Does science need to redefine the nature of humanity with the coming of AI?

[A speech to be delivered at the International Workshop**: "Human. Meanings and Challenges"** during the 29th General Assembly of Members of the Pontifical Academy for Life on **February 12-13, 2024**.]

I would like to begin by saying that I am just a humble quantum physicist. Compared with the complexity of the human condition, physics is very simple: one equation can describe the whole universe. But traditionally, physicists have not *always* exhibited humility; throughout the past century they (we) have arrogantly claimed we can understand all of chemistry and biology too; we've announced that philosophy is dead; and we even discuss the nature of reality itself.

Nothing, it seems, is outside of our remit.

It is then even more arrogant of me to claim to be able to comment meaningfully on, or have special insights into, the *human condition*. But, here we are at a meeting that includes philosophers, theologians, sociologists, psychologists, and neuroscientists; can any of *them* claim to have exclusive or unique insights into the human condition in the 21st century? I suppose I have as much right as anyone to share my thoughts.

I should add here that at least my background and interests are spread more widely than those of a traditional academic physicist. Alongside my theoretical physics research, I have spent three decades as a science communicator, public intellectual, broadcaster, and author – all of which have exposed me to a wide range of ideas and perspectives. However, I will try to restrict my comments to where I feel I can say something meaningful, and which might hopefully add to the discussion.

The subject I wish to focus on is the place of artificial intelligence in today's world – not only in terms of how the technology will impact humanity and the way we live our everyday lives, but also whether it is changing our views about *what it means to be human* (in the sense of us being living, thinking, sentient, self-aware entities).

When it comes to AI, and indeed to *any* new technology, humans have a remarkable ability to adapt very quickly. What was considered as science fiction a couple of decades ago is part of our daily lives today. My father is 92 and uses a smart phone; he does online banking; and he has Google Nest smart speakers in his apartment. Yes, he needs help and advice now and then, but he is not afraid, or suspicious, of these wonderous technologies that have only recently appeared.

Of course, we have always used technology to make our lives *easier*.. but not necessarily '*simpler'*. When we consider the societal problems created by social media, such as the polarisation of political views and the spread of extremist ideologies, or the concerns over fake news and misinformation, maybe claiming that our lives are 'simpler' now is certainly not true. But in general, we have always adapted to new technologies and quickly integrated them into our daily lives, from the steam engine to the light bulb, from the washing machine to the microwave, and from the internet to the iPhone. All these have made our life easier; and we adapt to them so quickly that we forget what it was like without it.

Crucially, none of them have made us any less 'human'. They've changed us, yes – and we might argue it has not always been for the better – but they have not altered our essence: what it means to be human.

Today, artificial Intelligence, machine learning, robotics and automated systems have the potential to change our world faster and more fundamentally than *any* previous technological revolution. AI has moved out of the realms of science fiction and into our everyday lives, working unnoticed, very often behind the scenes. And AI will grow to become *the* pervasive technology of the 21st century and beyond.

So, how worried should we be? We are familiar with the media's obsession with sensationalism, and the fact that 'fear sells', along of course with Hollywood's depiction of machines taking over the world. But while the *Terminator* movies are good fun, we do need to be more measured about the risks of AI, for it is both prudent and logical that we debate and invest in AI safely and ethically. Governments and policymakers have a responsibility to understand the capabilities and limitations of AI technology as it becomes an increasing part of our daily lives. This will require an awareness of when and where this technology is being deployed.

The current definition of AI is as follows: It is the broad range of technologies with the ability to perform tasks that would otherwise require human intelligence, like visual perception, speech recognition, and language translation. It often has the capacity to learn or adapt to new experiences or stimuli – what we call machine learning.

But this definition is already looking out of date. Pattern recognition in data, whether in images or speech, and decision making based on that recognition, is really about the rather basic AI we already have with us. It doesn't include logical problem solving, or even genuine creativity that we are already beginning to see.

But over the next few decades, AI technologies aim to reproduce, and even surpass, abilities that would require 'intelligence' if humans were to perform them. These include learning and adaptation; sensory understanding and interaction, reasoning and planning, optimisation, autonomy, even creativity and intuition.

Over the past decade, the most exciting development without doubt has been in artificial neural networks and machine learning, which are far more effective at a wide range of tasks because they are closer to replicating the way we human's think than the old so-called 'classical AI'. The famous example of DeepMind's AlphaGo program defeating the world grandmaster at the Chinese game of Go is remarkable. The AI made a move during one game that no human could understand – not, that is, until much later in the game when it turned out to be a tactically brilliant move. Unlike chess, the game of Go relies mostly on intuition rather than simply crunching through possibilities, making AlphaGo's achievement something of a landmark event in AI development.

We are also making advances in what are known as semantic technologies, the aim of which is to help AIs interpret and truly understand the data by encoding meaning separately from the application code. Semantic technologies differ radically from the way machines traditionally interpret data, whereby meaning and relationships are hard-wired into the coding by human programmers. Like machine learning, semantic technologies take us along the path towards artificial '*intelligence'* in the true sense of the word.

Often, the debates surround the notion of super-intelligent machines that far surpass our own cognitive abilities. But that is something for the future. No one, not even the AI specialists themselves know when we will reach what's called *artificial general intelligence* – when machines can do everything humans can, even potentially developing machine consciousness. I will return to that point shortly. However, the more immediate challenges, opportunities and risks of AI are of a far more mundane.

A point worth repeating is this: technology in itself is neither good nor evil. It's how we humans choose to make use of it. AI is no different. So, when it comes to, say, military applications like

autonomous AI killer drones, just as one hopes chemists and biologists would not wish to build chemical or biological weapons—and indeed, most countries around the world have signed treaties not to develop them—so most AI researchers have no desire to build AI controlled weapons. There is therefore an urgent need to be discussing international treaties on AI use in the military. The issue, as with any technology, is that those in power will always want to explore its capabilities, not always for good. AI is no better or worse in this sense than any other technology that has military applications. For example, it is difficult to stop the use of drones altogether just because they can deliver bombs when pretty soon we will all be expecting them to deliver our pizza.

And who is to blame when AI goes wrong? Are autonomous systems different from other complex controlled systems? Should accidents be treated like other kinds of mechanical failure? This raises important legal issues. Reaction to failures of autonomous systems is somewhat different to reaction to failures of human controlled systems. For example, a failure of a surgical robot that has genuinely learnt from past experiences cannot be attributed to any one person. The same applies with driverless, AI controlled cars. We are likely to be far less forgiving if they cause a fatal accident. Never mind that we've avoided vastly more people being killed in car accidents caused by human error. It's an interesting moral dilemma.

Of course, on a much deeper level, many people feel unsettled about AI developing true intelligence, and even consciousness. What does this mean for humanity on a metaphysical level. I believe I am in the appropriate venue today to be confident that there will be those who contemplate what it might mean on the level of the soul or the spirit, and I leave that subject to them. I am a physicist interested in studying the material world. In that sense, I am also a physicalist. Physicalism is a philosophical position that asserts that everything that exists is fundamentally physical or material in nature. So, all phenomena, including consciousness, I believe, can ultimately be explained by physical components or processes. In other words, there is nothing beyond the physical world.

However, I am not a reductionist. I do not think that processes as complex as the workings of the human brain, consciousness and intelligence can be reduced to the sum of their constituent physical parts. Consciousness is an emergent property that we could never predict by studying individual neurons in the brain – no more than the wetness of water could be predicted from the study of individual H_2O molecules.

But on the other hand, consciousness isn't magic – it is ultimately still the result of software running on the hardware of the brain. I have never heard a convincing argument stating that consciousness isn't ultimately a physical process. I am not a mind/matter dualist.

So, what of AI intelligence. Certainly, anyone who has used a Large Language Model like ChatGPT over the past year gets the sense that it somehow 'understands' what it is saying when we ask it a question. It can even sometimes fool us into thinking it is self-aware and might even pass the famous Turing test. But the reality is that ChatGPT has no 'understanding' of what it is saying or is aware of its own existence. If you stop mid-conversation with ChatGPT and go off on holiday for a week, it isn't sitting there impatiently wondering when you are coming back. It's just lines of code and algorithms. It is no more self-aware than a virus.

I should add in its defence that at least ChatGPT is honest. I asked it recently if it *was* self-aware. It replied: "No, I am a machine learning model. I don't possess consciousness, self-awareness, or personal experiences. I can generate responses based on patterns learned from diverse data during training, but I don't have emotions, intentions, or awareness like a conscious being." Even here, there was no 'thinking' that went into that answer.

Some commentators are claiming that these Large Language Models are already approaching true artificial general intelligence, or AGI – that we are getting closer to *thinking* machines. This is not true. We are still a long way from AGI, even though ChatGPT is indeed getting ever-more.. sophisticated. Over the past few months there has been talk of something called Q* that will arrive this year. It goes beyond what's called generative AI like ChatGPT because it uses something called model-free reinforcement learning too, like those algorithms that can learn to play chess or Go better than a grand master in just a few hours of being given the rules of the games. But Q* is still not "thinking" in the way we understand it.

So, don't believe all the hype. When experts on AI talk about these models being intelligent or that they can learn and reason, they are not taking into account what defines true sentience, which would include not just logical reasoning but, say, value, or morality.

When we measure AI intelligence we must remember that it isn't just about computation or logical reasoning. The celebrated Harvard psychologist Howard Gardner explained how there are different kinds of human intelligence, each representing different ways in which we process information: logical-mathematical intelligence, spatial intelligence, linguistic intelligence, interpersonal intelligence, and so on. So, asking when AI might develop human level intelligence is not simply a case of having more powerful algorithms.

And we will certainly need a better understanding of consciousness. Consciousness is not like a light switch that is either on or off. It's a continuum, like a dimmer switch. We also know that consciousness is not a uniquely human attribute. Who would argue that their dog is not conscious? Unlike ChatGPT, your dog *will* miss you when you are away; it *will* feel guilty if you scold it, and it *does* feel pleasure and pain. It may not be able to solve mathematical equations, but then neither can a human baby.

So then if a dog is conscious, is a mouse conscious? We would have to say yes. Is a spider conscious? Is a fly conscious? You can see there is no sharp boundary. As we move down to less complex life forms, consciousness becomes dimmer. It is in this sense that we say it is an emergent property that depends on the complexity of the neural processes.

But if consciousness is emergent it could also appear gradually in a machine as we increase its processing complexity, rather than suddenly switching on once the AI algorithms became complex enough. Maybe the question we should be asking instead, when contemplating whether AIs could one day be self-aware, is what is it that distinguishes life from non-life?

I mentioned earlier that I label myself as a physicalist – that everything that exists is fundamentally physical or material in nature. I also have sympathy for what is called organicism: the idea that while there's something mysterious about life, which we have yet to fully understand, it can, in principle, be explained by the laws of physics and chemistry. As such I see no reason why life could not one day be created artificially. And if so, what ultimate difference would there then be between us and machines?

Artificial general intelligence (AGI) is the representation of generalized human cognitive abilities in computational software, with the intention that an AGI system can perform *any* task that a human being is capable of. But does AGI necessarily imply sentience and self-awareness? A machine that exhibits true AGI will need to also exhibit the higher-level thinking processes that involve what we might call emotions – that is, to greater or lesser extent, those categories of intelligence that Howard Gardner described.

Some argue that AGI could potentially exhibit behaviours that *mimic* consciousness without actually *being* conscious. In this view, it might only *simulate* understanding, emotions, and self-awareness without possessing subjective experiences: a zombie. Others believe that true consciousness might be a *necessary* aspect of AGI, especially if it is intended to replicate human-like intelligence and decision-making.

Humans clearly have self-awareness, with the capacity for subjective experiences, feelings, and sensations. This implies an ability to perceive and be aware of one's surroundings, to experience pleasure or pain, and to have a subjective inner life. But this does not require human level sentience or consciousness, since all these attributes are present in many other animals.

Maybe then the question really is: 'what does it mean to be alive?' Here scientists still cannot agree. We might have simplistic conditions for life such as the ability to make copies of ourselves to follow Darwinian Natural selection, or to be able to metabolise, but these are all physical processes. They may be complex from a biochemical point of view, but philosophically they are simple. A bacterium is alive, but it is certainly not conscious.

And in any case, I don't require an AI to have to be able to procreate or metabolise before we would acknowledge that it is conscious and self-aware.

What does all this mean for us humans? There are many challenges, and potentially even existential threats, that we need to confront in the face of the rapid advances of AI. And we should certainly be prepared for the day that machines might develop true intelligence and consciousness – just as we should prepare for the day when we might discover life beyond Earth. Neither of these should give us an identity crisis.

But will AI ever *think* or *feel* like a human? I would say no. Why would it? And indeed, why should it? What makes us human is more than the neural connections in our brains. It is more than our intelligence, intuition, or creativity – all of which will likely one day be replicated in AIs. What makes us uniquely human is also about our behaviour and interaction with our physical surroundings, our relationships with each other within complex collective structures and societies; it is our shared cultures and beliefs, our history, our memories.

All these are deep and complex issues. So, we probably need the help of AI to address them!!!

Jim Al-Khalili

8 February 2024