al. [25] and Gosiewska and Biecek [20]. These practitioners found explanation methods to be unstable and did not generalize well to their contexts. Other interviewees (R2, R4, R10) desired explanations for non-technical stakeholders and users when uncertainty and overconfidence were concerns. But these explanations are costly to set up—requiring carefully crafted user interfaces and broader tooling—and are still very limited. R4 explained that the methods they had tried—LIME and feature importances—were insufficient at communicating uncertainty, reflecting recent work by Bhatt et al. [6]. This similarly aligns with prior research on uncertainty communication showing shown numerical representations are not effective at this, nor are confidence intervals [11, 27]. We suggest future work in explanations emphasize uncertainty communication and participatory design processes [38], as there is a desire for this work and rich prior literature in Data Visualization and HCI communities [39].

Onboarding for ML products and tools is an underexplored area. Cai et al. [10] detailed how clinicians benefit from onboarding covering the basic properties of a learned model, such as known strengths and weaknesses, its development point-of-view, and design objectives. R5's approach overlaps with these recommendations, but could benefit from increased transparency and focus on positioning the model's capabilities with respect to the downstream domain expert. Like Cai et al. [10] suggest, we believe the onboarding process for human-AI teaming should be continual and evolving, and should re-address use and understanding of tools over time. Onboarding should be designed based on studies of how users interact with the model, and should progress as users' mental models develop through use and exposure. In particular, any time an AI tool is updated—especially considering that updates may break human expectations of model performance without warning [3]—onboarding should be readdressed.

4.4 "Data Literacy Is Not a Silver Bullet." On Communication & Collaboration

As documented in prior work, ML practitioners do not work in isolation [41]. Instead, their projects typically require collaborations with domain experts and other cross-functional collaborations. Every interviewee raised effective communication as critical to their work, and R3 described this simply as "really tough." Despite the impactfulness of effective communication, best practices were *never codified*, and instead comprised of unwritten institutional knowledge and "water cooler talk" (R1).

4.4.1 Effective Communication. Interviewees cited communication as the principle mechanism to reduce and mitigate the risks of using ML (R4, R5, R8, R7, R15, R10, R11, R14, R1). Unsurprisingly, when communication was lacking, practitioners found they wasted time and money on disagreements (R8) and inappropriately targeted work (R11, R14, R1). Several interviewees discussed data literacy as a bottleneck to effective communication (R8, R10, R14). R8 discussed how mismatches between human expectations of business narratives and the analytics led to data insights being ignored: "A couple people in the company have a narrative in their head already, and if the analytics don't support that narrative, they don't want anything to do with those analytics." When stakeholders do not understand the metrics, they also do not *trust* the output—resulting in wasted effort (R8, R14). In response, R8 routinely engages in negotiations with business teams, often requiring external mediation to resolve disputes. Similarly, R14 lamented the "politics around the numbers" and how these so-called politics undermine their work. However, R14 also cautioned that a domain expert who was *too* technical was equally challenging to establish a productive working relationship with: instead of expressing their needs, these technical stakeholders spend time anticipating potential analytics pitfalls.

4.4.2 Institutional Knowledge. Over the life cycle of a project, ML practitioners become more familiar with their data, models, and domain. This evolving awareness affects data collection and model iteration but is typically shared person-to-person, not systematically. For example, R10 tracked large data and model changes through "informal discussions," while R8 ran scripts on an ad-hoc basis to check for data quality errors they noticed over time. Similarly, R17 described learning to intuitively assess the importance of errors based on past user feedback. These insights were rarely documented, but were often embedded in code and shared through conversations. We refer to this as *institutional knowledge*: information that exists across a team but is not recorded. Existing literature charmingly describes institutional knowledge as "the stuff around the edges" [32], considering the context, history, background, and knowledge entwined in artifacts like data and models. Yet, this raises questions of project permanence and responsible development: undocumented information will likely be forgotten. Only R14 described efforts to explicitly document this acquired knowledge, citing metadata as their primary tool for explicating changes to datasets. We speculate that this effort is a consequence of their team's positioning within a public municipality office.

4.4.3 Discussion: Communication and Collaboration. Developing stakeholder alignment is not a new challenge, but ML and data contexts introduce new gaps in vocabularies and experiences. Tools that assist non-technical stakeholders might bridge these gaps through by supporting data literacy and learning [26]. Practitioners currently struggle to bridge these gaps, but doing so is critically important, particularly when practitioners are not domain experts and so need stakeholder insight for new features [56].

Relying on institutional knowledge is also not unique to ML development. Nonetheless, we see this as an area which lacks sufficient investment both in research and in practice, and which could be improved through the introduction of comprehensive standards [4, 23, 34]. An added benefit of more consistent artifacts is the increased potential for internal auditing and quality checks during development [44]. Even beyond ML development, practitioners need careful consideration of the value of ML development to answer *is it necessary?* Future work should produce tools for comprehensive social-systems analyses encouraging stakeholders to examine the possible affects of models on *all* parties.

4.5 "Experiment, Iterate, See We're Getting Closer." A Model Is Never Finished

The life-cycle of a ML model requires continuous refinement. This is part and parcel of both an organization's growth and practitioners' increasing awareness of model limitations. Practitioners' response to challenges in planning, iterating, and evaluating are varied and often reactive. These responses are best characterized by what one interviewee called a "lack of best practices in training" (R11). Practitioners sought to improve their work despite resource constraints, operating under yet another *catch-22*: investing in reflection and refinement while staying within budget.

4.5.1 Data Quality: Planning, Ingesting, & Cleaning. In line with existing literature [22], we found that ML development is significantly hampered by challenges in data collection, management, and use. Nearly every company struggled to standardize data entry, collect enough data, and collect the *right* data to both mitigate bias and encourage robustness. R1 was the only exception, as they worked primarily on ML tooling and not immediate business applications. A major challenge in building ML models was predicting data requirements for good performance: interviewees complained their initial estimates were substantially incorrect, and explained that they actively seek to collect more and additional sources of data (R4, R12, R8, R3, R13, R6). Understanding data coverage needs by assessing real world variability and translating this to data requirements remains an open question in the ML community.

Because of this open question, practitioners often developed a reliance on a subset of *trusted data* (R12, R17, R9, R3, R14, R6). R3 explained that they scoped their recommendations based on the quality and consistency of the available data. They tried adding recommendations for brands with worse data cleanliness, but "it affects the models, because every bit of data you have is fragmented across hard-to-reconcile records." *They were not willing to take on the risks that messier data introduced to their recommendations.* Instead, they relied on data from a small set of familiar brands, explaining that while they considered promoting diverse brands, investing in data cleaning was ultimately too costly given the company's precarious position in entering the market. Lastly, companies found that even in labels provided by domain experts, labeling inconsistencies and

disagreements were common problems. In response, several companies developed complex routines for building consensus (R4, R5, R13, R8). These findings support work by Hohman et al. [22], emphasizing that challenges in data management are universal but are nonetheless often deprioritized when facing resource constraints.

4.5.2 Many Methods of Evaluation. In all of our interviews, we discussed the immense challenges of model evaluation extensively. No two companies had the same process, but every process involved multiple evaluation mechanisms. Above all else, extensive manual evaluation using hand-selected test cases was key to these strategies (R7, R16, R4, R8, R5, R17, R3, R13, R11R1). This first evaluation step was described as "weak but useful" by R3. Beyond manual evaluation, many companies implemented supplemental A/B testing (R3, R1) or beta tests with in-the-wild users (R16, R4, R17, R13, R14, R1, R6). Several interviewees discussed the tradeoff between using extensive in-house evaluation and relying moreso on user feedback (R13, R11, R17, R16, R2, R15, R7): user feedback is significantly cheaper and contains additional signal on whether errors are impactful to users, but mistakes cost user trust and engagement.

Interestingly, R4 compared beta testing models to their clinical trials with veterinarians. They considered this process necessary partly due to the black box nature of ML models, but also because they found it to be useful for assessing the broader impacts of deployment. They emphasized a desire to bring the rigor of clinical trials to their model evaluation—for example, through randomized control trials assessing the introduction of models to veterinary businesses. R10 also proposed adopting the scientific processes to assess ML by formulating and testing hypotheses. Both sentiments follow proposals to adapt the scientific study of behavior to "intelligent" computational models [43]. Lastly, none of the interviewees indicated that they had an effective methodology for evaluating fairness—though several expressed this as a desire (R14, R10, R13), and R10 was exploring Model Cards [34] as a step in this direction.

4.5.3 Model and Data Versioning. Many interviewees desired better model and data versioning (R2, R15, R10, R5, R8, R7, R3, R13). Some companies pointed to recent distribution shifts caused by COVID-19 as highlighting its importance (R7, R8, R4). Still, versioning remained elusive, and R17 explained their company was "too early" to invest in it. Four companies (R1, R4, R13, R8) did extensive model versioning. R1 included complete versioning of all the "data that went in, and the code as well" as a component of their evaluation pipeline. R13 and R8 version "everything," and R13 explained they're only able to afford this process because of a Google Cloud credit award. Were it not for these credits, they would not have the storage capacity needed to version their models and datasets.

Metadata is critical for evaluating data, directing modeling iteration, documenting changes, and retroactively incorporating data and model versioning (R14, R1, R12, R2, R3). R14 explained they relied on metadata to identify "what worked within a dataset" as metadata can reflect "things that are taken care of as the project is going on." Similarly, R1 relied on metadata to inform their work, even training overnight based on "metadata considerations." When creating groundtruth, R12 explained their central question is, "what metadata do I need to make sure this product matches the end description?" Metadata serves as documentation for institutional knowledge, yet remains an underutilized resource.

4.5.4 Discussion: A Model is Never Finished. Each interviewee adopted different processes for evaluating their models. This is not surprising: mechanisms for effectively testing ML models remain rudimentary, and recommendations inconsistent. We assert the research community should produce consistent testing recommendations, with an increased focus on test cases [7, 47] and model fairness assessments [34]. Prior work described comparing multiple models as *crucial* during evaluation [25], yet we found interviewees rarely implemented this process—while it is common to train models on different subsets of data, hyperparameters, or with different seeds *in parallel*, resource constraints can make training multiple models for comparison impossible (R17, R13, R8, R16). R4's proposal of using randomized controlled trials and other processes adopted from clinical trials to assess models is compelling. We believe the research community should further recommend best practices for adapting these evaluation mechanisms for ML contexts [59].

ML experience levels affected development practices. In cases where teams lacked ML experience, poor modeling decisions were followed by periods of indecision (R12). In some cases, these teams might end ML development entirely (R12). More mature teams (R6, R10, R14, R1) emphasized the fleeting lifespan of models and encouraged team members to prioritize frequent retraining, minimally complex models, and a willingness to "throw it out" (R6). In contrast, less experienced organizations might not *afford* retraining and replacing models, or lacked the experience to build modularity into models. As such, R3 characterized ML development as notably slow and defensive compared to other engineering tasks. We were reminded of parallels in software development best practices: guidelines made code cleaner and easier to debug or replace [40]. We believe future work adapting these practices for ML development and maintenance would be beneficial to practitioners.

4.6 "GDPR Doesn't Affect Us." Assessing Tensions Between Privacy & Growth

Machine learning necessitates the collection and management of large data collections. As a consequence, recent legislation such as GDPR and the CCPA have broad implications for the discipline. We asked interviewees about their relationships to these and other privacy legislation works to assess how practitioners are responding.

4.6.1 Government Regulation & Privacy Policy Impacts. The academic community continues to debate if GDPR encompasses the *right to an explanation* for an automated decision, but collectively agrees that GDPR encodes at least the *right to be informed* [19]. While the former interpretation has stronger implications for ML practitioners, both should have some affect. GDPR is complemented by emergent legislation seeking to protect user privacy—with many implications for data collection and handling [52]. We asked ML practitioners how this legislation affected their practices. Though R10 noted their company's legal team guided them on GDPR and other regulations, *no* interviewee indicated that they were directly concerned with the requirement to provide an explanation, despite using black box models extensively. By and large, interviewees expressed that GDPR and other legislation had not impacted their work in any substantive capacity (R2, R9, R16, R10, R4, R8, R3, R11).

A few interviewees explained that, to comply with GDPR, they leveraged their nature as platforms to avoid collection of personally

identifiable data (R4, R10, R8). Others lamented minor inconveniences relating to GDPR, such as increased latency from using remote servers based in Europe (R10) or vague concerns over future implications for cloud use (R11). One interviewee (R1) described how their company responded to GDPR by extensively changing their data handling practices, devoting over six months of engineering time to ensuring compliance. We should note that this company was both well-resourced and publicly listed. While they bemoaned that deletion requests do come in, and are "a pain in the ass," the interviewee explained that the process of adopting GDPR compliance in data handling had actually been immensely beneficial to the company. GDPR forced their company to develop better awareness of and practices for handling data, and this re-evaluation increased their overall data competencies. Despite implementing these extensive changes to data handling, R1 nonetheless remained unconcerned with any notion of a "right to an explanation."

4.6.2 Privacy Legislation is Insufficient. A common sentiment was that privacy legislation continues to be insufficient to protect users. Interviewees often felt it necessary to implement their own policies and tooling beyond any requirements (R17, R16, R13, R3, R6). In the absence of stronger privacy legislation, companies aspired to act "without malcontent whenever possible" (R3). Companies continue to internally assess their responsibilities to users' privacy, but find themselves attempting to balance these responsibilities with other desires. Many companies discussed managing the tension between user privacy and their desires to become ubiquitous, and to collect ever more extensive datasets (R16, R17, R13, R2, R3, R6).

4.6.3 ML to Satisfy Regulators. While we asked interviewees about how legislation changed their ML development and data practices, one interviewee explained that they instead used ML to respond to regulatory requirements (R7). In their real estate assessment business, regulations require reporting on properties' conditions and features. While their ML models had insufficient accuracy to meet consumer expectations, they found this accuracy rate to be acceptable for ensuring broad-sweeping regulatory compliance.

4.6.4 Discussion: Tensions Between Privacy and Growth. The relationship between privacy legislation and ML development is curious. From the researcher's perspective, the world is abuzz with chatter about the implications of GDPR for explanations of automated decisions. Yet, *every* practitioner we interviewed was unaware of these discussions, let alone the need to revise ML development practices in response. GDPR and other privacy legislation had started to affect data practices, but ambiguity abounds: while one company implemented extensive changes to their data management systems (R1), none of our interviewees considered the implications of these deletion requests for ML models. Should this deleted data also be deleted from any model training, test, and validation sets? Should the *model itself* be deleted in response to the request [18, 57]? These questions go unanswered in the research community, and unnoticed in these practitioners' realms. Organizations (R16, R17, R13, R2, R3, R6) continue to self-moderate ideas of "acting without malcontent" (R2)—analogous to Google's antiquated motto of "Don't be evil." These organizations experience tensions between their desires to sustain and grow their businesses, and to protect user interests.

5 CONCLUSION

When discussing ML practice with smaller and less visible organizations, we find these practitioners have many commonalities with their Big Tech counterparts: desires for explanation, lack of standardization, and unending difficulties communicating with stakeholders. We also uncover several new, divergent findings: for example, in contrast with past studies, the practitioners we interviewed expressed optimism for transparency through example, noted access differentials for bias mitigation, and experienced subdued implications from privacy legislation on ML development.

Most critically, resource constraints affect the development of responsible ML and amplify existing concerns about the challenges of responsible and fair ML development for these interviewed organizations. These constraints continuously affect the work that companies and organizations invest in; for example, while several companies wanted to invest in explanations for ML models, they found the costs of developing these techniques to be too high, especially given the "researchy" nature of these tools. Intuitively, the resource constraints of startups and small companies encourage increased caution in decision-making, but this requisite careful planning is untenable without sufficient experience and domain knowledge—both of which are difficult to acquire. Instead, organizations found predicting ethical and financial costs to be difficult, causing them to reconsider incorporating these methods. New (ML) product explorations were accompanied by exploding budgetary requirements. R6 suggested that, as with other forms of engineering work, predicting costs for ML requires experience and exceptionally large buffers of both time and money due to increased risk or complexity. But many of these companies lack the necessary expertise to assess cost and found the "cost of going external [to be] too high" (R8). These buffers of time and resources represent untenable costs for organizations actively seeking investment.

Finally, assessing the social implications and necessity of modeling is important but risky—in Big Tech teams, misunderstandings and oversights do not collapse the business. Big Tech has large teams with expertise in data and ML, as well as extensive investment in in-house tooling. In contrast, many interviewees lamented the challenge of hiring ML talent (R4, R8, R12, R15, R2, R17, R16), and the lack of accessible and comprehensive tools to assist in much-needed bigger picture analyses. One interviewee suggested less-resourced organizations' products are especially prone to "exhibiting this bias" *because* of limited resources and expertise (R11); all the more reason to center these practitioners as we consider the challenge of responsible ML development.

The potential benefits of ML technology must be spread beyond the agenda of Big Tech and into all corners of society, yet the vanguard of small organizations implementing ML struggle to realize the hype. We identify challenges across company and stakeholder expectations, bias, explainability and overconfidence, data literacy, model lifecycles, and privacy that lead to a sobering picture: At this point in time, opinion was mixed across our organizations on whether implementing ML was even a worthwhile exercise. Through our discussion, we highlight how and why implementing this promising technology can be especially fraught for resource-constrained organizations, and so draw attention to areas requiring further study from the broader machine learning community.

ACKNOWLEDGMENTS

Thanks to Betsy Peters for providing the impetus for this work, to our respective partners for proofreading and running commentary throughout the process, and to members of the MIT Visualization Group and the MIT Interactive Robotics Group for their feedback. SB is funded by an NSF GRFP. AH is funded by a Siebel Fellowship.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Cecilia Aragon, Clayton Hutto, Andy Echenique, Brittany Fiore-Gartland, Yun Huang, Jinyoung Kim, Gina Neff, Wanli Xing, and Joseph Bayer. 2016. Developing a Research Agenda for Human-Centered Data Science. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion* (San Francisco, California, USA) *(CSCW '16 Companion)*. Association for Computing Machinery, New York, NY, USA, 529–535.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[4] 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.

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[6] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2020. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *arXiv preprint arXiv:2011.07586* (2020).

[7] Serena Booth, Yilun Zhou, Ankit Shah, and Julie Shah. 2021. Bayes-TrEx: a Bayesian Sampling Approach to Model Transparency by Example. *Proceedings of the 35th AAAI Conference on Artificial Intelligence* (2021).

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[9] Andrea Brennen. 2020. What Do People Really Want When They Say They Want" Explainable AI?" We Asked 60 Stakeholders.. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[10] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[11] M. Correll and M. Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

[12] Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature* 538, 7625 (2016), 311–313.

[13] 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. *arXiv preprint arXiv:2011.03395* (2020).

[14] Stephanie Deitrick and Robert Edsall. 2006. The influence of uncertainty visualization on decision making: An empirical evaluation. In *Progress in spatial data handling*. Springer, 719–738.

[15] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[16] Robert Wall Emerson. 2015. Convenience sampling, random sampling, and snowball sampling: How does sampling affect the validity of research? *Journal of Visual Impairment & Blindness* 109, 2 (2015), 164–168.

[17] Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics* 5, 1 (2016), 1–4.

[18] Antonio Ginart, Melody Y Guan, Gregory Valiant, and James Zou. 2019. Making AI forget you: Data deletion in machine learning. *arXiv preprint arXiv:1907.05012* (2019).

[19] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[20] Alicja Gosiewska and Przemyslaw Biecek. 2019. iBreakDown: Uncertainty of Model Explanations for Nonadditive Predictive Models. *arXiv preprint arXiv:1903.11420* (2019).

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.

[22] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[23] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[24] 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.

[25] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[26] Aspen K Hopkins, Michael Correll, and Arvind Satyanarayan. 2020. VisuaLint: Sketchy In Situ Annotations of Chart Construction Errors. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 219–228.

[27] C. R. Johnson and A. R. Sanderson. 2003. A Next Step: Visualizing Errors and Uncertainty. *IEEE Computer Graphics and Applications* 23, 5 (2003), 6–10. https://doi.org/10.1109/MCG.2003.1231171

[28] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[29] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*. IEEE, 96–107.

[30] M. Kim, T. Zimmermann, R. DeLine, and A. Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions on Software Engineering* 44, 11 (2018), 1024–1038.

[31] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[32] Nina McCurdy, Julie Gerdes, and Miriah Meyer. 2018. A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 925–935.

[33] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[34] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[35] Michael Muller, Melanie Feinberg, Timothy George, Steven J. Jackson, Bonnie E. John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8.

[36] 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.

[37] Michael Muller, Christine Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. *Submitted for publication to CHI* (2021).

[38] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.

[39] Lace Padilla, Matthew Kay, and Jessica Hullman. 2020. Uncertainty visualization. (2020).

[40] David Lorge Parnas. 1979. Designing software for ease of extension and contraction. *IEEE transactions on software engineering* 2 (1979), 128–138.

[41] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 136 (Nov. 2018), 28 pages.

[42] Paula Pereira, Jácome Cunha, and João Paulo Fernandes. 2020. On understanding data scientists. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–5.

[43] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486.

[44] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[45] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *arXiv preprint arXiv:2006.12358* (2020).

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[47] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[48] Rwamahe Rutakumwa, Joseph Okello Mugisha, Sarah Bernays, Elizabeth Kabunga, Grace Tumwekwase, Martin Mbonye, and Janet Seeley. 2020. Conducting in-depth interviews with and without voice recorders: a comparative analysis. *Qualitative Research* 20, 5 (2020), 565–581.

[49] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[50] Tom Simonite. 2017. AI and 'Enormous Data' Could Make Tech Giants Harder to Topple. *Wired* (2017).

[51] Svenja C Sommer, Christoph H Loch, and Jing Dong. 2009. Managing complexity and unforeseeable uncertainty in startup companies: An empirical study. *Organization Science* 20, 1 (2009), 118–133.

[52] William Stallings. 2020. Handling of Personal Information and Deidentified, Aggregated, and Pseudonymized Information Under the California Consumer Privacy Act. *IEEE Security & Privacy* 18, 1 (2020), 61–64.

[53] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[54] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.

[55] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. 2017. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*. PMLR, 509–517.

[56] S. Viaene. 2013. Data Scientists Aren't Domain Experts. *IT Professional* 15, 6 (2013), 12–17.

[57] Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. 2018. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review* 34, 2 (2018), 304–313.

[58] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 211 (Nov. 2019), 24 pages.

[59] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine* 25, 9 (2019), 1337–1340.

[60] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact. CSCW* (Oct. 2020).