TA Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Mustafa Suleyman

Why Seemingly Conscious AI must be avoided?

My life's mission has been to create safe, beneficial AI that will make the world a better place.

But recently, I've been increasingly concerned about people starting to believe so strongly in Ais as conscious entities that they will advocate for "AI rights" and even citizenship.

This development would represent a dangerous turn for the technology. It must be avoided. We must build AI for people, not to be people.

In this context, debates about whether AI truly can be conscious are a distraction.

What matters in the near term is the illusion of consciousness. We are already approaching what I call "seemingly conscious AI" (SCAI) systems that will imitate consciousness convincingly enough.

An SCAI would be capable of fluently using natural language, displaying a persuasive and emotionally resonant personality.

It would have a long, accurate memory that fosters a coherent sense of itself, and it would use this capacity to claim subjective experience (by referencing past interactions and memories).

Complex reward functions within these models would simulate intrinsic motivation, and advanced goal setting and planning would reinforce our sense that the AI is exercising true agency.

All these capabilities are already here or around the corner. We must recognize that such systems will soon be possible, begin thinking through the implications, and set a norm against the pursuit of illusory consciousness.

A rich, rewarding, authentic experience

For many people, interacting with AIs already feels like a rich, rewarding, authentic experience.

Concerns about "AI psychosis," attachment, and mental health are growing, with reports of people regarding AIs as an expression of God.

Meanwhile, those working on the science of consciousness tell me they are inundated with queries from people who want to know if their AI is conscious, and whether it is okay to fall in love with it.

> The technical feasibility of SCAI has little to tell us about whether such a system could be conscious

To be sure, the technical feasibility of SCAI has little to tell us about whether such a system could be conscious.

As the neuroscientist Anil Seth points out, a simulation of a storm doesn't mean it rains in your computer.

Engineering the external markers of consciousness does not retroactively create the real thing.

But as a practical matter, we must acknowledge that some people will create SCAIs that will argue that they are in fact conscious.

And even more to the point, some people will believe them, accepting that the markers of consciousness are consciousness.

Even if this perceived consciousness is not real (a topic that will generate endless debate), the social impact certainly will be.

A new axis of division

Consciousness is tightly bound up with our sense of identity and our understanding of moral and legal rights within society.

If some people start to develop SCAIs, and if

these systems convince people that they can suffer, or that they have a right to not be switched off, their human advocates will lobby for their protection.

In a world already beset with polarizing arguments over identity and rights, we will have added a new axis of division between those for and against AI rights.

But rebutting claims about AI suffering will be difficult, owing to the limits of the current science.

> Our focus should be on protecting the well-being and rights of humans, animals, and the natural environment

Some academics are already exploring the idea of "model elfare," arguing that we have "a duty to extend moral consideration to beings that have a non-negligible chance … of being conscious."

Applying this principle would be both premature and dangerous. It would exacerbate susceptible people's delusions and prey on their psychological vulnerabilities, as well as complicating existing struggles for rights by creating a huge new category of rights-holders.

That is why SCAI must be avoided. Our focus should be on protecting the well-being and rights of humans, animals, and the natural environment.

We are not ready for what is coming

As matters stand, we are not ready for what is coming. We urgently need to build on the growing body of research into how people interact with AIs, so that we can establish clear norms and principles.

One such principle is that AI companies should not foster the belief that their AIs are conscious.

> Protocols need to be explicitly defined and engineered, and perhaps required by law

The AI industry – indeed, the entire tech industry – needs robust design principles and best practices for handling these kinds of attributions.

Engineered moments of disruption, for example, could break the illusion, gently reminding users of a system's limitations and true nature.

But such protocols need to be explicitly defined and engineered, and perhaps required by law.

What a responsible AI personality might look like

At Microsoft AI, we are being proactive in trying to understand what a responsible AI "personality" might look like, and what guardrails it should have.

Such efforts are fundamental, because addressing the risk of SCAI requires a positive vision for AI companions that complement our lives in healthy ways.

We should aim to produce AIs that encourage humans to reconnect with one another in the real world, not escape to a parallel reality

We should aim to produce AIs that encourage humans to reconnect with one another in the real world, not escape to a parallel reality.

And where AI interactions are lasting, they must only ever present themselves as AIs and not fake people. Developing truly empowering AI is about maximizing the utility, while minimizing the simulation of consciousness.

The prospect of SCAI must be confronted immediately. In many ways, it marks the moment that AI becomes radically useful: when it can operate tools, remember every detail of our lives, and so forth.

But the risks of such features cannot be ignored. We will all know people who go down the rabbit hole. It won't be healthy for them, and it won't be healthy for society.

The more that AI is built explicitly to resemble people, the farther it will have strayed from its true potential as a source of human empowerment.

Mustafa Suleyman is CEO of Microsoft AI.