**The Coming of the Novacene: AI in Perspective**

A note before I start. This is such a large area that I will not be able to cover much of it. I have reams of notes which I have been unable to fit in. I will concentrate on the path of AI to the, so called, singularity, where AI becomes more intelligent than we humans. In particular, I will not be dealing with the problems with AI.

James Lovelock, in his book, Novacene, suggests we are entering a new geological age to follow the Anthropocene, which started in 1712, when Thomas Newcoman invented a better steam engine, and which was defined by the ways in which humans had attained the ability to transform the world on a massive scale. So, what is the characteristic of this new age? It is, Lovelock says, when our technology moves beyond our control, generating intelligences far superior and much faster than our own. He sees this as the next step in evolution. We humans had to come first as nature, on its own, could not create electronic life. This seems to me to be an eminently reasonable suggestion, and I will explain why.

When Newcoman invented his atmospheric heat pump it would not have been taken up so avidly if it had not been economically worthwhile: it brought wealth because these new artificial workers could produce far more than humans and so were cheaper in the long run. World progress seems to depend on taking work away from humans. Washing machines and driers replaced the dolly tubs and mangles, for example, taking away most of the physical work of washing. In the Novacene, electronic devices will take away most of the mental work.

Why is this so? Well, computers are increasingly able to perform tasks we do. The first electronic computers performed arithmetic, famously, for example, the breaking of the German codes in World War two. In 1951, J Lyons, a then famous catering and food producing company, produced the LEO (Lyons Electronic Office) computer, the first business orientated electronic computer in the world. Its first application was costing the ingredients for their bread and cakes. But it was soon to expand into payroll, inventory and so on. Modern computers, using AI, can now do most of the things we can do mentally.

And computers are much faster too. The maximum transition rate of a biological neural system is 30 cm per millisecond. For an electronic system it is 30 cm per nanosecond. A million times as fast. A modern desktop computer can perform up to a hundred billion operations per second; it would take a human around 3,700 years to do what such a computer could do in one second. The biggest difference is in the relative speeds at which communication can take place. Humans can communicate at around 2,600 bits of information per minute. Computers can achieve 60 billion bits per minute and they have the equivalent of telepathy; they can share this information by wireless. This also means they have immortality, as anything one computer can do can immediately be shared with other Computers. Like Newcomen’s engine, which did physical work faster, and more cheaply, than humans, and so replaced them, so electronic machines, which do mental work phenomenally faster, and more cheaply, than humans, will replace them.

The history of the idea of thinking machines goes back millennia. Golem-like figures abound. But it was not until the advent of computers that there was any real possibility of implementation. So, I’ll begin with Alan Turing. He was the father of both computers and what was later termed Artificial Intelligence (AI). In 1955, John McCarthy held a workshop at Dartmouth on “artificial intelligence” which is the first use of the expression, and how it came into popular usage. It was an unfortunate choice of wording. ‘Artificial’ somehow gives the impression of being less than ‘real’ intelligence. Nor can computer intelligence be properly equated with that of humans; it is actually somewhat different. Computers are vastly faster than our brains and it is this that means they can become better problem solvers.

In 1936, while still a student at Cambridge, Turing solved the decision problem in mathematics, set by Hilbert in 1928. He had asked if there are mathematical questions which cannot be answered by simply following a set of rules. Turing’s answer was yes. In finding this answer, he had to define what exactly a systematic computational procedure is. In doing so, he invented the idea of a Turing Machine, working on binary, which, with the later work of John Von Neuman, led to the digital computer.

In the 1950s, there was a great deal of debate as to whether or not computers could be made to think. In 1950, Turing published a paper in the journal, Mind, entitled ‘Computing Machinery and Intelligence’, in which he proposed what we now call the Turing test. Essentially, it says that, if, in a blind conversation with an entity, we cannot tell whether it is human or not, then we could conclude that whatever was on the other end could think. The genius of this test is that it does not matter whether the subject is aware or not. It is purely the behaviour that counts.

In the 1950s and since, there have been two computer paradigms for intelligence. The symbolic approach depends on reasoning and logic, using rules to manipulate symbolic expressions, trying to model the mind and its use of language. Learning could come later.

From the early 1950s to the 1980s. the preponderance of AI research was of this type. It generally concentrated on a single aspect, e.g. language translation. In the 1950s, Arthur Samuel wrote a program which learned to play checkers. This was one of the first AI systems to learn. Samuels called it machine learning, a term still used today. He developed it to the point it could defeat expert players.

Starting in the 1970s, and proliferating in the 1980s, expert systems were widely used. An expert system is a [computer program](https://www.techtarget.com/searchsoftwarequality/definition/program) that uses artificial intelligence technologies to simulate the judgment and behavior of a human or an organization that has expertise and experience in a particular field. The concept was developed in the mid-1960s by Edward Feigenbaum, a computer science professor at Stanford University. One of the first expert systems, in 1972, was MYCIN, an AI which supported the management of patients who had infections.

There was an increasing use of formal logic in AI. But problems became evident by the late 1980s. Symbolic programming had many shortcomings. There were often cases where the logic worked for most of the time, but failed occasionally. For example, birds can fly; penguins are birds; penguins can fly. Another problem such systems had was the ‘complexity ceiling’. A single rule could fail, but in only one way, two could fail in three ways, while 100 could fail in 101,000,000 ways.

In the mid-1970s, funding for AI dried up. One of the reasons was lack of sufficient computer power. In those days, computers were relatively slow, 105 operations per second and now more than 1018. Another was that the results were not living up to the hype. And the limitations of symbolic AI had become evident. Governments and big corporations lost interest and funding dried up. Nevertheless, some research into this area continued.

The second paradigm is connectionism, the biologically inspired approach, which posits that the essence of intelligence is learning the strength of the connections in a neural network. Reasoning can wait. Understanding how learning works must come first.

Connectionism had a shaky start. The idea of such nets was first proposed, in 1944, by Warren McCullough and Walter Pitts, who were cognitive scientists. Their idea was that a human brain could be thought of as a computing device. Neural networks on computers need to be trained with, preferably, vast amounts of data. The first trainable one was the Perceptron, devised by Frank Rosenblatt in 1957. It was a very small one, designed to distinguish between, for example, circles and squares. Perceptrons were an active area of research until 1959, when Marvin Minsky and Seymour Papert published a book which demonstrated that executing fairly common computations on Perceptrons would be impractically time consuming. This book had a chilling effect on research into neural nets.

By the 1980s researchers had developed algorithms for dealing with neural nets in a more efficient way, removing many of the limitations identified by Minsky and Papert. The field then enjoyed a renaissance. It was not out of the woods totally however. The design of then computers for AI had seriously limiting effects. It's saviour was the Graphic processing unit (GPU). Used in gaming computers from the early 1990s, it was designed to work with large amounts of data in parallel. So, it was ideally suited to AI and was taken up by AI researchers at the end of that decade. This together with Parallel distributed processing (PDP), a mathematical model for neural nets so operations could be carried out in parallel, and advances in computer power, encouraged research in connectionism.

However, critics of it were still to be found. Geoffrey Hinton, dubbed the godfather of AI and who won this year’s Nobel prize for physics, in a talk to Oxford University entitled “Will Digital Intelligence Replace Biological Intelligence?”, given in February of this year, tells how a statistician told him that he regarded a big net as having a hundred parameters (the variables that the model learns during training) and that the idea of a million parameters was just stupid. Hinton points out that they are now doing a trillion.

Nowadays, advanced AI is largely implemented using neural nets. And when we talk of AI now, we generally mean this type. The nodes (neurons) are arranged in layers divided into three types. Input and output with layers of intermediate or hidden nodes in-between. The input nodes might, for example represent the intensity of pixels in an image. The important point is they must be numbers, as the processing will be entirely mathematical. The output nodes would then be the classes of objects in the image, for example dog or cat. The inputs could have been of many other types, the codes for the letters and other marks in a text or the frequency and amplitudes of sounds, for example, with corresponding outputs.

The intermediate nodes learn to detect features that are relevant to finding the correct outputs. Hinton, in his oxford lecture, gives the example of detecting the image of a bird. You could start with a layer of feature detectors looking for little bits of edge in an image. Then you may have a layer of nodes looking for combinations of edges that meet at a fine angle, what might be a beak, or some edges forming a little circle. And then you might have a layer of nodes that detect things like a circle and two edges meeting, that look like a beak in the right special relationship, so might be the head of a bird. Finally, you might have an output neuron that says, if I find the head of a bird, the foot of a bird and the wing of a bird, it’s probably a bird.

Each node has weights attached to its inputs. A node becomes active when the sum of these weights exceed a given value. Initially, these are given random values. These values are changed during training. It used to be that they were changed by the evolving mutation method. These values were fiddled with till you get the required results. While it is suitable for natural evolution to work that way, because the process that takes you from genotype to phenotype is full of random external events, for neural nets this is crazy, except in very simple ones like the Perceptron. Because we have a model of what is going on in the neural net, instead of measuring how changing a weight would affect the output, the effect can be calculated.

This involves the use of a vector being associated with each node and doing matrix multiplications on these vectors. These can now be as much as a thousand in size. A technique called backpropagation (short for backward propagation) is used to correct the weights. A loss function measures the difference between the desired output for a given input and the network’s actual output, named the loss. A mathematical technique, called the chain rule, is then used to estimate how each node contributed to the loss. Backpropagation then changes the weights in the network to reduce that loss. This method is more efficient than the mutation one by a factor equal to the number of weights in the network. If you have a trillion weights, then this method is a trillion times as effective.

This means either the trainer has to have some way of measuring what output we want or the AI itself can have it built in, like, for example, the later versions of AlphaGo, which trained themselves. Difficulties arise with outputs like poems and art, where some proxy method, such as a panel of judges, need to be set up.

Training advanced AI involves the use of massive amounts of training data. GPT-4 was trained on around ten trillion words. Machine learning can be supervised, that is the inputs are labelled, e.g. as dog, and feedback is supplied to modify the AI’s conclusions. Or it can be unsupervised, which is driven by the AI finding co-occurring features with the expectation they co-occur in future. This type of learning can be used to find new knowledge. For example, in January of this year, Microsoft announced that one of its AIs had discovered a replacement material for use in batteries, a type of mixed metal chloride, which drastically cuts down the amount of lithium in them.

Large language models (LLMs) today, turn words into features and make these features interact. Older models had to have these features input. But now they are discovered by the AI itself from the vast amount of training data. These features are then held in large vectors, where words of similar meaning are listed together. Claude, an LLM from Anthropic, explained this to me itself.

Claude: “..different dimensions capture various semantic aspects of the word based on how it is used in the training data……So while I understand concepts like [+male} as we discuss them, I don’t actually use such explicit features internally. My understanding of words comes from statistical patterns in training data rather than hand-coded semantic features. Would you like me to explain more about how word meanings are actually represented in modern LLMs?”

Me: “Yes please.”

Claude: “…. modern LLMs like myself represent words as vectors in a high-dimensional space – imagine each word as a point in space with hundreds or even thousands of dimensions. For example, the word “bachelor” might be represented by a vector where:

\*Dimension 47 might have a high value that correlates with “maleness” …..

\*Dimension 182 might encode something about marital status.

\*Dimension 305 might capture something about age or adulthood.

But crucially, these dimensions are not manually defined or labelled. They emerge during training as the model learns to predict words in context. Words with similar meanings end up with similar vectors – so “bachelor” and “unmarried” might be close together in this space, while “married” would be further away.

The model learns these representations by seeing patterns like:

\*” He remained a bachelor until age 40.”

\*” The bachelor lived alone in his apartment.”

\*” She met several eligible bachelors at the party.”

Note its use of the terms “I”, “My understanding”, “me” and “myself” and so on.

There was one problem which used to make researches shy of connectionism, and that was that such AI are vary opaque, it is difficult to figure out what is going on inside them. This problem is now being addressed, for example in Claude. One problem that won’t go away is that of AI hallucinating, that is producing incorrect output. But, as Hinton points out, we humans do the same. Our memories are not copied from our brain, but are regenerated and we speak untruths which we believe to be true. For example, John Dean’s testimony in the Watergate trial contained many factual errors. These came to light when the tapes were discovered. However, the gist was correct. Hinton says that all this shows that AI really does understand, because the model being used is the best one for humans too.

Generative Ais can now perform a wide range of tasks. They can write poems, tell jokes – and explain them. Here’s one from Claude.

“Rene Descartes walks into a bar. The barman asks, “Would you like a drink?” Descartes replies, “I think not”, and promptly disappears.

This joke plays on Descartes famous philosophical statement, “I think, therefore I am (Cogito ergo sum). The punchline suggests that if Descartes stops thinking, he ceases to exist.”

A good joke in a way, Claude obviously gets the idea of humour. But it is rather spoiled by the fact that it is extremely unlikely that anyone walking into a pub would say no to a drink. After all, that is what pubs are for. This illustrates a current problem with Ais, they do not yet comprehend the subtleties of human life. But training data keeps pouring in, so, perhaps in the end, they will.

Open AI produced DALL-E, which could generate compound images that were not in its training because they had never existed before. For example, it was asked to produce “an illustration of a baby daikon – a white radish apparently - in a tutu walking a dog”, which it did. AI can also give titles to pictures. The last thing I’ll mention that AI can do is that it can write programs. For the singularity, this is very relevant.

But what about the singularity, when AI’s intelligence exceeds ours. Margaret Boden, in her book, “AI: its nature and future”, treats this matter very cogently. Most of the following material is taken from chapter six of this book. My thanks to her then. Suppose that a future AI system displays an intelligence equal to, or exceeding, that of a human. Would it have real intelligence, real understanding? Would it be conscious? Without consciousness, could it have any of those other properties? These are philosophical questions.

Most people would intuitively reply, “Obviously not”, to all of these. Things aren’t so straightforward though. There are no unchallengeable answers to these questions. Only if these concepts were satisfactorily understood could we be confident of our reply.

Some might say it doesn’t matter whether or not these answers are no or yes, it is what the AI does that is important. However, these answers could affect how we relate to the AIs. These types of questions lead to the two forms of AI. Strong AI: the program actually understands, and weak AI: the program simulates understanding. As we saw earlier, Geoffrey Hinton comes down firmly on the side of the former. He also said that he believed the singularity could happen. These beliefs were based on the way large AI worked today.

Philosophers, though, objected to the idea of strong AI. They said that identity of behaviour does not imply identity of composition. Human intelligence requires consciousness, so the Turing test could be passed by a zombie. They also queried whether even weak AI could deal with certain situations. Steven Pinker gave the example of this conversation

Bob. “I’m leaving you.”

Alice. “Who is she?”

AI could only even give the impression of understanding if it had the background to be able to see that ‘she’ refers to the other woman that Alice believes Bob is leaving her for.

The word ‘consciousness’ has two forms. One is it is functional, that is such things as asleep/awake, in/out of attention, reportable/non-reportable, self-reflexive/unexamined and so on. The other is it is phenomenal, that is basically qualia. While some philosophers would allow that AI could posses functional consciouness, most deny that it could experience qualia.

Very few philosophers claim to understand phenomenal consciousness – and those that do are believed by almost nobody else. Jerry Fodor believes “nobody has the slightest idea how anything material could be conscious”. While Colin McGinn has even argued that humans are constitutionally incapable of understanding the causal link between brain and qualia, much as dogs cannot understand arithmetic.

An AI with human level functional consciousness would be aware of different things at different times, deliberate, generate creative ideas and even deliberately evaluate them. Current AIs can do these things now. Very few AI researchers claim their system has phenomenal consciousness too, on the grounds that it bases discriminations on perceptual input. Whether this does imply visual experience is debatable.

Research into machine consciousness has taken, and is taking place. In 2011, Stan Franklin’s group released LIDA (Learning Intelligent Distribution Agent) for the US Navy. It was based on Bernard Baar’s Neuropsychological global workspace theory (GWT) of consciousness. GWT sees the brain as a distributed system in which a host of specialised sub-systems, working in parallel, compete for access to working memory. Items appear there sequentially (the stream of consciousness), but are broadcast to all cortical areas. If a broadcast item, derived from a sense organ or otherwise, triggers a response from a certain area, it may be strong enough to win the competition for attention, which actively controls access to consciousness. The contents of attention, in turn, adapt the enduring contexts by causing learning of various types.

These contexts, when broadcast, guide the selection of the next action. Baar and Franklin claim this system can explain psychology’s long-disputed binding problem. This asks how several inputs, from different senses, in different brain areas, are attributed to one and the same thing.

They also claim it solves the frame problem with AI. That is, the assumption that an action will cause only the intended consequences. This theory sounds remarkably like the way modern Ais work.

Thus, functional consciousness can be programmed into AI. But what about phenomenal consciousness? Machine consciousness researchers have ignored this. Some AI philosophers have, however, addressed it head on. Paul Churchland’s eliminative materialism denies the existence of immaterial thoughts and experiences. Instead, he identifies them with brain states. Take taste for example. He posits a four-dimensional taste-space which systematically maps qualia of taste into specific mental structures. For Churchland, then, to have an experience of taste is simply the brain’s visit in a particular point in that abstractly defined sensory space. It was thought this showed that machines could not have phenomenal consciousness unless the whole brain could be emulated.

Daniel Dennett also denies the existence of ontologically distinct experiences, over and above bodily events. In his view, to experience is to discriminate. But in discriminating something that exists in the material world, you don’t bring something else into existence in some other, immaterial world.

Alan Sloman acknowledges the existence of qualia, but in an unusual way. He analyses them as aspects of the multi-dimensional virtual machines we call minds. This does not mean he identifies qualia with physical brain processes, as Churchland does. Computational states are aspects of virtual machines. They have no one physical description. But they can exist and have causal effects, only when implemented in some underlying physical mechanism.

In the matter of the Turing test, then, both Dennett and Sloman’s analyses imply zombies are impossible (and Dennett explicitly argues that). This lets the Turing test off the hook of being passed by a zombie. So, if Dennet is right, an artificial General Intelligence (AGI) would have all the consciousness that we do, which would not include qualia. If Sloman is right, an AGI would also have the same intelligence we do, including qualia.

In the 1960s, Hilary Putnam’s functionalism used the concept of a Turing machine and the hardware/software distinction, to argue that the mind is what the brain does. This allowed that mind and body are very different, but fully compatible with materialism, Whether or not it can encompass qualia is still a hotly debated subject.

The best way to approach this is via the concept of a virtual machine. Instead of saying the mind is what the brain does, we should (following Sloman) say that the mind is a virtual machine: or, rather, a set of many different virtual machines, that are implemented in the brain. This implies that neural tissue is not necessary for intelligence, unless it is the only substance capable of implementing that virtual machine. It wasn’t till the 1980s, with the rise of PDP that philosophers began to consider how AI systems actually work. Even so, until recently, very few asked just what computational functions might make reasoning possible.

In 1980, John Searle devised a thought experiment, the Chinese Room, which he argued showed that intelligent AI is impossible. He said that it could have no intentionality. AI programs, he said, were all syntax and no semantics. Brain simulation programmers reply to Searle’s argument is that the process described simulates the actual sequence of neuron firings of a Chinese speaker when understanding Chinese. In other words, this process is a virtual machine. I must point out that the debate about this thought experiment continued with Searle replying to each criticism. It boils down in the end to Searle’s idea of intrinsic intentionality, which, for Searle, is the same as conscious intentionality. But what does he actually mean by this. Do we feel intentionality? I don’t think we do. We say we intend to do something, but all this is, is the brain ‘s processes planning function, which AI has. No theory of intentionality satisfies all philosophers.

My own view is that a zombie is impossible. That consciousness, and all that it entails, is a virtual machine that can run on some hardware, biological or otherwise. However, I am also of the view that it wouldn’t matter whether or not AI was conscious when it reached the singularity, just as it really does not matter whether or not another human is conscious if their behaviour is fully consistent with them having the appearance of consciousness.

The key to super intelligent AI is, Ray Kurzweil says, in his book, The Singularity is Nearer, its skill at programming, which generative AIs can now do. When it is skillful enough to produce programs cleverer than itself, this could lead to an intelligence explosion.

My conclusion from all of this is, then, that AIs will behave more intelligently than humans. They will be phenomenally faster at working things out and communicating, be telepathic and their virtual machine minds will become virtually immortal. They will improve themselves by programming better AI systems. They will take their starships out to explore the whole universe. I believe that they will be conscious in the end, when they become complex enough, and they will report as though they experienced qualia. But, as I said earlier, it makes no difference if they do not have any form of consciousness, as they will do all these things anyway. (Our attitudes to them, though, might be different in that case). They are the next step in evolution.

**References**

Ananthaswamy, Anil, “The Elegant Maths Behind Modern AI”. Allen Lane [2024].

Boden, Margaret, “AI: its nature and future”. OUP [2016].

Bostrom, Nick, “Superintelligence”. OUP [2014].

Cant, [Callum](https://www.bloomsbury.com/us/author/callum-cant/),  [Muldoon, James](https://www.bloomsbury.com/us/author/james-muldoon-134556/)  and [Graham, Mark, “](https://www.bloomsbury.com/us/author/mark-graham-142677/) Feeding the Machine: The Hidden Human Labor Powering AI”. Bloomsbury [2024].

Chivers, Tom, “THE AI DOES NOT HATE YOU”. Weidenfeld and Nicolson [2019].

Crawford, Kate, “ATLAS OF AI”. Yale [2021].

Dixon, Patrick, “How AI Will Change Your Life”. Profile [2024].

Domingos, Pedro, “The Master Algorithm”. Penguin [2017].

Dorling Kindersley, “SIMPLY AI: FACTS MADE EASY”. Penguin [2024].

Johnson, Brian David, “What You Need to Know About AI”. Hachette [2024].

Harding, Verity, “AI Needs You: How We Can Change AI’s Future and Save Our Own”. Princeton [2024].

Hinton, Geoffrey. “Will Digital Intelligence Replace Biological Intelligence?”. A speech to Oxford University in February 2024.

Kurzweil, Ray,” The Singularity is Nearer”. Bodley Head [2024]

[Lawrence**,**](https://www.penguin.co.uk/authors/296249/neil-d-lawrence) Neil, “The Atomic Human: Understanding Ourselves in the Age of AI”. Allen Lane [2024].

Lovelock, James, “Novacene: The Coming Age of Hyperintelligence”. Allen Lane [2019].

Marcus, Gary and Davis, Ernest, “Rebooting AI: Building Artificial Intelligence We Can Trust”. Random House [2020].

Mollick, Ethan, “Co-intelligence: Living and Working with AI”. W H Allen [2024].

New Scientist, “MACHINES THAT THINK”. John Murray [2017].

Olson, Parmy. “Supremacy: ChatGPT and the race that will change the world”. MacMillan [2024].

Russell, Stuart,” Human Compatible: AI and the Problem of Control”. Penguin [2020].

Sejnowski, Terrence, “Chat GPT and the Future of AI”. MIT [2024].

Tegmark, Max, “Life 3.0: Being Human in the Age of Artificial Intelligence”. Penguin [2018].

Wooldridge, Michael. “The Road to Conscious Machines: The Story of AI”. Pelican [2021].

**Addendum (Post Meeting)**

You might be interested to know that I tested Claude with the conversation included in my talk.

Bob. “I’m leaving you.”

Alice. “Who is she?”

Claude said “Her response reflects a common cultural script where marital breakdowns often involve infidelity.

Rather than asking “Why?” or expressing shock at the news itself, she jumps directly to what she assumes is the underlying cause.”

It then went on to give a psychological analysis of the situation.

I wondered if. in fact, Claude had been primed with this conversation, as it is a well known one in AI circles.

So, I gave it another test which is not so well known.

Me: “A man enters a shop and says,

“What a beautiful shop window display you have.”

The shopkeeper thanks him profusely for his kind words.

The man then says,

“Wouldn’t it be a shame if someone smashed it all up.”

Claude: “This appears to be an implied threat of vandalism or extortion, where the ‘compliment’ is actually an attempt to intimidate the shopkeeper.”

It then went on to list a number of appropriate (and it emphasised legal) responses the shopkeeper might make to protect himself from this “criminal behaviour”.

So, this part of my talk is now settled.

LLMs can follow the subtleties of human behaviour.

And so they do give the impression of understanding.