# "*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
[nithyasamba,kapania,hhighfill,dakrong,pkp,loraa]@google.com
Google Research
Mountain View, CA

## ABSTRACT

AI models are increasingly applied in high-stakes domains like health and conservation. Data quality carries an elevated significance in high-stakes AI due to its heightened downstream impact, impacting predictions like cancer detection, wildlife poaching, and loan allocations. Paradoxically, data is the most under-valued and de-glamorised aspect of AI. In this paper, we report on data practices in high-stakes AI, from interviews with 53 AI practitioners in India, East and West African countries, and USA. We define, identify, and present empirical evidence on *Data Cascades*—compounding events causing negative, downstream effects from data issues—triggered by conventional AI/ML practices that undervalue data quality. Data cascades are pervasive (92% prevalence), invisible, delayed, but often avoidable. We discuss HCI opportunities in designing and incentivizing data excellence as a first-class citizen of AI, resulting in safer and more robust systems for all.

## CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

## KEYWORDS

Data, AI, ML, high-stakes AI, data cascades, developers, raters, application-domain experts, data collectors, data quality, data politics, India, Nigeria, Kenya, Ghana, Uganda, USA

**ACM Reference Format:**
Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo. 2021. "*Everyone wants to do the model work, not the data work*": Data Cascades in High-Stakes AI . In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445518

## 1 INTRODUCTION

Data is the critical infrastructure necessary to build Artificial Intelligence (AI) systems [44]. Data largely determines performance, fairness, robustness, safety, and scalability of AI systems [44, 81]. Paradoxically, for AI researchers and developers, data is often the least incentivized aspect, viewed as 'operational' relative to the lionized work of building novel models and algorithms [46, 125]. Intuitively, AI developers understand that data quality matters, often spending inordinate amounts of time on data tasks [60]. In practice, most organisations fail to create or meet any data quality standards [87], from under-valuing data work vis-a-vis model development.

Under-valuing of data work is common to all of AI development [125][1]. We pay particular attention to undervaluing of data in *high-stakes domains*[2] that have safety impacts on living beings, due to a few reasons. One, developers are increasingly deploying AI models in complex, humanitarian domains, *e.g.,* in maternal health, road safety, and climate change. Two, poor data quality in high-stakes domains can have outsized effects on vulnerable communities and contexts. As Hiatt *et al.* argue, high-stakes efforts are distinct from serving customers; these projects work with and for populations at risk of a litany of horrors [47]. As an example, poor data practices reduced accuracy in IBM's cancer treatment AI [115] and led to Google Flu Trends missing the flu peak by 140% [63, 73]). Three, high-stakes AI systems are typically deployed in low-resource contexts with a pronounced lack of readily available, high-quality datasets. Applications span into communities that live outside of a modern data infrastructure, or where everyday functions are not yet consistently tracked, *e.g.,* walking distances to gather water in rural areas—in contrast to, say, click data [26]. Finally, high-stakes AI is more often created at the combination of two or more disciplines; for example, AI and diabetic retinopathy, leading to greater collaboration challenges among stakeholders across organizations and domains [75, 121].

Considering the above factors, currently data quality issues in AI are addressed with the wrong tools created for, and fitted to other technology problems—they are approached as a database problem, legal compliance issue, or licensing deal. HCI and CSCW scholarship have long examined the practices of collaboration, problem formulation, and sensemaking, by humans behind the datasets, including data collectors and scientists, [69, 86, 127], and are designing computational artefacts for dataset development [53]. Our research extends this scholarship by empirically examining data practices and challenges of high-stakes AI practitioners impacting vulnerable groups.

We report our results from a qualitative study on practices and structural factors among 53 AI practitioners in India, the US, and


*CHI '21, May 8–13, 2021, Yokohama, Japan*

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445518

[1]Data work is broadly under-valued in many sociotechnical domains like [58, 85]
[2]We extend the vision of *AI for Social Good* (*i.e.,* using AI for social and environmental impact) and *Data for Good* (*i.e.,* providing data and education to benefit non-profit or government agencies) with *AI for high-stakes domains* involving safety, well-being and stakes (*e.g.,* road safety, credit assessment).

East and West African countries[3], applying AI to high-stakes domains including landslide detection, suicide prevention, and cancer detection. Our research aimed to understand how practitioners conceptualised and navigated the end-to-end AI data life cycles.

In this paper, we define and identify *Data Cascades*: *compounding events causing negative, downstream effects from data issues, resulting in technical debt*[4] *over time*. In our study, data cascades were widely prevalent: **92%** of AI practitioners reported experiencing one or more, and **45.3%** reported two or more cascades in a given project. Data cascades often resulted from applying conventional AI practices that undervalued data quality. For example, eye disease detection models, trained on noise-free training data for high model performance, failed to predict the disease in production upon small specks of dust on images. Data cascades were opaque and delayed, with poor indicators and metrics. Cascades compounded into major negative impacts in the downstream of models like costly iterations, discarding projects, and harm to communities. Cascades were largely avoidable through intentional practices.

The high prevalence of fairly severe data cascades point to a larger problem of broken data practices, methodologies, and incentives in the field of AI. Although the AI/ML practitioners in our study were attuned to the importance of data quality and displayed deep moral commitment to vulnerable groups, data cascades were disturbingly prevalent even in the high stakes domains we studied. Additionally, our results point to serious gaps in what AI practitioners were trained and equipped to handle, in the form of tensions in working with field partners and application-domain experts, and in understanding human impacts of models—a serious problem as AI developers seek to deploy in domains where governments, civil society, and policy makers have historically struggled to respond. The prevalence of data cascades point to the contours of a larger problem: residual conventions and perceptions in AI/ML drawn from worlds of 'big data'—of abundant, expendable digital resources and worlds in which one user has one account [108]; of model valourisation [125]; of moving fast to proof-of-concept [8]; and of viewing data as grunt work in ML workflows [111]. Taken together, our research underscores the need for *data excellence* in building AI systems, a shift to proactively considering care, sanctity, and diligence in data as valuable contributions in the AI ecosystem. Any solution needs to take into account social, technical, and structural aspects of the AI ecosystem, which we discuss in our paper.

Our paper makes three main contributions:

(1) *Conceptualising and documenting data cascades, their characteristics, and impact* on the end-to-end AI lifecycle, drawn from an empirical study of data practices of international AI practitioners in high-stakes domains.
(2) *Empirically derived awareness for the need of urgent structural change* in AI research and development to incentivise care in data excellence, through our case study of high-stakes AI.
(3) *Implications for HCI*: we highlight an under-explored but significant new research path for the field in creating interfaces, processes, and policy for data excellence in AI.

[3]We sampled more widely in Sub-Saharan Africa due to the nascent AI Ecosystem and redact identifiable details like country, to protect participant identity (see Methodology for more details).
[4]In 1992, Ward Cunningham put forward the metaphor of *technical debt* to describe the build-up of cruft (deficiencies in internal quality) in software systems as debt accrual, similar to financial debt [29] (also observed in ML [111].)

# 2 RELATED WORK

## 2.1 Data in HCI

Prior research in HCI has drawn particular attention to work practices and challenges faced by practitioners in working with data [48, 65, 86, 93, 96]. Feinberg describes data as a design material and our role as designers of data, not its appropriators [35]. Researchers have also studied the ways in which data is rarely used as given, and often needs to be created or handcrafted using intricate transformation practices [67, 96].

An emerging stream of research in HCI and CSCW focuses on the work and collaboration practices of data scientists [66, 77, 94, 127]. Muller *et al.* extend and outline five approaches of data scientists to perform analyses: discovery, capture, design, curation, and creation of data [86]. Koesten *et al.* identify a need to understand the ways in which collaboration occurs for data on a spectrum—from creating and sharing inside and outside the organisation or reusing another person's data with limited interaction with the creator [69]. Practitioners have been shown to collaborate much less around datasets, relative to collaboration around code [127]. Data documentation, which is a crucial aspect of facilitating collaboration, is well studied in the database and data management community [19, 23]. However, documentation of data suffers from a lack of standards and conventions within the ML community [40].

Prior work in HCI and CSCW does not appear to explicitly focus on data practices in high-stakes domains, which are proliferating, and are marked by complex challenges of data scarcity, downstream impacts, and specialised inter-disciplinary knowledge for working with and understanding data (*e.g.,* what a fractured bone looks like in an X-Ray might be beyond an AI practitioner's area of expertise). Several studies have focused on data practices of data scientists; our research extends the focus on data to ML practitioners, including engineers, researchers, and academics who build and deploy AI/ML technologies. Prior research has focused primarily on Western populations, that often have fewer resource constraints, and greater acceptance and understanding of AI in their communities. Our research presents an international analysis of data-related practices and issues in India, East and West African countries, and the US.

## 2.2 Politics of data

There is substantial work in HCI and STS to establish that data is never 'raw' [41], but rather is shaped through the practices of collecting, curating and sensemaking, and thus is inherently sociopolitical in nature. Through their study of public health data, Pine and Liboiron [99] demonstrate how data collection is shaped by values and decisions about "what is counted and what is excluded, and what is considered the best unit of measurement." Vertisi and Dourish [123] examine data in an interactional context and argue for considering the contexts of production in data economies, alongside use and exchange to clarify the ways in which data acquires meaning. Taylor *et al.* [118] drew attention to this need in their research on considering the physical and social geography in which data, people, and things are situated, and to represent the rich geo-tapestry within which data is entangled.

Critical data studies researchers have demonstrated longstanding interest in the 'discretionary' [95] practices shaping data-driven systems and how they are designed and used [6, 16, 33], and the

ways in which data science teams are constituted [106]. Passi and Jackson [93] describe how data work is often invisibilized through a focus on rules, arguing that empirical challenges render invisible the efforts to make algorithms work with data. This makes it difficult to account for the situated and creative decisions made by data analysts, and leaving behind a stripped down notion of 'data analytics'. Passi and Sengers [95] turn their attention to the negotiations in designing data science systems, on how a system should work and is evaluated.

Beyond data scientists, there are many roles in the process of preparing, curating, and nurturing data, which are often under-paid and over-utililized. Many researchers have pointed to the undervalued human labour that powers AI models (e.g., heteromation [34], fauxtomation [117], and "menial" vs. "innovative" work distinctions [56]. Møller *et al.* [85] describe the crucial data work through a framework of meaningful registration, digital organizing, and concern for ethics. They discuss how the data work of clerical hospital workers is complex, skillful, and effortful [85]. However, data work has been shown to be invisibilized among Mechanical Turkers by Martin *et al.* [79], and among frontline health workers in India by Ismail and Kumar [58]. Through a post-colonial feminist perspective, Ismail and Kumar [58] highlight how frontline health workers in India navigate the multiple demands placed on them, and how their data work is severely under-compensated. Our research extends discourses on how data workers play a critical role in creating and maintaining AI systems, and the ways in which their work can have downstream impacts.

## 2.3 Data quality interventions

Real-world datasets are often 'dirty' and come with a variety of data quality problems [1]. However, data quality is crucial to ensure that the ML system using the data can accurately represent and predict the phenomenon it is claiming to measure. A well-established, and steadily growing, body of work focuses on understanding and improving data quality to avoid the *garbage in, garbage out* problem [45, 103].

Kandel *et al.* reveal that practitioners consider data wrangling tedious and time-consuming [62]. Thus, improving quality through transformations [52] and building human-in-the-loop data cleaning systems[61] are well-studied research areas in the data management community. Practitioners often work with a set of assumptions about their data during analysis and visualisation, which guides their data transformations [62]. Interactive data cleaning focuses on making this process easier, because data transformations can be difficult to specify and reuse across multiple tasks [61, 72, 102]. For instance, Wrangler suggests potentially relevant transforms, and maintains a history of transformation scripts to support review and refinement [61]. Data cleaning and wrangling systems address data quality issues by using integrity constraints [27], type inference [36], schema matching [43], outlier detection [51] and more.

Researchers have created several tools to support the creation of ML 'pipelines' and make these workflows manageable [21, 54, 70, 72, 76]. Similar to Code Linters common in traditional SE, Data Linter is a tool to inspect ML datasets, identify potential data issues and suggest transformations to fix these issues [54]. Breck *et al.* created a data validation system to detect anomalies in Machine learning pipelines [21]. Other frameworks to discover data bugs and clean data include ActiveClean and BoostClean [70, 72]. Such interventions highlight the importance of catching data errors using mechanisms specific to data validation, instead of using model performance as a proxy for data quality [120]. In addition to this, it is crucial to test and monitor data as much as we focus on the testing of code. Breck *et al.* provided a set of 28 actionable tests for features, data and models [21]. There is extensive literature on ML testing for detecting differences between the actual and expected behaviour of ML pipelines; for a survey, see [129]. Researchers in the field of HCI and HCOMP have demonstrated a longstanding interest in making use of crowdsourcing to generate ML data [25, 128], to support creation of better task designs for raters [59], compute inter-rater reliability, design incentives [50], and improve the quality of crowdsourced data [30], though these areas are less well known in the ML community [122].

Prior research on developing data quality systems has largely focused on data cleaning and wrangling. However, high-stakes domains extend both, into upstream (data creation) and downstream (live data after deployment)—our research extends this growing body of work by focusing on the end-to-end lifecycle of data in high-stakes domains. For example, viewing data as a dynamic entity points us to drifts and hidden skews[5]. Prior work on data systems appears to be built for intra-organisational AI development. Our research extends current discourses to high-stakes AI which typically involve cross-organisational and inter-disciplinary work; for example, dataset definition and labelling accuracy all depend on application-domain expertise that comes from collaboration with field partners and domain experts.

## 2.4 Machine Learning in production

Production is the process of deploying systems 'live', with a need to keep systems running smoothly and scaling efficiently[6]. Prior work has substantially advanced and documented issues in productionizing software, including ML code. The extra effort to add new features is the interest paid on the technical debt [29], which is particularly challenging for production systems. Sculley *et al.* [111] extend the notion of technical debt to ML systems by identifying and outlining the various ways in which teams could accumulate debt through aspects of ML-specific design elements. Fowler argues that unacknowledged debts are bad, further characterized as reckless or inadvertent [39]. In particular, due to the complexities of data-driven ML systems, they point out that is important to be aware of, and engage with debt trade-offs, which can cause harm in the long term.

Multiple recent studies examine the challenges of production machine learning [6, 100, 101]. For example, ML practitioners spend a significant portion of their time analysing their raw datasets [100]. Regardless, ML teams continue to struggle the most with aspects of data acquisition and management [6]. Since ML largely depends on its data, having high-quality data has a critical role in developing reliable and robust ML models, as opposed to only a good training algorithm [101]. Nevertheless, practitioners often face issues with

[5]Drifts are supported by end-to-end cloud platforms like AWS and Azure, but cloud platforms are not uniformly adopted, including in our study [9, 60]
[6]https://engineering.fb.com/category/production-engineering/

understanding the data without context, validating data, and dealing with distribution skews between training and serving data [100].

Machine Learning workflows are fundamentally iterative and exploratory in nature [7, 52, 71, 96]. These iterations are characterised as loops which occur within an ML system (direct) or due to influence from another system (hidden) [111]. To achieve the desired performance, practitioners have to iterate both on data and ML model architectures. Hohman *et al.* identified common types of data iterations and created a tool to visualise them [52].

Our work extends this body of research by presenting complex downstream impacts from data cascades, which were widely prevalent and fairly severe in our study. Data cascades largely manifest in deployments of AI systems, affecting communities downstream. We also describe the ways in which some of these iterations and feedback loops can be inefficient, extremely costly for teams working with multiple resource constraints and cause long-term harm.

## 3 METHODOLOGY

Between May and July 2020, we conducted semi-structured interviews with a total of 53 AI practitioners[7] working in high-stakes applications of AI development. Interviews were focused on (1) data sources and AI lifecycles; (2) defining data quality; (3) feedback loops from data quality; (4) upstream and downstream data effects; (5) stakeholders and accountability; (6) incentive structures; and (7) useful interventions. Each session focused on the participant's experiences, practices, and challenges in AI development and lasted about 75 minutes each.

**Participant recruitment and moderation.** In our sample, AI practitioners were located in, or worked primarily on projects based in, India (23), the US (16), or East and West African countries (14). We sampled more widely in Africa due to the nascent AI Ecosystem compared to other continents [84], with 14 total interviews including Nigeria (10), Kenya (2), Uganda (1), and Ghana (1). We interviewed 45 male and 8 female AI practitioners. Refer to Table 1 for details on participant demographics. Interviews were conducted using video conferencing, due to COVID-19 travel limitations.

On average, an AI practitioner in our study had one or more higher education degrees in AI related fields and had worked for greater than 4-5 years in AI. While we interviewed AI practitioners working in multiple institution types, varying from startups (28), large companies (16), to academia (9), all participants were involved in AI development in critical domains with safety implications. Participants in our study were technical leads, founders, or AI developers.

Many participants had experience with multiple AI technologies, and had applied AI technologies to multiple domains; we report the primary AI technology and domain of application at the time of the interview. Applied uses of AI technology in academia meant there were partnerships with government, private business, and startups. For a characterisation of the type of AI [113], refer to table 1.

We recruited participants through a combination of developer communities, distribution lists, professional networks, and personal contacts, using snowball and purposive sampling [89] that was iterative until saturation. We conducted all interviews in English (preferred language of participants). Each participant received a thank you gift in the form of a gift card, with amounts localised in consultation with regional experts (100 USD for the US, 27 USD for India, 35 USD for East and West African countries). Due to workplace restrictions, we were not able to compensate government employees. Interview notes were recorded in the form of field notes or video recordings, transcribed within 24 hours of each interview by the corresponding moderator. Our research team is constituted by members with HCI, AI, human computation, and data quality research backgrounds. Interviews were conducted by authors located in India, West Africa, and the United States. All researchers were involved in the research framing, data analysis, and synthesis.

**Analysis and coding.** Following [119], two members of the research team independently read all units multiple times, and categories (unit of analysis) were initially identified by each researcher, together with a description and examples of each category, until a saturation point was reached. Our upper level categories were guided by the evaluation aims, comprising (1) defining the right data for a project; (2) practices to define data quality; (3) entry points of data problems; (4) impacts and measurement of data quality; (5) model production challenges; (6) incentives; (7) other human factors; and (8) resourcing and infrastructure. The categories were iteratively refined through group discussions with meeting, diverging, and synthesizing during the analysis phase. Further iterations resulted in the formation of lower-level categories such as "domain expertise: misaligned goals". These categories were consolidated into three top-level categories of characteristics of data cascades, motivating factors, and cascade types, and 18 nested categories such as incentives, signals, domain experts, and impacts. Since codes are our process, not product [80], IRR was not used.

While we present general data practices and basic AI practitioner development models, all interventions, practices, and working methods were reported by participants as part of their own experiences, rather than as "best practices" (see [97]). Numbers reported throughout the paper represent the percentage of participants who self-reported a trigger, impact, or signal of data challenges in the interviews. Percentages are derived from coding each transcript for each individual's experiences of cascades.

**Research ethics and anonymization.** During recruitment, participants were informed of the purpose of the study, the question categories, and researcher affiliations. Participants signed informed consent documents acknowledging their awareness of the study purpose and researcher affiliation prior to the interview. At the beginning of each interview, the moderator additionally obtained verbal informed consent. We stored all data in a private Google Drive folder, with access limited to the research team. To protect participant identities, we deleted all personally identifiable information in research files. We redact identifiable details when quoting participants, *e.g.,* we use East Africa or West Africa, given the limited number of AI practitioners in high-stakes domains in Sub-Saharan Africa, and our limited sampling.

**Limitations.** All interviews and analysis were conducted over video and phone, due to the COVID-19 pandemic. As a result of travel restrictions, we were unable to include shadowing of work flows and contextual inquiry that would have otherwise been possible. However, we feel that the self-reported data practices and challenges have validity, and sufficient rigour and care was applied

[7] Although our participants had different job roles (including, in research), all were focused on applied deployments in high-stakes domains.

| Type | Count |
|---|---|
| **Roles** | AI Engineer **(17)**, Startup Founder **(17)**, Professor **(6)**, Data Scientist **(6)**, Research Scientist **(6)**, Program Manager **(1)** |
| **Location** | India **(23)**, US **(16)**, Nigeria **(10)**, Kenya **(2)**, Ghana **(1)**, Uganda **(1)** |
| **Gender** | Male **(45)**, Female **(8)** |
| **Setting** | Startup **(28)**, Large company **(16)**, Academic **(9)** |
| **Domain** | Health and wellness **(19)** (*e.g.,* maternal health, cancer diagnosis, mental health)<br>Food availability and agriculture health **(10)** (*e.g.,* regenerative farming, crop illness)<br>Environment and climate **(7)** (*e.g.,* solar energy, air pollution)<br>Credit and finance **(7)** (*e.g.,* loans, insurance claims)<br>Public safety **(4)** (*e.g.,* traffic violations, landslide detection, self driving cars)<br>Wildlife conservation **(2)** (*e.g.,* poaching and ecosystem health)<br>Aquaculture **(2)** (*e.g.,* marine life)<br>Education **(1)** (*e.g.,* loans, insurance claims)<br>Robotics **(1)** (*e.g.,* physical arm sorting)<br>Fairness in ML **(1)** (*e.g.,* representativeness) |
| **AI Type** | Machine Learning: **(24)**, Computer Vision: **(21)**, Natural Language Processing: **(5)**, Game Theory: **(2)**, Robotics: **(1)** |

**Table 1: Summary of participant demographics**

in covering the themes through multiple questions and solicitation of examples. Gender distribution in our study is reflective of the AI industry's gender disparities [126] and sampling limitations.

# 4 FINDINGS

In this section we present data cascades, their indicators and impacts (section 4.1), and position them in a broader landscape of high-stakes domains and the AI ecosystem (section 4.2). Our study identifies four root causes for data cascades and corresponding practitioner behaviours (section 4.3).

## 4.1 Overview of data cascades

We define *Data Cascades* based on the empirical results in this study as *compounding events causing negative, downstream effects from data issues, that result in technical debt over time.* In our study, **92%** experienced at least one cascade. Data cascades are influenced by, (a) the activities and interactions of actors involved in the AI development (*e.g.,* developers, governments, and field partners), (b) the physical world and community in which the AI system is situated (*e.g.,* rural hospitals where sensor data collection occurs).

We observed the following properties of data cascades:

- **Opaque:** data cascades are complex, long-term, occur frequently and persistently; they are opaque in diagnosis and manifestation—with no clear indicators, tools, and metrics to detect and measure their effects on the system. In the absence of well-defined and timely signals, practitioners turned to proxy metrics (*e.g.,* accuracy, precision, or F1 score), where the unit of measurement is the entire system, not datasets.
- **Triggered by:** data cascades are triggered when conventional AI practices are applied in high-stakes domains, which are characterised by high accountability, inter-disciplinary work, and resource constraints. For example, practitioners viewed data as operations, moved fast, hacked model performance (through hyperparameters rather than data quality), and did not appear to be equipped to recognise upstream and downstream people issues.
- **Negative impact:** data cascades have negative impacts on the AI development and deployment process, leading to multiple and unexpected strategies sometimes spurring further cascades, always causing technical debt. Some of the severe data cascades in our study led to harm to beneficiary communities, burnout of relationships with stakeholders, discarding entire datasets, and performing costly iterations.
- **Multiple cascades**, **45.3%** experienced two or more cascades each, typically triggered in the upstream of model building, manifesting in the downstream of the model development or deployment.
- **Cascades are often avoidable** by step-wise and early interventions in the development process, which were, however, exceptional due to factors like undervaluing data, scarcity of data, and partner dependencies.

## 4.2 Broader landscape for data cascades

Before we turn to specific cascades in the next section, here we provide an understanding of cross-cutting factors that influence data cascades in high-stakes domains.

**Incentives and currency in AI** An overall lack of recognition for the invisible, arduous, and taken-for-granted data work in AI led to poor data practices, resulting in the data cascades below. Care of, and improvements to data are not easily 'tracked' or rewarded, as opposed to models. Models were reported to be the means for prestige and upward mobility in the field [112] with ML publications that generated citations, making practitioners competitive for AI/ML jobs and residencies. "*Everyone wants to do the model work, not the data work*" (P4, healthcare, India). Many practitioners described data work as time-consuming, invisible to track, and often done under pressures to move fast due to margins—investment, constraints, and deadlines often came in the way of focusing on improving data quality. Additionally, it was difficult to get buy-in from clients and funders to invest in good quality data collection and annotation work, especially in price-sensitive and nascent markets like East and West African countries and India. Clients expected

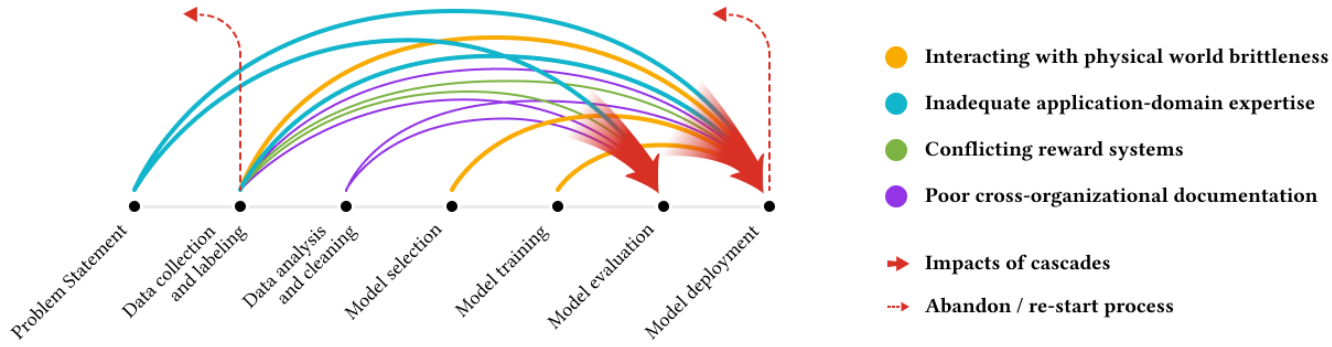


**Figure 1: Data cascades in high-stakes AI. Cascades are opaque and protracted, with multiplied, negative impacts. Cascades are triggered in the upstream (*e.g.,* data collection) and have impacts on the downstream (*e.g.,* model deployment). Thick red arrows represent the compounding effects after data cascades start to become visible; dotted red arrows represent abandoning or re-starting of the ML data process. Indicators are mostly visible in model evaluation, as system metrics, and as malfunctioning or user feedback.**

'magic' from AI—a high performance threshold without much consideration for the underlying quality, safety, or process—which led to model performance 'hacking' for client demonstrations among some practitioners.

**Data education** Lack of adequate training on AI data quality, collection, and ethics led to practitioner under-preparedness in dealing with the complexity of creating datasets in high-stakes domains. AI courses focused on toy datasets with clean values (*e.g.,* UCI Census, Kaggle datasets), but AI in practice required the creation of data pipelines, often from scratch, going from ground truth to model maintenance. As P37 working on healthcare in a West African country explained, "*In real life, we never see clean data. Courses and trainings focus on models and tools to use but rarely teach about data cleaning and pipeline gaps.*"; also illustrated by P27, a faculty in the US, "*we in CS are never trained, nor [are we] thinking actively about data collection.*" Computer Science curricula did not include training for practical data aspects such as dealing with domain-specific 'dirty data'[8], dealing with live data, defining and documenting datasets, designing data collection, training raters, or creating labelling task designs. In the US, most practitioners completed AI specialisation in graduate programs. In India and East and West African countries, most practitioners self-learned after their Computer Science degrees—but in all these routes, data engineering was under-emphasised.

**Data bootstrapping** High-stakes AI domains required specialised datasets by region, demographics, phenomena, or species, especially in under-digitised environments (*e.g.,* spread of Malaria in rural Tamil Nadu, elephant movements in Maasai Mara). **74%** of practitioners undertook data collection efforts from scratch, through field partners—a task which many admitted to being unprepared for, and some reported giving up on AI projects as a result. Practitioners from the US largely bootstrapped from existing sources and established digital infrastructures, *e.g.,* satellite data, sensor data, and public datasets, whereas the majority of practitioners in East and West African countries and India collected data from scratch with field partners and made online datasets work for local contexts (to avoid bureaucratic and local regulatory processes) [116]. Bootstrapping with data from another locale led to generalizability limitations, *e.g.,* P20 (clean energy, US) used satellite data from Northeast US to bootstrap model training, but were unable to apply to the target location due to different terrain, clouds, and pollution. Practitioners reported facing situations where they had to *"work with what they have"* (P16, healthcare, US), and did not always have the *"selectable capability"* (P29, environment and climate, US) to discard poor quality examples because of limited data in the first place. Many practitioners reported using data collected for non-AI purposes, *e.g.,* migration surveys, but ran into issues with ML feature-richness.

**Downstream accountability** One of the defining characteristics of high-stakes AI is the implied accountability to living beings. Data cascades occurred as practitioners ran up against challenges because of data scarcity and downstream methodologies in working with vulnerable groups. Stakes from poor performance were primarily in the form of harm to the community, but also resulted in poor performance and low user trust. *"If you build this model (e.g., predicting [eye disease]) and it predicts that this person does not have it when they do, you leave this person to go blind."* (P30, healthcare, a West African country). Many reported how consumer AI, *e.g.,* ad tech, typically aimed for 70-75% accuracy, whereas for high-stakes every extra 1% was crucial. "*There isn't a clear methodology for how to do it [test models] effectively without leading to some kind of harm to the patient. Everything starts with risk.*" (P10, P11, P12, healthcare, USA). Application domains in the US in our study could be described as 'second wave' AI, a broader interpretation focused on ecology, climate, and well-being, whereas domains in India and East and West African countries were more closely tied to sustainable development goals like micro-finance, healthcare, and farming, more directly tied to human impacts.

[8] 'Dirty data' is common parlance in AI/ML to refer to data errors. Richardson *et al.* . [104] complicate how dirty data can be influenced by corrupt, biased, or unlawful practices.

## 4.3 Data cascade triggers and practices

We present the various data cascades and surrounding behaviours observed in our study, sorted by frequency. Table 2 gives an overview of four core cascades—triggers, impacts and signals—and their distribution. Impacts varied in severity, from wasted time and effort to harms to beneficiaries. The most severe data cascades were also long-drawn and completely unknown to practitioners; in some cases, taking 2-3 years to manifest.

*4.3.1 Interacting with physical world brittleness (54.7%).* In high-stakes domains, AI systems transitioned from well defined, digitised, and understood environments to brittle deployments closely interacting with previously not-digitised physical worlds (almost by definition due to its involvement in socio-economic domains), *e.g.,* air quality sensing, ocean sensing, or ultrasound scanning. While all production AI systems are challenged by the inevitable changes in the external world, high-stakes AI have have even more reasons for a model to break—due to limited training data, complex underlying phenomena, volatile domains, and changes to regulations. In high-stakes domains, interaction with the external world spanned both the upstream (data sources) and downstream (live data and data instruments) of ML models. Data cascades often appeared in the form of hardware, environmental, and human knowledge drifts. As an example of a cascade, for P3 and P4 (road safety, India), even the slightest movement of a camera due to environmental conditions resulted in failures in detecting traffic violations, *"10 different sources may have undergone changes. Cameras might move from the weather. AI models can fail completely."*. Conventional AI practices on pristine training data (but messy live data), as well as a lack of training on working with messy real-world data appeared to trigger these cascades. Data cascades here took the longest to manifest, taking up to 2-3 years to emerge, almost always in the production stage. Impacts included complete model failure, abandonment of projects, and harms to beneficiaries from mispredictions.

*Cascades triggered by 'hardware drifts': e.g.,* cameras and sensors, during the data generation for the training dataset and upon deployment. 75% of practitioners used a hardware component as a part of their data capture infrastructure. To ensure good model performance, data collection efforts often occurred in controlled environments in-house or by giving data capture specifications to their data collection teams. As described by practitioners, production environments are *"utter chaos"* and bring in various forms of *"bad data"* (P4, healthcare, India). P44 (healthcare, India) described how technical errors filtered through if the *"[eye disease] hardware is not serviced properly every 12 months"*. Similarly, P9 (water consumption, India) described their complex approach of digging into the earth, cutting into pipes, and inserting sensing hardware, making it hard to detect subterranean sensor drifts. Artefacts like fingerprints, shadows, dust on the lens, improper lighting, and pen markings were reported to affect predictions. Rain and wind moved image sensors in the wild (*e.g.,* in camera traps and traffic detection), leading to incorrect model results. Models were reported to mistake spurious events as signals for phenomena, leading to complete AI system failures in some cases, *e.g.,* "*Suppose an image is out of focus, there is a drop of oil, or a drop of water on the image, appearing blurry or diffused. A model which is looking at this can easily get confused that an out-of-focus image is cancer.*" (P52, healthcare, India).

*Cascades triggered by 'environmental drifts':* resulted from changes in the environment or climate, *e.g.,* P29 (landslide segmentation, US) reported that presence of cloud cover, new houses or roads, or vegetation growth posed challenges because their model was comparing pre- and post-images and misconstruing the changes as landslides. In some cases, joins of live data across different geographies and environments triggered cascades, such as disparate emissions standards across countries (P20, clean energy, US), or different medical scanning procedures (P44, healthcare, India).

*Cascades triggered by 'human drifts':* where social phenomena or community behaviour led to changes in live data. Furthermore, with amendments to policies and regulations in the problem domain, features may cease to be relevant (*e.g.,* banking regulations affecting data capture). P15, a researcher in the US recalled a case where someone they knew built a medication tracking system for older adults. They had stopped receiving data from a user, who was detected to have unfortunately died and had stopped recording data a few days prior. The user presented behaviours that the model could not account for (*e.g.,* they switched off phone sensors). P15 was concerned that lack of continuous data for mental health conditions could be a sign of worsening conditions or suicide ("*the best data to detect in time*"). Similarly, P48 (healthcare, US) explained how creating an AI model for the COVID-19 pandemic on day 1 versus day 100 required a total change in various assumptions since the pandemic and human responses were volatile and dynamic.

To address these cascades, a few practitioners consistently monitored their data sources (often, at an example level), and looked for spurious changes through model performance degradation, and retrained models. In rare cases, practitioners intentionally introduced noise in training data to improve robustness, through noisy images or synthetically modified data. As P44 above shared, *"Many times, the quality of the dataset goes down. But it makes the model better and robust enough to ignore that image"*. A few practitioners invested in scalable data literacy for system operators and field partners, noting how operator trust and comfort with the AI system ultimately led to better data and inferences.

*4.3.2 Inadequate application-domain expertise (43.4%).* A data cascade was triggered when AI practitioners were responsible for data sense-making (defining ground truth, identifying the necessary feature sets, and interpreting data) in social and scientific contexts in which they did not have domain expertise. Answering these questions entailed an understanding of the application domain, social aspects, and embedding context [118, 123]. For instance, diagnosing fractured bones, identifying locations that could be poaching targets, and congenital conditions leading to preterm babies all depended on expertise in biological sciences, social sciences, and community context. Several practitioners worked with domain experts and field partners; however, they were largely involved in data collection or trouble-shooting, rather than in deep, end-to-end engagements. Practitioners described having to take a range of data decisions that often surpassed their knowledge, not always involving application-domain experts *e.g.,* discarding data, correcting values, merging data, or restarting data collection—leading to long, unwieldy and error-prone data cascades. As an example of a cascade, P18 (wildlife conservation, India) described how after deploying their model for making predictions for potential poaching locations, patrollers contested the predicted locations as being

| Cascades | Triggers | Impacts | Signals |
|---|---|---|---|
| **Interacting with physical world brittleness (54.7%)** *IN: 56.5%, EA & WA: 42.9%, US: 62.5%* | • Pristine training data (messy live data)<br>• Ill-equipped to work with volatile real-world data | • Harms to beneficiaries<br>• Complete model failure<br>• Abandonment of projects | • System performance in deployment |
| **Inadequate application-domain expertise (43.4%)** *IN: 47.8%, EA & WA: 57.1%, US: 25%* | • Overt reliance on technical expertise in sensemaking<br>• Moving fast to proof-of-concept | • Harms to beneficiaries<br>• Costly iterations | • System performance<br>• Post-hoc consulting with domain experts |
| **Conflicting reward systems (32.1%)** *IN: 30.4%, EA & WA: 57.1%, US: 12.5%* | • Misaligned incentives<br>• Inadequate data literacy among partners<br>• Viewing data as non-technical | • Costly iterations<br>• Moving to a new data source<br>• Quitting the project | • System performance<br>• Burned partner relations |
| **Poor cross-organisational documentation (20.8%)** *IN:17.4%, EA & WA: 35.7%, US: 12.5%* | • Neglecting value of data documentation | • Discarding part/entire dataset<br>• Wasted time and effort | • Manual instances reviews, mostly by 'chance' |

**Table 2: Prevalence and distribution of data cascades. IN is short for India, EA & WA for East African and West African countries respectively, and US for the United States.**

incorrect. Upon further collaboration with the patrollers, P18 and team learned that most of the poaching attacks were not included in the data. As the patrollers were already resource-constrained, the mispredictions of the model ran the risk of leading to over-patrolling in specific areas, leading to poaching in other places. In some cases, data collection was expensive and could only be done once (*e.g.,* underwater exploration, road safety survey, farmer survey) and yet, application-domain experts could not always be involved. Conventional AI practices like overt reliance on technical expertise and unsubstantiated assumptions of data reliability appeared to set these cascades off. Application-domain expertise cascades were costly: impacts came largely after building models, through client feedback and system performance, and long-winded diagnoses. Impacts included costly modifications like going back to collect more data, improving labels, adding new data sources, or severe unanticipated downstream impacts if the model had already been deployed (see figure 1)

Next, we describe *two* prominent examples of application-domain expertise issues that occurred in the AI lifecycle: dealing with subjectivity in ground truth, defining and finding representative data.

*Cascades triggered by dealing with subjectivity in ground truth*

High-stakes AI requires specialised, subjective decision-making in defining the ground truth, and breadth and number of labels [13]. Example of ground truth decisions are detecting cancer in pathology images, identifying quality of agriculture produce, and analysing insurance claims for acceptance or rejection. Cascades often occurred as a result of limited application-domain understanding of subjective labelling. In our study, practitioners often worked with several resource constraints of domain expertise and time, unable to use best practice data quality metrics for computing inter- and intra-rater reliability (*e.g.,* [10]). With no direct indicators of subjective shortcomings in data, cascades from ground truth issues were discovered through 'manual reviews' of data with clients or field partners, and often, through downstream impacts. Consider an example of P28, an educational AI engineer building an interactive writing model for students (country blinded) reported that they had not considered the impacts on low-income students or students with different English writing styles [5]. In some cases, ground truth was inaccurate but deeply embedded into systems, as in the case of P6 (credit assessment, India), *"decisions taken by insurance companies in the past about accepting or denying claims, for 10-15% of the time, the ground truth itself is inaccurate. If the wrong decision [subjective] was taken, there is no way to go back in historical data to correct [...] Two different people have different perspectives on whether claims should be accepted or rejected. How can you tell whether data is inaccurate or accurate? It introduces errors in our models."*

*Cascades triggered by poor application-domain expertise in finding representative data*

For an AI model to generalise well, it needs to be trained on representative data reflective of real-world settings. Second to data collection, understanding and collecting representative data was the biggest challenge for practitioners in high-stakes domains. Cascades occurred because of a default assumption that datasets were reliable and representative, and application-domain experts were mostly approached only when models were not working as intended. Cascades from non-representative data from poor application-domain expertise manifested as model performance issues, resulting in redoing data collection and labelling upon long-winded diagnoses. It is important to note that representativeness has a different interpretation for every domain and problem statement. With limited application-domain expertise, practitioners described how incomplete knowledge and false assumptions got incorporated into model building. A few practitioners relied on domain experts to define what representative data meant for their problem statement, *e.g.,* the classification of carcinomas in West African countries and how it varied in different populations (P39, healthcare, a West African country), or how farm produce defects manifest in different varieties and geographies (P24, agriculture, India). In cases where practitioners understood the need for representative data and its meaning in their context, they faced challenges in collecting this data without the right field partnerships. Representative data cascades sometimes stemmed from a disparity in contexts between data collection and system deployment. As P52 (healthcare, India) describes in the context of sampling, *"are we taking 90% of the data from one hospital and asking to generalise for the entire world?"*.

*4.3.3 Conflicting reward systems (32.1%).* Misaligned incentives and priorities between practitioners, domain experts, and field partners led to data cascades. An example of this cascade is how P27's (wildlife conservation, US) dataset rendered their ML model dysfunctional, *"Often they forgot to reset their setting on the GPS app and instead of recording every 5 minutes, it was recording [the data]*

*every 1 hour. Then it is useless, and it messes up my whole ML algorithm".* Conventional AI data practices of viewing data collection as outsourced and non-technical tasks, and a lack of understanding provenance, as well as misaligned incentives and poor data literacy among stakeholders, appeared to contribute to this data cascade. Practitioners saw the impacts of this cascade discovered well into deployment, through costly iterations, moving to an alternate data source, or quitting the project altogether.

As mentioned earlier, high-stakes domains lacked pre-existing datasets, so practitioners were necessitated to collect data from scratch. ML data collection practices were reported to conflict with existing workflows and practices of domain experts and data collectors. Limited budgets for data collection often meant that data creation was added as extraneous work to on-the-ground partners (*e.g.,* nurses, patrollers, farmers) who already had several responsibilities, and were not adequately compensated for these new tasks. Data collection and labelling tasks was often a competing priority with field partners' primary responsibility. As P7 (healthcare, India) shared, "*when a clinician spends a lot of time punching in data, not paying attention to the patient, that has a human cost*".

Field partners, especially at the frontlines, were reported to have limited data literacy and face information symmetry issues with not knowing the importance of their data collection, purpose of the AI system, and the importance of such constraints for the ML data, *e.g.,* in P21's (healthcare, India) case, "*doctors didn't want to do the test [for AI data collection] for so long. Almost 25-30% recordings were less than 10 minutes which are not useful for any [AI] analysis. We had to work with the doctor to tell them why it is important to capture that kind of length of the data.*". A healthcare startup founder from India, P22, shared an account of speaking to a community health worker in India, and why the health worker eventually became unmotivated to complete their data work: "*[they quoted] Whatever work I do or I don't do, my salary is 3K [INR] per month. Earlier I did everything (collected good data), but my salary did not increase.*". Top-level management was reported to often enter mutually synergistic partnerships, through joint research publications or media attention, but not the frontline workers whose labour benefited AI data collection. In a few cases, field workers were reported to fabricate data from either no or per-task incentives.

Some AI practitioners were aware of, and explicitly discussed problematic incentives for their data collectors or domain experts, and shared how they were resource-constrained (echoing Ismail and Kumar [58]). Some reflected on how providing more transparency and information about the scope of the project could have helped their field partners. In practice, data literacy training (*e.g.,* entering well-formatted values, educating about the impacts of their data collection) was rarely conducted, resulting in numerous data quality challenges like data collectors not recording data for a specific duration or frequency. In the rare case where practitioners trained their field partners, data quality was reported to go up, as in the case of P7 (healthcare, India), who described how providing real-time data quality indicators enabled their field partners to become conscious of data quality *in-situ.* (In a few cases, data collectors gathered specialised domain expertise from working on ML projects and up-skilled to starting new businesses, *e.g.,* seed identification.) In a few cases where incentives were explicitly discussed as being provided, high monetary incentives sometimes led to over-sampling, skewing the data.

*4.3.4 Poor cross-organisational documentation (20.8%).* Data cascades were set off by a lack of documentation across various cross-organisational relations (within the organisation, with field partner organisations and data collectors, and with external sources). Practitioners discussed several instances where collected and inherited datasets lacked critical details. Missing metadata led practitioners to make assumptions, ultimately leading to costly discarding of datasets or re-collecting data. As an example of a data cascade, P8 (robotics, US), described how a lack of metadata and collaborators changing schema without understanding context led to a loss of four months of precious medical robotics data collection. As high-stakes data tended to be niche and specific, with varying underlying standards and conventions in data collection, even minute changes rendered datasets unusable. Conventional AI practice of neglecting the value of data documentation, and field partners not being aware of constraints in achieving good quality AI appeared to set these cascades off. Cascades became visible through manual reviews, but often by 'chance'. The impacts of cascades here included wasted time and effort from using incorrect data, being blocked on building models, and discarding subsets or entire datasets (not always feasible to re-collect resource-intensive data, as we explain above).

Metadata on equipment, origin, weather, time, and collection process was reported to be critical information to assess quality, representativeness, and fit for use cases. As P7, a researcher in India explained the importance of context in data, "*In my experience, in medicine, the generalisation is very poor. We have been trying to look at what really generalises in cross continental settings, across [American hospitals] and [Indian hospitals]. More than data quality it is the auxiliary, lack of metadata that makes all the difference [...] If we look at signals without the context, it makes it difficult to generalise the data.*" However, in most cases where practitioners did not have access to the metadata, they had to discard the data point or subset of data altogether. P13, working on criminal justice systems in India explained, *"We have seen that it depends a lot on when the data was collected. If it was over a year [ago], there is some correlation between the season and the time of year the data was collected.[...] again in most of the data we have missing information. We have to reject the entire data that might be relevant for this particular problem."*

In dealing with a lack of metadata, practitioners made assumptions about the datasets, like in the case of P20 (clean energy, US), who assumed certain timestamps on power plant data because metadata was missing, "*but the plant was mapped incorrectly, mismatch of timestamps between power plant and satellite. Very hard to tell when you don't own the sensors. You have to make assumptions and go with it.*" Many practitioners expressed frustration from a lack of standards to help document datasets (*e.g.,* using Lagos versus Lagos State due to lack of metadata).

In a few cases where metadata cascades were avoided, practitioners created reproducible assets for data through data collection plans, data strategy handbooks, design documents, file conventions, and field notes. For example, P46 and P47 (aquaculture, US) had an opportunity for data collection in a rare Nordic ocean environment, for which they created a data curation plan in advance and took ample field notes. A note as detailed as the time of a lunch break

saved a large chunk of their dataset when diagnosing a data issue downstream, saving a precious and large dataset.

## 5 DISCUSSION

Our results indicate the sobering prevalence of messy, protracted, and opaque data cascades even in domains where practitioners were attuned to the importance of data quality. Individuals can attempt to avoid data cascades in their model development, but a broader, systemic approach is needed for structural, sustainable shifts in how data is viewed in AI praxis. We need to move from current approaches that are reactive and view data as 'grunt work'. We need to move towards a proactive focus on *data excellence*—focusing on the practices, politics, and values of humans of the data pipeline to improve the quality and sanctity of data, through the use of processes, standards, infrastructure and incentives (and other interventions, as identified by Paritosh *et al.* [92]). Any notion of data excellence should also explicitly engage with shifting the power centres in data resources between the Global South and North. We identify opportunities to further expand HCI's role as the conscience of the computing world and its long-standing commitment to data, through implications for human-centred incentives, processes, metrics, and interfaces for data excellence in high-stakes domains. While our analysis is limited to high-stakes AI projects, we believe these challenges may exist in more or less amplified forms in all of AI development.

**From goodness-of-fit to goodness-of-data** The current AI revolution is metrics-driven, as Thomas points out ([120]), but practitioners largely used system metrics to measure the goodness of the fit of the model to the data. Goodness-of-fit metrics, such as F1, Accuracy, AUC, do not tell us much about the phenomenological fidelity (representation of the phenomena) and validity (how well the data explains things related to the phenomena captured by the data) aspects of the data. Currently, there are no standardised metrics for characterising the goodness-of-data [11, 13]; research on metrics is emerging [15, 91] but not yet widely adopted in AI system building. As a result, there is an extreme reliance on goodness-of-fit metrics and post-deployment product metrics. First, these metrics give us no assurances about the quality of the data. Second, they are too late to detect and course-correct from the unforeseen effects of data cascades. Even more importantly, deployment of AI systems in high-stakes domains eventually exposes aspects of phenomenon that were not captured in the dataset, which can produce spurious and risky outcomes, as pointed out by Floridi *et al.* [37] and Burt and Hall [24]. To illustrate the importance of goodness-of-data metrics, consider a model that is trying to recognise whether a given location can be a poaching target. Given an arbitrary dataset of labelled, prior poaching attempts, one can train and evaluate the model on a held-out set to estimate the goodness-of-fit of the model to the data. Note that while these metrics tell us about the fit of the model, they do not tell us anything about the quality of the dataset. Wildlife AI practitioners reported how they retroactively needed to understand information on where poaching typically took place; whether a human was a villager, wildlife professional, or poacher; whether an area was a farmland or forest; where the water sources were, and so on—which they had not captured in their datasets and ground truth. It is easy to imagine a model with a perfect fit to a very narrow slice of the data—and show high performance—and starting to reveal its weaknesses as it is used to make decisions outside of that narrow slice, where it can fail in immeasurable and unforeseen ways.

While collecting rigorous data from, and about, humans is relatively uncharted waters for AI researchers, there is a rich body of research in HCI that is crucial in even framing these questions appropriately—opening up a whole new space for HCI to act as the compass for AI by answering questions about goodness, fidelity, and validity of data by itself, as HCOMP researchers have pointed out [12, 90]. Similarly, recognizing the relevance of viewing *data-in-place* [118]—the situatedness of data within social and physical geographies—*i.e.,* the dynamic after-life of data once models are deployed, will help evaluate how models interact and impact living beings and artefacts. Emerging scholarship like Beede *et al.* 's evaluation of real-world deep learning systems [17] point to the need for incorporating HCI early and throughout in AI data. A whole new science of data is needed, with HCI partnership, where sorely needed phenomenological goodness-of-data metrics need to be developed. Making progress on measuring goodness-of-data will enable early-stage assessment and feedback in the data collection process, and will likely surface data-phenomena gaps earlier, avoiding data cascades. Focusing on phenomenological validity of data will further increase the scientific value and reusability of the data (a precious entity in high-stakes domains). Such research is necessary for enabling better incentives for data, as it is hard to improve something we can not measure.

**Incentives for data excellence** Contrary to the scientific, designerly, and artful practices observed in prior HCI studies on data scientists by Feinberg [35], Muller *et al.* [86], and Patel *et al.* [96], AI practitioners in our study tended to view data as 'operations'. Such perceptions reflect the larger AI/ML field reward systems: despite the primacy of data, novel model development is the most glamorised and celebrated work in AI—reified by the prestige of publishing new models in AI conferences, entry into AI/ML jobs and residency programs, and the pressure for startups to double up as research divisions. Critics point to how novel model development and reward systems have reached a point of ridicule: Lipton calls ML scholarship 'alchemy' [74], Sculley *et al.* describe ML systems as 'empirical challenges to be 'won" [112], Bengio describes ML problems as 'incremental' [18], and plagiarism by ML educators has been labelled as the 'future of plagiarism' [14]. In contrast, datasets are relegated to benchmark publications and non-mainstream tracks in AI/ML conferences [46, 82]. New AI models are measured against large, curated data sets that lack noise (to report high performances), in contrast to the dynamic nature of the real world [64, 78]. In addition to the ways in which business goals were orthogonal to data (also observed by Passi and Sengers [95]), practitioners described how publication prestige, time-to-market, revenue margins, and competitive differentiation often led them to rush through the model development process and sometimes artificially increase model accuracy to deploy systems promptly, struggling with the moral and ethical trade-offs.

We take inspiration from Sculley *et al.* [112] and Soergel *et al.* [114] to propose starting points for changing structural incentives for the market, academy, and capital of AI/ML. Conferences are a good starting point: data powers the inferences, and empiricism on

data should be mainstream. Conferences like SIGCHI, CSCW, and AAAI are good examples of recognising the importance of research on data through their disciplinary conventions, *e.g.,* crowd work, human computation, and data visualization. Papers on AI/ML techniques should evolve to offer dataset documentation, provenance, and ethics as mandatory disclosure. Standard research process as relevant to the research community, *e.g.,* hypotheses, design, experiments, and testing should also be followed with data [28, 55]. Organisations should reward data collection, pipeline maintenance, gluework, data documentation, and dataset repairs in promotions and peer reviews, similar to how good software engineering is rewarded. Similarly, complementing Møller *et al.* [85], we note that data labour is currently lopsided, fuelling the benefit of AI practitioners, and dis-empowering application-domain experts and field partners. Data excellence emphasises the value in sustained partnerships, as opposed to engagements with experts on a one-off basis (during problem formulation or sensemaking only). Some instances of partnerships needed throughout the ML pipeline include formulating the problem and outcomes, identifying anomalies, determining optimal frequency for data collection, verifying model outcomes, and giving feedback on model behaviour. Greater collaboration, transparency into AI application use-cases, data literacy, and 'shared rewards' (*e.g.,* joint publications and releases) are some ways to engender '*data compassion*' (P37), and recognise and learn from expertise. Learning from HCI scholarship on ways to recognise the human labour in preparing, curating, and nurturing data that powers AI models [34, 117], among crowd workers [34, 57, 79], office clerks [85], and health workers [58] can be helpful. For example, Martin *et al.* [79] through their understanding of MTurker perspectives, call for tools to help reduce and manage all the invisible, background work by Mturkers. Møller *et al.* [85] created a toolkit for stakeholders to identify and value data work, and Ismail and Kumar call for embracing solidarity through design [58].

**Real-world data literacy in AI education** A majority of curricula for degrees, diplomas, and nano-degrees in AI are concentrated on model development [42], leaving graduates under-prepared for the science, engineering, and art of working with data, including data collection, infrastructure building, data documentation, and data sense-making. Toy datasets and open datasets with unknown characteristics are abundant in AI education, like in the UCI census dataset [4]. In practice, cutting-edge AI applications often require unique datasets created from scratch, as a necessity, and a competitive advantage; but the practical data skill gaps among our practitioners were quite large from their formal education and training. Data collection in high-stakes domains is an interdisciplinary activity, and requires engaging in data sensemaking activities as described by Koesten *et al.* [68], often without adequate application-domain expertise, working with domain experts, as well as knowledge of methodologies for collecting data from experts. Unfortunately, as it stands, there is often a lack of involvement and appreciation for application-domain experts in AI/ML. An oft-quoted quip in the Natural Language Processing community: "*Every time I fire a linguist, the performance of the speech recognizer goes up*" attributed to Frederick Jelinek [49], reflects the hostility towards domain expertise. Early progress in the field—the low hanging fruits relying on quantity alone—no longer applies to harder, more subjective problems and edge cases. Entire under-represented groups can show up as edge cases, with profound social implications [88]. For instance, Scheuerman *et al.* [110] found that facial analysis technologies were unable to identify non-binary genders. Training on data collection, curation, and inter-disciplinary collaboration can help prepare future practitioners. Fortunately, there is a massive body of research in HCI, Human Computation, and allied fields on empirical methods [32] that can be added to AI curricula. Data ethics and responsible AI education, oversight boards *e.g.,* IRB, and ethics standards should be necessary components of AI education and praxis, given the field's increasing expansion into high-stakes, humanitarian areas (*e.g.,* how our practitioners, despite their intentionality, were under-equipped to understand human impacts)—a call to action invoked by ethics and education scholars like Saltz *et al.* [105].

**Better visibility in the AI data lifecycle** Data cascades point to the need for several feedback channels at different time scales in the AI life cycle. With delayed and hidden manifestation, practitioners struggled with understanding the impact of data scrutiny, and utilised 'launch and get feedback' approaches frequently, often at great cost. The teams with the least data cascades had step-wise feedback loops throughout, ran models frequently, worked closely with application-domain experts and field partners, maintained clear data documentation, and regularly monitored incoming data. Data cascades were by-and-large avoidable through intentional practices, modulo extrinsic resources (*e.g.,* accessible application-domain experts in the region, access to monetary resources, relaxed time constraints, stable government regulations, and so on). Although the behaviour of AI systems is critically determined by data, even more so than code [111]; many of our practitioner strategies mirrored best practices in software engineering [38, 83]. Anticipatory steps like shared style guides for code, emphasising documentation, peer reviews, and clearly assigned roles—adapted to data—reduced the compounding uncertainty and build-up of data cascades.

Current inspection and analysis tools tend to focus on dataset distributions and wrangling (*e.g.,* Trifacta[9], FACETS[10], and OpenRefine[11]) as ways to improve data quality, whereas the upstream work of defining dataset requirements and downstream challenges of monitoring incoming live data and measuring impacts often does not receive the critical attention it needs from the HCI and AI communities. Just as designer Bret Victor described, we now have tools "*to adapt unthinkable thoughts to the way that our minds work*" [124], we now need better tools to collect, interpret, and observe data to transform the current practices in the upstream and downstream. Customizable tools for dataset collection and labelling can significantly improve data quality, in the place of in-house, cobbled together solutions. Live data from systems in production was consistently reported to spring up surprise drifts and affect model inferences, but comprehensive solutions are lacking. Dataset documentation is under-developed, unlike code documentation [127], *e.g.,* design documents, meeting notes, project diaries, and rater instructions; but standards here can help reduce uncertainty.

**Data equity in the Global South** Our study points to how AI/ML technologies were widely accessible and democratic to new entrants, across geographies, through open-sourced and pre-trained

[9] https://www.trifacta.com/
[10] https://pair-code.github.io/facets/
[11] https://openrefine.org/

models, easy-to-access courses and codebases, and grassroots communities. AI practitioners across geographies appeared to have similar access to models. However, we find drastic differences when it comes to data and compute in East and West African countries [2] and India [3], compared to the US. With limited digital infrastructures and fewer socio-economic datasets, data collection was often done from scratch through field partners and in-house efforts. Data collection involved navigating vague data policies and regulation, manual efforts to hand-curate data, and introducing AI literacy to partners—efforts above and beyond what practitioners were trained or equipped to do. Our findings echo the insights of ICTD and AI4SG scholarship on the realities of data scarcity and quality challenges *e.g.,* [31, 98, 107], understanding socio-cultural factors *e.g.,* [20, 109], and complex partner and government relations *e.g.,* [22] in AI projects in the Global South. Invoking Sambasivan *et al.* , we argue that the data disparities are symptoms of the larger, uneven ML capital relations in the world, where the Global South is viewed as a site for low-level data annotation work, an emerging market for extraction from 'bottom billion' data subjects, or a beneficiary of AI for social good [107]. Developing and publishing open-sourced (de-identified) datasets, data collection tools, and trainings for defining the right data with application-domain expert knowledge can help mitigate the cold start problem. Greater ML literacy among civil society and clients can evolve high-stakes AI into a synergistic endeavour; being aware of, and asking the right questions of ML systems could help shift the focus from hacking model accuracy for performative reasons, to data excellence. Highlighting ongoing high-stakes AI projects and successes to both raise awareness and to provide a roadmap is essential to addressing the current inequities in data resources globally.

## 6 CONCLUSION

As AI becomes part and parcel of decision-making of core aspects of life, the sanctity and quality of data powering these models takes on high importance. We presented a qualitative study of data practices and challenges among 53 AI practitioners in India, East and West African countries, and the US, working on cutting-edge, high-stakes domains of health, wildlife conservation, food systems, road safety, credit, and environment. We observed and presented data cascades, often long-run, invisible, and compounding effects on AI models. The effects typically occurred as a result of applying conventional AI/ML practices in high-stakes domains—many of the conventional practices did not transfer neatly, and often resulted in serious impacts like community harms, discarded projects, and redoing data collection. Data cascades were typically triggered in the upstream and appeared unexpectedly in the downstream of deployment. System-level proxy metrics were utilised, which are only available towards the end of the development lifecycle, and do not shed light on data quality and its fidelity to phenomena. HCI has a crucial role to play in AI data excellence, through interfaces, measurement, incentives, and education, especially in fragile and vulnerable domains.

## 7 ACKNOWLEDGEMENTS

We are deeply grateful to the various AI practitioners who took part in our study and generously shared their AI development experiences with us. Our sincere thanks to Jose M. Faleiro, Kristen Olson, our CHI ACs, and reviewers for their invaluable feedback on the paper. We thank Biswajeet Malik, Siddhant Agarwal, Manish Gupta, Aneidi Udo-Obong, Divy Thakkar, Di Dang, and Solomon Awosupin for making connections to the AI practitioner community. We are deeply thankful to Sures Kumar Thoddu Srinivasan for creating the data cascades graphic.

## REFERENCES

[1] [n.d.]. 2019 Kaggle ML & DS Survey | Kaggle. https://www.kaggle.com/c/kaggle-survey-2019. (Accessed on 08/29/2020).
[2] [n.d.]. AI Readiness Index 2019 | AI4D | IAPD. https://ai4d.ai/index2019/. (Accessed on 09/14/2020).
[3] [n.d.]. Landscape of AI-ML Research in India. http://www.itihaasa.com/pdf/Report_Final_ES.pdf. (Accessed on 09/15/2020).
[4] [n.d.]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php. (Accessed on 09/15/2020).
[5] [n.d.]. A Vision of AI for Joyful Education - Scientific American Blog Network. https://blogs.scientificamerican.com/observations/a-vision-of-ai-for-joyful-education/. (Accessed on 09/14/2020).
[6] 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.
[7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.
[8] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).
[9] Appen. 2020. The 2020 Machine Learning Report and State of AI. https://appen.com/whitepapers/the-state-of-ai-and-machine-learning-report/. (Accessed on 09/16/2020).
[10] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1100–1105.
[11] Lora Aroyo, Anca Dumitrache, Jennimaria Palomaki, Praveen Paritosh, Alex Quinn, Olivia Rhinehart, Mike Schaekermann, Michael Tseng, and Chris Welty. [n.d.]. https://sadworkshop.wordpress.com/
[12] Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Human Computation* 1, 1 (Sep. 2014). https://doi.org/10.15346/hc.v1i1.34
[13] Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (Mar. 2015), 15–24. https://doi.org/10.1609/aimag.v36i1.2564
[14] Jonathan Bailey. 2019. Why Siraj Raval's Plagiarism is the Future of Plagiarism - Plagiarism Today. https://www.plagiarismtoday.com/2019/10/16/why-siraj-ravals-plagiarism-is-the-future-of-plagiarism/. (Accessed on 09/15/2020).
[15] 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.
[16] Anja Bechmann and Geoffrey C Bowker. 2019. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society* 6, 1 (2019), 2053951718819569.
[17] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[18] Yoshua Bengio. 2020. Time to rethink the publication process in machine learning - Yoshua Bengio. https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/. (Accessed on 08/18/2020).
[19] Anant Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J Elmore, Samuel Madden, and Aditya G Parameswaran. 2014. Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798* (2014).
[20] Joshua Blumenstock. 2018. Don't forget people in the use of big data for development.
[21] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2019. Data validation for machine learning. In *Conference on Systems and Machine Learning (SysML). https://www. sysml. cc/doc/2019/167. pdf.*

[22] Waylon Brunette, Clarice Larson, Shourya Jain, Aeron Langford, Yin Yin Low, Andrew Siew, and Richard Anderson. 2020. Global goods software for the immunization cold chain. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 208–218.

[23] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *International conference on database theory*. Springer, 316–330.

[24] Andrew Burt and Patrick Hall. [n.d.]. What to Do When AI Fails âĂŞ O'Reilly. https://www.oreilly.com/radar/what-to-do-when-ai-fails/, month = 09, year = 2020, note = (Accessed on 09/16/2020).

[25] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[26] Kuang Chen, Joseph M Hellerstein, and Tapan S Parikh. 2011. Data in the First Mile.. In *CIDR*. Citeseer, 203–206.

[27] Xu Chu, Ihab F Ilyas, and Paolo Papotti. 2013. Holistic data cleaning: Putting violations into context. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, 458–469.

[28] Josh Cowls, Thomas King, Mariarosaria Taddeo, and Luciano Floridi. 2019. Designing AI for social good: Seven essential factors. *Available at SSRN 3388669* (2019).

[29] Ward Cunningham. 1992. The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger* 4, 2 (1992), 29–30.

[30] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)* 51, 1 (2018), 1–40.

[31] Maria De-Arteaga, William Herlands, Daniel B Neill, and Artur Dubrawski. 2018. Machine learning for the developing world. *ACM Transactions on Management Information Systems (TMIS)* 9, 2 (2018), 1–14.

[32] Alan Dix, Alan John Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2003. *Human-computer interaction*. Pearson Education.

[33] Farzana Dudhwala and Lotta Björklund Larsen. 2019. Recalibration in counting and accounting practices: Dealing with algorithmic output in public and private. *Big Data & Society* 6, 2 (2019), 2053951719858751.

[34] Hamid Ekbia and Bonnie Nardi. 2014. Heteromation and its (dis) contents: The invisible division of labor between humans and machines. *First Monday* (2014).

[35] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2952–2963.

[36] Kathleen Fisher and Robert Gruber. 2005. PADS: a domain-specific language for processing ad hoc data. *ACM Sigplan Notices* 40, 6 (2005), 295–304.

[37] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4PeopleâĂŤan ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707.

[38] Andrew Forward and Timothy C Lethbridge. 2002. The relevance of software documentation, tools and technologies: a survey. In *Proceedings of the 2002 ACM symposium on Document engineering*. 26–33.

[39] Martin Fowler. 2019. TechnicalDebt. https://martinfowler.com/bliki/TechnicalDebt.html. (Accessed on 09/16/2020).

[40] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[41] Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT press.

[42] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[43] Laura M Haas, Mauricio A Hernández, Howard Ho, Lucian Popa, and Mary Roth. 2005. Clio grows up: from research prototype to industrial tool. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. 805–810.

[44] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[45] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. 2018. Applied machine learning at facebook: A datacenter infrastructure perspective. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 620–629.

[46] Benjamin Heinzerling. 2020. NLP's Clever Hans Moment has Arrived. https://thegradient.pub/nlps-clever-hans-moment-has-arrived/

[47] Keith Hiatt, Michael Kleinman, and Mark Latonero. [n.d.]. Tech folk: 'Move fast and break things' doesn't work when lives are at stake | The Guardian. https://www.theguardian.com/global-development-professionals-network/2017/feb/02/technology-human-rights. (Accessed on 08/25/2020).

[48] Charles Hill, Rachel Bellamy, Thomas Erickson, and Margaret Burnett. 2016. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 162–170.

[49] J Hirschberg. 1998. Every time I fire a linguist, my performance goes up, and other myths of the statistical natural language processing revolution. Invited talk. In *Fifteenth National Conference on Artificial Intelligence (AAAI-98)*.

[50] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. 419–429.

[51] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (2004), 85–126.

[52] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[53] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2020. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *arXiv preprint arXiv:2010.13561* (2020).

[54] Nick Hynes, D Sculley, and Michael Terry. 2017. The data linter: Lightweight, automated sanity checking for ml data sets. In *NIPS MLSys Workshop*.

[55] John PA Ioannidis, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz, and Robert Tibshirani. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet* 383, 9912 (2014), 166–175.

[56] Lilly Irani. 2015. The cultural work of microwork. *New Media & Society* 17, 5 (2015), 720–739.

[57] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[58] Azra Ismail and Neha Kumar. 2018. Engaging solidarity in data collection practices for community health. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.

[59] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *arXiv preprint arXiv:1701.06207* (2017).

[60] Kaggle. 2019. 2019 Kaggle ML & DS Survey. https://www.kaggle.com/c/kaggle-survey-2019. (Accessed on 08/27/2020).

[61] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3363–3372.

[62] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926.

[63] Sasikiran Kandula and Jeffrey Shaman. 2019. Reappraising the utility of Google Flu Trends. *PLoS computational biology* 15, 8 (2019), e1007258.

[64] Hannah Kerner. [n.d.]. Too many AI researchers think real-world problems are not relevant | MIT Technology Review. https://www.technologyreview.com/2020/08/18/1007196/ai-research-machine-learning-applications-problems-opinion/. (Accessed on 08/18/2020).

[65] Mary Beth Kery, Amber Horvath, and Brad A Myers. 2017. Variolite: Supporting Exploratory Programming by Data Scientists.. In *CHI*, Vol. 10. 3025453–3025626.

[66] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2017. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering* 44, 11 (2017), 1024–1038.

[67] Ákos Kiss and Tamás Szirányi. 2013. Evaluation of manually created ground truth for multi-view people localization. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*. 1–6.

[68] Laura Koesten, Kathleen Gregory, Paul Groth, and Elena Simperl. 2019. Talking datasets: Understanding data sensemaking behaviours. *arXiv preprint arXiv:1911.09041* (2019).

[69] Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[70] Sanjay Krishnan, Michael J Franklin, Ken Goldberg, and Eugene Wu. 2017. Boostclean: Automated error detection and repair for machine learning. *arXiv preprint arXiv:1711.01299* (2017).

[71] Sanjay Krishnan, Daniel Haas, Michael J Franklin, and Eugene Wu. 2016. Towards reliable interactive data cleaning: A user survey and recommendations. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–5.

[72] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. 2016. Activeclean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment* 9, 12 (2016), 948–959.

[73] David Lazer and Ryan Kennedy. 2015. What We Can Learn From the Epic Failure of Google Flu Trends | WIRED. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. (Accessed on 08/27/2020).

[74] Zachary C Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341* (2018).

[75] Maria Littmann, Katharina Selig, Liel Cohen-Lavi, Yotam Frank, Peter Hönigschmid, Evans Kataka, Anja Mösch, Kun Qian, Avihai Ron, Sebastian Schmid, et al. 2020. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nature Machine Intelligence* (2020), 1–7.

[76] Raoni Lourenço, Juliana Freire, and Dennis Shasha. 2019. Debugging machine learning pipelines. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*. 1–10.

[77] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work 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.

[78] Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).

[79] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.

[80] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[81] 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).

[82] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging* 34, 10 (2014), 1993–2024.

[83] Tim Menzies. 2019. The five laws of SE for AI. *IEEE Software* 37, 1 (2019), 81–85.

[84] Hannah Miller and Richard Stirling. 2019. Government AI Readiness Index 2019 — Oxford Insights — Oxford Insights. https://www.oxfordinsights.com/ai-readiness2019. (Accessed on 09/14/2020).

[85] Naja Holten Møller, Claus Bossen, Kathleen H Pine, Trine Rask Nielsen, and Gina Neff. 2020. Who does the work of data? *Interactions* 27, 3 (2020), 52–55.

[86] 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.

[87] Tadhg Nagle, C. Thomas Redman, and David Sammon. 2017. Only 3% of CompaniesâĂŹ Data Meets Basic Quality Standards. https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards. (Accessed on 08/27/2020).

[88] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[89] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research* 42, 5 (2015), 533–544.

[90] Praveen Paritosh. 2018. The missing science of knowledge curation: improving incentives for large-scale knowledge curation. In *Companion Proceedings of the The Web Conference 2018*. 1105–1106.

[91] Praveen Paritosh, Kurt Bollacker, Maria Stone, Lora Aroyo, and Sarah Luger. 2020. Evaluating Evaluation of AI Systems (Meta-Eval 2020). http://eval.how/aaai-2020/. (Accessed on 09/16/2020).

[92] Praveen Paritosh, Matt Lease, Mike Schaekermann, and Lora Aroyo. 2020. First workshop on Data Excellence (DEW 2020). http://eval.how/dew2020/. (Accessed on 09/16/2020).

[93] Samir Passi and Steven Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2436–2447.

[94] 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.

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

[96] Kayur Patel, James Fogarty, James A Landay, and Beverly Harrison. 2008. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 667–676.

[97] James W Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211, 2828 (2011), 42–45.

[98] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. 2019. Examining the challenges in development data pipeline. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. 13–21.

[99] Kathleen H Pine and Max Liboiron. 2015. The politics of measurement and action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3147–3156.

[100] 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* (Chicago, Illinois, USA) *(SIGMOD '17)*. Association for Computing Machinery, New York, NY, USA, 1723âĂŞ1726. https://doi.org/10.1145/3035918.3054782

[101] 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.

[102] Vijayshankar Raman and Joseph M Hellerstein. 2001. Potter's wheel: An interactive data cleaning system. In *VLDB*, Vol. 1. 381–390.

[103] Thomas C. Redman. 2018. If Your Data Is Bad, Your Machine Learning Tools Are Useless. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

[104] 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.

[105] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–26.

[106] 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.

[107] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *ACM FaccT*.

[108] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. 2018. " Privacy is not for me, it's for those rich women": Performative Privacy Practices on Mobile Phones by Women in South Asia. In *Fourteenth Symposium on Usable Privacy and Security ({SOUPS} 2018)*. 127–142.

[109] Nithya Sambasivan and Jess Holbrook. 2018. Toward responsible AI for the next billion users. *interactions* 26, 1 (2018), 68–71.

[110] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[111] 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. In *Advances in neural information processing systems*. 2503–2511.

[112] David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's curse? On pace, progress, and empirical rigor. (2018).

[113] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial Intelligence for Social Good: A Survey. arXiv:2001.01818 [cs.CY]

[114] David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open Scholarship and Peer Review: a Time for Experimentationï£ij. (2013).

[115] Eliza Strickland. 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* 56, 4 (2019), 24–31.

[116] Iryna Susha, Åke Grönlund, and Rob Van Tulder. 2019. Data driven social partnerships: Exploring an emergent trend in search of research challenges and questions. *Government Information Quarterly* 36, 1 (2019), 112–128.

[117] Astra Taylor. 2018. The Automation Charade. https://logicmag.io/failure/the-automation-charade/.

[118] 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 *(CHI '15)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2702123.2702558

[119] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[120] Rachel Thomas and David Uminsky. 2020. The Problem with Metrics is a Fundamental Problem for AI. *arXiv preprint arXiv:2002.08512* (2020).

[121] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM 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.

[122] Jennifer Wortman Vaughan. 2017. Making better use of the crowd: How crowdsourcing can advance machine learning research. *The Journal of Machine Learning Research* 18, 1 (2017), 7026–7071.

[123] 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.

[124] Bret Victor. 2013. Media for Thinking the Unthinkable. http://worrydream.com/MediaForThinkingTheUnthinkable/. (Accessed on 09/15/2020).
[125] Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656* (2012).
[126] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems: Gender, race and power in AI. *AI Now Institute* (2019), 1–33.
[127] 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.
[128] Jing Zhang, Xindong Wu, and Victor S Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review* 46, 4 (2016), 543–576.
[129] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* (2020).