SPECIAL REPORT

The New Great Game:

Machines, Empires, and AI

INTELLIGENT INVESTOR

Intelligentinvestor.com.au
1300 880 160

About Us

With a track record of beating the market for 25 years, clear and straightforward language, and an 'open book' approach to stock research and analysis, Intelligent Investor offers actionable, reliable recommendations on ASX-listed stocks.

In 2014, Intelligent Investor became a part of the InvestSMART family, extending our expertise to even more Australian investors seeking quality analysis and advice.

Important information

Intelligent Investor
info@intelligentinvestor.com.au
www.intelligentinvestor.com.au
PO Box 744, QVB NSW 1230
1300 880 160

Contents

The New Great Game Part 1: Machines, Empires, and AI

Written by Angus Donohoo

Key Points

- **AI hype shares aspects of investment bubbles**
- **But transformational breakthroughs in AI models and hardware**
- **Race is on to own and control hyperintelligence**

New technologies tend to bring reality and possibility into conflict. Artificial Intelligence (AI) is no different. Billions are being invested into the promise of genuine technological transformation, yet the ensuing hype cycle features misplaced optimism, irrational investment, general confusion, and, at least for a while, sky-high stock valuations.

These are the same old trends that we saw in the dotcom boom and that our great grandfathers saw in the railway building frenzy. This Special Report acknowledges the hype cycle spinning in the background. But front and centre are potent underlying technologies that are sustaining it. The aim is to help you understand them and, when opportunity arises, profit from them.

This year, estimates indicate about US$200bn will be spent on AI. Some of it — perhaps a good deal of it — will be wasted, and yet the AI race will continue. The investments being made are an indicator of the potential profits that will accrue to the winners that harness this technology. The great game used to refer to the search for and exploitation of physical resources, and the battle for empire. Now it's about artificial intelligence.

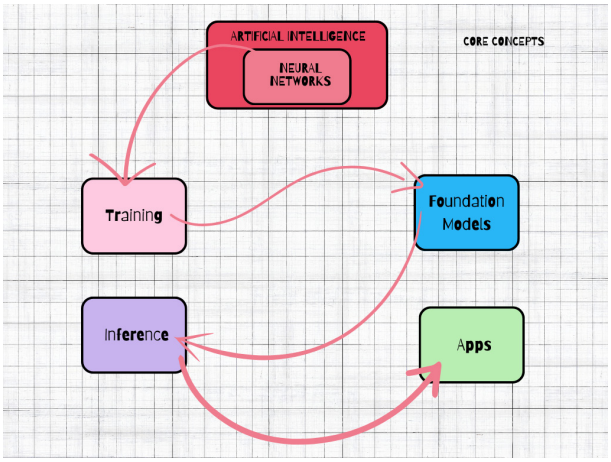
Over three sections, this Special Report will give you an understanding of what AI is all about. Don't expect any Buy recommendations though. Instead, it will help you understand the AI landscape and how the businesses and technologies that populate it fit together. Given the jargon-heavy nature of the subject, we've included a Glossary at the end.

Core Concept

The core AI technology is the **neural network**. These networks are **trained** using specialist semiconductor hardware called graphical processing units (GPUs). Once trained, neural networks, like the engine in a car, become the core component of **foundation models**. In a process called **inference**, these foundation models are deployed on mobile and personal devices, which people engage with via **applications**.

If this sounds complicated, don't worry. Each core concept will be explained in detail. For now, the key is to understand that AI is an intelligence that emerges from a combination of mathematical models and computer hardware. Together, the core concepts interact in a way similar to the figure below. Let's look at each in turn.

Figure 1: Core Concepts



1. Neural networks

Neural networks are a class of mathematical models that form the bedrock of the new breed of potent

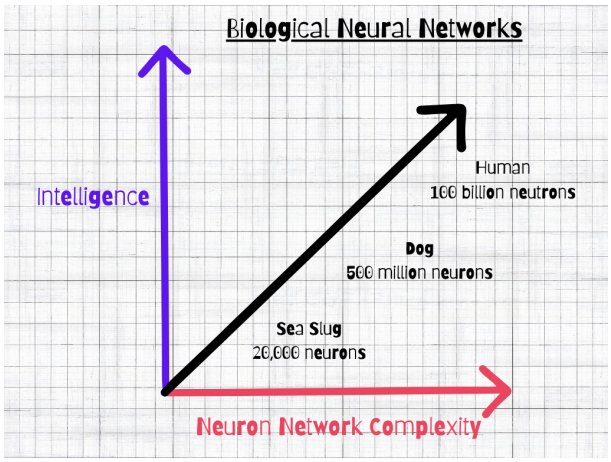
AIs. The science is decades old and emerged from concepts in biology. The human nervous system is a good example of a biological neural network.

The brain and nerves in the body are made up of neurons. There is a direct correlation between intelligence and the network complexity of neurons. Crudely, the more neurons, the smarter the organism. A sea slug is stupid, a human is smart, and neurons explain the difference (see Figure 2).

Beyond that, we don't know as much as you might expect. Science has a limited understanding of how consciousness and intelligence emerge from the human brain. As we shall see, AI has many mysterious elements, too.

Biological neural networks were first proposed in 1873 by Scottish engineer Alexander Bain. Bain made schematic diagrams of networked neurons, and proposed that these networks enabled human thinking and memory.

Figure 2: Biological Neural Networks



In the '40s and '50s, the neural network concept made the leap from biology to technology when artificial neural networks were proposed as computer architecture. At the time, this was largely conceptual, and the practical applications limited. History, however, proved the concept a useful one, although it took decades for computer hardware to catch up. No one anticipated how that would happen.

Traditional computer processing is performed by a central processing unit (CPU). CPUs are known as serial processors because they generally solve equations one at a time. These machines are very good at performing complex calculations, one after another, which is why every personal computer has one at its heart.

There is, however, a different kind of processor. If you're playing a video game with a character exploring a virtual world, CPUs are slow because lots of calculations — from changing shapes, lighting, angles and features of space — must be performed simultaneously. To do so requires a type of computation called parallel processing. This is what GPUs excel at.

Although built for gaming, GPUs began to get used for other tasks that benefited from parallel computation. Fields as diverse as cryptocurrency mining, weather forecasting, and AI all began exploiting the parallel capabilities of GPUs. It was found that by using GPUs, supercomputer performance could be unlocked for the cost of a few video gaming PCs.

But perhaps the most significant breakthrough occurred in 2017 when Google published a new type of computer model called a Transformer. While neural network architecture had been connected to GPUs as far back as 2004, Google's Transformer paper was AI's great leap forward.

To understand why, we must explain two concepts. The first — artificial neurons — change one number (the input) to another number (the output). Imagine a billion billion of these calculations every second (the speed of the cutting-edge Nvidia GPU) and you get a feel for how artificial neurons have begun to resemble their original biological predecessor.

The second concept is perhaps even more mind-blowing. The Transformer model connects the parallel computation of a GPU to a statistical weighting process that occurs in the 'hidden layers' of a neural network. Hidden layers describe the computational depth (the number of processing operations) of each 'artificial neuron'.

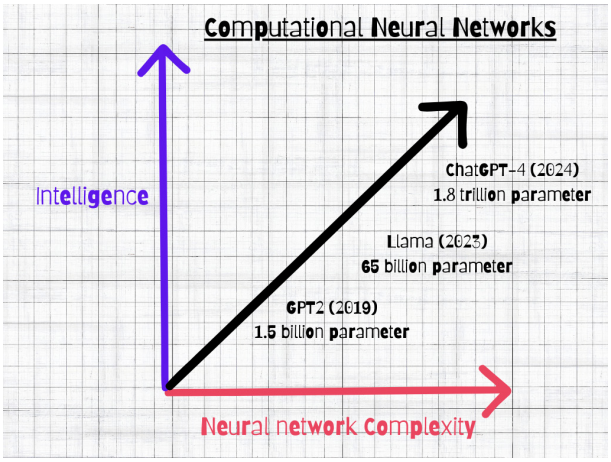
The intelligence of the system grows with the number of hidden layers as the artificial neurons consume more computation.

This process may be the origin of artificial intelligence. But, as with the human brain and as

we'll see in the next section, there's a lot of mystery under the hood. It is the act of training the neural network that draws the ghost out of the machine. We have algorithms and hardware but we don't fully understand what is happening.

This is a mystery that probably doesn't trouble Nvidia chief Jenson Huang, who is busy adding to his US$102bn fortune by solving AI's hardware problem. The more hidden layers, the smarter the AI and the more GPUs are needed to train it. The increasing model complexity shown in Figure 3 has lifted Nvidia's revenue from US$12bn to US$61bn in just five years.

Figure 3: Computational Neural Networks



Note the striking resemblance between a computer neural network in Figure 3 and the biological one in Figure 2. Intelligence increases with the complexity of the network.

There are different metrics to measure network complexity but, fundamentally, the more hidden layers the more parameters, and the more parameters, the more intelligent the AI.

This is why the AI models you might use on your personal computer are improving so fast. As Figure 3 shows, computer neural networks are becoming smarter at an incredible rate. ChatGPT-4 has 1,200 times more parameters than its 2019 predecessor.

This doesn't happen without the application of vast computing resources, though. By dialling up processing power, primarily via Nvidia chips, we've effectively been improving both halves of a mind/body problem.

To continue with the human/machine analogy, the complexity of the neural network model (the number of layers and parameters) is like the mind, but without training the body (GPU processing power) neither can effectively function. Increasing neural network complexity can loosely be thought of as an expansion of the neural network's mind, while the amount of GPU computation expands its physical body.

When neural networks are integrated with GPUs, the amount of processing power dictates the speed and quality of the mind. This process is called training and is the focus of the next section.

2. Training

By training a neural network, you teach it to perform a task. That might mean getting the AI to recognise an image, or to translate one language into another. This occurs by deploying the neural network across the GPUs and crunching lots of data.

Originally the approach was to choose a domain, such as photos of cats, and just run the neural network over the GPUs with lots and lots of cat images so that the AI would recognise cats. As it turned out, at least for the large neural networks known as 'foundation models', the intelligence is optimised if the neural network processes data 'unsupervised'.

This is one of the truly remarkable traits of advanced neural networks — the ghost in the machine if you like. The process that weights the neural network inside the layers and sets the parameters requires minimal human supervision or guidance.

The data sets are extremely large and sometimes unstructured and unlabeled. The works of Shakespeare are consumed alongside cat photos, agricultural statistics, and — possibly — your instant messenger conversation with your high school sweetheart from 1999. All of the internet is in play.

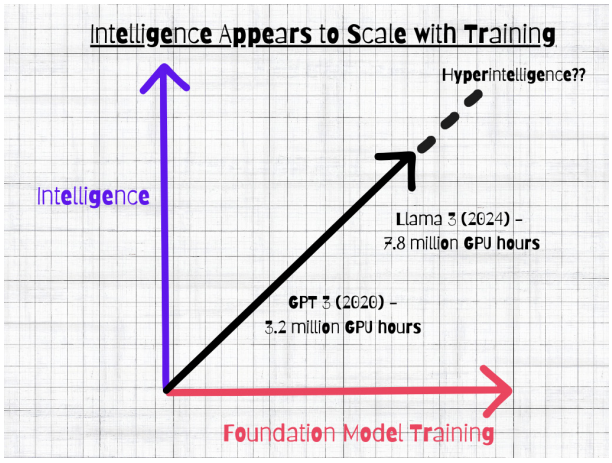
If that all sounds a bit spooky, it is. Perhaps it is also unavoidable if we are to secure its benefits. Some argue that no technology is either good or bad. Amazing developments, of which AI is certainly one, inevitably have embedded costs. Maybe there is not one without the other. Omelettes require broken eggs.

As far as model training is concerned, those companies best placed to do it need two things — data and processing power. And the more they have of both, the better. This is why companies like **Meta** and **Alphabet** have such a big advantage — they've been collecting these data sets for years and have the cash flow to sustain large investments in GPUs.

The so-called 'hyperscalers' — the large cloud computing providers like **Amazon**, **Microsoft**, **Alibaba**, and Google owner, Alphabet — are buying enormous numbers of GPUs and stacking data warehouses with them.

The great game that all are playing is based on a simple but extraordinary possibility — that the combination of data, neural networks and processing power will deliver a form of 'hyperintelligence' that will transform every area of life. Figure 4 depicts the game in play.

Figure 4: Intelligence Appears to Scale with Training



The end of the dotted line is the key bet that many large technology companies are making. By buying the most GPUs and training the largest neural networks, each is betting that they will own and control hyperintelligence.

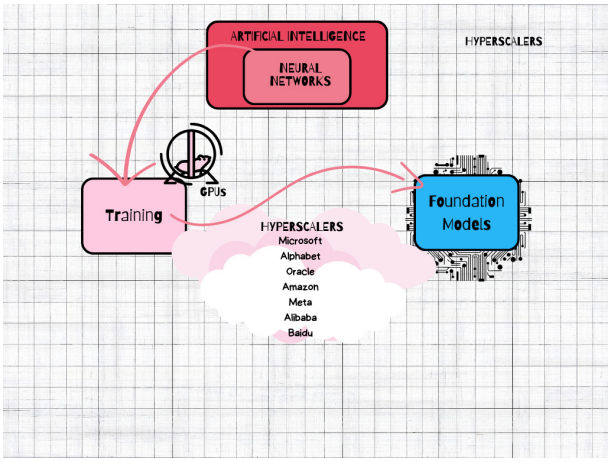
These nascent hyperintelligences are called 'foundation models'. Foundation models are deep neural networks that can perform novel tasks and generate text, imagery, audio and synthetic data. Generative AI, as it is known, will be covered in Part 2.

The scale is simply staggering. Microsoft and Meta are believed to have clusters of over 100,000 GPUs for training. The favoured GPU product is the Nvidia H100, and the cost of each training 'cluster' is believed to be over US$4bn with annual power consumption estimated at over US$120m.

The chief technology officer of hyperscaler **Oracle**, Larry Ellison, recently commented that to build a foundation AI model in the next three years, a business will need acres of GPU clusters, a nuclear power reactor or its equivalent, and to spend at least US$100bn. It turns out genesis is expensive.

Currently, hyperscalers are developing foundation models and offering model training as a service to enterprises. Figure 5. sketches out the landscape.

Figure 5: Hyperscalers



In Part 2, we'll go deeper on foundation models, then cover inference, and apps before addressing the stocks set to benefit from AI, and valuations.

Part 2: Foundations of Empire

In Part 1 we explored the core concepts, neural networks, and training. It's now time to turn to who's trying to profit from AI and how. There's a bit more jargon so as a reminder, we've attached a glossary at the end.

Key Points

- **Businesses have different AI strategies**
- **Generalised models threaten proprietary data**
- **Foundation model investment at risk**

3. Foundation Models

Businesses are taking a variety of approaches to profit from AI. Some aim to own the AI as if it were a brain in a box, others the hardware and/or services, and some are hoping to own the lot.

The most glittering prize is probably the brain in the box, or as we previously referred to it, 'hyperintelligence'. This is potentially different to previous speculative bubbles, when the profit accrued not necessarily to the product but to the tools used to extract it. As the old saying goes, in a gold rush, better to sell the picks and shovels rather than use them yourself.

This oft-repeated point may be apocryphal but it's true that **Bloomberg** has enjoyed better longevity selling financial data than most of the financial companies it has serviced, and airports are usually better investments than airlines.

So where's the gold in the AI rush and which companies are selling the picks and shovels? The most advanced artificial intelligences are embodied in so-called 'Foundation Models'. These are large neural networks that are 'generalised'. This means most are trained on a huge variety of data from images, to music, to text.

Most of the 'hyperscalers' discussed in **Part 1** are focused on building and owning these potent AIs. These are the companies aiming to strike gold.

Some foundation models are open source but many are proprietary. **OpenAI**, **Anthropic**, **Alphabet**, **Cohere**, and **Baidu** are each building out proprietary models, while **Hugging Face**'s models have been open source from the start.

Last July, **Meta** made its foundation model, Llama, open source. This dramatically altered the competitive landscape. Due to its core business in advertising, Meta does not need its AI models to be directly profitable. By open sourcing its model, Meta can harvest a much wider universe of developers to improve the AI that underlies its advertising model.

Meta's decision to open source its foundation model was a challenge to its hyperscale competitors. In response, OpenAI and Alphabet were forced to slash the prices of their competing models.

Other businesses are taking a different approach. **Amazon**, for example, appears to be looking more to the picks and shovels. Amazon is more focused on renting out data centres filled with GPUs, than on owning the best foundation model. Amazon is also helping clients build their own AI models.

Similarly, companies like **Nvidia** are building and selling the hardware. Nvidia is also dominating the software and establishing large GPU 'clusters' for rent. Nvidia is most commonly described as the 'pick and shovel play' of AI, but that potentially undersells the company's monumental significance to the field.

As we discussed in ***Nvidia in heaven or just hype?***, Nvidia has been the critical facilitator of AI, to the point where the company is almost the oxygen in the AI room. Yet the company is not beyond suffocating competition. Nvidia has been developing foundation models, and created similar disruption to Meta in September when it open sourced its highly generalised NVLM foundation model.

There has been much debate and confusion about the value of foundation models versus custom models for enterprise. The critical question for business owners

is this: should they deploy a pre-trained foundation model or spend a few million (or billion) dollars training a custom AI on proprietary data?

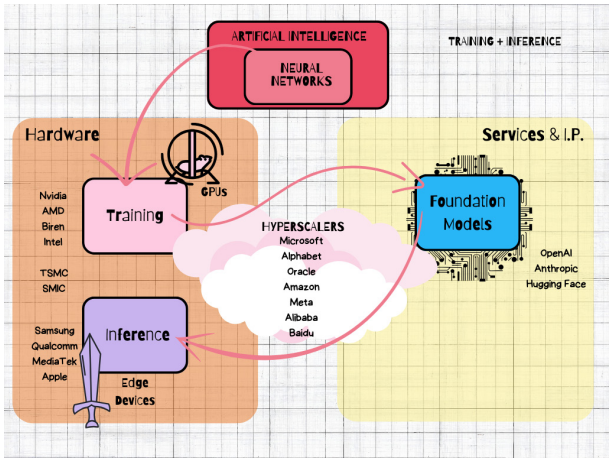
So far, using the smartest model possible seems like the best bet. The most powerful foundation models have tended to outperform custom AIs. BloombergGPT, a 50 billion parameter model released in 2023, was extensively trained on Bloomberg data to be the ultimate financial AI. Yet, on financial tasks, OpenAI's ChatGPT and GPT-4 both outperformed it.

While much of OpenAIs 'special sauce', or competitive advantage, comes from its secretive 'fine-tuning' of its AI models — a process where both humans and additional AIs improve the foundation model — wider evidence points toward significant outperformance by generalised models.

Generalised artificial intelligence draws the long-term value of proprietary data into question. The implication of generalised foundation models outperforming narrow proprietary AI models challenges the notion assumed by many businesses, that the decades of knowhow and data they have accumulated will protect them from competition.

The sheer intelligence of foundation models throws that into doubt. A clever enough AI does not need specific data on a particular industry or customer. The hyperscalers' colossal investments in foundation models appears partly based on this finding. Figure 6 shows their central role in the AI chain of association.

Figure 6: Training + Inference



4. Inference

Training is just one half of the story. The other is inference, the process of getting the trained AI to make decisions or predictions. For example, imagine a neural network trained on 100,000 images of cats. When shown an entirely new cat photo, the network can then infer that the new object is a cat.

Our favourite analogy is that while AI training is like spending years learning to play the piano, inference is akin to a skilled pianist playing a new piece of music.

Inference is far less computationally intensive than training. GPT-4, for example, requires 560 teraflops (trillion floating operations per second) of compute for an inference operation. A single GPT-4 inference operation on a medium sized laptop (costing around $600) would take around 93 minutes to run. To train GPT-4 on the same laptop would take almost seven million years.

What does this look like in dollar terms? The data is patchy, but it's estimated that each customer query to ChatGPT (an inference operation) costs OpenAI around US$0.35. This pales in comparison to the US$100m cost of training the model, but is still expensive compared to a traditional search query on Google, which costs about one US cent.

Inference queries can either be run on mobile devices, in the cloud, or close to the device over the internet (at 'the edge').

Advanced mobile AI devices like the latest **Apple** and **Samsung** phones have their own specialised GPU-type parallel processing chips that are specifically designed for running inference workloads. **Qualcomm** is a leader in designing these types of chips, as is Apple, Samsung and **Mediatek**.

TSMC continues to be the major fabricator of the silicon used in the hardware for inference and training. **Intel** has been making moves into GPU design and production but remains a laggard. Samsung and China's **SMIC** are the only other GPU fabricators. Samsung also makes high bandwidth memory, which is critical for GPU clusters to function.

SMIC has been fabricating the GPUs for China's Nvidia competitor, **Biren Technology**. Biren seems an essential company for China's push into AI, although its technology appears several generations

behind Nvidia. Nevertheless, the trade restrictions imposed on China relate to GPU and AI capabilities like those offered by Biren, and the business is blacklisted by the US Department of Commerce.

5. Apps

Once trained, the AI models can run on edge devices as slimmed down pieces of software or applications ('apps'). The apps will either be performing inference or using the results of it.

The apps that ultimately best leverage neural networks are works in progress, addressed by the many startups operating in the space. Many are likely to be forced into a fee paying relationship with one or other of the foundation models.

As for the foundation models themselves, these may well function more like an app store or cloud storage provider than an operating system like Windows. They will likely need to provide ongoing compute services to client apps. The large startups like OpenAI and Anthropic are deeply engaged in building out both end products as well as foundation models.

Foundation models will likely subsume a huge number of businesses and processes. Many are already capable of doing highly complex tasks from scratch. Microsoft's code repository GitHub has been transformed by AI assistance, and 77,000 organisations are using its Copilot AI to write partial and complete code. Some consider 2021's Copilot to have been AIs first so-called 'killer app'. The impact on coding has been a surprise, demonstrating AIs power and potential.

AIs are now outperforming humans in many tasks that just a few years ago would have seemed impossible. Along with coding, translation and various intelligence and academic tests are being impacted by AI.

Individual humans still meaningfully outcompete AI in some instances, but scale, cost, and speed are all now far beyond competition. The artwork accompanying this report and shown in Figure 7. was created on **MidJourney**. Its creation took about five minutes, provided dozens of options, and the service costs US$10 a month. A commercial artist would have charged a few thousand dollars.

Figure 7: Artwork created for report in MidJourney

OpenAI's services are other notable successes. The ChatGPT chatbot boasts 300 million users, while the Dall-E art and image generation product (like MidJourney) claims 1.5 million daily users.

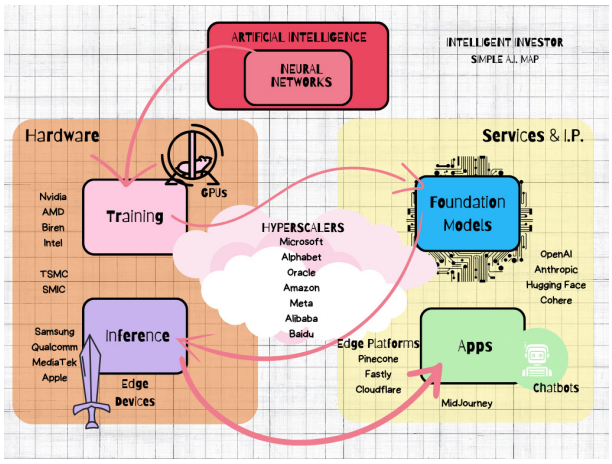
Microsoft's Copilot and Meta AI have also gained significant traction with chatbot products. As for China, **Baidu** looks to have the world's most widely used foundation model app, with over 300 million users reportedly using its Ernie chatbot.

Edge computing and vector databases are other notable areas to watch, as they provide the fulcrum where apps connect to the AI models. Edge networks like **Fastly** and **Cloudflare** have built high-quality edge platforms.

These can bring computationally intensive inference close to end users, at a far higher speed, without overburdening personal devices. Similarly, vector databases like **Pinecone** offer on-device AI models with edge storage. These are exciting but immature areas.

With all of these pieces, we can now see how our full AI map fits together in Figure 8.

Figure 8: Intelligent Investor Simple AI Map



The race to god-like machines?

This technology race is unlike any other. It is the new Great Game. If our chart in Figure 4 is even halfway correct, this isn't just a race for the next widget or better mousetrap, but a race to dominate a new form of corporate power.

The implications are simultaneously dizzying, terrifying and enchanting. If intelligence continues to scale with model complexity and the potency of GPUs used in training, then we may be well on the path to 'hyperintelligence', or 'god-like machines'.

On the other hand, if the rates of improvement slow and an additional $1bn in GPU investment results in an AI model that provides only $100,000 in additional value (or is only fractionally more intelligent), then the dream turns into a nightmare and hundreds of billions will have been wasted. No wonder analysts are watching with bated breath, asking whether there are limits to how smart these systems can get.

Will performance improvements continue to scale with GPU count, or will the intelligence gains taper off, so that improvements become incremental? If more compute drives merely marginal improvements in model performance, then the US$100bn investment in the next foundation model could be a capital expenditure that will never generate an adequate return.

Remember also that inference is a lot less compute-heavy than training, and that some models are now open source. If a business or individual can spin up an intelligence on their personal computer that's 80% as good as the best in the world, then we may be on a path to a new form of liberty and/or anarchy.

There is no question that the GPUs and foundation models will still be useful, it will just be the degree of how useful, and, critically, how profitable.

In **Part 3**, that's where we're heading.

Part 3: Paying to Play

This is Part 3 of our special report on artificial intelligence (AI). In Part 1, we covered the core concepts of AI, neural networks, and training. Part 2 looked at foundation models, inference, and apps.

Key Points

- **Bubble risk**
- **Valuations mixed**
- **Hyperscalers' profitability a cushion to spending**

Now in Part 3, we're focusing on valuation. As a reminder, the aim of this report is not to make specific Buy recommendations but to help you understand the AI landscape.

Our position? There is real value being created, but we also see the mania.

If you bought **Cisco** in early 2000 and held it, your investment would still be underwater today (see Figure 9). Great excitement in markets often leads to great destruction of capital. At the peak of the dotcom era, Cisco traded at 232 times earnings and 35 times sales.

Figure 9: Cisco historical share price

Those ratios weren't just about misplaced bets or foolish investors — they reflected honest optimism and the allure of new technology. Today, Cisco trades at 16 times earnings and 4.5 times sales.

Bubble, what bubble?

Valuations today are certainly enthusiastic for some tech stocks in certain areas, but the 'AI bubble' isn't as obvious as you might expect. Figure 10 compares the historical average price-to-earnings (PER) and price-to-sales (PS) ratios for most of the public companies we've covered so far. It looks at the period from 2009 to 2019 and then compares it to 2023 until today.

If the bubble was obvious, we'd expect a consistent valuation premium over the past two years. But, outside of GPU designers and high bandwidth memory manufacturers, results are mixed.

Figure 10: Historical versus recent average trading multiples for AI stocks

Technology	01/2009 - 12/2019	01/2023 - 01/2025	AI Premium?	01/2009 - 12/2019	01/2023 - 01/2025	AI Premium?
	Average trailing price to sales ratio	Average trailing price to sales ratio	% change	Average trailing price to earnings ratio	Average trailing price to earnings ratio	% change
GPUs						
Nvidia	5	31	520%	30	95	218%
Intel	3	3	-11%	16	68	325%
AMD	1	8	493%	62	340	451%
High Bandwidth Memory						
Micron	2	4	153%	13	28	124%
SK Hynix	2	2	35%	13	26	98%
Inference Devices						
Samsung	1	1	30%	11	16	45%
Mediatek	3	1	-75%	19	4	-79%
Qualcomm	5	4	-14%	23	19	-20%
Apple	4	7	103%	17	31	85%
Hyperscalers						
Microsoft	5	12	140%	23	33	39%
Alphabet	6	5	-11%	29	24	-18%
Meta	14	7	-51%	68	28	-59%
Oracle	5	7	49%	22	35	63%
Alibaba	13	2	-87%	42	23	-45%
Baidu	14	2	-85%	40	20	-50%
Amazon	3	3	12%	126	91	-28%
Fabricators						
TSMC	5	9	73%	16	22	40%
SMIC	2	5	260%	28	18	-37%

For hyperscalers, fabricators, and inference device companies, the signs of a valuation bubble aren't apparent.

That said, **Nvidia**'s average price-to-sales ratio over the last two years is 34 times, reminiscent of Cisco's peak of 35 times sales. Nvidia's recent peak was a multiple of 45 times sales.

Figure 10 is a coarse tool though. There are stories behind the numbers. Take the so-called GPU companies: Intel's PER spike, for example, can be explained by a broad collapse in profitability. Competitor **AMD** has grabbed much of Intel's market share, but that's mostly a CPU story, not a pure AI one.

Figure 11 helps clarify the AI market for the top AI GPU manufacturers, showing just how central AI has become to their business—especially for Nvidia.

The dominance of Nvidia in GPU design has not gone unchallenged, and the hyperscalers have been trying to break Nvidia's stranglehold in both software and hardware. On the software side, Meta's PyTorch and OpenAI's Triton have started to materially diminish the dominance of Nvidia's Cuda in AI software development for GPUs.

Figure 11: AI share of revenue and profit for GPU designers

GPUs & CPUs	Approximate AI % of Revenue	Approximate AI % of Operating Profit
Nvidia	88%	93%
Intel	26%	12%
AMD	48%	57%

In hardware, a major trend has been in 'custom silicon', and several of the hyperscalers have been building custom GPU equivalents under license to take on Nvidia. These custom chips, often called NPUs (Neural Processing Units) or TPUs (Tensor Processing Units), usually serve the same purpose as GPUs. Companies like **Broadcom**, **Marvell**, **TSMC**, **Alphawave**, and **Arm** are believed to be involved in designing **Alphabet**'s TPUs. The AI supply chain is deep, and if a bubble burst, there would be far-reaching consequence.

Don't forget memory

Figure 10 also shows that high-bandwidth memory manufacturers currently enjoy valuations far above their historical averages. But since memory companies report financials differently from GPU designers, disentangling the contribution of AI is less straightforward. Sales growth, however, offers a glimpse of how much AI is driving demand (see Figure 12). And it's a reminder that memory trades like a cyclical commodity.

Figure 12: Sales growth in memory

Sales Growth	2024	2023	2022	2021
Micron	62%	-49%	11%	29%
SK Hynix	106%	-27%	4%	35%
Samsung	19%	-14%	8%	18%

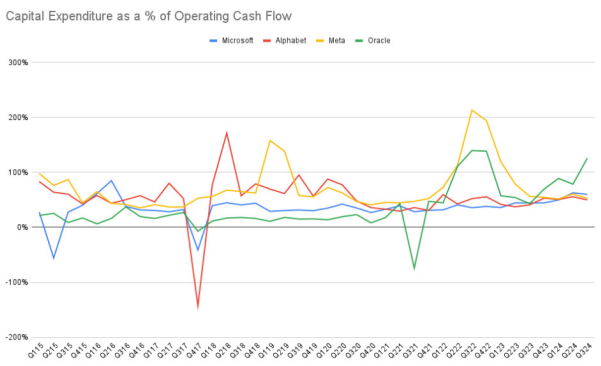
Both memory and GPUs are riding a huge industry cycle driven by demand from hyperscalers buying enormous amounts of hardware. Sales are booming for both, and suppliers are investing to meet that demand.

Hyper spending hyperscalers

Turning back to stocks where a bubble is harder to find, Figure 10 shows that, aside from Microsoft, most of the hyperscalers are not historically expensive. Yet much of the anxiety around an AI bubble focuses on these companies, likely because they're the ones pouring capital into AI. The hyperscalers are buying to meet a customer demand for AI services that either isn't obvious or that appears overmatched. In many instances, the hyperscalers don't even appear to have products.

Still, while these companies are spending heavily on AI infrastructure, they remain some of the most profitable businesses in history and most appear to be spending within their means. At least if we look at operating cash flow as a percentage of capital expenditures (capex) since 2015, Figure 13 tells the story.

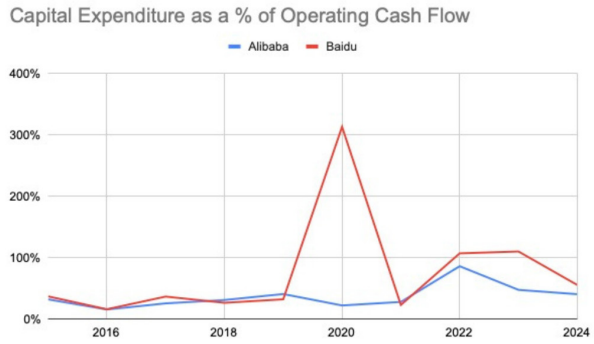
Figure 13: Hyperscaler capital expenditure



From the chart, only Meta looks like it's been overinvesting, though much of its capex has gone to virtual reality rather than AI. Oracle looks like it's playing catch-up and is doing so with debt.

Turning to the Chinese hyperscalers (and excluding Amazon, with its more capital-intensive, diverse business), only Baidu looks overstretched on these metrics. As a multiple of earnings, it's also the cheapest hyperscaler by far, with a forward PER significantly under 10.

Figure 14: Chinese hyperscaler capital expenditure



Fair weather friends

In Part 2, we explored potential scenarios. If the intelligence of foundation models keeps scaling with GPU investment, it makes sense for a hyperscaler to spend $100bn on the next foundation model. But if that relationship breaks, things could get ugly fast.

Below we take our best wild guess at some companies that might do well or poorly given the scenarios we've outlined.

Scenario 1: Neural networks stop getting ever smarter with ever more GPUs added to them.

Potential casualties: Nvidia, Micron, SK Hynix, AMD, TSMC, SMIC

Potential thrivers: Alphabet, Apple, Intel

Scenario 2: Open-source foundation models perform almost as well as proprietary models.

Potential casualties: Oracle, Amazon, Alphabet, Microsoft, Apple

Potential thrivers: Meta, Nvidia, Samsung, Fastly, Cloudflare, Qualcomm, Mediatek

Scenario 3: The trend of generalised foundation models getting smarter with more GPUs continues.

Potential casualties: Apple, Intel, Alphabet

Potential thrivers: Nvidia, AMD, Microsoft, Micron, SK Hynix, Samsung

This exercise is of course speculation, but it does outline the crucial competitive tensions. The new great game is finely balanced, and neither the pieces nor the shape of the board are clearly determined.

All we know for certain is that empires are being made and broken.

Disclaimer: The author owns shares in Alphabet, Microsoft, Samsung, Qualcomm, Intel, Alphawave, Fastly, Baidu, and Alibaba.

Glossary

- **Algorithms** - A sequence of instructions for performing a computation. An artificial neural network is a series of algorithms, or a mathematical model, that is itself built from a variety of probabilistic equations and linear algebra operations.
- **Applications** - Pieces of software that people can interact with.
- **Central Processing Unit (CPU)** - An integrated circuit fabricated in silicon, CPUs are the key logic processor inside most computers. Computation is generally performed in serial, or one at a time.
- **Fabricator** - The companies that specialise in fabricating silicon chips. The fabricators etch pieces of silicon with light and chemicals to make the extremely small, near atomic scale, features of modern silicon chips.
- **Foundation model** - A large AI model trained on a wide variety of data.
- **Graphical Processing Unit (GPU)** - An integrated circuit fabricated in silicon, GPUs are the key graphics processor inside most computers. Computation is generally performed in parallel, or all at once.
- **Generalised AI** - AI that performs a broad range of tasks and that demonstrates human-like general intelligence. Humans are not limited to performing one or a few tasks, but can perform a wide range of general activities from mathematics, to painting, to walking, to chewing bubble gum. Some AIs are beginning to excel in a wide range of domains.
- **Hidden layers** - The layers between the input and output layer in a neural network. As a network trains, the parameters adjust in the hidden layers. The number of hidden layers loosely corresponds to the complexity of the network and the number of GPUs needed to train it.
- **Hyperscalers** - Large cloud service providers. The term originally referred to the ability of cloud providers to scale any application with demand, so, for example a company could go from serving their application from a hundred people to a million.
- **Inference** - Deployment of a trained AI model. Less computationally intensive than training. The model infers something from new data on the basis of what it learnt from old data. Generally inference refers to the operations of the AI at 'the edge', or close to the end user.
- **Neural network** - A network of interconnected units called neurons. Can be biological or artificial and computational. Generally this report refers to the artificial kind. Artificial neural networks are the computational models that form the basis of the large AI models.
- **Parameters** - Weighted variables that adjust during the training process. Often used to proxy neural network complexity.
- **The edge** - 'The edge' usually refers to the network or compute space in close physical proximity to the end user. Because large files or compute intensive workloads (like AI) have extensive hardware and storage demands, data centres fulfil a critical centralised infrastructure role. Edge networks and devices fill the gap between data centres and end users and applications.
- **Training** - The process of feeding data into a neural network to weight its parameters. Generally, the more training, the smarter the AI.
- **Transformer** - A breakthrough neural network architecture invented by Google engineers in 2017. 'GPT' stands for Generative Pre-trained Transformer.

But is it really intelligent?

There's been a lot of bad theory of mind in the press, as people grapple with how smart or otherwise these AI systems actually are.

For decades the dominant test for establishing whether an AI system was truly intelligent was the Turing test. This test, devised by England's greatest ever code-breaker, Alan Turing, sought to establish how a machine might display human-type intelligence and qualify as conscious.

Put simply, the test was whether a machine could fool a person into believing that it too was another person. ***I trick therefore I am***.

The Turing test is now easily passed by various foundation models. So the trillion dollar question is, are these models definitely intelligent?

While the Turing test and a wide variety of academic and quantitative intelligence tests are now aced by these machines, it remains a thorny question.

Determining whether any other being or object is conscious is extremely difficult, and leads into various philosophical mazes.

It's fairly clear that our best science doesn't understand how the human brain and consciousness works. So making sweeping assessments of what machine intelligence ***isn't*** seems unjustifiably ambitious.

Some critics point to the probabilistic mathematics underpinning neural networks to define these AI as mere prediction engines. However, even if we accept the reduction to mathematical operations as the critical factor, we would still have to deny the possibility that humans may also be prediction engines. Some neuroscientists believe exactly that.

It seems reckless to equate neural network AIs with calculators. The Turing test remains useful. If AIs are materially smarter than us and active in the world, perhaps it's beside the point whether they weep or dream. Maybe it's none of our business.

INTELLIGENT INVESTOR

Intelligentinvestor.com.au
1300 880 160