

Everything You Need to Know About AI Data

Brought to you by Assemble You.

Emma was a project manager overseeing the implementation of AI-driven customer service tools.

At first, everything seemed to be running smoothly. The AI was efficient, processing customer queries faster than ever.

But soon, Emma began noticing issues.

The AI was making biased decisions, denying refunds for female customers at a much higher rate than male customers. Plus, a data breach revealed that the AI had been collecting more personal data than the company had disclosed, putting the company's reputation at risk.

Thinking back to a recent article she'd read about AI data, Emma took immediate action. She began a full audit of the AI system, uncovering that the data feeding the AI was not only of poor quality but also lacked transparency in terms of how it was collected.

Emma worked with her team to implement clear data usage policies and upgraded their data management strategies.

They ensured that only high-quality, ethically sourced data was used, and they communicated these changes transparently to their customers.

As a result, the AI system became more accurate and fair in its decisions, and the company regained the trust of its customers.

Emma's proactive approach not only resolved the immediate problems but also positioned her as a leader in ethical AI practices within her industry.

Great move for Emma!

Welcome to today's lesson on AI Data.

We're diving into a topic that's not just for tech enthusiasts. It's becoming increasingly relevant for everyone, regardless of the sector you work in.

By the end of this lesson, you'll have a solid understanding of how AI gathers and uses data, where this data comes from, and why transparency in data handling is crucial.

We'll also explore practical tips to navigate data concerns in your organisation.

So, why is AI data so important?

According to a study published by McKinsey & Company, 2023 was the breakout year for generative AI, and it's largely due to the data fuelling the models [1]. The more we understand about AI data, the better equipped we'll be to harness its power while managing its risks.

So, how do AI systems gather data?

Data is sort of like the fuel that powers AI engines. But unlike cars, AI doesn't run on petrol; it runs on information.

This information is collected from a variety of sources – everything from publicly available data on the internet, like social media posts and blogs, to more structured datasets like those found in databases and spreadsheets.

But where exactly is this data from? Well, it's everywhere.

According to science journalist Lauren Leffer, **your** personal information could be part of the dataset used to train AI models [2].

Companies often scrape the web or purchase datasets that include everything from your online shopping habits to your social media interactions. And this is more than just what you explicitly post online – AI can also infer details about you based on your behaviour, such as the time you spend on a website or the type of content you like.

Now, let's discuss how AI collects this data.

In most cases, AI systems are designed to automatically gather and process information. This is often done using algorithms that can sort through massive amounts of data quickly and efficiently.

These algorithms are programmed to search for specific types of data or look for patterns in the data they find. For example, an AI designed to recommend products might look for patterns in your past purchases to suggest new items you might like.

But this leads us to some concerns. One major issue is the lack of transparency around data gathering.

According to Vivek Jetley, EVP and Global Head of Analytics at EXL, many organisations fail to disclose what data they're collecting and how they're using it [3]. This lack of transparency can lead to many problems, including data breaches and misuse of personal information. Companies must be open about their data practices to comply with regulations and build trust with their customers.

Transparency isn't just a buzzword; it's a critical aspect of ethical AI development. Without it, we risk eroding public trust in AI technologies.

So, what can we do to ensure transparency?

One approach is to implement clear data usage policies that are easily accessible and understandable. Another is to regularly audit your AI systems to ensure they're using data responsibly.

Now, let's talk about how data is used and processed.

Once data is collected, it's not just put away for safe keeping – it's processed, analysed, and used to train AI models. This involves feeding the data into machine learning algorithms, which then use it to make predictions, generate content, or even automate decisions.

For instance, AI might use data to predict customer behaviour, generate marketing content, or optimise logistics in a supply chain.

The quality of the data used in AI is paramount [4].

Poor-quality data leads to poor-quality AI models, which can result in incorrect predictions or biased outcomes. Making sure your data is accurate, up-to-date, and relevant is crucial to the success of any AI project.

Lastly, where does AI pull information from, and what does it do with the data it gathers?

AI pulls information from its training data, which can include millions of pieces of information. Brian Eastwood at MIT Sloan reports that AI systems are increasingly being used in business decisions, making it more important than ever to understand the data behind these systems [5].

AI doesn't just store this data; it learns from it.

The more data an AI system has, the better it can become at making predictions or generating content. But this isn't without its challenges because the more data an AI system has, the more complex it becomes.

Managing this complexity requires robust data management strategies, which include not just gathering and storing data but also ensuring its accuracy and relevance.

We've covered a lot of ground in understanding AI data today. From how data is gathered and where it's sourced, to how it's processed and used in AI systems, it's clear that data is the backbone of AI.

But take it from Spiderman – with great power, comes great responsibility.

Transparency in data practices, ensuring high-quality data, and being aware of data concerns are all incredibly important for anyone working with AI. The success of AI, especially generative AI, is all dependent on the data it's fed [1].

Ensuring this data is of high quality and ethically sourced is key to unlocking AI's full potential.

Here's your takeaway for today: start by assessing the data practices in your own organisation.

Are you transparent about how you gather and use data?

Is the data you're using of high quality?

Take steps to audit your data sources and ensure that your AI systems are not just effective but also ethical.

Taking a proactive approach will help you stay ahead in the ever-evolving world of AI.

[1] *The state of AI in 2023: Generative AI's breakout year*, McKinsey and Company (2023)

<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>

[2] *Your Personal Information Is Probably Being Used to Train Generative AI Models*, Lauren Leffer, Scientific American (2023)

<https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/>

[3] *AI Reality Check: Why Data Is The Key To Breaking The Hype Cycle*, Vivek Jetley, Forbes (2024)

<https://www.forbes.com/councils/forbestechcouncil/2024/07/01/ai-reality-check-why-data-is-the-key-to-breaking-the-hype-cycle/>

[4] *Ensure High-Quality Data Powers Your AI*, Thomas C. Redman, HBR (2024)

<https://hbr.org/2024/08/ensure-high-quality-data-powers-your-ai>

[5] *Artificial Intelligence is Now Everyone's Business*, Brian Eastwood, MIT Sloan School of Management (2024)

<https://mitsloan.mit.edu/ideas-made-to-matter/artificial-intelligence-now-everyones-business>