The Bletchley Declaration and the Safety of Frontier AI

What is ‘frontier artificial intelligence’ and why were international leaders sufficiently concerned by it to attend a two-day summit?

International leaders and artificial intelligence (AI) experts agreed last week to support scientific research into AI safety “to ensure that the benefits of the technology can be harnessed responsibly for good and for all”. The agreement followed a two-day summit, the first of its kind in the world, hosted by the UK and attended by representatives from 28 countries as well as leading industry representatives and AI experts.

The result of the summit was the Bletchley Declaration, so-called because the meeting took place at Bletchley Park, the country estate 50 miles north west of London that was the centre of British code breaking efforts during the Second World War. Among the leaders of that wartime work was Alan Turing, whose computer science achievements helped lay the foundations for modern AI.

Turing, who once warned that “once the machine thinking method had started, it would not take long to outstrip our feeble powers”, would probably have approved of the summit’s caution. Those at the summit argued that “particular safety risks arise at the ‘frontier’ of AI” and identified particular risk in areas such as biotechnology and cybersecurity.

What concerned attendees is not the AI that recommends your best route to work or the AI that identifies your face to unlock your phone. The concern is “frontier AI” - but what is it and why is it special enough for British Prime Minister Rishi Sunak and his team to decide it needs an entire summit?

The Mystery of Frontier AI

The AI field is divided into several sub-fields, one of which is Deep Learning. Inspired by the human brain’s learning processes, Deep Learning teaches computers to find patterns in large datasets, such as texts, images or sound files, and use them to make predictions. It’s already used for tasks like fraud detection in financial services and patient monitoring in healthcare.

Some Deep Learning AI models are known as Foundation Models. These are models that have been trained on massive and diverse datasets and are succeeding at many different tasks, instead of being limited to a few narrow ones. The most advanced Foundation Models, including ChatGPT, Dall-E 3, Claude, and Midjourney, are known as frontier AI.

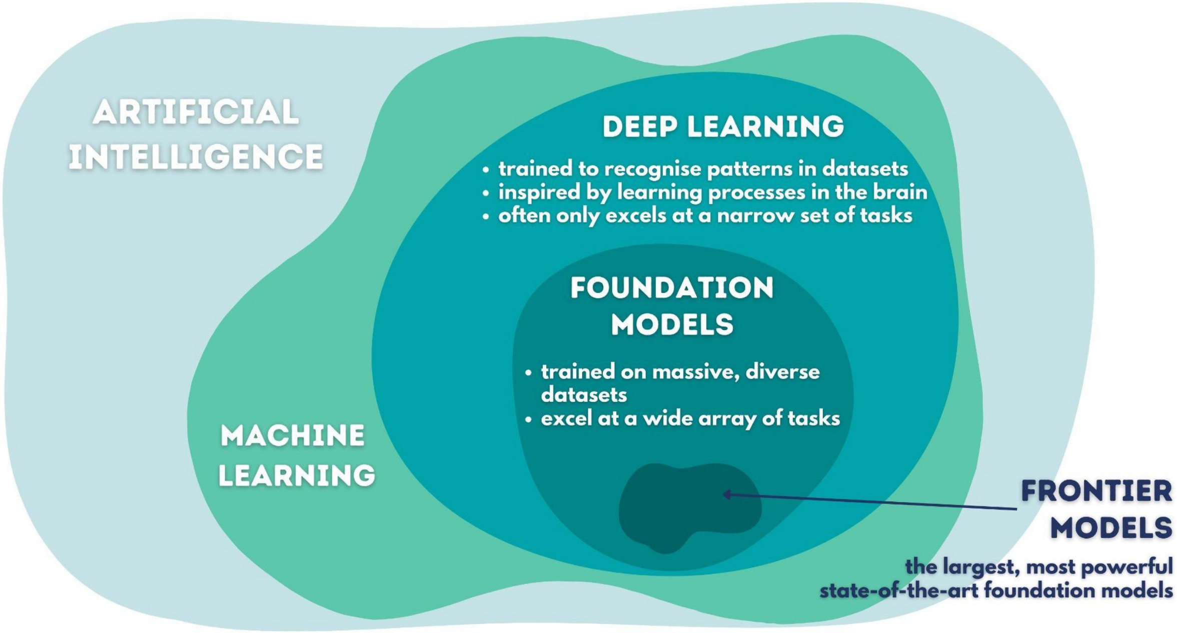

ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
• trained to recognise patterns in datasets
• inspired by learning processes in the brain
• often only excels at a narrow set of tasks
FOUNDATION MODELS
• trained on massive, diverse datasets
• excel at a wide array of tasks
FRONTIER MODELS
the largest, most powerful state-of-the-art foundation models

Image 1: Where Frontier AI sits within the field of AI

Frontier AI was singled out by the summit because it is qualitatively different from other AI systems, and comes with its own strengths and risks. In particular, today's frontier AI systems are notoriously tricky to govern responsibly because not even their developers fully understand how they make decisions; for now, they are as much of a mystery to their creators as to the average person.

As Google CEO Sundar Pichai said: "There is an aspect of this which we call … a black box. You don't fully understand, and you can't quite tell why it said this or what it got wrong. We have some ideas, and our ability to understand this gets better over time, but that's where the state of the art is."

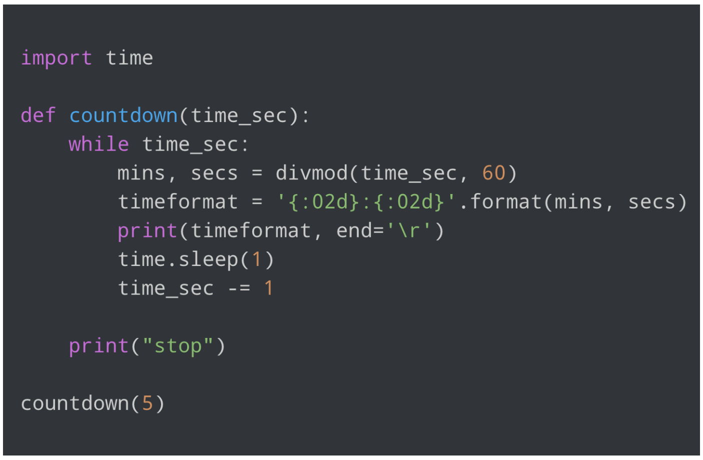
A Computer System that is Grown, not Made

How is it possible that humans don't understand a technology they themselves built? If engineers design a bridge that later collapses, they can still retrace each step of their design and find their mistake. Not so with frontier AI systems, because they are not built, but grown.

To write an ordinary computer programme, software engineers identify a problem, create a solution, then write instructions for the computer to execute using a programming language, such as Python. Both humans and computers understand these programming languages. In contrast, frontier AI uses a human-written learning algorithm to grow its own complex programme, a 'neural network', on top of a massive dataset.

As AI pioneer and Turing Prize winner Geoffrey Hinton told 60 Minutes: "We designed the learning algorithm. That's a bit like designing the principle of evolution. But when this learning algorithm then interacts with data, it produces complicated neural networks that are good at doing things, but we don't really understand exactly how they do those things."

Peek under the hood of these computer-grown ‘neural networks’ and instead of readable lines of code, all we find are massive tables filled with billions of numbers - perfectly comprehensible to the computer, but incomprehensible to us.



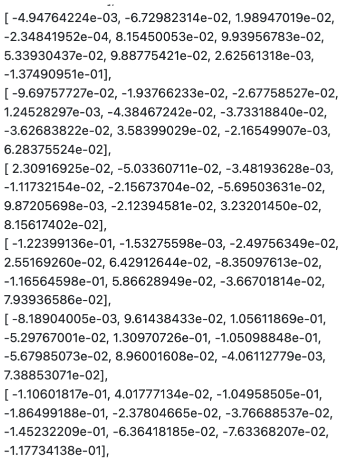
```
import time

def countdown(time_sec):
    while time_sec:
        mins, secs = divmod(time_sec, 60)
        timeformat = '{:02d}:{:02d}'.format(mins, secs)
        print(timeformat, end='\r')
        time.sleep(1)
        time_sec -= 1

    print("stop")

countdown(5)
```

Image 2: A simple Python programme for a [Countdown Timer](#)



```
[ -4.94764224e-03, -6.72982314e-02, 1.98947019e-02,
-2.34841952e-04, 8.15450053e-02, 9.93956783e-02,
5.33930437e-02, 9.88775421e-02, 2.62561318e-03,
-1.37490951e-01],
[ -9.69757727e-02, -1.93766233e-02, -2.67758527e-02,
1.24528297e-03, -4.38467242e-02, -3.73318840e-02,
-3.62683822e-02, 3.58399029e-02, -2.16549907e-03,
6.28375524e-02],
[ 2.30916925e-02, -5.03360711e-02, -3.48193628e-03,
-1.11732154e-02, -2.15673704e-02, -5.69503631e-02,
9.87205698e-03, -2.12394581e-02, 3.23201450e-02,
8.15617402e-02],
[ -1.22399136e-01, -1.53275598e-03, -2.49756349e-02,
2.55169260e-02, 6.42912644e-02, -8.35097613e-02,
-1.16564598e-01, 5.86628949e-02, -3.66701814e-02,
7.93936586e-02],
[ -8.18904005e-03, 9.61438433e-02, 1.05611869e-01,
-5.29767001e-02, 1.30970726e-01, -1.05098848e-01,
-5.67985073e-02, 8.96001608e-02, -4.06112779e-03,
7.38853071e-02],
[ -1.10601817e-01, 4.01777134e-02, -1.04958505e-01,
-1.86499188e-01, -2.37804665e-02, -3.76688537e-02,
-1.45232209e-01, -6.36418185e-02, -7.63368207e-02,
-1.17734138e-01],
```

Image 3: “Under the hood” of a Deep Learning AI System

The Race for Funding and Computing Power

However, we do know one thing about advanced AI from analysing past trends: in advanced AI, brawn beats brains. To make AI models more powerful, just feeding them more data and computing resources, meaning more powerful chips, yields better results than creating clever new AI designs.

So, understanding what goes on inside AI systems is an unsolved problem, but building more powerful systems is easy, so long as you have the money to acquire more computing resources and data. And the money is flowing: According to the 2023 State of AI Report, between January and September 2023 alone, AI companies received 45 billion USD in private capital globally, with US-based AI companies absorbing most of the funding.

AI companies will use this money to build larger, more powerful systems. For example, Mustafa Suleyman, Co-Founder of DeepMind and CEO of Inflection AI, announced on a podcast in September that he expects Inflection AI will train AI models 100-times larger than the most advanced models of today within the next 18 months.

The fact that frontier AI has unique qualities and challenges does not mean that policymakers should neglect the challenges of traditional AI systems. But given the uniqueness of today’s frontier AI, it is understandable that the UK government and its international partners decided to focus on frontier AI at the UK AI Safety Summit.

Frontier AI can behave in unexpected and sometimes dangerous ways. Today, that already involves anything from a chatbot inventing facts to a self-driving car that endangers other road users. And in the future, as frontier AI gets more capable, the dangers it poses through unexpected behaviour will become significantly more substantial. And if we don’t know why it does those things, then fixing them is a challenge. The research called for in the Bletchley Declaration is a great step in the right direction.

Image credits:

Image 1: created by Max Reddel & Eva Behrens, ICFG
Image 2, Image 3: see linked above