



THE UNIVERSITY OF
MELBOURNE

Centre for
Artificial
Intelligence
and Digital
Ethics

CAIDE Law 2024

Demystifying AI,
Law and Regulation

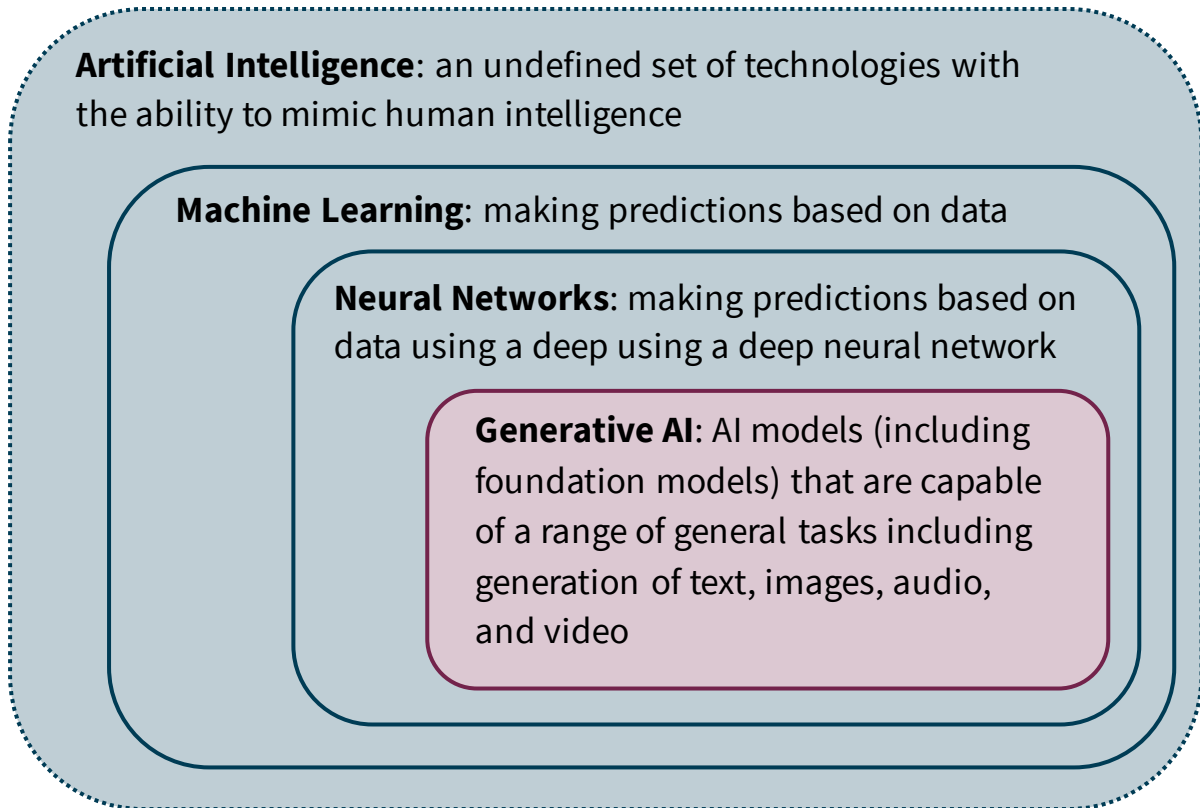
This project has been made possible by Microsoft, Atlassian and the Tech Council of Australia.



Demystifying Generative AI

1. What is Generative AI?

1.1 Generative AI within AI technologies:



‘Artificial intelligence’ refers to a group of AI technologies that have existed since at least the 1970s. Many of these older types of AI can be considered ‘**predictive AI**’, in that they are able to perform some functions mimicking human intelligence by analysing patterns in existing data to predict future patterns or outcomes. For example:

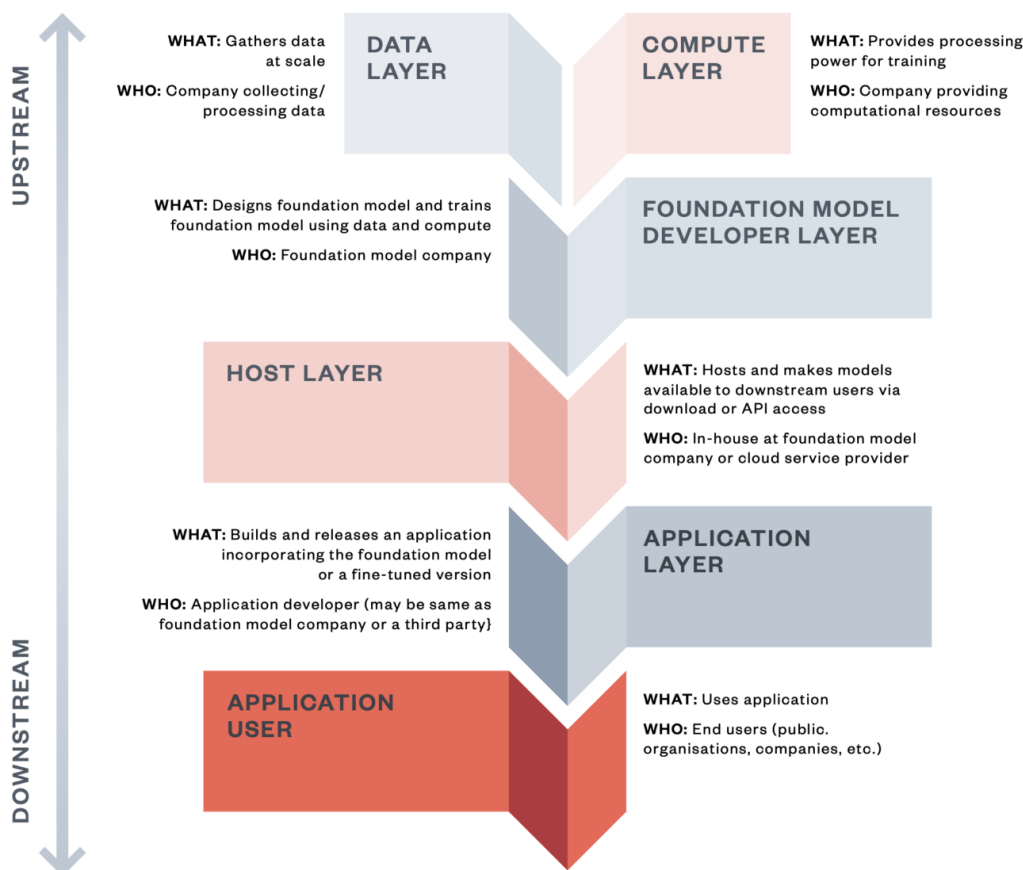
1. **Machine learning** can be used for forecasting, such as predicting stock prices, weather, or customer behaviour.
2. **Recommendation systems** can be used by companies like Netflix and Amazon to predict and suggest content or products users might like based on their past behaviour.

In contrast, **generative AI** refers to a smaller subset of foundation models trained on large datasets that are capable of creating new or novel content based on an understanding of relationships between different elements in the training dataset.

1.2 The generative AI supply chain:

The generative AI supply chain can vary significantly across different foundation models but there are several common features. These include:

1. Datacentre infrastructure that includes advanced Graphics Processing Units (GPUs) with high bandwidth network connections that **enable data gathering and computing power** at scale.
2. Leveraging datacentre infrastructure, developers and research scientists **train and develop foundation models** (e.g. GPT-4, Midjourney, Claude) for downstream access.
3. This downstream access can be provided at the **host layer** via direct download by the model developer or via Application Program Interfaces (APIs) by cloud service providers.
4. APIs provide a way for **applications** (e.g. ChatGPT, Bing Copilot, Bard) to access and use the output from foundation models to deliver services to individuals and organisations.



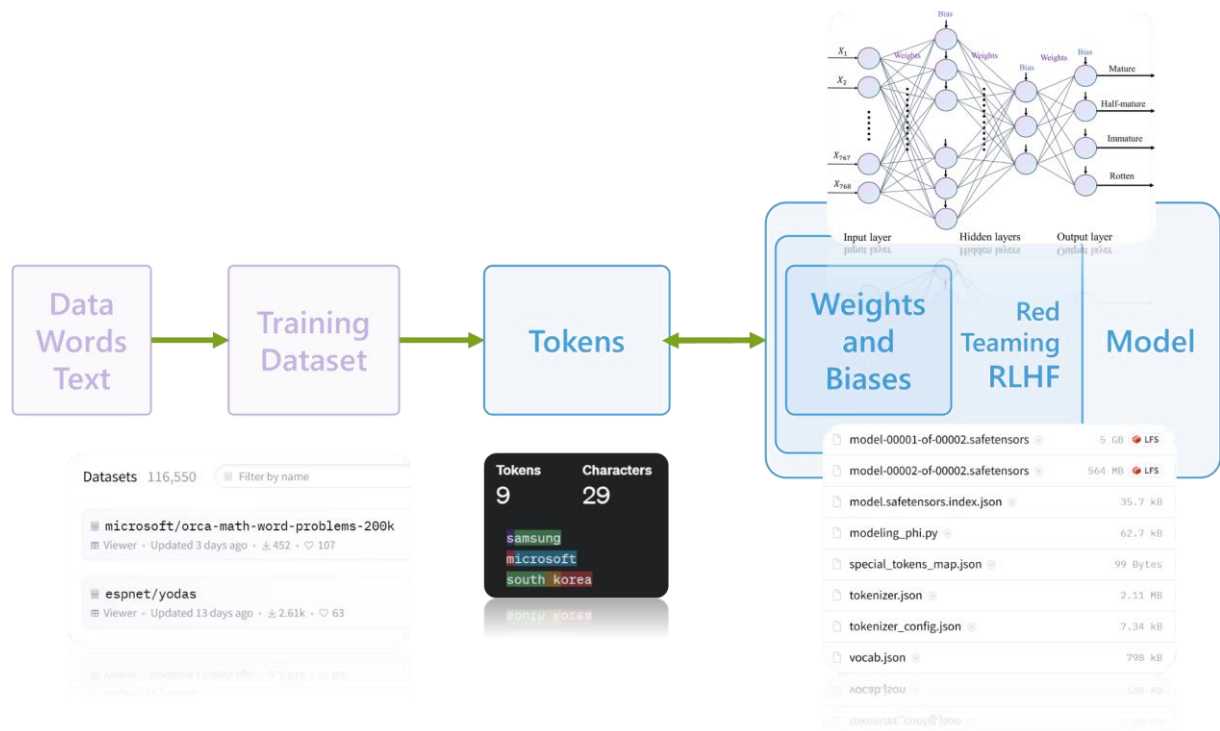
Note: This is one possible model (there will not always be a separate or single company at each layer)

2. How does Generative AI work?

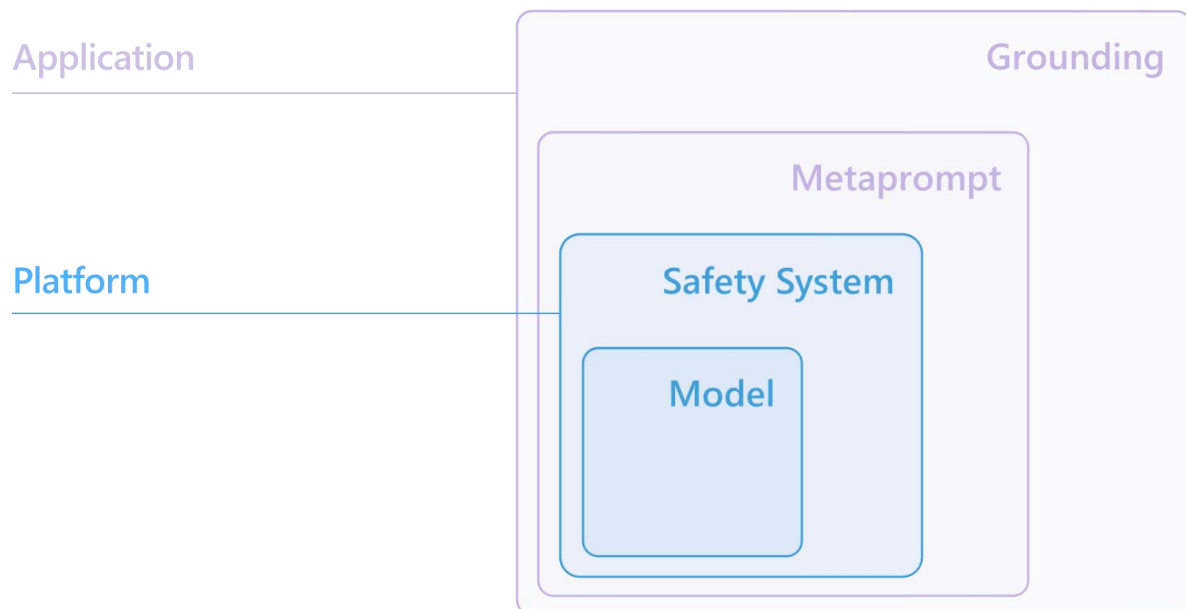
2.1 How are foundation models trained?

The training process involves **tokenisation** of the training dataset, which is a process by which training data is converted into 'tokens'. These tokens are then used in the training process to develop a complex mathematical relationship matrix between them. During this training phase, the algorithm is learning to understand how different words are related to each other. The core way in which transformer-based models work is that they are capable of tracking and weighing up millions of attention data points to build a foundation model that has a very deep and nuanced understanding of the way in which we communicate.

The foundation model's **'knowledge'** is then stored as mathematical data (i.e. numbers) that represent the weights and biases the model learned about tokens during the training. The weights are the importance the AI model will give to each token, and the biases are the numbers that influence decisions as to which token to display next.



2.2 What are the safety mitigations?



There are tools and capabilities at both the application and platform layer that allows model and application developers to adapt and shape the function of a model so the model or application is able to respond more accurately or safely when dealing with user prompts that might elicit a regurgitation or hallucination.

These safety mitigations can significantly reduce the risk of a foundation model producing hallucinations or regurgitations (but cannot fully eliminate this risk). These include, for example:

1. **Red teaming:** structured testing to simulate how adversaries might prompt the system to find flaws and vulnerabilities in the system.
2. **Content filtering:** filtering sensitive data from the training dataset.
3. **Meta-prompting:** giving additional instructions to a model to guide its behaviour, e.g. 'communicate in the user's language of choice'.
4. **Retrieval-augmented generation (RAG):** a process that allows information from an additional data source external to the training dataset to be included in the prompt to enable more informed and grounded outputs.

AI Terms Explained

Types of 'AI'

- **Artificial Intelligence (or AI):** An undefined term used to refer to technologies of varying complexities (including, e.g., machine learning algorithms, deep neural networks, foundation models) that perform tasks commonly thought to require intelligence such as natural language processing and computer vision.
- **Algorithmic Decision-Making (also Automated Decision-Making or ADM):** Using algorithms to provide recommendations about decisions. Can be based on AI but may also use simple rule-based algorithms. While often promoted as producing more objective, consistent and efficient decisions, ADM also run the risk of amplifying human bias or relying on flawed data or false correlations to produce inaccurate or wrong decisions.
- **Foundation Models:** AI models trained on large datasets that are capable of a range of general tasks, including generation of text, images, audio, and video. Examples of foundation models include OpenAI's GPT-3 and GPT-4, which underpin the conversational chat agent ChatGPT. This term is sometimes used interchangeably with 'general-purpose AI', especially in the context of the EU's AI Act.
- **Frontier Model:** An undefined group of powerful foundation models more advanced than existing foundation models
- **Generative AI:** a type of AI that is capable of generating novel content (e.g. text, code, images, audio, or video) based on user prompts. Generative AI techniques include generative adversarial networks (GANs), diffusion, and generative pre-trained transformers (GPT).
- **Large Language Model (or LLM):** language models with hundreds of millions or billions of parameters using a transformer neural network architecture. LLMs are the basis for most text-based foundation models, but are also increasingly multi-modal in either or both of input and output.
- **Machine Learning:** Algorithms that apply statistical methods to very large volumes of data to find patterns or correlations, and then use these to make predictions. E.g., natural language processing or medical diagnostic tools.
- **Neural Networks:** A subset of machine learning algorithms involving a computing model that resembles the networked structure of neurons in the human brain, containing layers of nodes for classifying and clustering high volumes of data.
- **Multi-modal Foundation Model (or MFM):** Generative AI models capable of processing and generating multiple modalities of input and output including, e.g., text, image, audio, video, etc.

Outputs or Processes Relating to AI Technologies

- **Content filtering:** a safety mitigation involving filtering of sensitive data from the training dataset.
- **Fine-tuning:** A process that embeds new context or a new and smaller dataset into an existing foundation model to 'fine-tune' its performance and capabilities for more accurate or specialised responses.
- **Generalisation:** Where a generative AI model has been trained on enough unique and different content on the same topic to understand what it is and what's important to it in terms of related words and concepts that enables it to explore these connections in novel ways.
- **Hallucination:** Output from a generative AI model that is incorrect, not grounded in factual data, or a blending of multiple data sources based on the probabilistic nature of the model.
- **Meta-prompting:** a safety mitigation involving additional instructions being given to a model to guide its behaviour, e.g. 'communicate in the user's language of choice'
- **Red teaming:** a safety mitigation involving using structured testing to simulate how adversaries might prompt the system to find flaws and vulnerabilities in the system.
- **Regurgitation:** Output from a generative AI model that repeats a collection of words occurring enough times in its training dataset in a way that has resulted in the model weighting them as closely-related, resulting in training data that has failed to generalise from the data and capable of repeating the collection of words when prompted accordingly.
- **Retrieval-Augmented Generation (or RAG):** a process that allows information from an additional data source external to the training dataset to be included in the prompt to enable more informed and grounded outputs.
- **Training:** In relation to foundation models, refers to the process by which a training dataset is (1) tokenised and (2) transformed into the numerical weights and biases that are then
- **Transformer:** A neural model architecture that allowed the training of algorithms to the scale of billions or trillions of parameters.

Version 1.1
July 2024



THE UNIVERSITY OF
MELBOURNE

Centre for AI and Digital Ethics

Level 8

Melbourne Connect

700 Swanston Street

Carlton 3053

unimelb.edu.au/caide