Schroders

Artificial Intelligence: A New Machine

Artificial intelligence (AI) may be one of the most important developments of our time. While most are familiar with the term, few understand the architectures and processes behind it, nor why it is finally to become relevant. Yet to comprehend AI is to gain an insight into the future. The impact will be felt across all industries and aspects of life itself. Its importance should not be underestimated.

James Gautrey, CFA

Portfolio Manager and Global Sector Specialist (Technology), Global & International Equities, Schroders

Us and Them

Historically, human brains and computers have been built differently. Our "beautiful minds" are made up of billions of neurons and trillions of synapses capable of processing vast amounts of small data packets simultaneously (parallelism). This makes image and pattern recognition, reading and writing, sensory perception and creativity second nature, yet finding the square root of 17,378,969,242,589 almost impossible.

In computers, we observe the opposite. Primarily, this is because they have been designed as "serial" processors i.e. extremely capable of performing single, pre-programmed tasks but completely overwhelmed by inference required to deal with new situations.

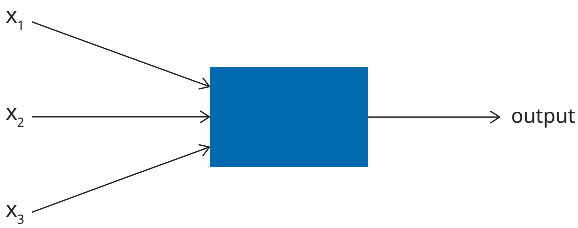
All of this is set to change, however, as computers are re-designed to become "neural networks" emulating the human mind. A key enabler of this has been the addition of Graphical Processing Units (GPU) to Core Processing Units (CPU), enabling machines to better execute both "serial" and "parallel" tasks.

Neural networks – Welcome to the Machine

Understanding neural networks is to understand AI. First described in 1943 by McCulloch and Pitts, an instructive example was developed in the 1950s by Frank Rosenblatt, tantalizingly named the "Perceptron".

In the perceptron model, multiple, binary inputs (xi) are used to produce a single, binary output, shown in Figure 1.

Figure 1



Source: Schroders.

The inputs could be anything, for example questions such as "will an Intel processor run faster than a human brain in 2019?" or "did Google's net income exceed £10 billion last year?". Each input is then "weighted" according to user preference. The product of these weights (wi) and inputs (xi) is measured against a "threshold" value (t) to determine the output, which might be a question such as "Should I buy stock in Intel?".

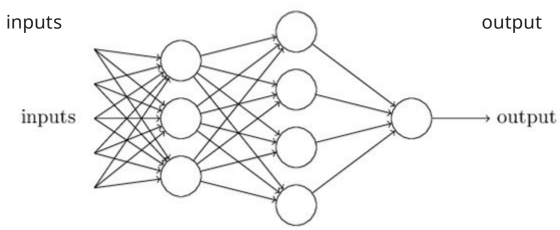
Mathematically:

$$\sum_{i=1}^{n} w_i . x_i > t$$

(A further example, "Should I go to a cheese festival this weekend?" is included in the appendix to aid understanding.)

Figure 1 showed a single-layer model i.e. one set of inputs and output. However, the model can be extended such that inputs relate to multiple outputs and, crucially, said outputs form inputs for another layer. This is known as the multi-layer model and simplified in Figure 2.

Figure 2



Source: Neural Networks and Deep Learning, Michael A. Nielsen, Determination Press (2015).

Furthermore, "Learning algorithms" allow the machine to dynamically alter the weights until the correct output is found. It is the combination of this process with multiple layer models that underpins the basis of AI and what is known as Deep Learning.

By now it should be clear how AI can combine positive attributes of the human mind (weighting evidence based questions) with modern technology (leading edge CPU + GPU) to produce unimaginable results.

Perhaps the best known is Google's Deep Mind beating Lee Sedol at strategy board game Go in March 2016. In the second game, the machine made a move leaving commentators speechless and forcing Lee Sedol to leave the room. It was a game-winning move later described by a tearful European champion, Fan Hui: "I've never seen a human play this move. So beautiful".

The Wall (part 1)

AI models require human input and large datasets in order to learn. This process is computer-intensive (GPU) and known as "training". Once trained, the systems can begin to answer problems – a process known as "inference" – and generate solutions to both programmed and un-programmed questions.

An excellent example is IBM's Watson for Oncology (Watson is the name of IBM's AI engine). The system was trained by oncologists from Memorial Sloan Kettering, one of the world's leading cancer hospitals, in combination with every academic paper and textbook on the subject (about 25 million in total).

Capable of reading both structured and unstructured data (e.g., a trial database and a doctor's handwritten note respectively), it also digests the additional 8,000 scientific papers written daily.

In a study of 1,000 cancer patients at the University of North Carolina (UNC), Watson recommended the same treatment as the doctors in 99% of instances. However, in 30% of instances Watson was able to recommend something new based on papers or approvals the human counterparts had not yet had time to read. Interviewed on CBS' "60 minutes", Dr. Ned Sharpless of the UNC said: "These were real things we would have considered actionable had we known about it at the time of diagnosis".

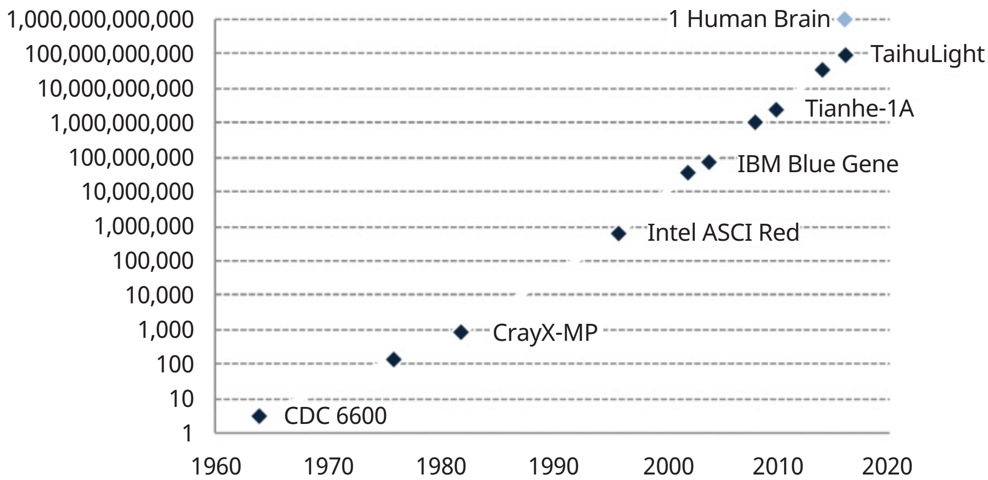
IBM will sell Watson for $250 per patient, opening the possibility of world-class oncology anywhere (not just the major metropolises) for a fraction of the cost and time to diagnose.

The Wall (Part II) – Moore's law a gating factor?

Multi-layer models – capable of auto-tuning, natural voice and image recognition – offer enormous scope of application. This will only widen if processors and technological architecture continue to progress as they have done in the past. Figure 3 shows the processing power of the world's faster supercomputer in the last 50 years, increasing at a compound annual growth rate (CAGR) of 59%. At this rate, an "exaflop" of processing power will be realized by 2021 – equivalent to our best estimate of that of a single human brain.

The observation made in 1965 by the co-founder of Intel Gordon Moore, known as Moore's law, that the number of transistors per square inch on integrated circuits has doubled every year since their invention, has gained increasing prominence in recent years.

Figure 3: Megaflops of world's fastest computers 1964 – 2016 (log scale)



Source: Schroders. Megaflop refers to a measure of computing processing speed, in terms of floating-point operations a second.

Critics, however, might point out that Moore's law has been slowing in recent years with the delay of Extreme Ultra Violet lithography (EUV) and that problems of quantum tunneling will impact further in the next few years. However, Moore's naysayers have been around as long as Moore's law itself, in fact it is sometimes quipped that the number of people predicting the end of Moore's law roughly doubles every 18 months.

Further, it is worth noting that while Moore has slowed for CPU, the addition of GPU (an architectural change to supercomputers) in 2010 has actually accelerated progress. Since 2010, the processing power of the world's faster computer has grown at a compound annual growth rate (CAGR) of 83% (vs. 59% for the past 50 years).

Investment implications – ex Technology

While natural to ask which industries and jobs will be affected by AI it seems more appropriate to ask which will not.

Initially, one of the gating factors is the availability of large datasets. AI models require these to produce optimal results hence their importance. It's no accident that early industry impacts are observable in video, image recognition and medicine – all fields with vast amounts of relevant, available data.

With sufficient data, successful early adopters will enjoy competitive advantage based on lower costs, time to market and insight. New entrants may even emerge if an industry is not moving fast enough. However, sustainable competitive advantage may need to come from proprietary solutions developed by in-house. If all solutions are merely bought from external vendors such as IBM or Google, speed of adoption would be the only differentiating factor. Internally developed or perhaps highly customized software may provide something more permanent. Winners could enjoy virtuous-circle benefits as the system improves with larger amounts of data. Data is scarce but Boeing, Toyota, Manulife and SEB (see below) are amongst a growing number of non-tech companies experimenting on their own in AI.

Nordic bank Skandinaviska Enskilda Banken's (SEB) adoption of Amelia – pitched by creator IP Soft (privately listed company in New York) as "Your First Digital Employee" – is an interesting example of the early adopter advantage. Figure 4 shows SEB's press release late last year:

Figure 4: Amelia to join SEB's customer service

06 Oct 2016 08:08

Amelia, a new digital employee at SEB

SEB would like to introduce Amelia, a new digital employee at the bank's internal IT support. By mimicking the human process of learning and conversation, the technology can be used to perform assignments and provide information. The next step is to let the bank's retail customers benefit from Amelias' skills as a complement to other services.

Source: SEB Corporate Website.

Amelia is a cognitive and emotive engine. This means "she" is capable of learning and reacting to different emotions in real-time (neural nets altering weights dynamically). If a customer is angry, her responses can change in real time to adapt accordingly. If she does not know the answer to a question, it is diverted to a human employee and the process learnt for future reference. It is estimated within three months Amelia will handle 60% of all incoming calls to the bank. Could she handle 99% within two years?

It will be instructive to see how Amelia benefits SEB in time. In theory, costs should fall considerably with similar or improving levels of customer satisfaction. Proven out, SEB would enjoy a material advantage forcing competitors to rapidly catch up. This dynamic is likely to recur throughout the world, especially in labor-intensive, white collar jobs with a high degree of standardization. Financial services would seem a strong candidate.

AI should not be thought of as the end of human labor, however, at least not yet. AI can free up human effort for more productive, creative uses which, combined, could materially enhance corporate and social prospects. IBM does not see Watson as a replacement for oncologists, for example, simply as "augmented intelligence" enabling doctors to spend more time on care and research.

The impact of AI, therefore, will not be generic. In many instances, implementation will enhance corporate profits and social impact. In other cases, the existing business model looks increasingly uncomfortable and the adoption process will be painful. An example might be the IT Services firms Cognizant and Infosys which both employ over 200,000 people each. While both are embracing AI and developing internal solutions of benefit, it is hard to believe the employee count is now optimal for the long term. As ever, we think they will face the perennial corporate problem of embracing change at the expense of the existing business.

Investment implications – Technology

Given industry nascence, it is hard to judge how software companies are positioned. We know Google was one of the earliest drivers dedicating significant resource (one of two quantum computers on the planet is dedicated to AI at Google) along with leading talent (Ray Kurzweil is Head of AI) and acquisition (Deep Mind). They would therefore seem very well placed though hard, tangible evidence beyond the Autonomous Vehicles and Alpha Go successes are not yet visible.

IBM appears to have the greatest lead in the corporate space, a position which largely remains unchallenged today with competitors focusing on the consumer instead. In time, IBM could own one of the most valuable assets in technology but we believe investors have two issues to consider:

- IBM will need to cannibalize some of its existing business – historically, its record here is poor
- IBM will spend a lot of cashflow buying expertise to train Watson. In 2015, it paid $2bn for the Weather Company and many more are likely to follow. While this may prove sensible investment in time, the amounts are significant and create a wide range of success/failure for the investor

Silicon offers investors perhaps the most tangible opportunity to participate. The addition of GPUs to server architecture for AI training is creating a significant boon for industry leader, Nvidia and its foundry partner, TSMC. Revenues from chips to data-centers grew 193% year on year (yoy) to $240 million last quarter and industry estimates put GPU penetration below 1%. In time, this figure will likely grow to at least 10% based on commentary from the hyper-scale web companies, implying > 10x unit growth opportunity.

Nvidia's momentum may therefore continue unabated for the next few quarters. With the stock on 24x (enterprise value, on net operating profit after tax for calendar year 2017), the stock can still attract buyers, assuming the gaming division (63% sales) maintains momentum.

While Nvidia and TSMC propel the GPU market with experts believing they have a one or two year advantage, Intel has formulated its response. The new Xeon Phi architecture is specifically designed for highly parallel workloads and in H2 2017 will integrate ASICs (Altera acquisition), 3D X-Point (proprietary memory) and Nervana (software allowing silicon programmability for AI). It has promised 10x out-performance of GPU solutions by 2020 based on this architecture.

With Intel stock currently unloved, a successful rebuttal of the Nvidia challenge could make it one of the best performing large cap tech stocks in the next three years. Intel is trading at the lower end of its 10-year relative PE (13x absolute PE) and margins are potentially depressed from reinvestment (Intel earns operating margins 10% points lower than peers, TSMC and Texas Instruments). History and most of the world's software (written for x86) is on your side, too.

It is hard to compare chips properly – all silicon companies have their preferred benchmarking which always shows their solution in a positive light. But it does seem like Nvidia may have left something of a pricing umbrella for Intel to compete with. GPUs are the only solution today but typically cost $6,000 each. Adding four to a server for AI therefore adds $24,000 to the cost compared to perhaps $2,000 for the CPUs. This pricing gap may explain why Nvidia is valued at $57 billion yet accelerator penetration is less than 1%.

Hyper-scale web companies may also chip away. Where cost will become prohibitive using third party solutions they have tended to invent their own (switches, routers, servers, storage, databases etc. etc.) Indeed, Google has already invented a processor called Tensor while Amazon bought Anapurnalabs (an Israeli chip designer) in 2015. Google has also begun optimizing machine learning in recognition of the training intensity phase.

Comfortably numb

AI heralds a new era as prior distinctions of human and computer expertise begin to fade. The impact will not be homogeneous or necessarily detrimental. In our view, early adopters can benefit where sufficient data already exists while some labor intensive business models may face serious challenges.

The Bank of England's Chief Economist, Andy Haldane, said the third machine age may result in 15 million UK job losses (50% of employment) and a further widening of inequality. His latter point is well taken and government interference at the will of the people should not be discounted. However, technological progress has generally resulted in higher standards of living across the population and AI shows remarkable promise in fields such as terminal disease.

Likely to become one of the defining themes of our age, investors can ill afford to ignore its impact on any and every industry which they assess.

The AI basket

Intel	13x PE, depressed margins, entrenched technology and software, may dominate AI anyway
Nvidia	24x EV NOPAT, trading idea as momentum continues for 2017, worried beyond that...
TSMC	Nvidia's partner, also likely to manufacture for Google, Amazon, Apple etc. etc. 12.4x EV NOPAT
Google	Frightening resource, willingness and ability, 20x EV NOPAT (incl. $3bn losses "other bets")
IBM	Controversial given a troubled legacy but look like dominating the enterprise AI software space, 12x EPS

Companies shown are for illustrative purposes only and do not serve as any recommendation to buy or sell.

Appendix

From “Neural Networks and Deep Learning”, Michael A. Nielsen, Determination Press (2015)

Suppose the weekend is coming up, and you’ve heard that there’s going to be a cheese festival in your city. You like cheese, and are trying to decide whether or not to go to the festival. You might make your decision by weighing up three factors:

- Is the weather good?
- Does your boyfriend or girlfriend want to accompany you?
- Is the festival near public transit? (You don’t own a car).

We can represent these three factors by corresponding binary variables x1 ,x2, and x3. For instance, we’d have x1=1 if the weather is good, and x1=0 if the weather is bad. Similarly, x2=1 if your boyfriend or girlfriend wants to go, and x2=0 if not. And similarly for x3 and public transit.

Now, suppose you absolutely adore cheese, so much so that you’re happy to go to the festival even if your boyfriend or girlfriend is uninterested and the festival is hard to get to. But perhaps you really loathe bad weather, and there’s no way you’d go to the festival if the weather is bad. You can use perceptrons to model this kind of decision-making. One way to do this is to choose a weight w1=6 for the weather, and w2=2 and w3=2 for the other conditions. The larger value of w1 indicates that the weather matters a lot to you, much more than whether your boyfriend or girlfriend joins you, or the nearness of public transit. Finally, suppose you choose a threshold of 5 for the perceptron. With these choices, the perceptron implements the desired decision-making model, outputting 1 whenever the weather is good, and 0 whenever the weather is bad. It makes no difference to the output whether your boyfriend or girlfriend wants to go, or whether public transit is nearby.

By varying the weights and the threshold, we can get different models of decision-making. For example, suppose we instead chose a threshold of 3. Then the perceptron would decide that you should go to the festival whenever the weather was good or when both the festival was near public transit and your boyfriend or girlfriend was willing to join you. In other words, it’d be a different model of decision-making. Dropping the threshold means you’re more willing to go to the festival.

Obviously, the perceptron isn’t a complete model of human decision-making! But what the example illustrates is how a perceptron can weigh up different kinds of evidence in order to make decisions. And it should seem plausible that a complex network of perceptrons could make quite subtle decisions.

Schroders

WP-AIJG2017