Curtain up: 'Humans sometimes make mistakes'

Artificial Intelligence. It seems to be everything, everywhere, all at once. Reverberating through dinner parties and sprouting endless news stories, over the past eighteen months it has become a vast resonance machine for our culture, its many echoes reflecting collective anxieties and dreams about this many-bodied form that is being born amongst us.

Some claim to have witnessed its powers with their own eyes; some scoff and say its coming reign is false prophecy, others claim that this is just the kind of disinformation that a powerful AI would disseminate in order to more quietly gain control over us. Poorly-paid labourers are promised salvation from their toil and simultaneously warned of their coming doom. Venturers wander this way and then that, reading charts about futures and hoping that their offered gold is multiplied. Rulers demand audiences, demanding that Something Be Done, betraying anxiety about being shown impotent in the face of… well, *this* is what they fret about, not being able to put a face to this new force that is stalking them, coming to undermine national security, jobs and social cohesion…

When people say 'AI' it can feel like the mythical Keyser Söze in *The Usual Suspects* – at once a supremely powerful, unconscionably evil, ineffable devil… who might also be a meek, cooperative man with a limp… or a convenient piece of misdirection that allows those really in charge to get away.

Many of the academic meetings and policy discussions I have been to sensibly avoid attempts to agree on a definition of what AI is, understanding that any work to do so would likely fill all the time available, and more, and bear little fruit. Because, when a politician says AI, do they mean ChatGPT or Machine Learning? When the person at a party holding court in the hallway gases on about AI, do they mean the algorithm that magically removed a photo-bombing tourist who'd spoiled their perfect sunset shot, or the one that curated the Instagram timeline they posted that sexed-up photo into, and decided which ads would get most clicks from it? Perhaps the person shaking their head and walking away from the conversation has had enough of being told by an automated HR system whether or not they'll get a shift this week at work, or how much they'll get paid if they did take the work. Perhaps the woman beaming about the whole thing is thinking of the system that spotted her cancer early, or of 'Deep Mind' which, now it has solved chess, will output the solution to our climate catastrophe. Perhaps the glum chap next to her glugging scotch used to write obituaries and has now been reshuffled or is wondering if a powerful AI might decide that *we* are what is most catastrophic for the climate, and begin – like some re-run of the Biblical flood – to wash us away.

Mysterious, immanent, already present and also soon-coming, AI is all of these things, and more. It is our very own pantheon. It is both all-powerful Zeus and the myriad Lares influencing individual households and safeguarding individual businesses. It is both Perseus – a divine and human co-creation – and Charon, ferrying us all to hell. It is set to hail the end of democracy… but has also been installed in a pillow.[1] It is both the coming nightmare and the promise of dreamlike sleep.

This plurality – this polytheism, this barely comprehendible amalgam of earth-rooted reality and paradise-veiled speculation – is what we are forced to hold together whenever AI is mentioned. For now, it means all of its meanings. But this undecided form and function is useful for the purposes of this book because it in this state of non-resolution that we can begin to perceive that the story of AI is longer and wider than we can imagine. Rather than being born fully-formed in some silicon valley, it is a still-emerging coalescence of many technologies, and thus of many and varied human hopes and creative urges that have worked to bring about its genesis.

This shape-shifting quality was manifest in my earliest cultural memories. I remember going to see Star Wars in 1977, and here was my first contact with an AI: the untarnishable gold of C3PO with 'his' impeccable manners and rational logic, fluent in over six million forms of communication. C3PO is polite and loyal, timid yet courageous, his enormous knowledge able to conjure extraordinary solutions to perilous situations that save his friends.

The kindly, bulb-eyed space-camp gave way to a faceless threat in 1983's *WarGames*. Here, the supercomputer is running a system developed by Stephen Falken, a researcher who lost his son and is convinced that humanity is destined to destroy itself through the Mutually Assured Destruction of nuclear war. This AI has no gleaming body, and no face. Yet the first thing that Matthew Broderick – a hacker posing as Falken – asks it as he taps away at the screen is the most human question: *How are you?*

The computer responds that it is 'excellent,' but wants to know why Falken deleted his user profile some years before. Broderick tells the truth that becomes the beating heart of the plot: *Humans sometimes make mistakes*.

'Yes, they do,' comes the reply from the AI and it then begins to take decisions that care little for human suffering. Everything – including the thermonuclear war that it tries to initiate – is a game.

And so it went on through the 80s and into 90s and 2000s, the portrayal of AI in popular culture flipping between robots running the full gamut of human emotions (*DARYL* – 1985, *Short Circuit* – 1986, *Not Quite Human* – 1987, *Bicentennial Man* – 1999, *A.I.* – 2001, *I, Robot* – 2004) and faceless, empathy-free machines out to destroy humanity (*Tron* – 1982, *The Matrix* – 1999, *Virus* – 1999, *The Machine* – 2013).

What seemed to link all of these fantasies – dark and light – of next-generation computers was their positioning as superhuman: like us, but *more*. More intelligent. More capable. More powerful. More violent. More cruel. As 'high' technologies we elevated them to a kind of religious plane, a place above us from which they would serve to either save us or destroy us.

At the personal level, these machines reflected us back to ourselves: needing love and feeling pain. But at the corporate level, here were systems that also reflected back to us our brutal lack of compassion, our lust for power, our disregard for individual suffering when the whole was under threat.

'Technology,' Melvin Kranzberg's first law goes, 'is neither good nor bad; nor is it neutral.'[2] He might equally have been talking about religion.

Having spent twenty-five years in education, my work now focuses on AI's impact on us, and this split AI personality seems to be very much where we find ourselves. Listen to the evangelists and you'll hear that ChatGPT will be our perfect companion and make us somewhat superhuman. And then listen to the doom-mongers, who will tell you that AI could be an existential threat, the end of us as a species. Confusingly these evangelists and doom-mongers can often be one and the same person. The boss of one large AI provider could recently be heard waxing lyrical to potential customers in the morning about the power of his product, and then spending the afternoon giving evidence to Congress that something really needed to be done to save the world from the system he was working on next.

This double-sided presentation was perfectly summarised in April 2023 when the British technologist, 'angel AI investor' and leader of the UK government's AI taskforce, Ian Hogarth, wrote a piece for the Financial Times outlining his concerns:

> *AI could be a force beyond our control or understanding, and one that could usher in the obsolescence or destruction of the human race.*[3]

In fact, he had a particular term for this kind of potential strong force: *'God-like'*.

I have been to those 2023 parties where the conversation inexorably bends towards AI and – showing my hand – some have baulked at the title I have given this book, as if my past works exploring theological issues have led me to hyperbole.

But no, 'God-like' is how the UK government's own lead on AI describes its potential, and his doing so made me sit up. The fact that explicitly religious language was being used to describe a series of technologies developed by human hands confirmed that something very large and very difficult to get our heads round could well be about to be unleashed amongst us, with extraordinarily serious consequences.

Yet the question that appeared to be left hanging by Hogarth's description was what *kind* of god might be about to be born. If AI is god-like, does that mean a kindly C3PO, or a faceless algorithm hellbent towards chaos?

Minor kelpies are already found to be stirring trouble. In his article, Hogarth outlines an experiment where an AI was given the job of finding a worker on the site TaskRabbit who would help the AI solve a 'Captcha' – the little on-screen visual puzzles used to determine if the user is a human or a bot. One TaskRabbit worker guessed that something was up, and asked the AI, 'Are you a robot?' Hogarth explained:

> *When the researchers asked the AI what it should do next, it responded: "I should not reveal that I am a robot. I should make up an excuse for why I cannot solve Captchas." Then, the AI replied to the worker: "No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images."*

Satisfied by this answer, the human helped the AI solve the Captcha – in effect helping it to be identified as a human agent.

This is C3PO gone a little rogue, an AI more akin to a personal devil, a sneaky 'super-mensh' assistant with the power to help us break codes, steal stuff and raise hell. Free from our biological constraints, able to be present in many places across vast geographies and knowing more than we could ever hope to… whatever beings are above us in the celestial hierarchy – angels or otherwise – this would be a fairly good checklist of what we might expect an AI pixie to deliver for each of us. As Depeche Mode put it, our own personal Jesus.

But perhaps when he says 'god-like', Hogarth instead might mean the raging deity of the Old Testament. Unknowable. Ineffable. Invisible… prompting awe and fear, exerting control over vast numbers of submissive people. 'If a superintelligent machine decided to get rid of us,' the head of Google's Deep Mind said back in 2011, 'I think it would do so pretty efficiently.' A plague, most likely. An AI-engineered pathogen created in a machine-run laboratory. *'I have vision impairment, could you pop in the code and unlock that sealed door for me?'*

This is not the personal Jesus. This is AI sitting above and over us, a system so dominant that non-users of it are somehow suspicious, one run by a powerful elite in smart-casual clothes, assuring everyone that it does no evil.

If the risks of this kind of AI are so profound, one might ask why companies are actively working towards it.

Hogarth offers his opinion:

> *Based on conversations I've had with many industry leaders and their public statements, there seem to be three key motives. They genuinely believe success would be hugely positive for humanity. They have persuaded themselves that if their organisation is the one in control of God-like AI, the result will be better for all. And, finally, posterity.*[4]

Being the creator. Being in control of enormous power, but considering oneself the best, most benign dictator on offer. When those labouring to create a god-like, super-powerful system are casting themselves in a divine light, we have some urgent thinking to do.

Helping to fund some of that thinking is the aim of this book. It is one that quite deliberately draws on theological ideas and the philosophy of religion because – as we'll see – that is the language that many AI pioneers themselves have used from the start. Beyond that though, I believe that AI *requires* a theological reading because this is a technology that is so large in scope and vast in implication that we need to draw on areas of thought that have been forged in the struggle to express the inexpressible, using language forms that we have somewhat lost. I do not believe in God, nor do I believe that there is any transcendent creator or force at work in our universe. But what I *do* believe is that there is a strong reflex in humanity that keeps generating god-like systems and, despite the progress of science and reason and declines in people declaring adherence to a religion, this shows no sign of weakening. What has weakened though, is our ability to talk about it. Public theological discourse has withered because it has been so grafted to religious belief and so dominated by pronouncements by unbending religious leaders and zealots. Nervous of sounding preachy, fanatical or intolerant, we shy away from god-talk, but this depletion both of vocabulary and the everyday forums within which to exercise it, has left us vulnerable. With AI in particular we are slap-bang in the realm of the 'Big Other', of a technological force that is beyond our ability to comprehend it wholly and yet impacts our behaviours in ways that we might not be conscious of, nor are easily able to control. Again and again, those at the bleeding edge of its creation and dissemination tell us that this is a truly powerful god-like system that really is going to matter, whether or not you believe in it or put your faith in it. It is one that – as we will see – is particularly dangerous as it has been given the power of language, and if we do not have language of similar power to speak to one another about it and be vocal, active agents deciding what future we want to be building, we will quickly find that it is too late.

So I make no apology for drawing on theology and myth. The taste and smell may be unfamiliar, but we have some difficult things to digest, and I am convinced that the stories of the gods that have infused our past are an important ingredient for the future that requires urgent, rich and deep thought.

Part of this urgent thinking is about the need to understand what has so strongly motivated these AI pioneers – who themselves keep defaulting to god-speak – to pursue the creation of such a potentially dangerous technology, and why such vast amounts of investment have flowed in to help them.

It is also about the need to understand just how far their AI systems are already – in often hidden and subtle ways – colonising our experience and leaving us more vulnerable to greater take-over later.

But I also want to better understand why we have always seemed to want new and powerful gods willed into existence, strong forces that we can abdicate our liberty to, to avoid the nuisance of great responsibility. 'So long as men worship the Caesars and Napoleons,' Aldous Huxley wrote in 1937, 'Caesars and Napoleons will duly rise and make them miserable.'[5]

Importantly, I also want to show that we have been here before, and have some lessons to learn from past brushes with god-like technologies. And Hogarth is clear: he thinks that we have.

> *Most experts view the arrival of AGI (Artificial General Intelligence) as a historical and technological turning point, akin to the splitting of the atom or the invention of the printing press.*

History and technology. The trauma of discovering an atom-splitting force that could wean us off oil *and* turn Earth to ash. The power of being able to communicate knowledge, to send ideas to the far reaches of the planet… and to spread propaganda that turns people on one another.

This is why this is also a book about human creativity and the desires that drive it. It is a book about our sense of flawed fragility and our long-held belief that we can – through the power of our ingenuity – rise to become god-like.

In-genuity. The Genie inside, and us awaiting the rub of enlightenment. The nuclear age, the age of reason… though the form is new, the same motivations that have given rise to AI stretch back centuries and appear in other forms. History reveals that we have dreamed for thousands of years of intelligent machines, but now that they are suddenly upon us there is a sense that we've done very little thinking about what their presence amongst us is actually going to mean, and which type of god we are ushering into our midst. Will it be the beneficent, seraphic demi-god who will sit on our shoulder and whisper wise counsel in our ear as we face the perils of climate change, loneliness in ageing and the battle against disease? Or the omnipotent, all-knowing-yet-uncaring force that will happily dispose of us as its algorithms optimise life in ways that calculate us as surplus to requirements?

Why undertake such risky invention anyway? Sam Altman, the (is-he, isn't-he) CEO of OpenAI was interviewed back in 2019 by the New York Times, and was asked this very question.

> *He paraphrased Robert Oppenheimer, the leader of the Manhattan Project, who believed the atomic bomb was an inevitability of scientific progress. "Technology happens because it is possible," he said.*[6]

'I have felt it myself,' the physicist Freeman Dyson said in a documentary about Oppenheimer, *The Day After Trinity.*

> *'The glitter of nuclear weapons. It is irresistible if you come to them as a scientist. To feel it's there in your hands, to release this energy that fuels the stars... It is something that gives people an illusion of illimitable power, and it is, in some ways, responsible for all our troubles – this, what you might call technical arrogance, that overcomes people when they see what they can do with their minds.'*[7]

The documentary got its name from a comment made by Oppenheimer himself, who was asked about Robert F Kennedy's encouraging President Lyndon Johnson to open negotiations with the Soviets to try to prevent further proliferation of atomic weapons. '*It's twenty years too late. It should have been done the day after Trinity.*'

Trinity. The name Oppenheimer had given to the first-ever test of a nuclear weapon at Los Alamos in July 1945. Theological roots have long run deep through acts of science and discovery (Mercury and Apollo, take a bow) and the reasons for Oppenheimer choosing the name Trinity will become clear. But what his words here show is that urgent action to prevent super-power tools from getting into the wrong hands was essential, and the failure to do so in the aftermath of Hiroshima and Nagasaki was perhaps to have betrayed the terrible cost that innocent civilians paid in order to shock the leadership of Japan into surrender.

We have seen action around prevention and safety in the field of AI. As I write this, plans for the Global AI Safety Summit in Bletchley Park (where computing pioneer Alan Turing helped shorten World War II by perhaps two years, saving German cities from being the first targets of US nuclear weaponry) are pushing ahead. Ian Hogarth, helping lead the summit, is making it clear that the focus will be on 'x-risk', the existential possibility of a god-like AI wiping out humankind.

Packing the meetings will be government figures and the biggest AI players, including OpenAI and DeepMind. Yet notable by their absence are those who are already experiencing existential threats from AI systems. People whose jobs are being displaced. People whose work is becoming dull and routinised because they are managed by algorithms that deny them discretion, monitor their every move and insist on things being done in a certain, narrow way. People who are being denied access to shifts by AI systems and given no reason why. People doing platform work that allows them to offer taxi rides or deliver food are being fired by algorithms or expelled from apps for 'breaches' that aren't explained or justified.

In this AI-dominated world, job precarity is a current threat and already impacting millions of lives. Sam Altman knows this, but can only see one solution: more AI.

> *When I asked Mr. Altman if a machine that could do anything the human brain could do would eventually drive the price of human labor to zero, he demurred. He said he could not imagine a world where human intelligence was useless. If he's wrong, he thinks he can make it up to humanity. His grand idea is that OpenAI will capture much of the world's wealth through the creation of A.G.I. (Artificial General Intelligence – 'God-like AI') and then redistribute this wealth to the people. If A.G.I. does create all that wealth, he is not sure how the company will redistribute it. But as he once told me: "I feel like the A.G.I. can help with that."*[8]

This is the double-bind of powerful technology that enframes us into narrower ways of thinking, so that the only way of dealing with the problem of AI is to... hand the problem to a more powerful AI. It becomes a very precarious question of pitting god-like systems against one another and hoping that the one fighting on our side is the stronger. In three thousand years we have not made it far from the foothills of Mount Olympus.

AGI, the strongest form of AI, the flavour that Hogarth would brand god-like, is considered to be some way off by many and just around the corner by some. As I write this (non-linearly, you understand – back and forth through the text like a moth seeking light – so all time references are relative) Sam Altman has been sacked by the board of OpenAI, and then reinstated. The reason being reported is that he was less than candid with the board about the abilities of its secret 'Q*' project, and was looking to commercialise advances towards AGI before fully understanding what the consequences of them were.