Human-Centricity as an Enabler of High Data Quality for AI

by Nadia Morozova

Image Credit | Cavad

Human-centricity provides business leaders an easy way to ensure high data quality of their AI tools and solutions.

☑ **INSIGHT** | FRONTIER   26 Jun 2025

Understanding the Context

Companies continue to face challenges with implementation of AI strategies in their organizations. As research shows, these challenges usually spread across technological, organizational, and cultural domains. Many of the technological challenges "... can be traced back to data: data not being available, data not being structured for the desired use, and data definitions not being unified across the company."[4]. This paper is based on an interpretive three-month case study of AI strategy development and deployment in one of the world's largest professional services organizations and 6 unstructured expert interviews with the industry-leading professionals working in the field of AI with backgrounds in pharmaceuticals, financial services, and retail. Thematic analysis was conducted to identify the key themes and patterns in the collected data. The main themes were developed around underlying success factors for data-driven organizational culture change and deployment of AI-enabled tools and solutions.

RELATED CMR ARTICLES

Rebecka C. Ångström, Michael Björn, Linus Dahlander, Magnus Mähring, and Martin W. Wallin, "**Getting AI Implementation Right: Insights From a Global Survey**," California Management Review 66/1 (2023): 5–22.

The author's observations of the conversations happening in various C-Suite leadership rooms show that the leadership teams still consider AI as a "black box" that does some magic. Data quality issues get some attention only when something goes wrong. Is this the right approach? On one hand, this approach allows business professionals to save time and resources on in-depth conversations about the tools and how they were built. And let's acknowledge, quite often all these long calls don't go anywhere! On another hand, this lean-in approach, willingness to get seemingly low hanging fruit can lead to very tangible negative business outcomes. We don't talk only about reputational risks, we are talking

about a substantial decrease in customer retention rates, higher employee attrition rates and broken logistical and supply chain networks. Do these outcomes seem like a fair risk to take?

Asking the Right Question

How to ensure then that AI brings value and minimizes risk? I'm sure you've heard data leaders in one of your leadership meetings passionately saying 'garbage in - garbage out'. Yes, their point may be valid, but it's not always easy for you to differentiate between what is good, what is bad, and what is good enough. After all, given all the various pressures these days, we are having to deal more and more with data that is just good enough. **One of the most straightforward approaches to take in considering data quality is to ask yourself the question: 'Does the data used for building this AI solution represent reality?'.** Let's take an easy example. I'm sure you experienced in your career a moment or heard from your peers how your marketing or consumer insights department were saying something like '80% of consumers agree that they are willing to buy the product in the next 6 months'. Sounds very promising! So, full of enthusiasm you decide to invest in your next product line or a new product launch, which ends up quickly failing. "Bad data" - you would say, "bad question" - they would answer[7]. The question definitely has a lot of limitations, but had you asked your teams who those '80% of consumers' were and how many consumers were asked this question?

Applying the Framework

This situation represents the first building block in ensuring that you are building your machine learning algorithms with data that represents reality - **accurate data sampling** (refer to Figure 1).

Human-centricity as an enabler of high data quality for AI

Represents current status
via accurate data sampling

Envisions change
via being future-proof

Gives different perspectives
via data connectivity

Fills in missing gaps
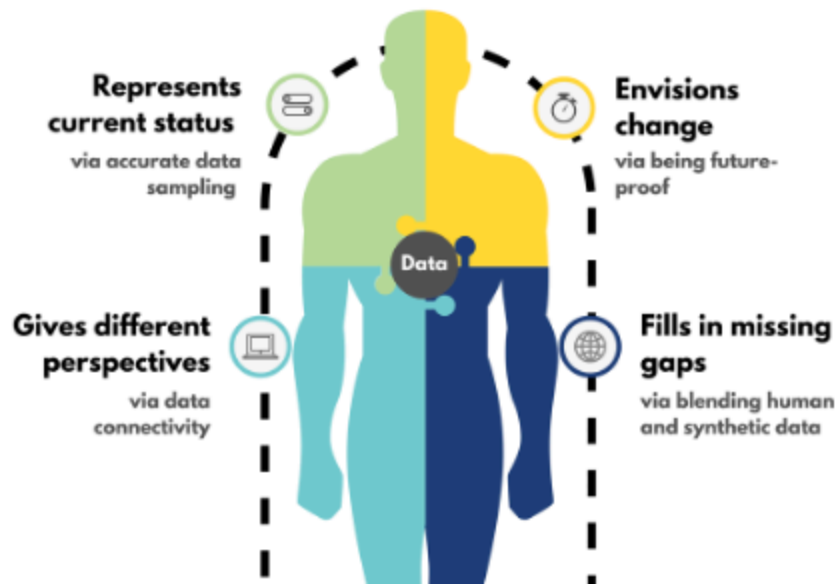via blending human and synthetic data

Data

Figure 1. How human-centricity enables high data quality for AI

In a nutshell, it means that you have a sufficient amount of data that represents your population[6]. For instance, in the example of that willingness to buy question, your answer will be accurate only if your sample represents the reality of your market in terms of socio-demographics, behavior and attitudes. Let's say, your customers mostly buy your products in small kiosks, and you ask the question of consumers in an on-line store, obviously your sample might quickly become biased. This is an example of selection bias. The same happens with models if you have data only from a fraction of your customers. It won't be able to predict behavior for the whole population of your clients, let alone prospects. Therefore, **as a data or business leader you need to ensure that you avoid biases in your data.** The most critical ones are selection bias, outlier bias, survivorship bias and omitted variable bias.

This brings us to the second important building block in ensuring that machine learning algorithms represent reality - **being future proof**. Let's come back to our example of consumer willingness to buy. Today, your consumers are very interested in your new proposition, they like the product design, new experience it provides and although it might cost a bit more, consumers see a clear value for money. Tomorrow, several companies announce a series of layoffs. This totally changes the equation, and your consumers will

not be open anymore to pay for new design features. Your survey data quickly becomes obsolete. The same happens with the large machine learning algorithms, which are largely reliant on the past data. They stop representing reality very quickly. As a result, **data and business leaders need to sufficiently broaden the scope of variables, including the ones which might help to predict changes in the overall landscape, e.g. economic situation; find the ways of incorporating real-time data in their models.**

The third important component of ensuring that your machine learning algorithms which build the foundation of your AI solutions are representing reality is **ensuring data connectivity**. The first step is to develop a framework which will include all aspects of consumer behavior as we are thinking human-first and map them to the sources of data you have for each of these behavior indicators. In our example of the new product line launch we might want to collect data on what products our customers are buying, in what stores, in what occasions, with which mindset and of course what marketing touchpoints they were exposed to during their journey. **All this data will be coming from very different sources, and it might take your organization a lot of time to identify ways to connect these data sources together, taking only data you need following the data minimization principle and ensuring that the datasets are free of noise and unstructured and unconnected data points**.

One of the most common challenges, which quite often is disregarded, is how to best define your measures and metrics[4]. For example, conversion from your site might mean a very different thing from conversion from a digital platform like Instagram. This will eventually help you to avoid future AI hallucination, which means that AI algorithms produce incorrect or misleading results. Therefore, **data and business leaders should constantly review and supervise their models.**

The final important area which allows companies to ensure that their machine learning algorithms are representing reality is a **responsible blending of human and synthetic data**. As we have seen so far, models which are used to develop AI solutions require a lot of clean, transparent and reliable data, which is definitely difficult to obtain, especially under strong time and resource pressure. Therefore, nowadays it becomes critical to find the right balance between human and synthetic data. In a nutshell, synthetic data is artificially generated data and according to predictions made by Gartner in 2023, by the end of 2024,

60% of data for AI would be synthetic. For context, only 1% of data in 2021 was synthetic[2]. Although synthetic data has very clear benefits like flexibility and scalability, it requires substantial resources for data verification, as it is prone to biases, especially an outlier bias and it has strong reliance on human data, which as we discussed might get obsolete very quickly[1]. Coming back to our previous example on launching a new product line, synthetic data could help to provide missing information on competitor sales and their market share growth. However, that data might come from a quarter with low seasonal sales. This can skew the overall predictions of competitor performance and can underestimate their customer loyalty. Thus, **as a data or business leader consider the overall cost of AI and especially the cost of production of reliable synthetic data. Start gradually with the continuous focus on data accuracy and manage biases and risks.**

Concluding Remarks

As the examples have shown, it is really difficult to make sure that AI is fully representing reality, and especially at the current stage of AI development. However, with the rapid development of AI humans will continue being interested in other humans, which will eventually lead to the situation that we'll need to employ a human-centric approach in leveraging and understanding the data behind "the black box" of machine learning algorithms. As humans, we also won't stop valuing other humans' attention and time, which provides us the on-going opportunity to leverage both big and small data[3]. It's important to remember that AI can make the same mistakes as humans make (i.e. both machine learning algorithms and humans are biased), and we should remember that according to McKinsey (2021) survey only 15% of machine learning projects succeed[5].

References

1. Altay Ataman, "**Synthetic Data Vs Real Data: Benefits, Challenges in 2025,**" AIMultiple, (2025).
2. Gartner Press Release, "**Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning**," (2023).

3. Martin Lindstrom, "Small Data: The Tiny Clues That Uncover Huge Trends," St. Martin's Press, 2016.

4. Rebecka C. Ångström, Michael Björn, Linus Dahlander, Magnus Mähring, and Martin W. Wallin, "**Getting AI Implementation Right: Insights From a Global Survey**," California Management Review 66/1 (2023): 5–22.

5. Rohit Panikkar, Tamim Saleh, Maxime Szybowski, and Rob Whiteman, "**Operationalizing Machine Learning in Processes,**" McKinsey & Company, (2021).

6. Roger W. Hoerl and Thomas C. Redman, "**What Managers Should Ask About AI Models and Data Sets**," MIT Sloan Management Review, (2023).

7. Yash Raj Shrestha, Shiko M. Ben-Menahem, and Georg Von Krogh, "**Organizational Decision-Making Structures in the Age of Artificial Intelligence**," California Management Review 61/4 (2019): 66–83.

Nadia Morozova (Follow)

Dr. Nadia Morozova is Chief Analytics & Insights Officer at Enriched Insights and Strategic Industry Advisory Board Member at Warwick Manufacturing Group – University of Warwick (United Kingdom). Her research is focused on data-driven organizational culture change and consumer neuroscience.