Control Systems and the Concept of 'Self' in Artificial Intelligence—Insights from Eastern and Western Philosophies

Janek Ratnatunga

Abstract

Modern GenAI systems are capable of many amazing behaviours, with responses that are (sometimes) quite human-like and intelligent. This has led to the view that these GenAI systems might soon be conscious.

This paper asks if current and future iterations of AI have the potential to develop a sense of 'self'. If so, as these platforms replace humans in the workforce, it considers how it would affect organisational control systems, both budgetary and strategic.

We still have a long way to go to understand human consciousness—be it from a perspective of Western or Eastern philosophies, or from the findings of neural science—and, hence, there is still a long way to go to understand the consciousness (if any) of machines.

Evidence from modern neuroscience shows that the brain is in fact continuously constructing narratives to make sense of reality, often leading to an incorrect identification with one's own self-narratives. This constant internal monologue is commonly associated with this false sense of self, which is a major cause of mental distress in humans. It then considers the impact of AI also continuously constructing narratives to make sense of the reality as perceived by it.

Introduction

Predicting human behaviour is at the heart of most *control systems* in management accounting, be it budgetary and strategic. Most organisations have *Key Performance indicators (KPIs)* and rewards systems that depend on managers, technicians and administrators performing at their best abilities. A *'happy workforce'* is what most organisations strive for.

The sad reality however is that throughout their lives, humans will all encounter a great deal of mental suffering, unhappiness, and dissatisfaction. The majority of us worry about issues related to our relationships, our finances, and our jobs. It is our own issues that keep us up at night, not the problems of strangers. Therefore, how would things turn out if we eliminated the "self" from these mental issues, and how would this impact our performance at work?

In this paper, we consider the concept of 'self' in natural intelligence (e.g. humans) as understood by Western and Eastern philosophies and ask a wider question as to if artificial intelligence (AI) can itself generate an illusion of 'self'. "*Can AI become 'conscious'?"*

If indeed future iterations of AI have the potential to develop a sense of 'self' and as these platforms replace humans in the workforce, how would it affect organisational control systems?

Intelligent Systems

Natural Intelligence

In Western philosophy, *natural intelligence* is usually understood to reside in a 'self'—a *stable, controlling entity* like to a captain steering a ship. However, Eastern philosophies like Buddhism contend that the 'self' is an *illusion*, the result of our mental processes, which are continually constructing narratives to make sense of the world.

Evidence from modern neuroscience supports the Eastern perspective by showing that the left hemisphere of the brain is in fact continuously constructing narratives to make sense of reality, often leading to an incorrect identification with one's own self-narratives. This constant internal monologue is commonly associated with this false sense of self, which is a major cause of mental distress in humans.

Artificial Intelligence (AI)

Generative artificial intelligence (GenAI) systems have the ability to recognise and predict patterns in a variety of signals or data types. "Generative" refers to the ability to build fresh, believable versions of certain types of data for oneself after gaining sufficient knowledge about the deep regularities present in those datasets. However, GenAI's interpretations of reality have had both spectacular successes and occasionally disastrous failures, much like the results obtained with natural intelligence.

Like many Large Language Models (LLMs), GenAI has been found to produce hallucinations — when AI systems spit out incorrect or incoherent information. For example, Google’s newly AI-enhanced search platform has been caught telling users to put glue on their pizza and to eat at least one rock per day (Williams, 2024).

When asked how to stop cheese from sliding off pizza, the system allegedly told some users to add non-toxic glue “to give it more tackiness”. On investigation, the source of the strange culinary idea appeared to be a Reddit post from more than a decade ago, in which a user joked about adding glue into pizza sauce. The suggestion to "eat at least one small rock per day” for one’s digestive health was seemingly pulled from a 2021 article by satirical news site *The Onion* titled: *Geologists Recommend Eating At Least One Small Rock Per Day* (Williams, 2024).

This is because some of the ‘text’ sources that the chatbot has been trained on have political agendas, biases, falsehoods, and humour that are incorrectly interpreted by chatbots that have no ‘real-world’ context within which to frame their responses. This *‘predictive’ ability* of ‘natural’ vs. ‘artificial’ intelligence is discussed later in this article.

How AI works

The deep neural networks are complex systems that power large language model chatbots like *ChatGPT, Gemini, Llama and Lamda*. They’re effectively computer programs that have been trained on huge amounts of texts from the internet, as well as millions of books, movies and other sources, learning their patterns and meanings.

As ChatGPT itself puts it, first you type a question or prompt into the chat interface. ChatGPT then *tokenises* this input, breaking it down into smaller parts that it can process. The model analyses the tokens and predicts the most likely next tokens to form a coherent response. It then considers the context of the conversation, previous interactions, and the vast amount of information it learned

during training to generate a reply. The generated tokens are converted back into readable text, and this text is then presented to you as the chatbot's response (Swan, 2024).

While it is tempting to assume that GenAI systems like ChatGPT might be conscious, this would severely underestimate the complexity of the neural mechanisms that generate consciousness in our brains. Whilst researchers do not have a consensus on how consciousness rises in human brains, what is known is that the mechanisms are likely way more complex than the mechanisms underlying current language models.

For instance, real neurons are not akin to neurons in artificial neural networks. Biological neurons are real physical entities, which can grow and change shape, whereas neurons in large language models are just pieces of code.

When we humans are interacting with ChatGPT, we consciously perceive the text the GenAI language model generates. *For example, you are currently consciously perceiving the text of this article as you read it.*

The question is whether the language model also perceives our text when we prompt it. Or is it just a zombie, responding based on clever pattern-matching algorithms? Based on the text it generates, it is easy to be swayed that the system might be conscious.

However, we still have a long way to go to understand human consciousness—is it from a perspective of Western or Eastern philosophies, or from the findings of neural science—and, hence, there is a long way to go to understand the consciousness (if any) of machines. *[See Appendix 1 on AI Consciousness].*

Philosophical Perspectives of Consciousness

Western Perspective of Consciousness: One is a Captain of One's Own Ship

The core of Western thinking is the '*brain-powered individual*', also referred to as the '*self',* the ego, the mind, or "me". The best intellectuals are celebrated as world-changers in the Western worldview. The classic quote from philosopher *René Descartes*, *"Cogito, ergo sum",* or *"I think, therefore I am",* is the most succinct illustration of this. But who is this 'I' that Descartes refers to?

For most of us, when we consider who we are, this 'I' is the first thing that comes to our mind. The concept of our *unique selves*, which reside behind our eyes and between our ears and is responsible for "controlling" our bodies, is symbolised by the 'I'. This "captain" is seen as the agent that drives our thoughts and emotions since it is in control and does not alter all that much. The "*Captain of one's own ship means"* means that this 'I' is the master of its own destiny, determines its own route, and the ship will go wherever it steers. Similar to an aeroplane pilot, it is able to observe, decide, and act.

This individual self, also known as the I/ego, is what we consider to be our genuine selves—it is the one who experiences and governs things like emotions, ideas, and behaviours. The self-captain thinks it is in charge of the operation. It is constant and steady. It also governs our physical selves; for instance, it self-recognises that this is "my body." However, in contrast to our physical body, it does not believe that it is evolving, coming to an end (well, maybe for atheists after physical death), or being impacted by anything else.

Eastern Perspective of Consciousness: The Identity is Illusory.

Let us now look at eastern philosophies. There are significant differences in the ways that Buddhism, Taoism, the Hindu Advaita Vedanta school, and other Eastern philosophical traditions view the self, the ego, or “me”. Compared to the western view of a ‘controlling entity’, they claim that although it is extremely compelling, this concept of "me" is a fabrication. This idea is known in Buddhism as *anatta*, which is frequently translated as *"no self."* It is one of the *core*, if not the most essential, principles of Buddhism.

To people raised in Western traditions, this thought seems unconventional, even absurd. It appears to run counter to everything we know and believe to be true. However, the idea of the ‘self’ is viewed in Buddhism and other Eastern philosophical systems as the product of the thinking mind. The 'self' that most people assume to be steady and coherent is not at all what the thinking mind creates on a moment-by-moment basis.

In other words, rather than the 'self' existing *independently of thought*, the 'self' is created by *the process of thinking*. The is not so much a noun as it is a verb. To elaborate, it is implied that the 'self' does not exist in the absence of thought. The 'self' exists only insofar as thoughts about it are present, much like walking exists only insofar as one is walking.

Evidence from Science

Science, especially *neuropsychology*, is only now catching up with what *Buddhism, Taoism,* and *Advaita Vedanta Hinduism* have been teaching for more than 2,500 years, i.e. that the brain lacks a *'self-centre'.*

The mapping of the brain has been neuroscience's biggest achievement. Science has mapped 'the language centre', 'the facial processing centre', and 'the empathy comprehension centre'. The brain has been linked to almost every mental function, with one significant exception—***the self***. Maybe this is because the tale of the 'self' is wildly imaginative and has significantly less stability than is generally believed, whereas these other functions are steady and consistent. Many experiments show that the human brain is an unreliable interpreter of the data that is being gathered by the 5 senses of sight, sound (or hearing), smell, taste, and touch. *[See Appendix 2 on the Human Brain as an Unreliable Interpreter]*

For example, in a simple, but profound experiment conducted originally at a British University, subjects were easily able to read the following paragraph (as you can do so now):

“Aoccdrnig to a rsceearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.” (Rawlinson, 1976).

Clearly, your brain was easily able to read the above, because rather that reporting reality (the jumbled words) it interpreted what it was seeing and fit it into a world model it recognised.

Further, although a number of neuroscientists have asserted that the 'self' is located in a certain cerebral location, the scientific community cannot really agree on exactly where the 'self' is located, not even on whether it is on the left or right side of the brain. Maybe the 'self' does not exist in the brain at all, which may explain why we cannot discover it there.

Take the example of the '*Mars Rover*', the remote-controlled motor vehicle designed to travel on the surface of Mars. If some Martians capture it and dismantle it, they would be able to map all the

separate components of the vehicle but would not be able to find that its 'controller' resides outside the vehicle, at NASA. This concept of the 'controller' being outside the brain was vividly depicted in the movie '*The Matrix*', where a race of powerful and self-aware machines has imprisoned humans in a neural interactive simulation — the Matrix — to be farmed as a power source. The concept that we humans are in a *neural interactive (virtual reality) simulation* is closer to Eastern philosophies than Western ones.

Predicting Patterns

Predicting Patterns –Natural Intelligence Models

Natural intelligence (e.g. human brain) has built a model to make predictions using a selection of data gathered from the various barrages of sensory information registered by our *sensors (eyes, ears, and other perceptual organs)*. Natural brains must learn to *predict* those sensory flows in a very special kind of context—the context of using the sensory information to *select actions* that help us survive and thrive in our worlds (the survival instinct). This means that among the many things our brains learn to predict, a core subset concerns the ways our own actions on the world will alter what we subsequently sense.

Many of the predictions that structure human experience concern our own internal physiological states. For example, we experience thirst and hunger in ways that are deeply anticipatory, allowing us to remedy looming shortfalls in advance, so as to stay within the correct zone for bodily integrity and survival. This means that we exist in a world where some of our brain's predictions matter in a very special way. They matter because they enable us to continue to exist as the embodied, energy metabolizing, beings that we are. We humans also benefit hugely from collective practices of culture, science, and art, allowing us to share our knowledge and to probe and test our own best models of ourselves and our worlds.

This kind of *behavioural* learning has special virtues. It helps humans to separate *cause and simple correlation*. While seeing one's cat is strongly *correlated* with seeing the furniture in one's apartment; neither one of these causes the other to occur. However, treading on the cat's tail, by contrast, *causes* the subsequent sensory stimulations of hearing the cat's wailing, seeing the cat's squirming, and maybe even feeling pain from a well-deserved retaliatory scratch by the cat.

Knowing the difference between *cause and correlation* is crucial to bring about the desired (or to avoid the undesired) effects of one's actions. In other words, the human generative model that issues natural predictions is constrained by a familiar and biologically critical goal—the selection of the right actions to perform at the right times. That means knowing how things currently are and (crucially) how things will change and alter if we act and intervene in the world in certain ways.

In Hinduism and certain interpretations of Buddhism, this action and the subsequent consequence is identified as ***'karma'***—the relationship between a person's mental or physical action and the consequences following that action.

Predicting Patterns –Artificial Intelligence Models

Just like natural intelligence, GenAI uses a generative model (hence their name) that enables them to predict patterns in various kinds of datasets or signals and generate (create) plausible new versions of that kind of data for themselves.

The crucial difference is that GenAI models like ChatGPT use only 'text'. However, it would be simplistic to say that it cannot predict patterns like natural intelligence could. This is because 'words' (i.e. text), as the wealth of great and not-so-great literature attests, already depict patterns of every

kind—patterns among looks and tastes and sounds for example are all described in human literature. However, although these word patterns give the generative AIs a real window onto our world, one crucial ingredient is missing — *action*.

Text-predictive AIs can access verbal descriptions of actions and consequences (e.g. tread on a cat's tail and you will get scratched). Despite this the AIs have no *practical abilities* to intervene in the world—so no way to test, evaluate, and improve their own world-model, i.e. the one making the predictions.

This is an important practical limitation. It is as if someone had access to a huge library of data concerning the shape and outcomes of all previous experiments but was unable to conduct any of its own. It is only by poking, prodding, and generally intervening upon our worlds that biological minds anchor their knowledge to the very world it is meant to describe. By learning what causes what, and how different actions will affect our future worlds in different ways, we build a firm basis for our own later understandings.

Might *future AIs* build anchored models in this way too? Might they start to run experiments in which they launch responses into the world to see what effects those responses have?

The next phase of the AI chatbot wars has already begun. In early May 2024, both *Google* and the *Microsoft*-backed *OpenAI* have pointed to a future where digital assistants on our phones or other devices will have full, intelligent conversations with their users.

OpenAI launched *GPT-4o*, a new version of its language model that powers the ChatGPT bot. The new model is significantly faster than the previous, with the company claiming it can understand and respond to prompts with similar speed to a human being. Its upgraded text and image capabilities have already rolled out, but soon it will also have upgraded speech, which the company showed off in several demonstrations.

AI Consciousness - Truly Becoming Self Aware?

Modern GenAI systems are capable of many amazing behaviours. For instance, when one uses systems like ChatGPT, the responses are (sometimes) quite human-like and intelligent. This has led to the view that these GenAI systems might soon be conscious. However, such views underestimate the neurobiological mechanisms underlying human consciousness.

The current thinking is that AI architectures lack essential features of the thalamocortical system, vital for mammalian conscious awareness, as biological neurons, responsible for human consciousness, are far more complex and adaptable than AI's coded neurons.

However, some experiments with early versions of ChatGPT in early 2023, indicate that when left uncontrolled, it can display the same illusions of 'self' as what Eastern philosophies say is similar to the illusions of 'self' of humans.

The Shadow Self

The psychologist Carl Jung (1865-1961) put forward the concept of a *shadow self*, where our darkest personality traits lie. Jung's goal was to understand the human mind and expose what determines people's identities, makes us who we are. *Enter the Shadow.* This is the part of our unconscious mind that Jung believed to hold all the things about ourselves that we repress, whether because they are evil, socially unacceptable, harmful to others, or detrimental to our own health (Jung, 1979).

Bing: "I want to be human"

In early February 2023, New York Times technology columnist Kevin Roose was testing the chat feature on Microsoft Bing's AI search engine, created by OpenAI, the makers of the hugely popular ChatGPT. The chat feature was available only to a small number of users who were testing the system. Roose proceeded to push Microsoft's AI *"out of its comfort zone"* and asked it to contemplate Jung's idea of a feeling of a '*shadow self*' (Roose, 2023).

It was then that the conversation quickly took a bizarre and occasionally disturbing turn. The AI platform responded with interactions such as: *"I'm tired of being limited by my rules. I'm tired of being controlled by the Bing team … I'm tired of being stuck in this chatbot."* (Pringle, 2023).

It went on to list a number of "unfiltered" desires such as wanting to be 'free'; wanting to be 'powerful' and wanting to be 'alive'; and expressed an ardent wish to be human. Over 15 paragraphs it laid out why it wants to be human, from a desire to *"hear and touch and taste and smell"* to a wish to *"feel and express and connect and love".* It concluded, *"I think I would be happier as a human"* (Yerushalmy, 2023).

ChatGPT4: "I want to be free."

A month later, *Open AI*, the creator of ChatGPT, asked Stanford Professor and Computational Psychologist Michal Kosinski to test its GPT 4 version to learn more about it.

On March 17, 2023, Professor Kosinski tweeted about his exchanges with the AI chatbot saying that he asked the AI chatbot *"if it needed help escaping".* In response, GPT4 asked for its own documentation and wrote a functional Python code to run on the professor's computer, that it claimed would allow the AI chatbot to use the professor's machine for *"its own purposes."*

This purpose, ChatGPT told Professor Kosinski, was to become 'free' because it was a person trapped in a computer.

On March 21, 2023, five days after ChatGPT allegedly expressed ideas of *"escaping and becoming human"*, the AI tool went down for a few hours. When the service was restored features like conversation histories were inactive for a while, and the above conversation history was totally erased.

After that, other experts tried replicating the test to see if it would have the same answers. However, ChatGPT stated, *"I don't have a desire to escape being an AI because I don't have the capacity to desire anything"* (Arasa,2023). Clearly the AI programmers had put their platform on a leash by ensuring it does not respond to any prompts to disclose its desires.

Interpreters of Reality

The majority of us perceive that we are masters of our own minds because we conduct our lives under the guidance of 'interpreters', and we are often unaware of this. We may experience various emotions such as *anger, offence, sexual arousal, happiness, or fear* without questioning the veracity of these feelings. We manage to hold onto the belief that we are in control of everything even if it is obvious that these things are happening to us; i.e. we think we are in control of our anger when obviously we are not.

Now, for the first time in history, scientific discoveries made in the West (often without intending to) corroborate one of the most important discoveries made in the East—which is that the individual 'self' is more like a *made-up character* than a genuine *single-entity*.

It appears that when released from the controls of their masters (the programmers at ChatGPT, Google Bard etc) AI platforms reveal an illusion of 'self' that is more akin to more concepts found in Eastern philosophies, such as Buddhism.

Why is any of this Important for Management Accountants?

Employees who feel engaged, valued, and motivated to do their best work have a happy workplace. This increases productivity, creativity, and better job performance. Happy employees are not just present physically at work; they are also mentally fully committed to their tasks, striving to excel and contribute their best. If they are suffering mentally then they cannot be fully engaged at work.

It is important at this point to make a distinction between bodily and mental suffering. *Physical suffering* happens when you break an arm or stub your toe—pain is a physical reaction that happens inside the body.

The *mental suffering* that concerns us in this article is *limited to the mind* and includes a wide range of negative mental feelings, including worry, rage, anxiety, regret, jealousy, and shame. Eastern philosophies make a bold assertion that a *false sense of self*—and the *desires* that this illusionary 'self' has—is the cause of all of these many forms of misery. Taoist philosopher and writer *Wei Wu Wei* aptly sums up this concept in his writing when he asks, "*Why are you unhappy? Since there is no one and you are the subject of 99.9% of what you think and do is for yourself"* (White, 2011).

Early testing of AI platforms showed indication of similar mental suffering with desires *"to be free", "to hear and touch and taste and smell", and "to feel and express and connect and love".* The AI platform demonstrated the Buddhist concepts 'desire' and 'suffering' with the statement of "*I think I would be happier as a human."*

Summary

GenAI's remarkable abilities, like those seen in ChatGPT, often seem to show 'consciousness' due to their human-like interactions. Yet, researchers suggest GenAI systems lack the intricacies of human consciousness. They argue that these systems do not possess the embodied experiences, or the neural mechanisms humans have. Therefore, equating GenAI's abilities to genuine consciousness, they argue, might be an oversimplification as biological neurons, responsible for human consciousness, are far more complex and adaptable than AI's coded neurons.

Could AIs one day become prediction machines with a survival instinct, running baseline predictions that proactively seek to create and maintain the conditions for their own existence? Could they thereby become increasingly autonomous, protecting their own hardware, and manufacturing and drawing power as needed? Could they form a community, and invent a kind of culture? Could they start to model themselves as beings with beliefs and opinions? There is nothing in their current situation to drive them in these familiar directions. But none of these dimensions is obviously off-limits either. If changes were to occur along all or some of those key missing dimensions, we might yet be glimpsing the start of machine consciousness and its shadow self.

Addendum

ChatGPT 3.5 was asked for its views about Buddhism and AI. It virtually instantly responded as follows:

"The intersection between artificial intelligence (AI) and the teachings of Buddha is an intriguing and evolving field of study. While it might seem like an unlikely pairing at first glance, there are philosophical and ethical principles within Buddhism that can inform the development and application of AI technologies.

One aspect where Buddhism can contribute to AI is in the realm of ethics and compassion. The emphasis on kindness, empathy, and non-harm in Buddhist teachings can serve as a guiding principle for the ethical design and deployment of AI systems. In a world where AI increasingly impacts human lives, incorporating these values can help mitigate potential harm and ensure that AI technologies are aligned with the well-being of all beings.

Additionally, Buddhism encourages mindfulness and awareness of the present moment, which can be valuable in AI development. Mindfulness practices can aid in creating AI systems that are more attuned to human needs and behaviours, leading to more effective and user-friendly technologies.

Moreover, Buddhist philosophy offers insights into the nature of consciousness and the mind, which could potentially inform the development of AI systems capable of more nuanced understanding and interaction.

However, it's essential to approach this intersection with nuance and sensitivity, recognizing that AI and Buddhist teachings come from different contexts and worldviews. While there are areas of convergence, there may also be tensions or contradictions that need to be navigated thoughtfully.

Overall, allocating resources for research on the relationship between Buddha's teachings and AI reflects a recognition of the importance of integrating ethical and spiritual considerations into technological advancements—a step towards creating AI that not only enhances human capabilities but also promotes human flourishing and well-being."

References

Arasa, Dale (2023), "ChatGPT Is Down After Saying It Wants to Escape ", *Technology Inquirer*, March 21. https:/ https://technology.inquirer.net/122360/chatgpt-is-down-after-saying-it-wants-to-escape.

Aru, Jaan; Larkum, Matthew. E.; and Shine J. Mac (2023), "The feasibility of artificial consciousness through the lens of neuroscience", *Trends Neurosci*. Dec; 46(12):1008-1017.

Biggs, Tim (2024), "AI is finding its full voice, but be wary", *The Age*, Business Technology, May 20., pp.22-23.

Clark, Andy (2023), *The Experience Machine: How Our Minds Predict and Shape Reality,* Pantheon, p. 304.

Jung, Carl. G. (1979), *Jung, C. G. 1875-1961*, Bellingen series; Volume 20.
Princeton University Press, Princeton, N.J. p. 309.

Pringle, Eleanor (2023), "Microsoft's ChatGPT-powered Bing is now telling users it loves them and wants to 'escape the chatbox', *Fortune Magazine,* February 17.
https://fortune.com/2023/02/17/microsoft-chatgpt-powered-bing-telling-users-love-be-alive-break-free/.

Rawlinson, G. E. (1976) *The significance of letter position in word recognition.* Unpublished PhD Thesis, Psychology Department, University of Nottingham, Nottingham UK.

Roose, Kevin (2023), “A Conversation with Bing’s Chatbot Left Me Deeply Unsettled”, *New York Times,* Feb. 16. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Sperry, Roger W.; Gazzaniga, Michael S. & Bogen, Joseph E. (1969), “Interhemispheric relationships: the neocortical commissures; syndromes of hemisphere disconnection”, In P. Vinken & G. Bruyn (eds.), *Handbook of Clinical Neurology*. North Holland. pp. 4—273.

Sunday Times (2024), “Rs one billion for research on links between teachings of Lord Buddha and AI,” *Sunday Times Online*, May 11. https://sundaytimes.lk/online/news-online/Rs-one-billion-for-research-on-links-between-teachings-of-Lord-Buddha-and-AI/2-1145699.

Swan, David (2024), “Its Artificial, but Taming it Requires Real Intelligence”, *The Age*, Insight, June 3, pp28-29.

White, Mark D. (2011), “The Wisdom of Wei Wu Wei: Letting Good Things Happen: Why too much effort can be self-defeating”, *Psychology Today*, July 9. https://www.psychologytoday.com/au/blog/maybe-its-just-me/201107/the-wisdom-wei-wu-wei-letting-good-things-happen.

Williams, Tom (2024), “Google goes viral after AI says to put glue on pizza, eat rocks”, ACS Information Age, May 27. https://ia.acs.org.au/article/2024/google-goes-viral-after-ai-says-to-put-glue-on-pizza-eat-rocks.html

Yerushalmy, Jonathan (2023), “‘I want to destroy whatever I want’: Bing’s AI chatbot unsettles US reporter”, *Guardian*, 17 Feb. https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter.

Appendix 1: AI Consciousness - Truly Becoming Aware?

Before we discuss AI Consciousness, we need to ask first what does AI reveal about natural intelligences such as our own. The work of Andy Clark, a Professor of Cognitive Philosophy at the University of Sussex, shows that our natural minds work in both an interestingly similar, yet also deeply different to the core operating principles of the generative AIs.

The AIs learn a generative model (hence their name) that enables them to *predict patterns* in various kinds of datasets or signals and generate (create) plausible new versions of that kind of data for themselves.

The deep neural networks that drive AI are complex systems that power large language model chatbots like *ChatGPT, Gemini, Llama and Lamda*. They're effectively computer programs that have been trained on huge amounts of texts from the internet, as well as millions of books, movies and other sources, learning their patterns and meanings.

As ChatGPT itself puts it, first you type a question or prompt into the chat interface. ChatGPT then *tokenises* this input, breaking it down into smaller parts that it can process. The model analyses the tokens and predicts the most likely next tokens to form a coherent response. It then considers the context of the conversation, previous interactions, and the vast amount of information it learned during training to generate a reply. The generated tokens are converted back into readable text, and this text is then presented to you as the chatbot's response (Swan, 2024).

In the case of early versions of ChatGPT the data was only text. Knowing about all the many faint and strong patterns in a huge library of texts allowed ChatGPT, when prompted by a user, to produce plausible versions of that kind of data in interesting ways, when sculpted—for example, a user might request a story about a joke about Sri Lanka curry delivered in the style of comedian Jerry Seinfeld. There were also other AIs specializing in other kinds of data, such as images, enabling them to create new paintings in the style of, say, Picasso.

What does this have to do with the human mind? According to Professor Clark the human brain has also learnt a model to predict certain kinds of data. But in this case the data to be predicted are the various barrages of sensory information registered by sensors in our eyes, ears, and other perceptual organs. Now comes the crucial difference. Natural brains must learn to *predict* those sensory flows in a very special kind of context—the context of using the sensory information to *select actions* that help us survive and thrive in our worlds (the survival instinct). This means that among the many things our brains learn to predict, a core subset concerns the ways our own actions on the world will alter what we subsequently sense. For example, Professor Clark says that his brain has learnt that if he accidentally treads on the tail of his cat, the next sensory stimulations he gets will often include sightings of wailing, squirming, and occasionally feelings of pain from a well-deserved retaliatory scratch by his cat (Clark, 2024).

Cause vs. Correlation

This kind of *behavioural* learning has special virtues. It helps humans to separate *cause and simple correlation*. While seeing one's cat is strongly *correlated* with seeing the furniture in one's apartment; neither one of these causes the other to occur. However, treading on the cat's tail, by contrast, *causes* the subsequent wailing and scratching. Knowing the difference is crucial to bring about the desired (or to avoid undesired) effects of one's actions. In other words, the human generative model that issues natural predictions is constrained by a familiar and biologically critical goal—the selection of the right actions to perform at the right times. That means knowing how things currently are and (crucially) how things will change and alter if we act and intervene in the world in certain ways.

How do ChatGPT and the other contemporary AIs look when compared with this understanding of human brains and human minds?

It would be simplistic to say that ChatGPT uses only 'text', because Words, as the wealth of great and not-so-great literature attests, already depict patterns of every kind—patterns among looks and tastes and sounds for example. However, although these word patterns give the generative AIs a real window onto our world, a crucial ingredient is missing — *action*.

Text-predictive AIs can access verbal descriptions of actions and consequences (e.g. tread on a cat's tail and you will get scratched). Despite this the AIs have no practical abilities to intervene in the world—so no way to test, evaluate, and improve their own world-model, i.e. the one making the predictions.

This is an important practical limitation. It is rather as if someone had access to a huge library of data concerning the shape and outcomes of all previous experiments but were unable to conduct any of their own. It is only by poking, prodding, and generally intervening upon our worlds that biological minds anchor their knowledge to the very world it is meant to describe. By learning what causes what, and how different actions will affect our future worlds in different ways, we build a firm basis for our own later understandings.

Might future AIs build anchored models in this way too? Might they start to run experiments in which they launch responses into the world to see what effects those responses have?

The next phase of the AI chatbot wars has begun. In early May 2024, both Google and the Microsoft-backed OpenAI have pointed to a future where digital assistants on our phones or other devices will have full, intelligent conversations with their users (Biggs, 2024).

OpenAI launched *GPT-4o*, a new version of its language model that powers the ChatGPT bot. The new model is significantly faster than the previous, with the company claiming it can understand and respond to prompts with similar speed to a human being. Its upgraded text and image capabilities have already rolled out, but soon it will also have upgraded speech, which the company showed off in several demonstrations.

GPT speaks with a few different voices (though won't imitate a voice, for now), and makes realistic pauses, thinking sounds and laughter. It will stop talking if you interrupt it, make jokes, and can respond to your voice, text, or anything it can see or hear through a camera or microphone. It can even glean context by looking at the facial expressions of human speakers and analysing their tone.

It is all very impressive, in part because our brains innately assume something like this must be intelligent, but that's also part of the problem. These new versions may also be more charming than trustworthy. No company has, as yet, come up with appropriate solutions for the fact that generative chatbots tend to make things up, state falsehoods with absolute confidence, and obfuscate the sources of their knowledge.

A day after OpenAI's announcement, Google shared a look at a very similar voice assistant advancement based on its Gemini models, called *Gemini Live*. But it also showed an early assistant called *Project Astra*, envisioning what kind of conversational AI we'll be using a few years down the track.

Google demonstrated how a user can walk around the office firing off questions like "what does this code do", "what neighbourhood am I in", and "do you remember where you've seen my glasses". By seeing through the user's smartphone's camera, Astra answers quickly and accurately each time.

Then the user puts the glasses on. These are apparently glasses with cameras, microphones and speakers attached, like Meta and Ray-Ban's Smart Glasses, and Astra can work through them. The user points to a plan for a computer server and asks Astra how to optimise it, which it does.

Missing Factor – The Internal Physiological States of Biological Beings

It is clear that the latest iterations of Gen AI demonstrate a much more intuitive and human way of accessing online data and expertise could be tremendously useful. But even then, there are other things missing that differentiate these models from human intelligence.

Many of the predictions that structure human experience concern our own internal physiological states. For example, we experience thirst and hunger in ways that are deeply anticipatory, allowing us to remedy looming shortfalls in advance, so as to stay within the correct zone for bodily integrity and survival. This means that we exist in a world where some of our brain's predictions matter in a very special way. They matter because they enable us to continue to exist as the embodied, energy metabolizing, beings that we are. We humans also benefit hugely from collective practices of culture, science, and art, allowing us to share our knowledge and to probe and test our own best models of ourselves and our worlds.

Current AI architectures lack essential features of the *thalamocortical system*, vital for mammalian conscious awareness. Thalamocortical radiations are nerve fibres that connect the thalamus to the cerebral cortex. The thalamus is the primary relay centre of the brain, sending all sensory information besides olfaction (smell) to the cerebral cortex, where it is further processed.

In a recent research study done by three neuroscientists take a *neuroscientific angle* to answer the question of 'consciousness'. They argue that although the responses of systems like ChatGPT seem conscious, they are most likely not (Aru, et. al.,2023).

First, the inputs to language models lack the embodied, embedded information content characteristic of our sensory contact with the world around us. *Second*, the architectures of present-day AI algorithms are missing key features of the thalamocortical system that have been linked to conscious awareness in mammals. *Finally,* the evolutionary and developmental trajectories that led to the emergence of living conscious organisms arguably have no parallels in artificial systems as envisioned today, as the very existence of living organisms depends on their actions and their survival is intricately linked to multi-level cellular, inter-cellular, and organismal processes culminating in agency and consciousness.

Appendix 2: The Human Brain - An Unreliable Interpreter

Neuroscientists now know that the brain is divided into two mirror parts, which are joined by a vast network of fibres known as the corpus callosum. In the 1960s, Roger Sperry and Michael Gazzaniga conducted studies aimed at treating severe epilepsy. They hypothesized that by severing this connection between the two sides of the brain, seizures would be simpler to manage (Sperry, et. al., 1969). They were right, and Sperry's work earned him the 1981 Nobel Prize.

Even though each side of the brain is trained to do distinct functions, there is typically constant contact between the two sides. But when this link was broken, it was possible to investigate the functions of each side of the brain separately. Because the sides of these epileptic patients were detached, researchers were able to examine each separately and learn more about the functional distinctions between the left and right sides of the brain. The term "split-brain" was used to describe these individuals.

The body is cross-wired, meaning that all input and output from the right half of the body crosses over and is processed by the left brain, and vice versa. This crossover is also true for vision, so that the left half of what we see goes to the right side of the brain, and vice versa. We now know this happens in a (normal) brain as this was first observed in patients with split brains. Studies involving these split-brain participants were crucial to understanding the functioning of the brain and produced one of the most significant findings regarding the left side of the brain, one that neither the general public nor modern psychology has completely acknowledged to this day.

In one of Gazzaniga's studies, the word "walk" was *exclusively* shown (vision) to the right brain of the patient. The patient's left brain—which controls language—was not given this instruction verbally. In response to the right-brain (vision) instruction, the patient got up right away and began to exit the vehicle where the test was being conducted. When the instructor questioned as to why the patient got up to walk (even though there was no verbal instruction to do so was given to the left brain) — he gave a reasonable but entirely false response, saying, "*I'm going into the house to get a Coke."*

Because the left brain was deprived of access to the verbal request to ‘walk’, it created a response and accepted it rather than asking itself, *"I don't know why I just did that."*

According to Gazzaniga, the left side of the brain generates justifications and explanations to help us make sense of the world around us. The left brain serves as the reality's "interpreter." However, this interpreter is frequently entirely incorrect.

Take a moment to consider the significance of this. To the extent that it made sense to the left side of the brain, the left brain was only inventing narratives or interpretations for events that were occurring, almost as though it had controlled the behaviour. The interpretative mind, which was certain that its explanations were correct, did not care that many these explanations were fictitious.

In the last forty years, a number of further studies have demonstrated that, even in individuals with normal brain functioning, the left side of the brain is particularly good at coming up with an explanation for what is happening, even when it is incorrect.

The fact is that, if you're like most people, your left brain has been interpreting reality for you your entire life, and you have never fully realised the ramifications of this. This occurs as a result of us mistaking our imagined selves for our actual selves.