oxford public philosophy

Prof. Nayef Al-Rodhan

Sentience, Safe AI and The Future of Philosophy: A Transdisciplinary Analysis

Around the world, Artificial Intelligence (AI) is seeping into every aspect of our daily life, transforming our computational power, and with it the manufacturing speed, military capabilities, and the fabric of our societies. Generative AI applications such as OpenAI's ChatGPT, the fastest growing consumer application in history, have created both positive anticipation and alarm about the future potential of AI technology. Predictions range from doomsday scenarios describing the extinction of the human species to optimistic takes on how it could revolutionise the way we work, live and communicate. If used correctly, AI could catapult scientific, economic and technological advances into a new phase in human history. In doing so it has the potential to solve some of humanity's biggest problems by preventing serious food and water scarcity, mitigating inequality and poverty, diagnosing life-threatening diseases, tackling climate change, preventing pandemics, designing new game-changing proteins, and much more.

AI technology is rapidly moving in the direction of Artificial General Intelligence (AGI), the ability to achieve human-level machine intelligence, with Google's AI Chief recently predicting that there is a 50% chance that we'll reach AGI within five years. This raises important questions about our human nature, our sentience, and our dignity needs. Can AI ever become truly sentient? If so, how will we know if that happens? Should sentient machines share similar rights and responsibilities as humans? The boardroom drama at OpenAI in late November 2023 also deepened the debate about the dangers of techno-capitalism: is it possible for corporate giants in the AI space to balance safety with the pursuit of revenues and profit?

As AI advances at a breakneck speed, ethical considerations are becoming increasingly critical. Sentient AI implies that the technology has the capacity to evolve and be self-aware, in doing so feeling and experiencing the world just like a human would. According to the British mathematician Alan Turing, if the human cannot distinguish between whether it is conversing with an AI or another human, then the AI in question has passed the test. However, given AI's sophisticated conversational skills and ability to give the impression of consciousness, the Turing Test is becoming too narrow and does not grasp all the nuances of what makes us sentient and, more broadly, human. To stay on the front foot of technological progress, we need to supplement the Turing Test with transdisciplinary frameworks for evaluating increasingly human-like AI. These frameworks should be based on approaches rooted in psychology, neuroscience, philosophy, the social sciences, political science and other relevant disciplines.

We do not yet have a full understanding of what makes a thing sentient but transdisciplinary efforts by neuroscientists, computer scientists and philosophers are helping develop a deeper understanding of consciousness and sentience. So far, we have found that emotions are one of the important characteristics needed for sentience, as is agency or intrinsic motivation. A sentient AI would need to have the ability to create autonomous goals and an ability to pursue these goals. In human beings, this quality has evolved from our intrinsic survival instinct, while in AI it is still, for now, lacking. According to recent studies, a sense of time, narrative, and memory is also critical for determining sentience. A level of sentience comparable to humans would require autobiographical memory and a concept of the linear progression of time. In current AI systems, these capabilities are limited - but recent developments raise uncomfortable philosophical questions about whether sentient AI should share similar rights and responsibilities in the event that it becomes a reality. And if so, how does one hold the technology accountable for their actions? And how will we define - legally and ethically - sentient AI's role in society? We currently treat AI technology and machines as property, so how will this change if they are granted their own rights? There is no clear-cut answer, but as I argued in 'Transdisciplinarity, neuro-techno-philosophy, and the future of philosophy', we should attribute agency to machines whenever they appear to possess the same qualities that characterise humans. I also believe that machines ought to be treated as agents if they prove themselves to be emotional, amoral, and egoist.

These debates, however they unfold, will clearly have deep implications on the future of philosophy itself. In 'Transdisciplinarity, neuro-techno-philosophy, and the future of philosophy' I make the case that it is a short step from AI's present capabilities to its potential future use developing novel philosophical hypotheses and thought experiments. It is therefore not unthinkable that future AI systems could break new ground in the field of normative ethics, helping pinpoint moral principles that human philosophers have failed to grasp. However, we should be mindful that their conception of morality or beauty, for example, may have nothing in common with ours, or it may supersede our own capacities and reflections. This could limit the ability of sophisticated artificial agents to answer long-standing philosophical questions, however superior they may be to the most advanced human intellectual output. We should consider how these developments are likely to impact how we understand the world around us, both in terms of the subject matter and of the theorising entity involved. Artificial agents will no doubt be put under the microscope and will be studied alongside the human mind and human nature: not just to compare and contrast, but also to understand how these artificial entities relate to - and treat - one another, and humanity itself. There is also the question of how human philosophers will react if and when AI-steered machines become superior philosophical theorisers. Will flesh and blood philosophers be forced to compete cognitively with entities whose intellectual abilities vastly supersede our own? Will AI systems overtake our limited human reasoning and reflective capacities? If this happens, what does this mean for our own human agency, the control we have over our lives and the future of our societies?

It is still unclear if and when AGI will become sentient. But even if it doesn't, it will still be possible to build a highly intelligent machine without it being sentient: AI does not have to be AGI-level smart to take control of our human minds, societal norms, lives and transnational futures. That is why we need to be alert to the potentially catastrophic effects of AI technologies if dignity needs and civil liberties are compromised, regardless where or for whom. I believe that humans are emotional, amoral, and egoistic beings, and dignity is the most fundamental of our human needs. I define dignity holistically, to mean much more than the absence of humiliation but also the presence of recognition, through a set of nine dignity needs: *reason, security, human rights, transparency, justice, accountability, opportunity, innovation, and inclusiveness*. AI will impact these needs in complex and unpredictable ways, by reinforcing or endangering them - and sometimes, both. Fundamental freedoms, such as the right to privacy and civil liberties, are threatened by intrusive surveillance as well as new policing and profiling methods powered by AI systems. AI tools can also target individuals online with tailored content, in doing so reducing exposure to different points of view and reinforcing biases and vulnerabilities. It is essential that biases embedded in algorithms do not amplify discriminatory practices and deepen social injustice, alienation, and discrimination.

Powerful AI technologies will progressively increase our capabilities, for good or ill. We therefore need to be clear-sighted about the AI governance frameworks urgently needed to futureproof the safe use of AI. The recent high drama at OpenAI, whose founding mission is “to ensure that artificial general intelligence benefits all of humanity”, gave us a glimpse of the main rift in the AI industry, pitting those focused on commercial growth against those uneasy with the potential ramifications of the unbridled development of AI. However well-motivated AI governance schemes might be, they are less robust than one would hope. At the same time, self-regulation by global tech companies is becoming increasingly difficult given the large sums at stake and the economic and political influence of these companies.

With this in mind, we must keep an open mind not just about the immediate man-made dangers of AI technologies but also their potential to redefine what it means to be human. They will shape how we understand and engage with the world, in doing so making us reevaluate our place in it. Our chances of survival as a species and the likelihood of our existence in a free, independent, peaceful, prosperous, creative and dignified world will depend on the future trajectory of AI. Our historical yearning for longing and belonging hangs in the balance. To protect citizens from potential harm and limit the risks, AI should be regulated just like any other technology. We must also apply transdisciplinary approaches to make sure that the use and governance of AI is always steered by human dignity needs for all, at all times and under all circumstances. AI’s trajectory is not predetermined, but the clock is ticking and humanity may have less time than it thinks to control its collective destiny.

about the author

Professor Nayef Al-Rodhan is a philosopher, neuroscientist, geostrategist and futurologist. He is an Honorary Fellow of St. Antony’s College, Oxford University; Head of the Geopolitics and Global Futures Programme, Geneva Center for Security Policy (GCSP) in Switzerland; Senior Research Fellow, Institute of Philosophy at the University of London; and a Member of the Global Future Council on the Future of Complex Risks at the World Economic Forum. He is also a Fellow of the Royal Society of Arts (FRSA). He holds an MD and PhD, and was educated and worked at the Mayo Clinic, Yale University and Harvard University in the United States.

He is a prize-winning scholar who has written more than 300 articles and 25 books, including most recently 21st-Century Statecraft: Reconciling Power, Justice And Meta-Geopolitical Interests, Sustainable History And Human Dignity, Emotional Amoral Egoism: A Neurophilosophy Of Human Nature And Motivations, and On Power: Neurophilosophical Foundations And Policy Implications. His current research focuses on transdisciplinarity, neuro-techno-philosophy, and the future of philosophy, with a particular emphasis on the interplay between philosophy, neuroscience, strategic culture, applied history, geopolitics, disruptive technologies, international relations, and global security. His books and articles may be found at www.sustainablehistory.com.