What do we mean by “without data, there is no AI”?

A short paper by Ben Snaith - Senior Researcher, the ODI
21st December 2023

Open Data Institute

Contents

About

This short paper was researched and produced by the Open Data Institute and published in December 2023. Its author is Ben Snaith. If you want to share feedback by email or would like to get in touch, contact ben.snaith@theodi.org

What do we mean by ‘without data, there is no AI’?

This short paper shows how focusing on the data can help us understand AI better and build a healthier AI ecosystem.

Introduction

"Without data, there is no AI" has become a mantra for the ODI, prominently featured in the framing of the ODI Summit event in November 2023. However, it's a concept that has long circulated within the realm of Artificial Intelligence (AI). For us, it refers to the data infrastructure of AI – including data assets, tools, standards, practices, and communities. It is a call to look at data and other socio-technical foundations of AI to better understand their design, outcomes and implications.

This short paper is designed to unpack this phrase and answer the related question of how and why the ODI research team will be investigating data-centric AI in a new programme of work.

Data and the AI lifecycle

If we look at an AI lifecycle – even in the abstracted diagram below – many parts centre on the data! Data is foundational to AI models. It provides the information that a machine learning model is trained on and learns from. It is collected, wrangled, curated, aggregated and then used in the model. Data is used to test and benchmark the model's success. And data is inputted for utilisation once the model is operational.

Building an AI system typically involves sourcing large amounts of data and creating data sets for training, testing, validating, and deploying. This process is iterative in that it may require several rounds of training, testing and evaluation until the desired outcome is achieved and data plays an important role at each step.

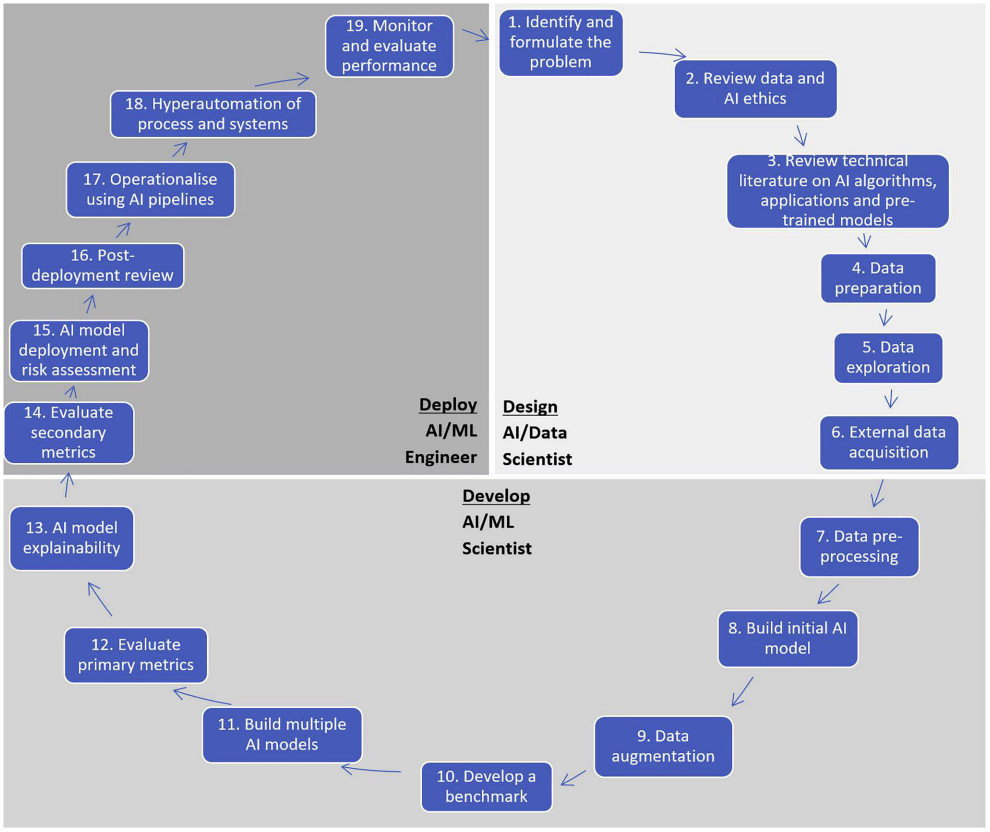

Figure 1. An artificial intelligence life cycle: From conception to production - ScienceDirect

The datasets used in AI development are increasingly huge in scale – 'We have now entered the era of trillion parameter machine learning models trained on billion-sized datasets scraped from the internet.' This has necessitated a rise of companies such as snorkel.ai and data.world to make the data more useful for training AI models; reducing the need for large training datasets, and adding context and explanations.

Where does this data come from?

This depends on what type of AI we are talking about, but we will focus on the latest generation of foundational models due to the centrality of their massive data needs where typically data is crawled, obtained as enterprise data, or a mixture of both. This is an evolving landscape, as data transparency is not even close to being a default within AI usage. A Washington Post investigation concluded that 'many companies do not document the contents of their training data - even internally'. Transparency – in terms of both model design and data – is a recognised challenge. The training of AI models 'depends on the availability of public, scrapable data that leverages the collective intelligence of humanity, including the painstakingly edited Wikipedia, millennia's worth of books,

billions of Reddit comments, hundreds of terabytes' worth of images, and more'. Many different types of datasets feed into AI models, for example:

- **Textual data –** The CommonCrawl dataset is often used as a source for training large language models. The organisation has been regularly crawling the worldwide web and generating archival snapshots. It is an open-source database of over 250 billion pages collected over 16 years. The sheer scale of this dataset has an enduring allure in the AI community and has been used as a seeding dataset in training pipelines of high-profile projects such as GPT-3.

- **Visual data** – you may have seen text-to-image generation tools like Midjourney, DALL-E and Stable Diffusion. Stable Diffusion was trained on five billion human-generated images that were scraped from the internet without the consent of the original creator. LAION created the datasets used by Stable Diffusion by using Common Crawl to identify images of sufficient resolution and alt-text details so that the algorithm could be accurately trained. LAION has today - 20th December 2023 - pulled their open-source datasets following investigations by Stanford researchers who identified over 3000 instances of child sexual abuse material.

- **Synthetic data** - not all data used is 'real' data. For various reasons – including ease of collection and privacy concerns – computer-created data or 'synthetic data' is being used. For example, relying on historical data may prevent an algorithm from being useful within new contexts, like when new diseases are detected. Therefore, in medicine and healthcare, accurate synthetic data can be used to increase diversity in datasets and increase the robustness and adaptability of AI models.

- **Additional types of data** – we have only scratched the surface here. For example, we haven't covered training, evaluation and test data; nor human feedback during model training; nor data labelling typically used for fine-tuning and safety testing; or knowledge graphs.

Issues of access and use of data

To get to the point where they are useful for a machine-learning model, datasets used to require labelling, which allows the model to learn – for example, 'select all the squares with traffic lights'. Increasingly foundational models are 'unsupervised', meaning labelling is no longer in play as they self-learn.

There are also concerns that costs are becoming prohibitive as publishers lock down datasets further and require larger sums for access, as well as the increasingly prohibitive costs of the processors needed to run the models. Access to data is being explored within a different ODI research programme.

The scale and complexity of the use of data within artificial intelligence, combined with obfuscation - i.e. so-called black-box algorithms - can make AI unknowable. Researchers are concerned that 'we don't really know what they're doing' in any deep sense. 'If we open up ChatGPT or a system like it and look inside, you just see millions of numbers flipping around a few hundred times a second, and we just have no idea what any of it means.' Therefore, explainability and openness are vital in order to connect the investigation of data with the investigation of models.

With the scale of datasets used, there is a concern about potential 'model collapse' where AI models are trained on synthetic data rather than human-generated data and therefore become divorced from 'real' data and 'real' events. To prevent this, datasets will require better labelling, such as data nutrition labels.

Towards understanding, accountability and better AI

A directional shift towards the study of AI datasets leaves us with work to do - to help bridge this gap in understanding and prevent further obfuscation of data and impacts:

1. **Investigate data** - in harmony with other considerations like better and more safe model design, focusing on the data - and the models - opens up the opportunity to analyse the data sources, spot and test for bias, and identify data quality or collection issues.

    - Researchers, auditors, regulators, policymakers, and other AI stakeholders can start to analyse and study these datasets, leading to a better understanding of their capabilities, limitations, risks, and any harm they may cause or exacerbate.

    - We can look at data gaps in machine learning which affect the quality of datasets. We can query the volume of data actually needed – which Andrew Ng suggests is far less than companies might think – and consider archiving or deleting less useful data for the protection of privacy and to reduce the environmental impact of data stewardship

2. **Make data AI-ready –** whether in response to issues highlighted or as a systematic attempt to prevent AI harm – steps can be taken to ensure data is ready for application in AI systems.

    - Improving the quality and contextual applicability of datasets, making data collection more ethical, stopping collecting data, utilising differential privacy, making use of synthetic data, utilising transparency approaches such as data cards and expanding responsible AI practices across the ecosystem.

    - Improving the quality/addressing poor datasets, including the retraction of the MS Celeb and TinyImages datasets, blurring of the images of people and filtering out of constituent images to create a sanitised version of the original dataset.

    - Existing technical tools such as HuggingFace's data measurements tool, Snorkel.AI and Google's Know Your Data (KYD) tool are already being used within the sector. However, there is no guarantee that the users of these tools are also literate in responsible AI.

    - Such steps must be embedded within governance and technical processes to ensure that AI users see data curation, data improvement, and retraining of the AI system on updated data as part of an ongoing cycle.

3. **Set frameworks and benchmarks for AI safety.**

    - If existing benchmarks have been found to be inaccurate or harmful, they need to be improved. Yet this isn't happening. For example, ImageNet has been shown to 'contain consequential biases' but is still considered a 'research standard.' Critiques of the dataset culture itself focus on the overemphasis on benchmarking to the exclusion of other evaluation practices, legal and ethical issues in data management, distribution, reuse, and labour practices in data curation.

    - There may also be a need to have public audits of datasets or mandatory reports of training data sources as included in the EU's AI Act.

4. If risks are too big or uncontrollable, **consider whether we should stop developing AI.**

- There is a brewing - and controversial - fear of Artificial General Intelligence (AGI) being developed, which could outpace current safety requirements. There are some calls for non-proliferation pacts for AI. Chatham House called for an assurance model similar to the US Food and Drug Administration (FDA), which would consist of a scaled launching model alongside robust auditing requirements and comprehensive risk assessments to evaluate both the direct and indirect implications of the product in question.

5. Across these issues, there is a strong push to **enhance governance models.**

    - The newly agreed EU AI Act demonstrated an appetite to tackle the complex regulatory and governance questions of AI. But the narratives surrounding the attempt show it is not an easy field to get right; campaigners have criticised the act for stepping back from an earlier commitment to ban facial recognition technologies.

    - In the US, the Federal Trade Commission (FTC) ordered OpenAI to document all data sources used to train its large language models. A group of the world's largest media organisations published an open letter urging lawmakers to require transparency of training datasets through new legislation.

    - The ODI previously reviewed the Bletchley Declaration and the US Executive Order. There is talk of an 'uncertain path ahead' as states try and negotiate the future that benefits their interests.

6. **There needs to be accountability mechanisms for harm suffered due to bad data.** Accountability is often considered a cornerstone of the safe development of AI and is included in the OECD's AI principles: 'Actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.'

7. All of the focus - so far - has been on improving the data within current development models. There also needs to be opportunities to learn from these investigations to ***change*** **how the development takes place.**

    - For example, some may accept that data can never have all of the answers and so develop models that combine data with other forms of knowledge input, for example, through

user-centred AI that utilises UX design principles to change how AI is developed and tested

Conclusion

This is a sprawling space where it can be difficult to unpack language, technologies and intentions. The ODI will be utilising its new Data-centric AI programme to explore AI data infrastructure as a means of creating a safer and more responsible AI ecosystem. Many recent regulatory announcements omit transparency altogether but focus on national security and the harms of frontier models. A focus on data infrastructure is missing within these global dynamics; hence the ODI will be breaking our work down into three levels: **Make data AI-ready, Make AI data accessible and usable** and **Make AI systems use data responsibly**.

This is the first in a series of pieces that will originate from the ODI's new Data-centric AI programme. In 2024, we will dive deeper into AI Incident databases, the AI lifecycle and other topics.