Data Desiderata: Reliability and Fidelity in High-stakes AI

Shivani Kapania, Nithya Sambasivan, Kristen Olson, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
Google Inc.
{kapania, nithyasamba, kristenolson, hhighfill, dakrong, pkp, loraa} @google.com

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 to living beings. 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 the USA. We define and report on the challenges faced by practitioners with two *essential* desiderata within AI high-stakes domains: data reliability and fidelity. We discuss thoughtful strategies that practitioner teams can adopt to improve data reliability and fidelity, and avoid data disasters—resulting in safer and robust systems for all.

As models become more commoditized, AI is increasingly expanding into consequential, *high-stakes domains* where the safety impact for living beings and the environment can be significant and immediate, such as cancer diagnosis, credit assessment, and poaching detection. The impact of data quality and representativeness in high-stakes AI domains is even greater due to the direct downstream implications on human lives and ecology.

While data largely determines performance, fairness, robustness, safety, and scalability of AI systems (Halevy, Norvig, and Pereira 2009), (Mehrabi et al. 2019), it is one of the biggest bottlenecks to AI development (Chu et al. 2016). Acquiring usable and high quality data is often the messiest and least predictable part of the AI engineering process. For AI researchers and developers, data is often the least incentivized aspect, viewed as 'operational' relative to building new models and algorithms. Intuitively, AI developers understand that data quality matters, often spending inordinate amounts of time on data-related tasks (up to 80% of time (Kaggle 2019)). However, in practice, most organisations fail to create or meet any data quality standards; in part, because attention to data quality is not valued (Nagle, Redman, and Sammon 2017).

In this work, we explicitly focus on two desiderata of data excellence— the reliability and fidelity of data due to the critical nature of ensuring these two properties for efforts of data collection in high-stakes domains. (Paritosh 2012) defines reliability as a general guarantee that the data obtained are independent of the measuring event, instrument or person. *Reliability* of data is a measure for its consistency, and stability which allows for replicability. It should not be confused with correctness or validity (*e.g.,* validity of a test). Metrics of reliability based on properties of the data collection process can be used to demonstrate it. Examples of reliability metrics are variance in human annotations, or variance of test results if people take it multiple times.

While reliability has been defined and used across multiple communities such as statistics, quantitative research and more (Krippendorff 1970), fidelity, on the other hand, seems to be less well-defined for AI. Fidelity is closely related to validity of the data— ensuring fidelity is a preemptive approach during data collection to ensuring validity. According to (Joppe 2000), validity determines whether the research truly measures that which it was intended to measure or how truthful are the research results. Thus, *fidelity* of data is a measure for its 'goodness'— whether the data has parity with the phenomena it seeks to represent. However, practitioners may have to make multi-objective trade-offs between two essential desiderata: reliability and fidelity which can be conflicting and mutually antagonistic (Paritosh 2012), that is, if we try to improve reliability by over-simplifying our definition of the phenomenon, it might compromise the fidelity of the data.

We present results from a qualitative study of AI data practices in high-stakes domains in India, Sub-Saharan Africa and the US. Through interviews with 53 AI developers, researchers, and founders working in AI application areas like landslide detection, suicide prevention, regenerative farming, and eye disease prediction, *our research aims to understand their mental models, practices, and challenges in working with data quality in the end-to-end AI life cycle*. Practitioners in our study struggled to define and measure the two most important desiderata of data excellence: *reliability* and *fidelity*. 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.

Type	Count
Domain	Health and wellness **(19)** (*e.g.,* maternal health, cancer diagnosis, mental health) Food availability and agriculture health **(10)** (*e.g.,* regenerative farming, crop illness) Environment and climate **(7)** (*e.g.,* solar energy, air pollution) Credit and finance **(7)** (*e.g.,* loans, insurance claims) Public safety **(4)** (*e.g.,* traffic violations, landslide detection) Wildlife conservation **(2)** (*e.g.,* poaching and ecosystem health) Aquaculture **(2)** (*e.g.,* marine life) Education **(1)** (*e.g.,* loans, insurance claims) Robotics **(1)** (*e.g.,* physical arm sorting) Fairness in ML **(1)** (*e.g.,* representativeness)

Table 1: Summary of participants' domains

Methodology

Between May and July 2020, we conducted semi-structured interviews with a total of **53** AI practitioners 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. 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.

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 (Miller and Stirling 2019), with 14 total interviews including Nigeria (10), Kenya (2), Uganda (1), and Ghana (1). We interviewed **45** male and **8** female AI practitioners.

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.

We recruited participants through a combination of developer communities, distribution lists, professional networks, and personal contacts, using snowball and purposive sampling (Palinkas et al. 2015) that was iterative until saturation. We conducted all interviews in English (preferred language of participants) using video conferencing. Each participant received a thank you gift in the form of a gift card. Interview notes were recorded in the form of field notes or video recordings, transcribed within 24 hours of each interview by the corresponding moderator.

Findings

A challenge which may limit reliability and fidelity in high-stakes domains is the need for domain expertise. AI practitioners were often 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 expertise. Answering these questions entailed an understanding of the application domain, social aspects, and embedding context (Taylor et al. 2015), (Vertesi and Dourish 2011). 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 make 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. Participants in our study saw a wide range of impacts— from wasted time and effort to downstream effects in deployment— which 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.

We first summarise characteristics of high-stakes AI initiatives that took part in our study. We then turn to a description of the challenges that practitioners faced in defining and measuring reliability and fidelity of their data.

Characteristics of AI high-stakes domains

AI is increasingly applied to needs that governments, nonprofits, and private industry have historically struggled to meet, such as human development, environmental sustainability, and wildlife conservation (Tomašev et al. 2020). Historically, these domains have been resource-constrained (Weber and Toyama 2010); the aspiration in applying AI to these domains is to make limited resources scale in areas like healthcare, education, and poverty alleviation. In our study, we interviewed AI practitioners working on projects ranging from cancer diagnosis, premature birth detection, insurance claim analysis, micro-loan provisioning, and more. Data practices in high-stakes AI are especially tenuous due to the nature of these domains:

Lack of existing data: High-stakes AI domains have a pronounced lack of readily available, high-quality datasets, due to various constraints of novelty, specificity, and complexity of application areas.

Well-rounded datasets: Datasets in high-stakes AI require not just volume, but also diversity, heterogeneity, and comprehensiveness. *e.g.,* credit assessment requires diverse subgroup data and feature-rich data to make assessments for different users.

Inter-disciplinarity: High-stakes AI is often at the combination of two or more disciplines; *e.g.,* AI and maternal health, requiring stakeholders across multiple organizations

to come together to define both the problem statement and the datasets required to build effective AI solutions.

Resource constraints: Many AI applications in social and economic domains are often situated in academia, non-profits, public sector and fledgling startups, characterised by typically fewer resources, including inability to source large and high-quality datasets.

Upstream and downstream: Most AI tools assume pipelines starting with readily available datasets and ending in inferences, whereas high-stakes AI fundamentally expands into both the upstream—data is almost always newly created, gathered or merged—and downstream—the models can have serious impacts on the general public.

Data Reliability

While reliability of the data should be a minimum requirement in developing and deploying AI in high-stakes domains due to the critical consequences— several practitioners in our study struggled to define what reliable data meant for their use case, and often had to prioritise time-to-market, revenue margins, and competitive differentiation over data reliability. Reliability required a form of *data robustness* which was difficult to control due to the structural practices within ML and the inherent nature of most high-stakes AI domains. We describe two ways in which data reliability was affected (or compromised): conflicting reward systems between AI practitioners and domain experts/data collectors, and subjectivity in decision-making for defining ground truth.

Imperfections in the data due to conflicting reward systems. As mentioned earlier, high-stakes domains lacked pre-existing datasets, so practitioners often had to collect data from scratch. Data collection and labelling efforts by field partners and domain experts were almost always extraneous tasks on top of their primary responsibilities, which is for example, to ensure the well-being of their patients. Most data collection was time- and resource-intensive. ML dataset collection practices were reported to conflict with existing workflows and practices of domain experts and data collectors. We observed a larger issue of limited data literacy and information symmetry with field partners, especially at the frontlines. For many teams, the awareness of poor data quality only came after months of making progress with model training iterations or when they had deployed their system for use in the real world. P41, a researcher working on agriculture in East Africa described their experience of working with vehicle operators, who carried out data entry for P41's AI system in addition to their own internal systems. "*They were good in their system, but not in this one. They just did the bare minimum to enrich the data. [...] It really slowed down the project. 2 months later we discovered that the survey tables weren't as reliable as they should've been.*".

Understanding and handling subjectivity in ground truth. High-stakes AI requires specialised knowledge, subjective decision-making in defining the ground truth and breadth and number of labels necessary (Aroyo and Welty 2015). Examples of ground truth decisions are detecting cancer in pathology images, identifying quality of agriculture produce, and analysing insurance claims for acceptance or rejection. In each of these areas, decision-making is influenced by factors including, but not limited to, the decision maker's expertise, educational background, years of experience, opinions, biases, and threshold of caution (*e.g.,* have their past experiences moulded them to become stringent, and err on the side of caution to have a false positive rather than false negative in detecting health issues?). The reliability of the data was affected 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.,* (Aroyo et al. 2019). With no direct indicators of subjective shortcomings in data, issues of reliability were discovered through 'manual reviews' of data with clients or field partners, and often, through downstream impacts. In some cases, ground truth was highly inaccurate but deeply embedded into systems, as in the case of P6, running credit and insurance assessment, who referred to how decisions taken by insurance companies in the past were wrong 10-15% of the time, but there was no way to correct historical archives.

Data Fidelity

Practitioners had to make assumptions about their data (*e.g.,* meanings of different features, if the data is representative) due to insufficient application-domain expertise, and inadequate cross-organisational documentation. We now describe these two ways in which data fidelity was affected in high-stakes domains.

Insufficient 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. Non-representative data from poor application-domain expertise resulted in model performance issues, resulting in re-doing 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. Issues due to a lack of representative data 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?"*.

Inadequate cross-organisational documentation. Prac-

titioners discussed several instances where collected and inherited datasets lacked critical details due a lack of documentation across various cross-organisational relations (within the organisation, with field partner organisations and data collectors, and with external sources). Missing metadata led practitioners to make assumptions, ultimately leading to costly discarding of datasets or re-collecting data. 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 off impacts such as 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. 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.*"

Thoughtful Data Practices

In this section, we discuss thoughtful data practices that practitioners or their teams can adopt to improve data reliability and fidelity, and avoid data disasters. We acknowledge that these may not be an exhaustive list, but represents what was reported to us in exceptional cases. We would also repeatedly emphasise the importance of allocating resources to engage with and consulting domain experts early on, but also maintaining sustained engagements throughout the project lifecycle as a step towards data excellence.

Well-designed incentives for data collectors. 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. In a few cases where incentives were explicitly discussed as being provided, high monetary incentives sometimes led to over-sampling, skewing the data. Key considerations and questions to think about when designing incentives for data collectors:

- Incentive amount: Consult with your domain expert on what an appropriate incentive would be for people collecting the data.
- Time to Data: Document and communicate expectations around how long and how many data entries you want people to collect to avoid rushed, error prone collection.
- Speaking up: What incentives are there for collectors who speak up about discrepancies or interesting insights beyond the scope of the task?

Training for data collectors. 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. 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. 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.) Key considerations and questions to think about when training data collectors or raters:

- Onboarding: Develop and conduct a data collection training session to address questions and concerns your collectors have about the process.
- Instructions: How does the wording of your collection/labelling instructions affect the data? For example, asking if a topic could be used to describe an article instead of if it's the best topic for the article.
- Errors: Is there a high risk that data will be collected or labelled incorrectly due to issues like boredom, repetition, or lack of appropriate tools?
- Review: Meet with your collectors periodically throughout the process to review any challenges they are encountering collecting the data and the quality of the data.

Improved planning and documentation. In a few cases where effects due to incomplete or a lack of metadata 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.

Key considerations and questions to think about when documenting your dataset (for others see (Gebru et al. 2018)):

- Milestones: Has your team documented the success metrics that will determine the end of the data collection effort?
- Over communicate: Does your team have scheduled check-ins for monitoring collected/labelled data?
- Artefacts: Does your team have a shared document where data decisions are regularly reviewed and updated?

Conclusion

As AI becomes part and parcel of decision-making of core aspects of our planetary existence, the sanctity and quality of data powering these models takes on high importance. Through a qualitative study with 53 AI practitioners in India, East and West African countries, and the US, we shed light on the data practices and challenges of working on cutting-edge, high-stakes domains of health, wildlife conservation, food systems, road safety, credit, and environment. We present a set of (rare) thoughtful data practices that we saw in our study and some key considerations for practitioners as they adopt those practices to improve the reliability and fidelity of their data.

References

Aroyo, L., and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36(1):15–24.

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 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.

Chu, X.; Ilyas, I. F.; Krishnan, S.; and Wang, J. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, 2201–2206.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Halevy, A.; Norvig, P.; and Pereira, F. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24(2):8–12.

Joppe, M. 2000. The research process. retrieved february 25, 1998.

Kaggle. 2019. 2019 kaggle ml & ds survey. https://www.kaggle.com/c/kaggle-survey-2019. (Accessed on 08/27/2020).

Krippendorff, K. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement* 30(1):61–70.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Miller, H., and Stirling, R. 2019. Government ai readiness index 2019 — oxford insights — oxford insights. https://www.oxfordinsights.com/ai-readiness2019. (Accessed on 09/14/2020).

Nagle, T.; Redman, C. T.; and Sammon, D. 2017. Only 3% of companies' data meets basic quality standards. https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards. (Accessed on 08/27/2020).

Palinkas, L. A.; Horwitz, S. M.; Green, C. A.; Wisdom, J. P.; Duan, N.; and Hoagwood, K. 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):533–544.

Paritosh, P. 2012. Human computation must be reproducible.

Taylor, A. S.; Lindley, S.; Regan, T.; Sweeney, D.; Vlachokyriakos, V.; Grainger, L.; and Lingel, J. 2015. Data-in-place: Thinking through the relations between data and community. CHI '15. New York, NY, USA: Association for Computing Machinery.

Tomašev, N.; Cornebise, J.; Hutter, F.; Mohamed, S.; Picciariello, A.; Connelly, B.; Belgrave, D. C.; Ezer, D.; van der Haert, F. C.; Mugisha, F.; et al. 2020. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications* 11(1):1–6.

Vertesi, J., and Dourish, P. 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.

Weber, J. S., and Toyama, K. 2010. Remembering the past for meaningful ai-d. In *2010 AAAI Spring Symposium Series*.