Faustian Pacts and Digital Tyranny

Christopher Summerfield talks to **Fergus Byrne** about the promise and the peril of intelligent machines and the trade of convenience for control

If an algorithm can inflate a book's price to 23 million dollars, while another wipes more than a trillion dollars off the value of stocks in one morning, what can we expect from future advances in artificial intelligence (AI)?

Christopher Summerfield

Is AI destined to save humanity by solving impossibly complex problems, or will it take over and destroy the planet? Right now, choices seem quite stark. Amid the endless hype over the pros and cons of AI, it sometimes feels like we are forced to pick a side—either cheerful optimism or gloomy foreboding.

In his new book, *These Strange New Minds*, Professor of Cognitive Neuroscience at the University of Oxford, Christopher Summerfield, contends that to truly understand AI, we need to move beyond such polarities. He suggests we carefully examine what these powerful models actually do, what they can achieve, and, perhaps most importantly, find ways to regulate them without restricting their potential.

These Strange New Minds explores the rapid advances in AI, particularly large language models (LLMs), which are the methods by which information is generated for AIs such as Gemini and ChatGPT. The book examines how AI systems learn, reason, and communicate, comparing their capabilities to human cognition. Christopher also discusses ethical, social, and philosophical implications such as trust, bias, and safety. Through historical context and engaging examples, he offers an in-depth view of how AI has developed.

While speaking with him ahead of his visit to the Dorchester Literary Festival this year, I mentioned how I was surprised to discover that many of my children's generation now use ChatGPT as an internet search engine instead of the traditional Google, Bing, or Yahoo. He explained the difference between the two methods of accessing information. 'A search engine basically finds top-ranked websites and returns them to the user in an ordered list,' he says. Whereas an LLM is based on a very sophisticated computer program trained on vast amounts of text from the internet, books, and other written sources.

A notable difference, which has raised some concerns, is that when someone uses ChatGPT to perform an internet search, they receive a reply in clear English instead of having to sift through various websites for the answer. Which, of course, makes it much more appealing than trawling through different websites. This also makes it less likely that a user will bother to check for other opinions or sources.

In *These Strange New Minds*, Christopher delves into the surprising capabilities of modern AI, highlighting some of the benefits, especially in medicine and education, but he also raises critical questions about privacy, potential societal impacts, and what he calls the 'Faustian pact' we're making with technology as we trade convenience for an ever-increasing reliance on algorithms.

A common concern is what happens if AI begins to think independently. In his book, Christopher explores this idea in depth, and when I speak to him, he shares experiences from colleagues and friends. 'I have a lot of friends who use language models for advice about social scenarios,' he says. 'They might, for example, put in a loose but anonymised description of office politics and ask, "What should I do about this?" I've not tried this myself, but apparently, the model is very good. So, in a sense, from a sort of third-person perspective, if you ask it objectively about things that relate to human social interaction, it's as knowledgeable about those as it is about physics and medieval history.'

But is it thinking for itself? LLMs don't "think" in the way humans do. Christopher points out that they don't have consciousness, emotions, or personal experiences. Most importantly, they are missing what he says are the two vital aspects of human existence: "they don't have a body, and they don't have any friends." Their function is to process information and generate text based on the vast amount of data they have been trained on. This process is complex and can appear thought-like, but it's fundamentally different from human cognition. LLMs don't have intentions, beliefs, or self-awareness.

However, Christopher has concerns about some of the current developments in AI. 'These tools are very powerful,' he says, 'and as I say in the book, we should obviously be concerned about the impact on

individual psychology, such as what they do to the user, particularly vulnerable people. But I think the thing that we should be thinking more about is the complex system dynamics. What are the secondary effects that are hard to anticipate?'

'We have devolved power to the digital tools and the organisations that build them, and we've often done that willingly, because it is productivity-enhancing or efficiency-gaining for us'

He gives the example of the suggestion that 'the algorithms which power social media have led to a fragmentation of political discourse and an increase of political polarisation.' Admitting that some political scientists don't believe this, he suggests that since it's a popular idea, let's say for the moment that it's true. 'When social media came along, everyone thought it would be a net positive for democracy.' He recalls how 'people were super excited' that protest groups were mobilising on Facebook. 'It was like social media democracy,' he says. 'However, it turned out that that prediction went the other way! So, most people now think that social media is a negative for democracy.' This raises the concern that with AI, it is impossible to predict what's around the corner. 'We just don't know what's going to be the wider societal impact.'

This same concern applies to Agentic AI, systems that can autonomously set goals, plan, and execute multi-step tasks by utilising tools and adapting their actions to achieve desired outcomes with minimal human intervention. He describes them as digital agents deployed by businesses, private individuals, public sector organisations, or even governments, who are doing things on our behalf. Their processes, he says, 'potentially cut us out of the loop of making a lot of our choices and could have a lot of weird secondary effects that are hard to anticipate, and I'm not sure that they would be positive.'

Christopher feels we may struggle to regain control from the automation we've handed over to technology. 'We have devolved power to the digital tools and the organisations that build them, and we've often done that willingly, because it is productivity-enhancing or efficiency-gaining for us.' As AI becomes more efficient and can do more for businesses and ourselves, Christopher points out that 'when we want to retake control of processes, it's going to be a lot harder, because we've devolved, or delegated a lot of that autonomy to these AI models.'

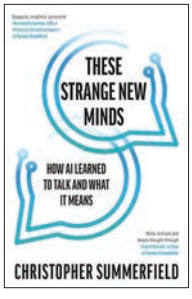
Personalisation is another potential concern. Christopher highlights that if algorithms can already assist with our location, tastes, and even political views, it is a small step for an LLM to be trained to imitate the social behaviours we value in others. This may seem harmless, possibly even useful, but as we place our trust in an AI model, it grants "AI systems unnerving levels of power over our lives" while at the same time insulating us from ideas and opinions that may not fit in with what we already believe.

So what can we do to mitigate some of these potential issues? He suggests that many of these concerns should in some way be the responsibility of those developing AI models. For example, AI companies employ 'raters' to review outputs and determine if they are straying into territory that doesn't chime with societal norms. Christopher points out that the teams building these models are 'pretty large and sophisticated.' There will be a rubric or set of rules that raters who are recruited are expected to abide by. He says, 'that rubric will specify things that are non-negotiable for the company, for example, hate speech is illegal in the UK, so the model definitely shouldn't be generating hate speech.'

Christopher believes that many of the policies are probably a mix of common sense, avoidance of legal jeopardy, and reputation management. Although the raters will have some discretion, he thinks they will 'broadly stick by the rubric, especially for edge cases where there might be something that's tricky politically or socially. So that there is relative compliance with the policy that's dictated by the company.'

So hopefully, the person running the company and setting the rubric has society's best interests at heart.

In *These Strange New Minds*, Christopher Summerfield covers a vast amount of information, helping us to understand how LLMs work, how they were created, and where they might go. He explains that the idea behind AI development is that we can distil all knowledge into a single system, which he describes as "a monolithic oracle" that offers universally palatable replies to everyone. Aside from voicing concerns, he also compares it to something that would be "as slick and bland as a career politician" or as exciting as going on a date with Wikipedia. These may not exactly be descriptions that could inspire fear of world domination by artificial intelligence, but AI is evolving, and we are already within its reach.



Christopher Summerfield will be speaking at Dorchester Literary Festival on 25th October. To book tickets or for more information about the Festival visit: www.dorchesterliteraryfestival.com.