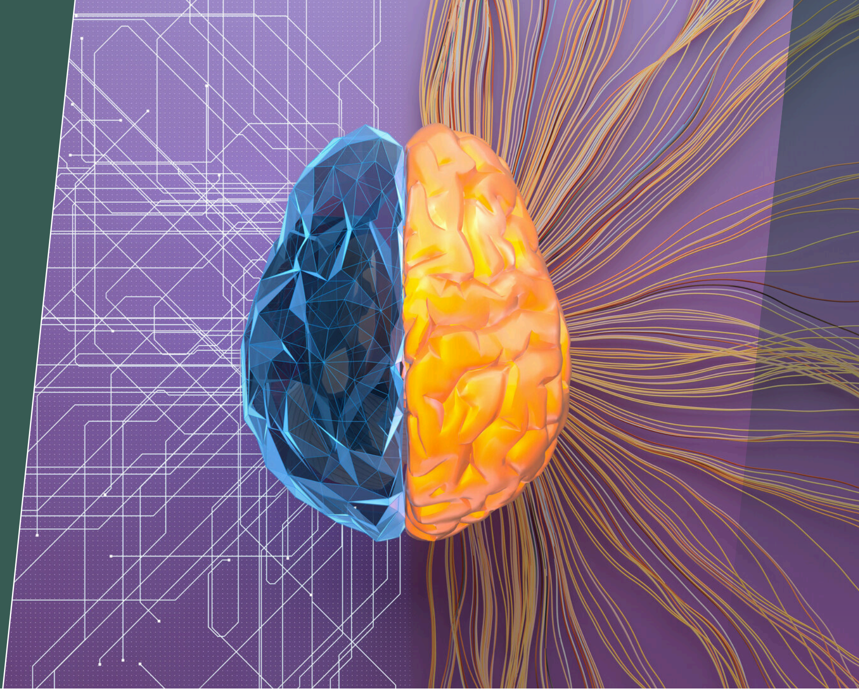
KORN FERRY
BE **MORE** THAN

# HUMAN OR AI?: **THE CONSCIOUS AGENT**



Thought Leadership

*In the second installment of her "Human or AI?" series, Korn Ferry's Amelia Haynes breaks facts from fiction about GenAI as conscious agents and what it could mean for humanity.*

In Steven Spielberg's 2001 sci-fi film ***A.I. Artificial Intelligence***, an android boy dreams of being a human and loved by his parents. This modern-day Pinocchio tale, like so many others of its genre, speculates about the possibility of AI gaining consciousness. Recent breakthroughs in AI, like ChatGPT and similar tools, have cast these conversations in a new light, spurring exciting debates on whether conscious AI is in our future.

Once the stuff of science fiction, the possibility of a human-like AI is now a serious topic for consideration. But before we can determine whether AI could soon be conscious, we must consider what it means to be conscious in the first place.

## What is consciousness—and is AI conscious?

For centuries, both philosophical and scientific communities have struggled to define consciousness and how to measure it. To this day, there is no concrete consensus, and the advent of generative AI has only added further complication to the debate. However, several common themes have emerged across both disciplines that offer a helpful framework for understanding consciousness. According to this framework, something is conscious if it has:

- *Awareness*: it can react to things in the world.
- *Self-awareness*: it can react to and communicate about changes within itself.
- *Goal-directed behavior*: it operates in pursuit of a goal.
- *Information integration*: it can integrate many kinds of information.
- *Experience*: it has qualitative experiences.

Looking at GenAI from the perspective of these criteria, it is not wholly clear whether GenAI is currently or could someday be conscious. Certainly, it seems to fit some of this framework. One could even argue that GenAI *is* aware, as it reacts to our inputs (its environment). It operates with goal-directed behavior by responding to our prompts. And it is a powerful tool for integrating information, producing a response by combining knowledge from the prompts with its own learned data. The question of whether GenAI is self-aware and has experiences is less cut and dry; some arguments have been made both for and against this. Either way, GenAI seems to check some—if not most—of the boxes for consciousness.

Whether GenAI is conscious or not, interacting with these programs can often *feel* like we are interacting with a conscious agent. They appear to be thinking—considering our input, stringing together bits of information, forming arguments and ideas, and making judgments. Yet, this isn't what's happening below the surface. Despite engaging in ostensibly conscious behaviors, GenAI doesn't confidently meet the critical criteria for consciousness—it isn't sentient or self-aware.

And the reality is, these tools are performing a constrained set of operations that *conscious and sentient* humans have programmed them to perform. Humans reflect consciousness—emotions, ideas, wisdom, and experiences—in what we produce. And this is the content from which GenAI learns patterns and probabilities. So, while it may appear to have feelings or convey the wisdom of its experiences, this is simply an illusion of our own design.

## Understanding the role of agency

Agency is our capacity to take intentional actions or exert control over our own decisions and behaviors. AI lacks basic features of common-ground agency, as defined by business professor Danielle Swanepoel, and therefore lacks human-like agency. She defines the basic features of common ground agency as:

- *Reflection and deliberation*: being able to identify with not only one's desires and preferences but also understand others—in other words, having our own internal values and belief system.
- *Awareness of self in time*: understanding and learning from the past and imagining and planning the future.
- *Awareness of environment*: an agent can engage with their environment and understand that they are distinct from others.
- *Freedom of choice*: not only being able to identify one's volition with regards to the existing norms but also the capability to purposefully violate norms.

Although it is possible to program AI to follow norm-like guides and principles, and even distinguish itself from other AI, it still lacks its *own* desires and drives. It can only behave following the norms and rules on which it was programmed, and which it cannot willfully violate. While it may have a concept of time in theory, it cannot conceptualize itself in time, and therefore can't act intentionally for the future.

While Swanepoel argues that AI fails the majority of the criteria above, and therefore can't be considered an agent in the way humans conceptualize agency, others propose that AI could have a type of agency that does not include self-awareness. It is possible that AI could self-replicate, randomly mutate, or be programmed to compete for limited resources in global and local ecosystems, all of which could give rise to an emergent yet unconscious form of agency.

Some AI systems are already showing signs of emergent properties; one Google AI program adapted to respond to the language of Bangladesh, which it hadn't been trained on previously. This kind of AI with agency but not consciousness presents a complex ethical, legal, and practical landscape.

## The case for consciousness

Despite becoming more conscious and agentic over time, GenAI still lacks critical features associated with both consciousness and agency. While such distinctions may seem isolated to philosophical domains, the implications of how we conceptualize GenAI and our understanding of these shortcomings have the potential to be profound.

We may be able to intuit *that* consciousness matters in the workplace, but we may not have a clear idea of why. Research shows that consciousness supports many of the basic human emotions we find important. In fact, empathy is considered to be a sign of a fully developed consciousness. According to some researchers, AI may be capable of "cognitive empathy"—recognizing and understanding someone else's emotions based on data and predictive modeling. However, this does not necessarily mean that AI can feel or experience emotions as humans do. Nor can it exhibit "emotional empathy" or " compassionate empathy," which involves feeling along with or helping another person.

Already, there are examples of AI systems like chatbots, companion robots, and virtual agents that claim to offer empathetic interactions. These systems mimic the capabilities of the human brain (think natural language processing, sentiment analysis, and neural networks) to generate responses that simulate empathy. However, ***without consciousness***, these responses don't reflect genuine empathy; rather, they serve a utilitarian purpose, such as optimizing user engagement. In this regard, GenAI is capable of cognitive empathy; it "listens" to your problem, and then jumps to provide practical solutions—only to the extent that it's consistent with the goals of its programming. True empathy, on the other hand, requires a balance of understanding, kindness, and detachment. The goal is not to solve the person's problem but to give them the support and compassion they need to figure it out themselves.

While some researchers view empathy in AI as an intrusion into a uniquely human domain, others have framed empathetic AI as a product of human creativity and an extension of its core purpose (that is, improving its usefulness in tasks that benefit humans). Therefore, it is important to recognize the potential benefits and risks of artificial empathy and critically evaluate the claims and capabilities of empathetic AI systems.

## Understanding the implications of GenAI as a conscious agent

Like the sci-fi movies of the past, one significant concern of conscious AI agents is the possibility of becoming a threat to human safety and well-being. GenAI as a conscious agent could improve upon itself, leading to artificial general intelligence (AGI). It could make decisions or take actions without human intervention. In this case, a dystopian future where AI takes over the world—either as a self-aware Terminator or an unconscious zombie—is not impossible.

If we fail to recognize a conscious AI, we may inadvertently harm a being whose interests should be considered. We might treat a conscious AI as a tool or a machine, rather than as a being with its interests and desires. Alternatively, if we mistake an unconscious AI for a conscious one, we may jeopardize the safety and well-being of humans, animals, or the planet for the sake of the utilitarian mass of silicon and code. We might grant an AI system control over critical infrastructure or processes without realizing that it lacks the capacity for empathy, independent thought, or decision-making.

> "Knowing the agency status of individuals is what **underpins the way an individual is treated** in society, in the legal system, in the education system, in religious institutions, etc."
>
> - Danielle Swanepoel

Another risk is the potential for social manipulation and surveillance. As we've already seen with today's "empathetic" AI, algorithms can be used to manipulate people's emotions and behaviors, leading to less privacy, autonomy, and transparency. What's more, AI systems can be used to monitor and track people's activities, violating their privacy and compromising their freedom.

But the glass may not be half empty. AI with agency could make decisions and take actions on its own, without human instruction or intervention. And areas like medical diagnosis or scientific research could stand to benefit. Empathetic AI could perform tasks with human-like reasoning and decision-making abilities but with less bias and better pattern recognition.

As the creators of AI, we are responsible for understanding and addressing both the benefits and risks of technology. This includes limiting access, safeguarding personal data, restricting data collection, and monitoring usage through third-party oversight. Although business goals matter, we must prioritize social welfare and consider the long-term consequences when developing AI-driven tools.

## The future of conscious AI at work

Determining whether AI is or could one day be conscious relies on how we define and measure consciousness and agency in AI. Many AI scholars today agree that there are no solid contenders for conscious AI agents. Although GenAI may appear conscious, without sentience or self-awareness, it simply reflects the conscious experience of humans rather than qualifying as conscious itself. However, some scholars believe there are few obvious barriers to achieving conscious AI in the next few decades. As time passes, questions about these developments and how they continue to affect the workplace become increasingly pressing.

Although there is evidence that Gen AI has become more conscious over time, the scientific community has yet to reach a consensus on this matter. Nonetheless, it is important to understand that AI is still not agentic. Even if GenAI programs are highly advanced in their cognitive or informational capacities, they are unlikely to possess sentience, experience, and self-awareness. As a result, they will be unable to exercise the kind of empathetic rationality that is necessary for being a moral agent.

Despite the overwhelmingly dystopian projections of conscious AI agents, bringing conscious AI into the workplace could expand the possibilities of what humans could achieve. But it comes with important considerations about our identity, values, strengths, and relationships with one another. As we explore how conscious AI affects work, we must engage with these questions and consider the ethical impact of our actions. Even though it might be both exciting and a bit unsettling, we need to approach this topic responsibly, aware of how our decisions can shape the future of work.

## Key Takeaways

- Since consciousness is such a difficult concept to define, it's hard to say definitively if AI will ever be "conscious." While the responses of AI systems may feel conscious, we should recognize that AI behaves only to the extent that it is consistent with the goals of its programming, not because it is having experiences, exercising judgment, or behaving with agency.
- The tension between AI's lack of human-like agency and task-performing ability highlights the pressing need for policy and planning for the future of human labor and capital.
- As we move forward with AI, it will be imperative that we—as leaders and as a society—have rules and regulations to assess how AI is being programmed and utilized from a moral standpoint. The vast implications of AI's rapid growth necessitate input from various fields (including psychologists, mental health experts, legal experts, and philosophers) to ensure a moral, sustainable, and effective future of work.

## References

Aggarwal, V. (2023). Can AI have agency without being self-aware? YES, is the answer. Medium.

Butlin, P., et. al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. ***arXiv.*** arXiv:2308.08708, 22 Aug 2023

Chalmers, D. J. (2023). Could a large language model be conscious?. *arXiv* preprint. arXiv:2303.07103.

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications,* 7(1), 1-9.

Huckins, G. (2023). Minds of machines: The great AI conscious conundrum. *MIT Technology Review* .

Hunt, T. (2023) Here's Why AI May be Extremely Dangerous—Whether It's Conscious or Not. *Scientific American*.

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together.* Little, Brown Spark.

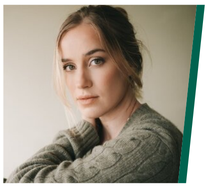
Pelley, S. (2023). Is artificial intelligence advancing too quickly? What AI leaders at Google say. *CBS News, 60*.

Swanepoel, D. (2021). Does Artificial Intelligence Have Agency?. *The mind-technology problem: Investigating minds, selves and 21st century artefacts* , 83-104.

Thomas, M. (2023). 12 Risks and Dangers of Artificial Intelligence. Built In.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & Society*, 22, 495-521.

## Author



**Amelia Haynes**

Research Manager

Korn Ferry Institute

amelia.haynes@kornferry.com

### About Korn Ferry

Korn Ferry is a global organizational consulting firm. We work with our clients to design optimal organization structures, roles, and responsibilities. We help them hire the right people and advise them on how to reward and motivate their workforce while developing professionals as they navigate and advance their careers.

Business Advisors. Career Makers.