# Artificial Intelligence (AI)

**Xuesong Zhao**
Partner

### Natural language

Natural language processing (NLP) remains a key battleground for tech giants fighting for supremacy in artificial intelligence (AI) and one of the most impactful events of recent years has been the astonishing pace the industry has developed.

The technology is incredibly advanced – using transformer-based, large-parameter models – though the tasks they are performing, in language terms, are basic. Microsoft and Nvidia, two such giants, collaborated to launch the successor to their own generative language models, Megatron-Turing Natural Language Generation model (MT-NLG), featuring 105 layers with 530 billion parameters, twice as many parameters as the existing largest model.

The successful implementation of general purpose language models will most drastically help those applications designed to be task-agnostic using limited data sets, just as humans perform a new language task from very few samples. The breakthrough of this new type of large AI model that can generalise the language training process marks the beginning of the next generation of AI. The initial results are encouraging as the model demonstrates unmatched accuracy in a broad set of natural language tasks.

Behind the scenes, numerous innovations and breakthroughs were achieved to help even the largest, slowest-changing language model to be trained. One barrier was that computational resources were the bottleneck for the development of large parameter, natural language generation models. The parameters of transformer models have increased 275-fold every two years, compared to 25x for the parameters of all AI models, while Moore's Law only expects transistors to double every two years (more likely every three years due to the increasing complexity in scaling). Training a 530 billion parameter model requires over 10 terabytes of memory and activating each transformer block would require 16.9 terabytes of memory. To put this into context, the most powerful graphics processing unit on the market today has 80GB of memory.

Google, another AI pioneer, introduced a successor to its widely popular Bi-directional Encoder Representation from Transformers (BERT). BERT has a few distinct elements which allow machines to understand the meaning of text in languages and soon after its launch in 2018, it gained popularity to become one of the most trained NLP models. After extensive training, BERT achieved superhuman accuracy in various reading comprehension benchmark tests and Google used it to power almost every single English-based query on Google Search in 2020, marking the beginning of the transition from keyword-based searches to more complex search queries.

At January's annual I/O developer conference, Google announced BERT's successor. One thousand times more powerful, MUM (a Multi-task Unified Model) can generate as well as understand language. It can also help remove barriers as this should allow Google Search to learn from sources that are not in the language users submit to the search engine. MUM will also allow Google to significantly improve its visual search capability by incorporating its capabilities and enable users to discover more videos and images that have not been explored before.

### AI-enabled digital twin

Humans have been using machines to predict future events for thousands of years. The Antikythera Mechanism, considered the first analogue computer humans invented, was made by ancient Greeks 2,000 years ago to determine and visualise the position of the sun, moon and planets. It is thought to have been used to predict solar and lunar eclipses. More recently, another planet-based example is NASA using simulators to match conditions on the fated Apollo 13 mission in the 1970s, marking one of the earliest examples of what would become known as a 'digital twin'. The simulator guided the crew with exact steps to safely return to Earth after two of the three fuel cells in the service module exploded 330,000km from Earth.

A digital twin is a digital representation of a physical object, which has one-to-one unique identity and mirrors the physical object with same metadata, structure, functional and system model via real-time data transmission. Until recently, AI's involvement in digital twins has been limited, with simulation playing a more pronounced role. While digital twin adoption has been increasing steadily, particularly in aerospace and high-value manufacturing such as automotive, proliferation has been slower than expected because of the amount of operational data, including frequent online and offline measurements and predefined outputs that many industries simply do not have.

Once integrated with AI, digital twins can predict the performance and state of individual components and internal processes and thus their combined impact on the overall system being tested. Its adoption has seen a sharp increase thanks to the ongoing improvement of machine-learning (ML) models that support transfer learning, such as BERT, which can effectively solve the data limitation issue. Siemens Energy is an early adopter of Nvidia's Ominiverse solution to perform predictive maintenance for their power plants. Not only did the AI-infused digital twin improve the thermodynamic efficiency of the power plant by more than 60%, but it also reduced planned downtime by 5.5 days per year, saving $1.7bn per year.

The digital twin market is expected to reach $86bn by 2028, a CAGR of 42.7% from 2021 to 2028.

**Environmental impact**

The computing demand driven by AI has increased significantly, and the impact of this on energy consumption and the environment has been a major concern for investors. Even as the parameters of machine-learning models have been increasing exponentially, the energy consumption and carbon footprint of AI implementation have not increased proportionally. The social and economic benefits AI enables outweigh the cost of training and delivering AI. For instance, training Google's GShard multilingual translation model generated 4.3 tCO2. Assuming two billion people benefit from this service, the amortised per user CO2 impact is estimated at less than the equivalent of sending one text message.

In 2020, data centres and their related data transmission networks consumed the equivalent of 2-2.5% of the world's electricity use, excluding energy consumed by cryptocurrency mining. While these power consumption numbers are large, cloud data centres are demonstrably more energy efficient thanks to much higher average server utilisation and purpose-built infrastructure designed to minimise energy consumption. Also, importantly, hyperscale server intensity enables industrial-scale renewable energy procurement which plays a key role in further reducing the environmental impact. In fact, information and communications technology (ICT) companies accounted for half the global corporate renewable energy procurement in the past five years. Technology companies continued to dominate clean energy purchases during 2021. In 2021, for the second consecutive year, Amazon was the largest corporate buyer of clean energy globally and, alongside Microsoft, Meta and Google, these four accounted for 48% of global clean energy purchases last year. As a result, shifting a workload from on-premise to, say, Amazon Web Services (AWS) would reduce the carbon footprint by 88%.

Following a muted 2020, AI funding in the private market almost doubled in 2021 to $57bn. The top three AI funding events were Databricks ($1.6bn funding round), Momenta.AI ($1.3bn Series C) and 4Paradigm ($700m funding round). Only 28% of AI funding went into late-stage companies as most AI innovations remain early stage. Transactions in this area saw a record total of 130 exits, including several high-profile IPOs.

AI is likely to play a pivotal role across multiple industries over the coming years, to convert information (data) into knowledge thereby increasing business productivity and efficiency as well as helping society through, for example, reducing greenhouse gas emissions. It has certainly come a long way from playing board games.

**Xuesong Zhao, Partner, Polar Capital Technology Team**

**18 May 2022**

## Find out more

### Client Services

**E** investor-relations@polarcapital.co.uk
**T** +44 (0) 20 7227 2700
**F** +44 (0) 20 7227 2799