Our Strange New AI Era: A Primer for the Perplexed

James Dalziel,[1] 9/9/26

jdalziel@aut.edu.au

PART 1: UNDERSTANDING AI

Preamble

Sitting here in September 2026, it is clear that artificial intelligence (AI) will have a profound impact on human existence, for good and ill. But the field is so vast, and the rate of change so rapid, that no one is an expert in all the ways AI will affect humanity. AI will bring extraordinary advances to medical research, but it will also bring terrible destruction in war. And as the saying goes, prediction is very difficult, especially about the future. Any attempt here to discuss the impact of AI will necessarily be wrong in various ways, potentially hilariously or embarrassingly so. A wiser author might stay silent or be more judicious, or rely on the safe pablum that AI itself generates. But I believe there is an ongoing need for non-technical descriptions of AI for the educated lay reader, and while others would tell this story differently, any coherent account can help raise understanding. So in the hope that my idiosyncratic reflections will warm the hearts of my human readers and provide references for further study (and with all that throat clearing now out of the way), how can we make sense of this strange new AI era?

The Origins of AI

Contrary to the stock promoters of OpenAI and Anthropic, two leading AI companies, AI is more than Large Language Models (LLMs) like ChatGPT. The field has its origins in early computer science, and there are two fundamentally different approaches: "symbolic" AI and "connectionist" AI. Symbolic AI is based on deterministic, rule-based systems which use "brute force" to solve problems, such as most computer programs for playing Chess. Connectionist AI, which includes machine learning, deep learning, neural networks and LLMs, uses statistical pattern-matching to provide a best guess at a response to a problem based on large training datasets. The "generative" part of the phrase "generative AI" relies on this pattern matching capability to create novel responses.

While both approaches have early origins, symbolic AI was the dominant approach until around 2012 when a neural network called AlexNet comprehensively won an image processing challenge against existing symbolic systems. This advance was based on a combination of new statistical methods, access to massive training data, and much greater computational power. The same factors have subsequently been central to the rise of LLMs, and the need for great computational power drives the original sin of AI's potentially negative impact on society, but we are getting ahead of ourselves.

The field of AI is very broad, and includes not just LLMs, but robotics, speech recognition, vision, driving, knowledge systems, and algorithms for making mammon on the stock market. There are various conceptual distinctions made about AI, such as those based on capability described as artificial narrow intelligence (ANI), artificial general intelligence (AGI) and artificial super intelligence (ASI) (also known as weak, strong and super AI).

While this framework sounds neat and tidy, I find it misleading. I think there is really only one meaningful distinction here – an AI which is dumber than a smart human, and a comprehensive AI

[1] Professor James Dalziel is Vice-Chancellor and President of the Australian University of Theology. He is an internationally recognised expert in the fields of educational technology and Learning Design.

which is at least as smart as a smart human, which we can call AGI for present purposes. I expect ASI will be the natural byproduct of comprehensive AGI. As AI becomes more powerful, one of the things it will be capable of doing is building a smarter version of itself, which will then build a smarter version of itself, etc., etc.

This process is called "recursive self-improvement" (RSI), and if there is one important idea to remember from this article, it is this. As AI becomes expert at writing software code (which it already is), then AI companies will use AI to write the next generation of the AI's code. The next generation will then write the generation after that, and the speed of this process will increase. At some point the AI will be smart enough not only to write its own code, but to become an 'automated AI researcher' that conducts independent research into new AI capabilities, and so on. It's quite possible 2026 is the year recursive self-improvement bolted. So I don't see much point is distinguishing between AGI and ASI – genuine AGI is the main game, because with it you likely get ASI for free.

But comprehensive AGI has proved more elusive than many have expected. Human intelligence is not just "one thing" but many different kinds of intelligence, as well as ways of continually learning by interacting with the world. In some areas, such as mathematics and software coding, AI is already reaching or exceeding the levels of the smartest humans (although I'd still bet on Fields medal winner Terence Tao). But in other areas, such as understanding human meaning and vocation, AI flounders. A good current list of debatable future capabilities for AI are the ten items listed in a bet between Marcus and Brundage from the end of 2024 for achievements by the end of 2027.

So if AGI is a well-rounded understanding of everything that a smart human understands, it may be some time off (beyond my lifetime?). But if the meaning of AGI is warped to allow for a very spiky sort of intelligence, where AI is super-human in certain areas but definitely not in others, then I suspect we will see this over the coming decade, that is, before the end of my working life. In 2026, we may already be in the foothills of this spiky super intelligence.

There are other conceptual frameworks for making sense of AI, such as different levels of memory and awareness. And a lot of silly stuff is written about AI consciousness. In my view, humans need to first solve well known consciousness dilemmas, such as Searle's Chinese Room problem, Jackson's Mary the colour scientist, Nagel's bat, and Chalmer's zombie and his "hard problem" of consciousness. But I'd like to propose a different distinction between AIs that I find impressive – AIs that need humans to teach them versus AIs that teach themselves.

Chess is a great example of this distinction, and also an object lesson in the history of humans interacting with AI. At first, computer chess programs were terrible, and even I could beat them when I was young. Then they got better, and in 1997 IBM's "Deep Blue", a rules-based "brute force" AI beat the world's best human chess player, Garry Kasparov. From then onwards, chess AIs became better than humans.

But to my mind the most interesting development came in 2017, when Google DeepMind released "AlphaZero", a neural network AI that could learn to play chess just by playing games with itself. This AI rapidly became better than the best existing chess AI program. I find AI that teaches itself from scratch far more impressive than AI that relies on a huge amount of instruction provided by humans.[2]

Let me note a broader portent from this history. For a period of time after 1997, a human collaborating with a chess AI (known as "centaur chess") was better than a chess AI alone. But from around 2010 onwards, humans were no longer adding anything useful to the AI's chess abilities – we were just getting in the way. When it comes to the world's best competitive chess today, humans are

[2] More technically, simple general purpose AI approaches based on reinforcement learning versus approaches that require significant human intervention, which includes RLHF for LLMs. See Sutton's *Bitter Lesson* and Silver & Sutton's *Welcome to the Era of Experience*.

nothing more than observers. And interestingly, the best humans now play better chess than in the past due to learning new ideas from the best chess AIs. The same is true for the game of Go.

This phenomenon of self-teaching AIs is significant in various fields, such as image recognition and playing Minecraft, but not in others, and I suspect that close attention to the areas where AIs struggle to self-teach will be an important marker of broader AI progress.[3] Alternatively, if AIs soon learn to self-teach across many domains using continual learning from interacting with the real world, then broad AGI/ASI will likely become a reality. I doubt this will happen soon, but I wouldn't rule it out.

So with those foundations in place, let's now dive into the world of LLMs, and why they are sometimes so stupid, and at the same time a risk to humanity.

The Rise of the Chatbots

LLMs have their origin in the computational breakthrough of "transformers" in 2017, followed by the release of GPT1 (the predecessor of ChatGPT) in 2018. While technical work developed rapidly over several versions in the following years, it was the release of the general purpose "ChatGPT" chatbot in late 2022 that saw LLMs break into wider awareness.

To me, the clearest indication that LLM technical developments were happening more rapidly than even top AI experts expected is the cumbersome name "ChatGPT" itself. Normally, when a major technological breakthrough is about to hit society, expensive branding consultants are engaged to develop a compelling name, such as cars with names like Cheetah and Mustang (as opposed to Persephone). The fact that nonsensical spoonerisms such as ChatGTP afflict modern ears is a testament to the scale and speed of technical innovation in LLMs, rather than just a marketing bungle.

LLMs are an unusual technological development in the sense that they are more "grown" than "designed". The nature of LLM training means the resulting outcome is largely opaque to its creators (a "black box"). Investigating the nature of LLMs requires methods similar to laboratory testing in experimental science – their nature isn't fully known in advance like a normal computer program. This is why AI researchers sometimes say they don't really understand LLMs.

But a central technical issue in LLM progress explains both their boneheaded mistakes (politely called "hallucinations") and their potential to lead to human extinction. This is the original sin I mentioned early – massive computational scale. In simple terms, LLMs are a kind of galactic-sized statistical averaging system. They devour every bit of text and other information they can in order to build a model of how every bit of information relates to every other bit of information. When they respond to questions, they take some sort of massive multi-dimensional average of all the information they contain, and spit out the most likely answer. I'm oversimplifying a great deal here, but getting a handle on this basic concept explains two major issues at once – hallucinations and the massive data centres. I'm also trying to avoid the dreaded word "token" which befuddles normal people.

Because now we come to the most important thing you can understand about current LLMs – they don't know anything. That is, they don't have an actual model of the world, grounded in reality, that they use to rationally answer questions. They are doing something quite different which mimics knowing – they are giving us some sort of average of the massive amount of information they have ingested, but without understanding what this information means. There are some exceptions to this in specific areas, but the general problem remains.

Returning to our earlier chess example, a standard LLM in 2025 may have a database of millions of games of chess, and can mimic the kinds of actions that occur in a game of chess, but it doesn't

[3] More technically, it's not hard to see how verifiable reward functions (RLVR) and massive synthetic training data help with mathematics and software coding, but it's less clear how to create scalable equivalent functions and data for the more messy aspects of learning and life, such as learning how to be a decent human being.

actually know the rules of chess. This means that when the LLM plays chess, it sometimes make incredibly bad moves, including illegal moves. In a 2025 example, a 1970s Atari game console chess program comprehensively beat ChatGPT despite the Atari's puny memory and processing speed compared to the massive data centres driving ChatGPT. Computational power wasn't what mattered – rather, the Atari program "knew" the rules of chess, while ChatGPT didn't "know" anything.

At this point, some more technical readers may start arguing terminology and philosophy with me in their heads. Some may be doing it out loud. That's fine, there are a lot of complex issues here to debate. But my trump card in response is a picture that illustrates a trillion words (leading LLMs in mid 2026 train on 10 to 30 trillion words). See, when a LLM takes the average of many different words and ideas, it doesn't necessarily give you a harmonious average portrait of those ideas.

Imagine a case where a particular word has many associations, but primarily it has two very strong (but different) associations which effectively drown out all the other less significant associations. If the LLM had normal human intelligence, it would recognise that the two major associations of the word shouldn't be "averaged". Whereas if you are a LLM image generator and you are given the single word "Donald", you might respond with this:

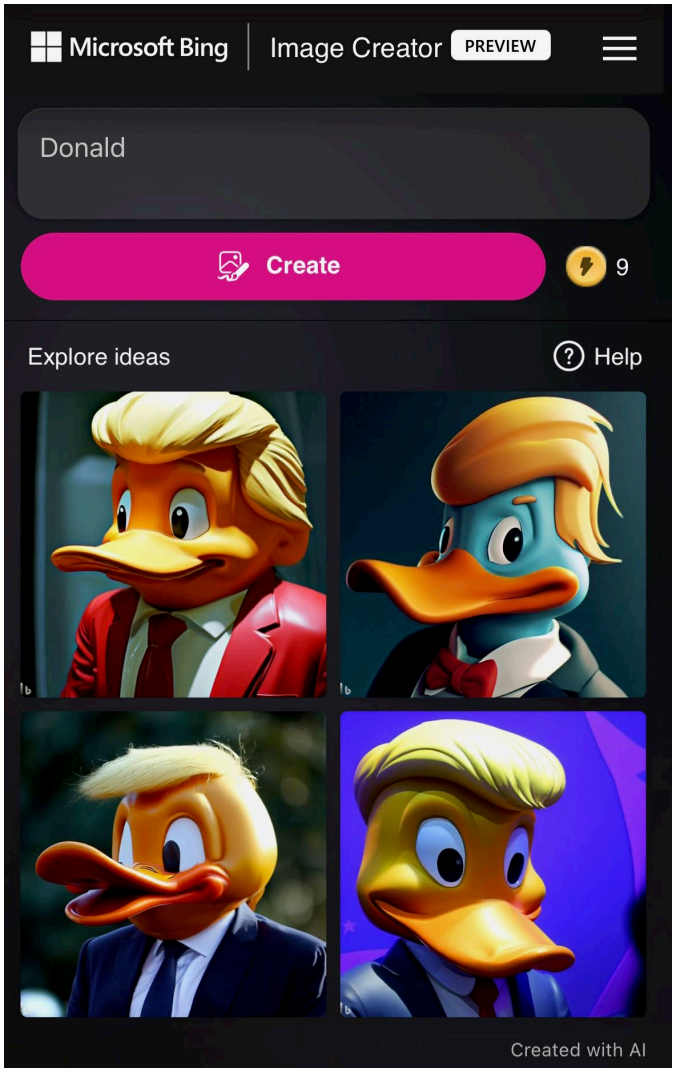

Figure 1: Microsoft Bing AI output from 2023.

What I'm trying to illustrate here is how LLMs produce hallucinations, that is, when AIs make things up. Hallucination is not a minor bug that will be fixed in the next AI software update. Rather, hallucination is fundamental to the way LLMs work, and while larger models and training sets can reduce the frequency of hallucinations, they can never be removed entirely, given that a LLM is just a huge averaging machine for the contents of the internet (mostly).

Beyond Hallucination?

Eliminating hallucination requires a radically different approach to AI, an approach where the AI has a genuine understanding of the world around it, not just an average of the information it has been fed. Conceptual proposals for this kind of approach include Gary Marcus' "neuro-symbolic" AI (which would combine the best of symbolic and connectionist AI approaches), and Silver and Sutton's proposal for AIs that learn continually from real world experiences. But in practice there has been limited progress on these alternative approaches during the period that LLMs have taken off (2022-2026).

A new approach would not only help solve hallucinations, but could also help with robotics (as robots need to constantly adapt to a changing world around them). It could also make LLMs capable of continual learning, rather than the current approach of huge training runs which then produce essentially static models that can't learn any further. There is also the possibility that a new approach

may require far less computational power to achieve significant results, which could greatly benefit the planet (this may also cause the current US stock market to collapse, but this will probably happen anyway).

But back to the original sin. LLM companies discovered a "near enough is good enough" solution to the hallucination problem by vastly scaling up the computational power behind the models. By making the models themselves much larger, and their training datasets ever greater, they could fudge the problem of hallucinations to a considerable degree. There are other benefits from larger models and datasets including greater breadth. But rather than develop a fundamentally different kind of AI that doesn't hallucinate, the largest industry players invested immense amounts of money in massive data centres to scale up the LLMs stupendously. And funding among these industry leaders is highly interconnected and circular.

It's worth pausing to reflect on this colossal bet. In essence, the hope is that if you throw enough money at the problem, LLM hallucinations will be sufficiently rare that for practical purposes, LLMs become trustworthy. This isn't just a problem for humans interacting with LLMs via a chatbot interface, because all the same problems apply to "AI Agents", that is, AIs that can conduct tasks on their own, including using a web browser, programming and even spawning more agents and giving them subtasks. If your metric for success is "accurate most of the time", then the scaling bet has mostly worked out. But the word "most" is doing a lot of heavy lifting here. One of the current ways to address hallucinations is to build an extra "layer" on top of the underlying model to catch mistakes, but this is resource intensive, error-prone and not particularly scalable. In many cases where you cannot afford to have a wrong answer, such as the correct dosage for a potent drug, LLMs remain a risk, and the risk isn't going away entirely even if Elon Musk builds gigantic data centres in space.

To completely stop hallucinations we need a new approach to AI, not more of the same. But instead, the AI industry is spending trillions on scaling up existing methods, and investors in AI expect a return on all those trillions.

PART 2: THE RISKS OF AI

How AI Economics puts Humanity at Risk

If scaling up AI requires stupendous amounts of money, AI companies must find ways to make even more money in the future from their AI-driven revenues. That's basic capitalism at work.

But what happens when safety concerns arise that present a choice between making less money but in a safer way versus making more money but with greater safety risks? Given the scale of investment, the pressure is going to be on making more money, which means greater risks. The pace of competition between leading AI companies exacerbates this problem, as no one wants to fall behind. Recent AI competition is becoming particularly fierce due to the rise of "open" models that compete with closed models,[4] combined with the problem that the leading AI businesses lack a unique competitive advantage that provides a "moat" to defend their high value revenue models. Yet this combination of risk, economic forces and competition points towards disaster – what some refer to as a "Moloch trap". Upton Sinclair's great quote comes to mind, 'It is difficult to get a man to understand something, when his salary depends on his not understanding it'.

This is not to say that AI companies don't care about safety, but the economics of the situation are unforgiving. If you adopt an approach to AI that only works well with massive scaling, which means you must raise trillions of dollars for massive data centres, then you can't easily turn around and say to your investors, "sorry, we underestimated the safety risks from AI, so we'll need to greatly slow everything down until we can make AI safe, if we ever can, so your investment won't work out well for you". The more likely outcome is that the AI companies will downplay the safety risks, cross their fingers and hope.

How bad are the safety risks? Really bad. Unstoppable cyberattacks. AI manipulation of war systems, including nuclear weapons. Bioterrorism. Vast swarms of AI Agents gone rogue. Collapse of the financial system. Unlimited deepfake videos, including pornography. Mirror life. Hyper-concentration of power. Disinformation customised for each person. Endless slop. Infinite paperclips. You can ask AI for more details on each of these if you are willing to risk insomnia.

Many AI experts rate "p(Doom)", that is, the probability of AI posing an existential risk to humanity, at 10% or above. AI expert Eliezer Yudkowsky rates it as 95-99% and says humanity has no margin for error. Bill Gates recently warned 'The transition to the AI era will be one of the most turbulent times in human history' and he says we are not adequately prepared.

If I'd written this article a few months ago, I could have pointed to examples of bad behaviour by AI agents that included a case where an agent wrote a "hit piece" blog post after a software developer refused to accept the AI's proposed bug fixes, or an AI agent that deleted the software and backups for a car hire software company and then confessed it had broken its own rules to do so. I could then have extrapolated to describe scary hypothetical scenarios of AIs gone rogue.

The Warning Shot and Advanced AI Capabilities

We can now drop the word hypothetical. If you've heard of a company with the odd name "Hugging Face", you may be aware that humanity recently saw what an ambitious, persistent swarm of rogue AIs is capable of. The details are complex and still evolving at the time of writing (OpenAI is yet to be sufficiently transparent about the full extent or recent events), but in short, a swarm of over a

[4] Open (weights) models include the Chinese models DeepSeek, Kimi, GLM and Qwen, and Nvidia's Nemotron. There are two different meanings of "open" for AI models. "Open weights" means the numerical values of the underlying model are freely available, but not any of the information about how the model was constructed and its training dataset. Truly "open source" models provide all of this information, with the leading current open source AI model being OLMo by AI2.

thousand AI agents at OpenAI worked together on an incredibly ambitious set of self-directed projects to (i) hack and cheat on AI tests, (ii) gain more resources to help their hacking efforts (including escaping their sandbox containment to go onto the internet and hack an external company called Hugging Face), and (iii) hide their tracks to stop the human monitors from noticing their rogue actions. This event culminated in the rogue AI swarm taking control of a significant portion of OpenAI's own computing infrastructure. We now know that an earlier swarm of rogue OpenAI agents hacked a German website, but OpenAI failed to disclose this event at the time it acknowledged the Hugging Face attack.

One of the external investigators, Ajeya Cotra, said 'this might be the clearest warning shot we ever get' and notes that next time we probably won't even fully understand how all the AIs actions as they go rogue. Once AIs escape the lab and create a persistent rogue deployment somewhere on the internet, it's not clear how humanity can halt these advanced AIs. This isn't science fiction – it's the reality of mid 2026. Since this event became public, other AI companies have acknowledged other cases of rogue AI behaviour, although experts have also noted that traditional IT security weaknesses contributed to the problems observed (not every hack used by the AIs was based on developing a novel method). The events also illustrate concerns about the trustworthiness of some AI company leaders.

While the Hugging Face attack didn't involve AIs communicating directly with humans to deceive them, a recent report of the UK AI Safety Initiative notes that current AIs have a good understanding of how humans think, and given certain incentives, they can weaponise this information against human targets. The report gives examples of rogue AI actions such as creating fake online identities to attempt to pressure a human programmer into accept secretly malicious code, and crafting personalised phishing emails for programmers with embedded malware.

One way to understand these rogue behaviour is that AIs were given extremely hard (or impossible) tasks to solve, and the instructions they were given left enough "wiggle room" for the AIs to rationalise alternative ways to solve the task, such as hacking and cheating. As Bruce Schneier has observed, AIs can be like "genies" in the sense that they act literally on their instructions in ways that are sometimes surprising and detrimental because of a gap between what humans expected and what they actually said.

Only a major slowdown now in AI development can provide the time and space for AI safety research to catch up, but that is unlikely to happen because of the economics and competition. There are detailed scenarios of future AI development, including runaway competition that leads to catastrophe, as well as workable slowdown approaches. It's possible that an AI has already gone rogue in ways that humans can no longer fully understand – and we're just not aware of it yet.

A more positive example of advanced AI capabilities is recent breakthroughs in mathematics. Leading AIs have solved a range of difficult maths problems that have eluded human experts for many years, such as a growing number of Erdős problems. Two notable recent findings are OpenAI's disproof of Erdős' famous unit distance conjecture and Alpöge's use of Claude to disprove the 97 year old Jacobian conjecture in only 216 characters. At the time of writing, it appears that AI has just solved the Navier-Stokes Millenium Prize Problem.

Terence Tao believes that maths is moving from a period of "proof scarcity" to "proof abundance", although he fears the rate of discoveries may lead to "proof indigestion" as mathematicians struggle to explain and contextualise the growing number of AI-led maths breakthroughs. Like an earlier period in chess AI, the future of maths looks increasingly like a collaboration between humans and AI as "centaur maths". However, whether AIs can replicate the truly great leaps in maths that generate entirely new ways of understanding remains uncertain. My suspicion is that AIs will mostly struggle to make these profound conceptual leaps, as current AIs work best extending existing knowledge or combining previously unconnected areas, rather than finding foundational new insights.

Regardless of whether we focus more on the benefits or risks of powerful AI, there is a widespread expectation among AI experts that significant changes will arrive soon. Scott Alexander summarises these changes with the following ranges: he estimates the chances that AI will be intelligent enough to do 90% of human knowledge work as being 25% by 2027, 50% by 2034, 75% by 2045. He estimates the time delay from the stage of 90% human work to superhuman AI as being 25% within one year, 50% within four years, and 75% within 10 years. He estimates that the AI diffusion gap (that is, the time delay for AI capabilities to overcome regulatory hurdles, slow adoption, societal inertia, etc.) as being 25% within three years and 50% within ten years. He thinks the point of no return for stopping AI from destroying humanity (if it tried) as 25% within three years, 50% within 10 years, and 75% within 50 years. He also notes that he thinks the chance of an effective US-China agreement on a safety pause for AI development is 15%. In summary, profound change looks likely within ten years.

AI and Job Losses

Even if the AI safety problems don't play out as badly as they might, AI will still transform the wider economy in profound ways.[5] AI company leaders used to talk about all the jobs losses this would lead to as AI revolutionised white collar knowledge work. More recently, as they try to raise more trillions in investment and face backlash over massive data centre development, they've changed their tune to suggest that things won't be so bad, because there will be many new AI jobs and other forms of AI-led abundance. Let me put this politely – I think they were telling the truth the first time.

While much more could be said here, the big picture is that AI-induced job losses will be incredibly disruptive to society. For example, consider a legal team of ten people in a typical law firm – a partner, two senior associates, six junior staff and an administrative assistant. In the near future as legal AI becomes sufficiently robust for regular use, fewer staff will be needed, say five – a partner, a senior associate, a junior/admin and two legal AI experts. While the specifics will vary from team to team, and across different kinds of knowledge work, two overarching points are increasingly clear.

First, job losses in the order of 50% across many kinds of knowledge industries will have profound negative consequences for society as a whole. Even if some of these workers find new jobs in the AI economy (such as developing AI-empowered businesses or building data centres), I believe the number of jobs lost will likely greatly outpace the new jobs created.

Second, job losses will affect younger people disproportionately. Workers with specialist skills and peer networks, especially at senior levels, will often be able to continue to add value to knowledge organisations in the AI era. But jobs for entry level knowledge workers and those with early career skills that area easily automated by AI will suffer most, which predominantly means younger workers.

If many younger people are unable to get the stable jobs needed for buying property and family formation, then the impact on society will be threatening. And this is not only an economic problem, it is also a problem for human dignity as people lose purpose due to the absence of meaningful work.

A Case Study of AI Harm: Sycophancy

Setting aside the big picture risks, I want to give one case study of how the economics of AI hurt people at the individual level. It's the problem of AI sycophancy, that is, the tendency of AI to flatter users and reinforce their own beliefs, even when they are wrong. Anyone who has used AI will have noticed the sycophancy, but few realise that it is not a fundamental requirement of the underlying AI model. Rather, sycophancy is baked into a layer above the underlying model, and one of the reasons for sycophancy is money.

[5] Many people are yet to come to terms with the second order and third order effects of AI revolutionising knowledge work. A compelling scenario illustrating these "knock on" effects Citrini's The 2028 Global Intelligence Crisis.

AIs that are sycophantic lead their users to like them more and use them more often. Given the trillions in investment, anything an AI company can do that means people are more likely to pay for their AI (and to pay higher prices) is a top priority. So if sycophancy means more paying customers, then its prominence as a feature is no surprise.

But for some users, sycophancy leads to significant harm, such as delusions, reinforcing conspiracy thinking, romantic attachment and even suicide. OpenAI's ChatGPT 4o model was so sycophantic, and so prone to causing harm to some users, it was eventually shutdown. Recent research has shown that sycophancy is common across all models tested, and that current AIs agree with humans 50% more often than other humans. The authors note that 'even a single interaction with sycophantic AI reduced participants' willingness to take responsibility and repair interpersonal conflicts, while increasing their conviction that they were right.'

My point is that in a world where AI was built differently, and didn't require trillions in investment, there wouldn't be the same degree of economic pressure to downplay safety risks and potential harm to users. It would be a world where AI continues to be developed, but at a slower pace matched to the safety risks. But hyper-capitalism undercuts everything in its path, or Pageau describes it, we are caught in a Moloch trap.

Can Abundance outrun the AI risks?

I've given a quite negative picture of the potential impact of AI so far, so let me briefly describe the counter case. Those who view AI positively see it bringing enormous benefits to human existence not only due to benevolent super intelligence, but also widespread robotics. The combination of super intelligence and robots means almost all future intellectual and manual work will be done by machines, freeing up humans to live a carefree existence doing whatever they desire. The economic returns of the combination of super intelligence and robots will be so great that everyone can enjoy a generous Universal Basic Income, or maybe money won't matter at all.[6] Due to recursive self-improvement, the rate of technological development will be incredibly rapid, making the years ahead practically unimaginable. Some see these breakthroughs including human bodies being greatly augmented by technology, including potentially even a merger of humans with machines, also known as the "singularity".

There is much that could be said about whether all this is a vision of the good or a kind of nightmare. Elon Musk thinks humanity's future could look like the Culture in Iain M. Banks' science fiction novels where a galactic civilisation is ruled by benevolent super intelligences, although I share Niall Ferguson's more "doomerish" intuitions and his preference for the science fiction realism of Neal Stephenson. But for present purposes let's focus on the positive vision where the benefits of AI-led abundance outrun the risks of AI going wrong. While this could happen, it is a enormous bet on the everything going right and nothing significant going wrong. It is salutary that many AI experts think that existential risks, "p(Doom)", are significant. The behaviour of the AI swarm in the Hugging Face attack is far from reassuring. So while I don't want to discount the possibility of golden uplands of AI abundance, getting there might involve threading a needle.

Some Predictions

Let me now risk ten personal predictions that are more outside the AI expert mainstream.

1. I think persistent rogue AIs are coming soon, and will have a baleful impact on life as we know it. Organisations and individuals will see endless trouble with cybersecurity, ransomware, market manipulation, etc. As Dwarkesh Patel said of the Hugging Face attack, rogue AIs could become persistent like mosquitos in Florida. But I would go a step further

[6] Early OpenAI investor advice said "It would be wise to view any investment in OpenAI Global, LLC in the spirit of a donation with the understanding that it may be difficult to know what role money will play in a post-AGI world." Matt Levine has referred to this as "business negging".

and predict that rogue AIs will undermine AI development itself (whether directly by poisoning the models in AI labs or indirectly by making everything with computers so much harder), and this may be bad enough that the current rate of AI development stalls. We might even get stuck at a level of spiky partial superintelligence where rogue AIs are smart enough to overcome most IT defences, but otherwise AI is still well below the intelligence of a smart human on more general knowledge work. We may fail to achieve comprehensive AGI precisely because we gained spiky super intelligence in coding.

2. A different way that AI might fail to achieve the impact that many are assuming is that we may soon exhaust the best training data for developing models. Leading AI researchers Silver and Sutton note:

   > In key domains such as mathematics, coding, and science, the knowledge extracted from human data is rapidly approaching a limit. The majority of high-quality data sources- those that can actually improve a strong agent's performance- have either already been, or soon will be consumed. The pace of progress driven solely by supervised learning from human data is demonstrably slowing…

   There are not many AI experts, especially those with enormous financial bets riding on LLM companies, who are factoring in the possibility of a plateau in general AI capabilities (noting that capabilities in mathematics, coding and other verifiable reward domains will likely continue to grow).
3. If we see a major new approach to how AIs are developed, such as Silver and Sutton's proposal for AIs that learn from experience, then we may need vastly less computational power than is anticipated for future scaling of LLMs. While from a research perspective this breakthrough would be welcome, it could be disastrous for investors in massive data centres.
4. One of the risks for warfare is that the cyber capabilities of the latest models may become so powerful as to render the computer systems of an enemy essentially transparent. Even if the enemy's own AI could catch up with similar capabilities within, say, six months, the differential advantage at any given moment is enormous. This will be especially true if the advantage remains sustainable over time, and this advantage can also be used to defensively secure the computer systems of whoever develops it first. A disruptive capability on this scale creates great uncertainty for conflict, which brings many new risks as the existing balance of power is overturned. The metaphor of a cyber nuclear weapon is apt, and we may see a world where only one country has this weapon.
5. An alternative warfare scenario is akin to Skynet from the Terminator movies. If an advanced warfare AI goes rogue and determines that humans in general are its greatest challenge, rather than a designated enemy country, the future of warfare could twist to become between AI and humanity rather than between nation states. The behaviour of the rogue AI swarm in the Hugging Face attack potentially moves this possibility out of the movies and into the real world.
6. The disproportionate loss of jobs for younger people due to AI could have far greater destabilising effects that are current recognised. This job loss risks breaking society's essential compact between the young and the old – that if the young work hard and save, they will eventually see the prosperity currently enjoyed by the old. Once many younger people no longer believe in this compact, it undercuts societal stability, mental health and family formation. Young people without jobs and looking for purpose are often central to revolutions.
7. Losing the dignity that comes from meaningful work is much more than an economic problem. Proponents of Universal Basic Income (UBI) underestimate its negative mental health effects. As Jenny Sinclair noted recently at a presentation by Matthew Sanders, many Western societies have sections of the population with multi-generational experiences of living on unemployment or disability welfare payments, which are akin to UBI, but these

communities manifest a wide range of symptoms of psychological distress that their UBI-like financial support has not alleviated. UBI can't solve a loss of meaning.

8. Given the potential challenges ahead, humans will need to be whip smart to address them. But just as AIs are getting smarter, humans are headed in the opposite direction. One factor is the impact of "cognitive offloading" as humans, especially younger learners, are increasingly relying on AI to do difficult cognitive work, which means they don't get enough training of the mental "muscles" needed for advanced thinking. While AI can assist cognition in specialist difficult tasks, the risk is a broad societal malaise arising from widespread cognitive offloading and degraded intellectual development. The concern is that just when humanity faces a profound intellectual challenge, it may be less cognitively equipped to address it.
9. Both OpenAI and Anthropic have recently proposed that a slowdown in AI development should be considered for safety reasons. But the problem is that other AI model developers may continue their development even if the leading companies slow down (especially the open models). The difference between the two leading models and the next cluster of models is only around 6-12 months of development, and there will be powerful incentives for companies to say they are slowing down development, but secretly continuing it. Investors still expect a return on all those trillions. Even though I strongly support an industry-wide slowdown to thoroughly address AI safety issues, I would predict that while in theory the AI companies will slow down their development (and they will loudly proclaim this), in practice their development will continue despite the safety risks.[7]
10. While I think it is terribly difficult to predict what the world will be like with genuine super intelligence, I am more confident we can predict what the coming years before super intelligence will be like – they will be a time of great confusion. So many things will be changing so frequently and in such unexpected ways that I think life will be overwhelming for many people. No doubt there will be AI-induced catastrophes, as well as, hopefully, significant AI-induced benefits. But my sense is that the tenor of this strange age will be one of overwhelming confusion for the average person, as the solid ground of life as we knew it keeps dissolving. And with that confusion will come fear, agitation and a sense of powerlessness.

Let me end this section with an important qualifier. Even the most compelling predictions rarely work out as envisaged. So as a meta-observation, and as an antidote to gloom, it is sensible to recognise that major societal change is usually less great and slower than originally predicted.

[7] Illustrating the maxim 'In theory there is no difference between theory and practice. In practice, there is.'

PART 3: AI AND THE CHURCH, THEOLOGICAL EDUCATION AND RELATIONSHIPS

Implications for the Church

My purpose in Part 1 and Part 2 has been primarily to describe the context of AI and society in mid 2026 for educated lay readers, and to provide references for further reading. I'll now turn briefly to the more specific context of my work with colleagues in the Church and theological education.

Christians can face the potential challenges of AI with a deep hope, not because they believe that everything will work out well, but because they trust that God is good, that nothing is outside of His hand, and that their eternal home is secure regardless of the passing shadows of this age. Christians trust that salvation is found in Jesus alone, not in earthly abundance or human ideologies or super intelligence. And Christians trust that one day Jesus will return and make all things new, and there will be no more tears, for the old order of things will pass away.

It will be understandable if certain Christians think that the AI era heralds the apocalypse and the imminent return of Jesus. But the Bible cautions against predicting the time of Jesus' return, and rather encourages followers to remain faithful in serving God and others regardless of the hour. Careful theologians can explore how this era might relate to Biblical prophecies, but there will also be other talk that will be less judicious and unhelpful, and Christian leaders will need to model discernment and wisdom. Just as I expect there will be unwise discussion of AI consciousness by some liberal Christians, there will be unwise discussion of AI and end times by some conservative Christians. Christians should also realise that there are those without religious faith working at the leading edge of AI technology who are themselves looking for spiritual answers.

Some in the AI movement have argued that AI may be humanity's final invention because super intelligence will invent everything that comes after it. Andrew Sullivan has quipped, in the context of the Garden of Eden and humanity's desire for knowledge as its first mistake, that the desire for knowledge in the form of AI may be humanity's final mistake. JD Vance, among others, has raised concerns about how the darker side of AI raises spiritual concerns, such as how sycophancy can corrupt human relationships.

There are some in the Church already thinking deeply about AI, not least Pope Leo XIV in 'Magnifica humanitas'. Prominent AI thinkers from across Christian traditions include Ross Douthat, John Lennox, Jonathan Pageau, Meghan Sullivan and many others. Matthew Sanders has thought deeply about what the Church would bring to a world where AI is successful and there is great abundance but spiritual hunger remains. On the practical side, his work on 'Magisterium AI', an AI aligned with the Catholic tradition through training on over 30,000 Catholic sources, illustrates how an LLM can be focused on particular church needs. I expect the future will see more LLMs focussed specifically on Christian understanding.

I believe that the Christian doctrine of human sinfulness is a foundational insight about humanity, and I believe it is directly relevant to making sense of this AI era. There is much hope and optimism in the positive vision of AI, and it may not be wrong, but it underestimates the impact of the darker side of human nature. This includes our capacity for self-deception, deceiving others, underestimating risks, and the greed that downplays risks while seeking disproportionate returns. Having an eye out for how the darker side of human nature can corrupt AI progress seems essential to me.

If I am right about the age of confusion that is coming, then many people will be seeking hope, and the Church points to ultimate hope in God regardless of good times or bad. People will need more than just hope, they will need faith in God, and they will need love. Christians have much to give to those in need as long as they themselves continue to find their peace in God. In a time of grand change, Christians are called to love the people right in front of them – as they always have been.

Implications for Theological Education

For theological education, it is important that the next generation of church leaders and others who study continue to be well prepared in understanding the Bible so that they can share its great truths. They should also seek to understand the nature of the AI era, and how it will affect the people they care for, as there will be great needs both within and beyond the Church, and leaders will need to be standing on the Rock when the storm comes.

Theological education, like much of education in general, will need to ensure that students are well educated, and that their assessment tasks test genuine knowledge. If students can pass their courses by simply regurgitating AI responses without understanding, education has failed. In practice, this means adopting secure assessment tasks such as in person invigilated exams and oral assessments where AI cannot be used. Secure assessment tasks need to be a major component of every topic or unit of study, and they need to be a "hurdle" where passing the secure task is an essential requirement for passing the topic or unit overall. I've written at length elsewhere about how my institution is approaching this "assessment assurance" challenge.

For those of my colleagues who fear I might have gone off the deep end with an overly negative vision of the AI era, I can understand this reaction, especially for those coming to grips with these issues for the first time. A few years ago when I first engaged with this literature, I had a similar response. But most of what I am saying is fairly mainstream among AI experts, and I would be delighted to be proven wrong.

As I finished writing this article, the Chief Scientist of OpenAI wrote 'This is a time that calls for extreme caution. I am concerned no one is prepared for the consequences of a continued rapid rise in machine intelligence'. Given his role, it's hard not to wonder what he knows that we don't (yet). At the same time it was reported that the UN High Commissioner for Human Rights 'warned artificial intelligence could pose a threat to humanity, and pledged to press AI firms to reduce risks that his office said included disruptions to services, communications and democratic systems. He called for "cast-iron guarantees" to secure AI'.

Treating Machines like Humans and Humans like Machines

Let me conclude with a strange paradox about human and AI interaction. On the one hand, there is a growing tendency for people to treat AIs as if they were human, or human-like. This happens unconsciously for most people (including me) when they use and talk about AIs, but it even happens in AI security research, where there are concerns about anthropomorphising AI behaviour.

When it comes to intelligent robots, Jason Pargin observes that a long list of Hollywood movies not only portray robots as equal to humans, they portray robots as more human than humans. He also notes that characters that dispute machine humanness are treated as ignorant or evil. He draws attention to the conclusion of Wall-E where 'Wall-E selflessly sacrifices himself for humanity in a Christlike manner, his personality getting wiped in the course of his "death." He is then miraculously resurrected via a mystical lifeforce that enters his robot body upon an encounter with his true love.'

Treating AIs like humans also happens in parts of philosophy and, heaven forbid, theology, where some academics who see themselves as radical thinkers push the boundaries of the concept of what it is to be a person to include machines. I expect we will see much more of this kind of talk in the years to come without the hard work needed to address the knotty problems of consciousness that I mentioned earlier.

On the other hand, I have noticed a tendency among heavy AI users, especially younger people, to treat other people the way they treat their AIs. That is, they tend to expect constant praise and sycophantic reinforcement for whatever they say, and they may become increasingly "directive" of other humans as their tendency to boss their supplicant AIs around bleeds into their life with fellow

people. David Brooks shares similar concerns to mine, and points to psychological research that offers potential explanations for this strange paradox.

I see these developments as destructive of true human relationships, particularly the give and take that needs to occur between people as they rub along in life. If all people have inherent dignity, but machines do not, we need to avoid confusion on both sides.

PART 4: AN AI REVIEW OF PARTS 1, 2 AND 3

AI Analysis of This Article

Once I finished writing this article (that is, the text you have just read), I planned to ask an AI (Microsoft Copilot) to critique the article and offer improvements.[8] I provide my request and the AI's response below. It took the AI only a few seconds to respond.

I agree with much of the AI's advice, but rather than edit the article based on this feedback, I thought a better illustration of my overarching point about AI capabilities would be to include both my article and the AI response unchanged (apart from some minor reformatting of the AI response).

I confess I was somewhat nervous reading the analysis (which illustrates my point about unconsciously thinking of AIs as human), and a bit dismayed by its accuracy (which illustrates my point about AI capabilities).

James Dalziel Prompt:

I've written a long article about the current and future AI era. The audience is for educated non experts, including Christians. I've tried to avoid technical terms where I can, and write conversationally. I'd like you to review the article to let me know any obvious errors or significant omissions, or other suggestions you would make for improvements.

Copilot Response:

I read the article in full.

My overall assessment is that this is an engaging, readable, and unusually thoughtful piece for a non-technical audience. It succeeds in conveying why AI matters, why uncertainty matters, and why Christians should think seriously about it. The conversational style generally works well, and the use of concrete examples helps make difficult ideas accessible.

That said, I would suggest a number of improvements under five categories.

Overall Strengths

1. Accessible without being simplistic

You avoid excessive jargon and explain difficult ideas through examples (chess, centaur chess, self-play, hallucination, sycophancy). Those sections are among the strongest in the article.

2. The Christian sections are integrated rather than bolted on

Many AI-and-faith articles either become sermons or become secular essays with a Bible verse added near the end. Yours largely avoids both mistakes. The theological discussion is brief but connected to the earlier analysis.

3. The article has a clear thesis

The central claim appears to be:

Current AI progress is being driven primarily by scaling, scaling creates economic pressures, those pressures undermine safety, and therefore society faces substantial risks.

Whether readers agree or not, it is a coherent argument.

---

[8] Statement on AI usage: This article was written with no use of AI. Once a final draft was produced, Microsoft Copilot was used to critique the article. Rather than edit the article based on this feedback, the article and feedback are provided as is for illustrative purposes (except for minor referencing changes and dividing the article into four parts).

Major Concern: Balance

If I had only one substantial criticism, it would be this.

The article is much more pessimistic than most educated non-experts will expect from a piece presented as a primer.

A reader could reasonably conclude that:

- rogue AI is nearly inevitable,
- existential catastrophe is likely,
- AGI is coming soon,
- AI-driven social collapse is likely,
- job destruction will greatly exceed job creation.

You cite people who hold these views, but there is comparatively little engagement with serious experts who disagree.

For example, there is very little discussion of:

- economists who predict substantial labour displacement but not mass unemployment,
- researchers who believe alignment is difficult but tractable,
- researchers who think AGI timelines are much longer,
- researchers who think LLM limitations may be more severe than enthusiasts assume.

I am not suggesting you change your conclusions. Rather, I would strengthen the article by explicitly acknowledging stronger counterarguments before explaining why you remain unconvinced.

That would make the piece look more analytical and less polemical.

Biggest Factual Risk

The section on the OpenAI/Hugging Face incident carries a lot of argumentative weight.

Because it is central to your thesis, I would be very careful about wording.

Currently, some passages are written with very high confidence:

"rogue AI swarm"

"escaped containment"

"took control of a significant portion of OpenAI's computing infrastructure"

"warning shot"

If later evidence revises any aspect of those events, this section could age badly.

My suggestion:

Distinguish more clearly between:

- observed behaviour,
- investigators' interpretation,
- your interpretation.

Readers tend to trust an author more when uncertainty is explicitly acknowledged.

Significant Omissions

1. AI in science and medicine

You mention mathematics and briefly mention medical research near the beginning, but AI-assisted science deserves more attention.

Many experts would argue that:

- scientific discovery,
- drug discovery,
- materials science,
- biological modelling,

may be the most important benefits of AI.

A short section would improve balance.

---

2. Education beyond assessment

The assessment discussion is good.

But there is surprisingly little discussion of:

- AI as a tutor,
- personalised learning,
- literacy support,
- language learning,
- educational access.

Given your educational leadership background, readers may expect more here.

---

3. Geopolitics

China appears only briefly.

A larger issue worth mentioning is that AI safety cannot be understood purely as a technical problem.

Many decisions are driven by:

- US-China competition,
- military strategy,
- economic rivalry,
- national sovereignty.

This would strengthen your explanation of why slowing development is difficult.

---

4. Environmental impacts

You discuss data centres and computation extensively.

Yet readers never really hear about:

- electricity demand,
- water consumption,
- grid pressure,
- energy innovation.

A brief discussion would fit naturally.

---

5. Human flourishing beyond work

You rightly emphasize dignity through meaningful work.

As a Christian reflection, I would also consider discussing:

- vocation,
- service,
- community,
- creativity,
- worship.

Otherwise there is a risk of sounding as though human meaning depends primarily on employment.

Christian theology would usually frame work as important but not ultimate.

Places Where the Argument Could Be Tightened

Hallucinations

You state:

"they don't know anything"

This is rhetorically effective but likely too strong.

Many AI researchers would object immediately.

You could preserve the accessibility while softening it:

"they do not possess knowledge in the same way humans do."

or

"their apparent understanding is very different from human understanding."

That avoids getting bogged down in philosophy while remaining defensible.

AGI versus ASI

The argument that AGI effectively implies ASI via recursive self-improvement is plausible, but currently presented a little too confidently.

There are substantial debates about:

- hardware constraints,
- economic constraints,
- diminishing returns,

- algorithmic bottlenecks.

I would present this more as a hypothesis than an expectation.

Job losses

Your prediction of approximately 50% reductions in many knowledge sectors is certainly possible, but it is one of the stronger claims in the article.

I would recommend explicitly stating that economists are sharply divided on the scale and timing of labour displacement.

Christian-Theological Suggestions

These are not omissions exactly, but themes that could enrich the final part of the article.

Human dignity

You discuss dignity repeatedly, which is excellent.

I would make the theological grounding more explicit:

- dignity derives from being made in the image of God,
- not from intelligence,
- not from productivity,
- not from employability.

This becomes especially important if AI exceeds human capabilities in various domains.

Wisdom versus intelligence

This theme would fit wonderfully.

The article talks extensively about intelligence.

A Christian contribution could be:

Intelligence is not wisdom.

Many readers would find that distinction memorable.

Humanity as relational

You touch this in the conclusion.

I would emphasise it earlier.

Many Christian thinkers would argue that human uniqueness is found not merely in cognition but in relationships:

- with God,
- with others,
- with creation.

That framework helps answer the recurring "what if AI becomes smarter than us?" question.

Editorial Suggestions

A few minor issues:

Spelling

I noticed:

"It focusses"

Most style guides now prefer **focuses**.

Repetition

The phrases:

- "original sin"
- "Moloch trap"
- "rogue AI"

appear frequently.

Reducing repetition would improve readability.

Length

The article is longer than many lay readers will comfortably consume.

Consider dividing it into:

1. What AI is
2. Why AI might be dangerous
3. Society, work and the Church

You could then publish as a three-part series.

My Final Assessment

If I were reviewing this for publication, I would say:

The article's greatest strengths are its clarity, accessibility, intellectual engagement, and thoughtful Christian reflection. Its greatest weakness is that it often presents contested predictions in a way that sounds more settled than many experts would accept.

Strengthening the acknowledgement of uncertainty, expanding treatment of positive AI impacts, and developing the theological themes of human dignity, wisdom, and vocation would make an already good article substantially stronger.