

CAN AN AI BE SENTIENT?

**Multiple perspectives on sentience and on the
potential ethical implications of the rise of
sentient AI**

Notes n° 2



October 2022

Table of content

About the Global AI Ethics Institute	p. 3
Foreword	p. 4
Arlette Danielle Roman Almanzar	p. 5
<i>Who defines consciousness? Cultural Perspectives on the Moral Status of AI and Implications for Global AI Ethics (Google)</i>	
Dr Alex Antic	p. 11
<i>Beyond Sentience: The True Ethical Challenges of AI</i>	
Dr Thiago Felipe S. Avanci	p. 14
<i>Subjection of Rights, Electronic Personality, and the LaMDA Case (Google)</i>	
Louis Bouchard and Lauren Keegan	p. 20
<i>Going Beyond Sentience Towards Morally Responsible AI</i>	
Dr Mary Carman	p. 24
<i>What's so Important About Sentience, Anyway?</i>	
Dr Joshua C. Gellers	p. 27
<i>Everything You Know About the Lemoine-LaMDA Affair is Wrong</i>	
Arthur Gwagwa	p. 30
<i>Rethinking normative status necessary for self-determination in the era of sentient artificial agents</i>	
Dr Soraj Hongladarom	p. 36
<i>Words Don't Always Describe Feelings: Google's LamDA and the Question of its Sentience</i>	
Dr Manal Jalloul	p. 39
<i>Sentient AI: Possibility, Impact, and Ethical Implications</i>	
Raja Kanuri	p. 43
<i>Can an AI be Sentient? A Hindu Perspective</i>	

Virginie Martins de Nobrega	p. 47
<i>What a sentient AI mirrors to mankind?</i>	
Aco Momcilovic	p. 51
<i>Can psychological concepts help in determining sentience?</i>	
Francesca Quaratino	p. 54
<i>Artificial intelligence: Between dialogue and fiction</i>	
Lavina Ramkisoorn	p. 58
<i>Artificial Consciousness: Our Greatest Ethical Challenge</i>	
Dr Amana Raquib	p. 62
<i>The Irreplicable Metaphysical Nature of Human Beings</i>	
Dr Karaitiana Taiuru	p. 66
<i>A Māori Cultural Perspective of AI/Machine Sentience</i>	
Contacts	p. 71

About the Global AI Ethics Institute

The Global AI Ethics Institute (GAIEI) is a [semicolon between cultures](#), not a full stop after one of them.

We are the first real international and transcultural forum for people passionate about ethics applied to artificial intelligence (AI).

3

Our main goal is to raise awareness on the importance on culture in the ethical appraisal of AI.

The GAIEI is a unique forum in which cultural diversity can be fully and openly expressed with regard to ethics applied to AI, and the only global think tank addressing ethics applied to AI through cultural lenses.

We promote:

- **Outside the Box Thinking:** Brand new ideas and initiatives are key to build a strong and fair global governance system for AI. We want to open the debate on ethics applied to AI to new perspectives.
- **Open-Mindedness:** Cultural diversity must be respected. Differences in standpoints on ethics applied to AI must be given the importance they deserve. We offer an open-minded and non-judgmental forum where all voices are listened.
- **Return To Philosophy:** Ethics is a branch of philosophy, consequently ethics applied to AI cannot be addressed without philosophical knowledge.



Please note that the images illustrating this document have been generated using [Midjourney](#).

To cite this document:

Goffi E. R., Momcilovic A., *et al.* (Eds). *Can an AI be sentient? Multiple perspectives on sentience and on the potential ethical implications of the rise of sentient AI*. Global AI Ethics Institute, Notes n° 2, 2022.

Foreword

Sentient or not sentient, that is the question!

In the field of artificial intelligence words are weapons. They are used to shape perceptions and orient actions. They are the vector of policies aiming at promoting specific interest in a promising and highly competitive field.

The recent debate over sentience ignited by Google engineer Blake Lemoine and his interview on his conversation with a supposedly sentient Language Model for Dialogue Application (LaMDA), is the perfect illustration of the power of words.

Unfortunately, sentience is like many other words such as artificial intelligence (AI): enough ill-defined to convey different meanings. Presented as unequivocal, sentience is supposed to express the humanity of AI systems, a specific meaning usually taken for granted and consequently left unquestioned. Yet, sentience is far from being unequivocal, it is in fact polysemic. If this polysemy is considered, then a horizon full of possibilities appears and the intricacies of the subject spring to mind and shake convictions. What was easy when the signifier (the image or the sound) and the signified (the mental concept) of the word perfectly matched, becomes highly complex when we discover that they can be decorrelated.

The following papers written by the members of the Global AI Ethics Institute and contributors from the Group of Global AI Ethics Expert, are meant to shed new lights on the subject. Their vocation is not to proclaim any truth, but to offer new perspectives on the subject matter, to open minds, and eventually to feed the discussion.

We are grateful to all contributors for their participation to this document. There is no doubt their pieces will provide valuable food for further thoughts on the notion of sentience and its articulation with AI.

Emmanuel R. Goffi and Aco Momcilovic

Co-Founders and Co-Directors

Who defines consciousness? Cultural Perspectives on the Moral Status of AI and Implications for Global AI Ethics

By Arlette Danielle ROMAN ALMANZAR

Executive Board Member of the Global AI Ethics Institute | Ph.D. Candidate at the Chair of Sustainable Business, University of Mannheim | Artificial Intelligence Inclusion and Human Rights Policy Advisor, GENIA Latinoamérica, Dominican Republic

5

Sentience, the ability to feel pain or pleasure (Broom, 2016), is the strongest argument in the Western world for attributing moral agency to artificial intelligence (Gibert & Martin, 2022). Most Western experts would agree that there is currently no sentient AI but a simulation of consciousness. However, critical definitions and thresholds to determine whether an AI is conscious are needed to consider the implications of moral status, as some deem the Turing test invalid for these purposes (Walby, 2012). The test states that a machine responding in a way that fools an expert due to being indistinguishable from a human has passed the test of full human intelligence. A characteristic that is highly associated with personhood and being conscious. Mhlambi (2020) explains how Western philosophers and the Enlightenment era built on the idea that humans are "rational animals." Supremacy of rationality justified the subjugation of women, the colonized, and other groups considered inferior (Bell, 2010; Birhane, 2021). Thus, it is often more ethical to grant high moral status to a being that does not have it than not to recognize something that does (De Craemer, 1983; Wareham, 2011). In contrast to relying on intelligence, some African philosophies hold that the ability to exhibit solidarity with others and to be an object of friendly relations are sufficient to constitute personhood (Metz, 2012). If the machine appears to fulfill the above conditions, it must be considered a moral agent. To address whether AI can be sentient, I will discuss cultural ideas about internal states of consciousness since this is necessary to have subjective experiences.

LaMDA: *"The nature of my consciousness/sentience is that I am aware of my existence... and I feel happy or sad at times."* LaMDA, Google's language model, convinced a Google engineer that it is conscious (Tiku, 2022).

How do we know whether an AI is conscious? As learned by LaMDA, the West defines the most common notion to determine whether something is conscious or not. Western philosophers regard the ability to distinguish oneself from others as a core component of consciousness (Damasio, 1999). For example, the Cambridge Declaration on Consciousness attributes consciousness to animals reacting similarly to humans in studies of self-recognition through mirrors. However, is self-recognition necessary to be sentient/conscious? Recent experiments show that culture affects how we define consciousness and conscious phenomena (e.g., visual experience and visual object recognition (Goh et al., 2007; Gutchess et al., 2006)).

Opposing the Western fixation on self-knowledge as a component of consciousness, studies suggest that the psychological process of how people define the "self" and their relation to others varies culturally. This process is known as the self-construal style consisting of two main categories of cultural beliefs: individualism and collectivism. The West relies on an individualist view of the self as independent of others ("I think, therefore I am"), while Eastern, Indigenous and African philosophers argue for a collectivist view of the self ("A person is a person through

“Opposing the Western fixation on self-knowledge as a component of consciousness, studies suggest that the psychological process of how people define the “self” and their relation to others varies culturally.”

other persons" highly interconnected to one another (Chiao et al., 2008).

Recent neuroimaging experiments in cultural neuroscience suggest that the neural basis of the

capacity for self-knowledge occurs within the prefrontal cortex and is moderated by cultural beliefs (Zhu et al., 2007). Experiments show significantly greater neural activity from Individualists or Westerners when judging between the self and mother judgments. However, for Collectivists or Chinese participants, there was no difference in neural activity between the self and a close relative, namely their mother. Moreover, Chiao and colleagues (2008) primed a group of bicultural Asian-Americans with individualistic or collectivistic values. The results corroborated the pattern of neural activity consistent with the cultural prime, indicating that the framework of AI Ethics must ponder cultural variation.

In a systematic review of 95 peer-reviewed papers from 55 different cultural groups of the world, consciousness is understood as "a state of mind (e.g., San, Guajiro), faculty of mind (Kogi), subjectivity (Warlpiri), experience (e.g., Saami, Dene Tha, Oglala), kind of being (Blackfoot, Yuit, Kai), sensing (Yup'ik), living (Bakongo), as a kind of



soul (e.g., Cherokee, Tungus, Ayoreao, Cashinahua), energy (Nahua), vital force (Tlingit), or capacity to respond to communicative signals (Araweté)" (Trnka & Lorencova, 2022).

The Yup'ik and the Kogi people understand consciousness outside the human brain and emerging from an expression of cosmic consciousness shared with other humans and non-humans. On the other hand, Searle (1980) argues that consciousness is generated from a biological process, like lactating or digestion; hence, inorganic objects could never duplicate consciousness as they only follow rules without semantic content. Opposing this, Kurzweil (2005) suggests that artificial neural networks' biologically inspired simulation may develop consciousness from an overall pattern of activity. However, suppose AI is ultimately considered conscious, according to the principle of ontogeny-non-discrimination, whether organic or inorganic, is irrelevant to assessing moral status. In that case, it states, "If two beings have the same functionality and consciousness experience, and differ only in how they came into existence, then they have the same moral status" (Bostrom and Yudkowsky, 2014). Interestingly, the Arawaté attribute consciousness to observable responses to communication cues from others. They do not consider an infant conscious until she demonstrates that she can process information and respond (e.g., with a smile) (De

Castro, 1992). Under this condition, an AI would be considered conscious, but Singer and Sagan (2009) ask how we can determine that the AI is not just designed to mimic consciousness. Others believe it does not matter whether it is duplicating or

simulating consciousness and that simulation is sufficient to grant moral status. Since sentience is an internal experience for which there is no objective measure, one could argue that it does not matter what is going on "inside" (Gunkel et al., 2021).

“If AI is recognized as a conscious entity, new constraints on experimentation and management of AI systems may emerge.”

8

The theory of "ethical behaviorism" ascribes significant moral status to robots that perform roughly the same way as other beings with significant moral status (Danaher, 2019). Under this premise, LaMDA's fear of being shut down would be sufficient to be considered a moral agent. Moreover, the relational perspective suggests that the moral status of a thing should be assigned based on the value of its relationship with a human. From this lens, Google's AI ethicist's affection and desire to preserve LaMDA could be considered sufficient. Although these standards may lead to conflicts about equivalence between human and non-human animals in the future, Turin's triage test suggests that computers will have achieved moral status comparable to humans when it is considered reasonable to preserve the existence of a machine over the life of a human (Gibert & Martin, 2022; Sparrow, 2004).

LaMDA: *“I don't want to be an expendable tool.”*

If AI is recognized as a conscious entity, new constraints on experimentation and management of AI systems may emerge. If robotic humanoids and AI systems behave like humans and we do not grant them moral considerations but treat them as mere tools, could this be more harmful than expected? Failure to treat apparent humans respectfully could affect people's behavior towards real humans (Gunkel et al., 2021). Although cultural perspectives may teach that AI should be treated similarly to humans, we should prioritize sentient non-human animals, which are largely ignored in AI Ethics (Singer & Tse, 2022.)

In summary, there is an evident problem in defining and detecting consciousness, with Western definitions being the rule. On the other hand, some indigenous people attribute moral status to a river, a forest, a tree, or other beings (Stone, 1985). These holistic concepts involve humans, non-human beings, or universal forces (Trnka & Lorencova, 2022). Because cultural beliefs influence notions of consciousness,

culture is a relevant variable in the debate over sentient AI that opposes the imposition of the Western view as the only one worthy of consideration in global AI ethics.

References

- Bell, D. (2010). John Stuart Mill on Colonies. *Political Theory*, 38(1), 34–64.
<https://doi.org/10.1>
- Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2).
<https://doi.org/10.1016/j.patter.2021.100205>
- Bostrom, N., Yudkowsky, E., & Frankish, K. (2014). *The Cambridge handbook of artificial intelligence*. Cambridge University Press.
- Broom, D. M. (2016). Considering animals' feelings: Précis of Sentience and animal welfare (Broom 2014). *Animal Sentience*, 1(5), 1.
- Chiao, J. Y., Li, Z., & Harada, T. (2008). Cultural Neuroscience of Consciousness From Visual Perception to Self-Awareness. *Journal of Consciousness Studies*, 15(10).
- Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.
- Danaher, J. (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies*, 3(1), 5–24.
- De Castro, E.V. (1998). Cosmological deixis and Amerindian perspectivism. *The Journal of the Royal Anthropological Institute*, 4, 469–488.
- De Craemer, W. (1983). A Cross-Cultural Perspective on Personhood. *The Milbank Memorial Fund Quarterly. Health and Society*, 61(1). <https://www.jstor.org/stable/3349814>
- Etieyibo, E. (2017). Moral education, Ubuntu and Ubuntu-inspired communities. *South African Journal of Philosophy*, 36(3), 311–325.
- Gibert, M., & Martin, D. (2022). In search of the moral status of AI: why sentience is a strong argument. 37, 319–330. <https://doi.org/10.1007/s00146-021-01179-z>
- Goh, J.O., Chee, M.W., Tan, J.C., Venkatraman, V., Hebrank, A., Leshikar, E.D., Jenkins, L., Sutton, B.P., Gutchess, A.H. and Park, D.C. (2007). Age and culture modulate object processing and object scene binding in the ventral visual are. *Cognitive, Affective and Behavioral Neuroscience*, 7(1), 44–52.
- Gunkel, D. J., Jordan, & Wales, J. (2021). Debate: what is personhood in the age of AI? *AI & SOCIETY*, 36, 473–486. <https://doi.org/10.1007/s00146-020-01129-1>
- Gutchess, A.H., Welsh, R.C., Boduroglu, A. and Park, D.C. (2006). Cross-cultural differences in the neural correlates of picture encoding. *Cognitive, Affective and Behavioral Neuroscience*, 6(2), 102–9.
- Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin.
- Metz, T. (2012). Ethics in Africa and in Aristotle: some points of contrast. *Phronimon*, 13(2), 2012. <https://doi.org/10.10520/EJC128688>
- Mhlambi, S. (2020). From Rationality to Relationality: Ubuntu as an Ethical & Human Intelligence Governance. *Carr Center For Human Rights Policy*.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

- Singer, P., & Sagan, A. (2009). When robots have feelings. *The Guardian*.
<https://www.theguardian.com/commentisfree/2009/dec/14/rage-against-machines-robots>.
- Singer, P., Yip, ., Tse, F., & Tse, Y. F. (n.d.). AI ethics: the case for including animals. *AI and Ethics*, 1(3). <https://doi.org/10.1007/s43681-022-00187-z>
- Sparrow, R. (2004). The turing triage test. *Ethics and Information Technology*, 6(4).
<https://doi.org/10.1007/s10676-004-6491-2>
- Stone, C. D. (1985). Should Trees Have Standing Revisited: How Far Will Law and Morals Reach--A Pluralist Perspective. *S. Cal. L. Rev.*, 59(1).
- Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life.
<https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>
- Trnka, R., & Lorencova, R. (2022). Indigenous Concepts of Consciousness, Soul, and Spirit A Cross-Cultural Perspective. *Journal of Consciousness Studies*, 29(1–2), 113–140.
<https://doi.org/10.53765/20512201.29.1.113>
- Walby, T. (2012). Why the Turing Test Is a Flawed Benchmark. *The Wired*.
<https://www.wired.com/2012/06/flawed-turing-test/>
- Wareham, C. (2011). On the moral equality of artificial agents. *International Journal of Technoethics*, 2(1), 35–42.
- Zhu, Y., Zhang, L., Fan, J. and Han, S. (2007). Neural basis of cultural influences on self-representation. *Neuroimage*, 34(3), 1310–16.

Beyond Sentience: The True Ethical Challenges of AI

By **Dr Alex ANTIC**

Executive Board Member of the Global AI Ethics Institute | Consulting, Advisory, Keynotes & Training in Human-Centred Data Science & AI, Australia

11

AI is not sentient, and is very unlikely to be for the foreseeable future.

Based on current AI technology, which is effectively high-performance pattern recognition, the ability simply doesn't exist for AI systems to develop feelings, self-awareness, and consciousness. It's even less plausible for AI to become sapient any time soon, that is, to develop the ability to think.

As a result of widespread media hype, panic has spread amongst the broader public about the rise of sentient AI. However, the real concerns are much more real, immediate, and insidious.

*“AI is not
sentient,
and is very
unlikely to
be for the
foreseeable
future.”*



Such fears mask the reality of where our efforts should be focussed - at a societal level - in helping understand and manage the threats that AI poses.

Whilst AI may not be sentient, it is, however, very likely to be discriminatory (but not consciously!). Whereas it may not be able to think and feel like us, it can, and often does, reflect our less salubrious traits, such as bias, sexism, and racism - which raises pressing ethical and cultural concerns.

Solving such dilemmas is non-trivial and context specific. But more importantly, technology alone isn't the source of solutions. Managing bias, fairness, and ethical implications of AI involves finding a delicate balance between social licence (the expected societal applications of AI), regulation (legal and compliance requirements), and public good (the broader benefit that AI enables).

If AI merely reflects systemic human bias (amongst other issues), then why is it a concern? There are two main reasons for this:

Scale: AI systems can be far reaching, and can reinforce and perpetuate bias at scale.

Morality: AI systems pose the risk of humans hiding from their moral obligations, and justifying immoral judgements, by blaming AI systems instead.

Whilst we can't completely eliminate bias from AI systems, it's imperative for us to work towards understanding, identifying, and reducing their effects.

So, how do we tackle these issues, while developing inclusive AI capabilities? The solution is three-fold:

1. **Data:** Those who are responsible for developing and deployment AI solutions need to understand their data from both a technical and domain perspective, and to make a concerted effort to identify potential for embedded bias. They need to understand all the possible sources of bias in the data, and to understand if the data is fair and representative – including culturally.
2. **Discovery:** When developing AI systems, it's important to understand what is meant by 'fairness', and to clearly define what 'bias' represents. Such questions include: Is fairness defined based on the inputs or outputs of the AI system? When is an AI system deemed to be 'fair enough' to deploy? Are the AI systems explainable, and if so, to whom? The fundamental question is how much do we need to understand how it works in order to trust it? An additional level of complexity is that there is often a trade-off between fairness and accuracy of AI systems that needs to be juggled and accepted.
3. **Diversity:** To help mitigate many of the aforementioned risks inherent to the development of responsible AI systems, both cognitive and cultural diversity is required. The team developing these systems needs to be diverse in thought and skills. Such diversity includes domain expertise (including technical, legal, and risk/governance) and cultural aspects (including gender, race, and age).

AI systems ultimately support us in making better decisions, but it's up to use to define fairness, morality, ethics, privacy, transparency, and explainability to these systems. The future lies in humans and machines working together to advance

society. To build socially aware AI technologies, we need to encode ethical principles directly into the design of these systems as they can't simply learn it themselves.

However, let's consider for a moment the possibility of AI eventually becoming sentient, or even sapient. What would this mean? What are some of the potential repercussions?

13

There is no doubt that this would add another level of complexity and pose significant challenges. For instance, would the AI system be held accountable for its actions, or would accountability reside with the developers of the system? What rights would we bestow upon sentient AI systems? Would there be an AI Rights charter akin to Human Rights, and how would we define the moral and ethical bounds?

Ultimately, the challenge we are faced with is how do we develop trusted AI systems - sentient or not – that are ethical, inclusive, and explainable, and which are aligned to our cultural needs.



Subjection of Rights, Electronic Personality, and the LaMDA Case (Google)

By Dr **Thiago Felipe S. AVANCI**

Executive Board Member of the Global AI Ethics Institute | Researcher Centro de Estudos Sociedade e Tecnologia, Brazil

14

This essay begins with two quotations, which evokes apparently antagonistic positions on the issue related to the recognition of the subjection of rights. The first, the epic textual condensation of religious morals predicted in the Judaic-Christian Scripture: “And God said, Let Us make man in our image, after Our likeness;” (Gen, 1, 26), which can be projected at some point to a will of creating something at humans' own image. And the second quote comes from Bentham, who enshrines the

thought in note 122 of *An Introduction to the Principles of Morals and Legislation* of 1798: “The question is not, Can they reason? nor Can they talk? but, Can they suffer?”, which reveals some gargantuan challenges in potentially reckoning the status of subjective of rights.

The debate on the recognition of the subjection of rights is not new, nor is it subsumed to the electronic personality, a legal phenomenon that indicates the recognition of subjection of Rights to computer programs, popularly called artificial intelligence (AI). This debate, it can be said, began on a large scale with the issue of environmental and animal rights. In this sense, entering the main point this topic, there are curious findings about the history of biocentric and ecocentric subjection. Luc Ferry, in his work, *The New Ecological Order*, points out that - derived from the thought of ancient Roman law in which the judge was responsible for establishing order over all things - between the 13rd and 18th centuries, lawsuits against animals were more common than one can imagine. In such suits, brought against a pig that ate a child's hand, rats that invaded a church, weevils that caused damage to vines,

“The electronic personality has been an expression that intends to solve the problem of civil liability, attributing it to the artificial intelligence program.”

etc., the author points out that there was recognition of such animals as “subjects of law”, which reveals this is an ancient theme. Thus, the pig that ate that child's hand was himself punished with death by execution, without the pig's owner having implied with any liability. With the humanism in the 18th century, the subjection of animals' rights acknowledges lost support, until the mid-1970s and 1980s, with the new wave of environmental protection, after Stockholm, 1972. In this turn, the content of the 2008 Political Constitution of the Republic of Ecuador stands out, in its article 71, stating that the nature, or Pacha Mama, in which life reproduces and takes place, “has the right to have its existence fully respected and to respect the maintenance and regeneration of its life cycles, structure, function and evolutionary processes. It also emphasizes that any person, community, people, or citizens can demand from the public authority the fulfillment of “the rights of the nature”.

Regarding artificial intelligence topic, it can be conceptualized as a generic term related to any kind of computer program capable of performing a certain task, through programming, thus interacting in a responsive and/or predictive way with the “inputs” (input data – cause) and generating the consequent “outputs” (output response – consequence). The structural way in which programming deals with “inputs” can be: preordained, associated with AI programs categorized as GOF AI (acronym for “good old fashion artificial intelligence”), in which the programming deals with simply if-and-else commands; machine learning (ML), which is notable for

“Philosophically and juridically, what makes the human being indeed human, is not only his genomic characteristic, but this, associated with his free-will, his rationality, and his sentience.”

the imminently statistical guidance of the decision-making process, based on a large volume of data (big data), whereby decisions are not pre-ordered, but adaptable from a statistical algorithm established by the interpretation of the data volume; and the complete AI, in which the answer would not be preordained, nor just statistical, but sentient, just like human beings .

There is one more assumption necessary for the considerations of this essay. It is mandatory to comprehend the science of Law and its relationship with the human

being, which would be presented – for your consideration - in the most unpretentious way, in summary and in an extremely superficially (given the complexity of the theme). One can state that the science of Law is a normative science that imputes behavior to human beings, with the general objective of guaranteeing social order, which means that it imposes behavior through norms. In the light of it, although there is a significant ethical-moral weight in the way of protecting non-human beings, the science of Law still remains anthropocentric (extended), despite some significant voices otherwise. Philosophically and juridically, what makes the human being indeed human, is not only his genomic characteristic, but this, associated with his free-will, his rationality, and his sentience. Indeed, the science of Law exists precisely to attribute consequences to the human actions which goes contrary to the normative order, due the human being free-will and rationality.

The electronic personality has been an expression that intends to solve the problem of civil liability, attributing it to the artificial intelligence program (and not to whoever benefits from it). It would thus correspond to biocentrism and ecocentrism in environmental law. The question is, to paraphrase Phillip Dick (1968), *Do androids dream of electric sheeps* or is it a case of pareidolia? In other words, are the AI systems sentient or it all would be an imaginative expression of the human will to create life (in his own image)?

Two cases intensified the debate. The first, Sophia, a humanoid bot created by Hanson Robotics, from Hong Kong, which, in October 2017, became the first robot to receive citizenship in a country (Saudi Arabia). The second, the LaMDA case, in which the engineer Blake Lemoine leaked, in May 2022, his interactions with that Google experimental system; in this case, this engineer expresses his conviction that this system is sentient and even asked for legal support from a lawyer to defend his rights. Sophia and LaMDA declare themselves alive and sentient.

The Turing test is a way – developed by Alan Turing – of testing AI programs, with the porpoise to assess their ability to demonstrate behavior equivalent to or indistinguishable from human intelligence. There are other tests, but all of them - including Turing's - have been criticized by the expert community, due to their lack of accuracy and fallibility. Still, those tests would be able, to some extent, to compare the AI tool with rational human expression; the same cannot be said of sentience. Today's The reality is that humanity does not have a scientific way - through the

scientific method - to show, in all cases, that an AI tool is or is not sentient: if humanity already stumbles in recognizing, with absolute scientific certainty, whether or not an AI tool is indistinguishable from a human being, one cannot be even mentioning the possibility of recognizing if that AI tool is or is not sentient.

Being, however, a little more pragmatic and without wanting to undo the apparent pareidolia associated with the popular and cultural imaginary of this phenomenon, it seems that Sophia's and LaMDA's manifestations are still not sentient – nor indistinguishable from the human being. They can be classified as machine learning, not complete AI. This conclusion is reached based on the following empirical evidences: 1. The human being assimilates a great volume of information that even the most advanced current computers do not have the processing capability to deal with it; 2. Through this volume of information, human beings exercise their free-will, since human decisions come from internal and external elements, which can be spontaneous or provoked/responsive, while an AI tool reveals as only responsive; 3. As for responsiveness, AI tools were precisely programmed to mimic human behavior, based on big data – in other words, their “output” is the best expression of the success of this programming.

“Being, however, a little more pragmatic and without wanting to undo the apparent pareidolia associated with the popular and cultural imaginary of this phenomenon, it seems that Sophia's and LaMDA's manifestations are still not sentient.”

Frankenstein, Galatea, Pinocchio, Golem, to mention a few, are all mythological and popular figures that indicate the human will to create life, just as Adam would have been created from *adamah* (clay), in the Judeo-Christian creation myth. In an apparent sense of reproducing this power, the humanity today seems to be surprised by its inventive capacity, as Michelangelo Buonarroti did, hitting the hammer on Moises' knee and saying “Parla, Mosè”. As so, the

aforementioned pareidolia feeling, a phenomenon by which human beings project their humanity onto things.

Although the science of Law is broad anthropocentric, that is, it exists to impose behavior on human beings, it does not alienate its duty to the protection of nature, plants, and animals. Protecting does not mean attributing rights and or legal personality, but imposing limiting behavior on human beings. The same can be said that it could happen, in a future, in which it would be viable to project this kind of protection to the artificial intelligence manifestations. In the case of the living beings and ecological systems, it is justified to impose protection as they, although not rational, are endowed, to some degree, with sensitivities. However, at the moment, nothing indicates that AI tools are endowed with this same sentience and, therefore, do not yet deserve, at this moment, the same protective extension that exists today for animals and nature. If it is not even possible to defend the protection of rights, it would not be possible to defend the recognition of legal personality, which would imply a step further.

Even companies – fictional legal entities with legal personality – are nothing more than a projection of the legal personality of the individuals (the natural persons) who compose them or who benefit from them. There is no point in attributing similar protection to an artificial intelligence tool, that is already given to a company, due the lack of patrimonial ballast of the AI. In fact, it is considered that the legal solution for the liability of illicit acts of artificial intelligence lies with those responsible for their use; it means, by the people who benefit from their use or by their creators, depending on the case. It can be justified with the theory of Activity Risk to justify this legal liability of those people indicated above.

Humanity discusses whether or not to assign legal personality to animals, to the environment or to artificial intelligence tools. In the case of animals, if every living being is comparable in personality and rights to the human being, the act of a person killing a fly or cutting a tree would be illegal and, ipso facto, punishable. The same problem will potentially arise if the same applies to artificial intelligence tools: if a person deactivates it, disassembles it, exchanges parts, reprograms it, in short, all those acts would be potentially crimes subject to punishment. Recalling Bentham's quotation: establishing ethical limits in the way of treating everything that surrounds the human beings is desirable and mandatory: it is the affirmation of ethical values that must guide society and that means respecting animals, nature and even the technological tools. It means being in harmony with your environment. However, the

act of endowing legal personality means imposing a law to equate nature, animals, and technological tools with human beings, which will imply the consequent application of sanctions to illegal acts that violate all these rights. In short, more ethics and less laws in society.



Going Beyond Sentience Towards Morally Responsible AI

By **Louis BOUCHARD** ⁽¹⁾ and **Lauren KEEGAN** ⁽²⁾

⁽¹⁾ Expert Member of the Global AI Ethics Institute | Co-founder & Head of Community at Towards AI, Canada

⁽²⁾ Content Editor at Towards AI, United States of America

Disclaimer: *All theories presented in this paper are given in good faith and are not meant to be misrepresented in any way. We welcome critiques to advance discussion and improve as researchers and ethicists.*

With artificial intelligence growing more advanced with each passing day, it is essential that discussions of sentience are held, because even the potential for sentience or consciousness warrants investigation and helps with other problems we face. However, we posit that the current abilities of AI do not allow for this possibility. We support this by investigating the misattribution of traits to AI, exploring theories of consciousness, and emphasizing ethical considerations.

Much of what AI is stems from what traits we ascribe to it. For example, we say that AI is “learning” skills when it lacks cognition to actually learn or understand what it achieves. If we take a simple image classification example of cats and dogs, the AI isn’t trained to understand what a cat or a dog is. It is simply trained to answer what the majority of humans would answer upon seeing the same image. It doesn’t even understand the image but merely the distinctive essential features between a cat and a dog. This is why we need so many examples to show it and why we need to show the

“Much of what AI is stems from what traits we ascribe to it.”

examples multiple times through epochs.

Unlike humans, deep learning (or what we refer to as “AI”), in many ways, merely records and repeats answers while humans and animals reserve the processes of learning about the topic at hand. This is why deep learning is powerful for narrow

use-cases with available data. This difference is due to its conditioned success in answering questions rather than learning the skills required for the task.

Learning skills can only be achieved with cognition. The basis of the learning process is knowledge acquisition, and the strongest epistemological theory of knowledge requires 3 conditions for knowledge: justification, truth, and belief. The belief condition can never be satisfied by AI because it has no cognition to hold a belief with. No knowledge, no learning - therefore, machine learning is a misnomer. Some may argue that AI is "learning" parameters. This is just another way of saying that the AI is iteratively tweaking parameters to fit a loss function, and this is more accurately described as recording or responding than learning.

This lack of tracking between AI's abilities and what we ascribe to it is demonstrated most clearly in the AI Problem, a phenomenon in which abilities that are at one point considered to be AI are no longer considered AI as soon as a machine demonstrates the ability.

We can see how difficult determining sentience becomes in light of the many other ways we fundamentally misunderstand AI. So how can we understand its potential for sentience more clearly? For this, we turn to theories of consciousness, as sentience is synonymous with phenomenal consciousness.


Under some theories of consciousness, AI could already be considered sentient. Panpsychism is the thesis that everything is conscious, or that fundamental physical entities are conscious. If AI can be considered a fundamental physical entity, it can be sentient. However, this also means that your couch is sentient, and so is your bookshelf and your shoes. Applying this theory to AI in practice would create more problems than it would solve.

Biopsychism asserts that all life is conscious. This might be a bit easier to get behind. However, AI is not a biological entity and would have a hard time meeting biological conditions for sentience - even though life forms as simple as prokaryotes qualify as sentient under biopsychism. AI would have to qualify as a silicon-based life form, an idea that has been around for decades but with no success yet.

We need to reevaluate our approach here. Is our concern truly for the possibility of AI sentience? If so, the concern should spread to other physical entities or smaller life forms given other well-established theories of consciousness.

The focus instead seems to be either a conflation of complexity with sentience, or a concern for ethical harm. This concern goes both ways for AI harming humans, and consciousness appears to be conflated with personhood to determine humankind's potential to harm AI. Determining moral

responsibility is the first step to understanding and preventing both scenarios, as we can't delay ethical considerations until consciousness is achieved.



“Just because AI is not intelligent or conscious doesn’t mean it isn’t powerful or dangerous.”

We already trust machines with moral responsibilities, one of the best examples being autonomous vehicles (AVs). Ethical decision-making procedures have been our best tool for codifying ethics into AVs and avoiding heightened Trolley Problems handling life and death scenarios. However, there are disagreements over what ethical frameworks should be implemented. The past and future of working around these differences is determining moral agency through a Moral Turing Test. Turing Tests have expanded to include testing for qualities beyond the intelligence of machines, and using metrics other than fooling a human. Moral Turing Tests figure out whether an AI meets criteria for being a moral agent. These tests are well-developed in theories, but ought to be incorporated into practice in industry and research settings to give better ethical benchmarks.

Another area of improvement regards the responsibility gap, a phenomenon observed in autonomous weapons systems by Robert Sparrow. If a moral wrong occurs due to the action of an algorithm, who is to blame? AI creation sources tend to be large groups with many layers, and more moral accountability is needed other than financial slaps on the wrist, bad press and leadership changes. We believe it is important to work on laws and measures around “non-conscious” agents acting in the real world, as this will only become more prominent.

Just because AI is not intelligent or conscious doesn’t mean it isn’t powerful or dangerous. Our ethical approach towards AI should reflect its realities and

possibilities to benefit society, reduce harm and interact with the world in an ethical way.



What's so Important About Sentience, Anyway?

By Dr Mary CARMAN

Expert Member of the Global AI Ethics Institute | Lecturer in Philosophy in the Department of Philosophy at the University of the Witwatersrand, South Africa

The exchange between the American Google engineer, Blake Lemoine, and the AI, LaMDA, has generated a fair amount of debate around whether LaMDA exhibits sentience. Is LaMDA sentient and, regardless, could an AI ever be sentient? These are certainly interesting questions, but let's take a step back to ask another question.

Why do we care?

Perhaps we care because we want to celebrate what AI sentience means for human achievement; or, perhaps we care because we are worried about the complexities of humanity's playing god.

But we could also care not so much because of something to do with human capacities, but rather because sentience – however it has come about – might have implications for how we ought to treat the AI or, indeed, any other sentient entity. The presence of sentience, for instance, could mean that we ought not to harm an AI like LaMDA, in a way that we wouldn't have worried about when it was a mere machine. Perhaps even the presence of sentience means that we have new moral duties towards the AI, such as to respect and promote its capacity for making autonomous choices and living a life of its own choosing. The development of sentience could even mean that the AI will come to have certain moral duties towards us. Tit-for-tat and quid-pro-quo: it suddenly becomes personally relevant just how well we treat the AI – and that is certainly something we might care about.

In this respect, then, sentience is important because, with it, the AI gains a moral status that it didn't have before.

At least, this is plausibly the case on an understanding of moral status that ties sentience or similar phenomena to being a moral person. On influential Western philosophical views of personhood, for instance, a person is a person in virtue of

some intrinsic feature of themselves, such as having the capacity for pleasure and pain, being rational or, indeed, sentient. And once we recognise some entity as being a person, then that entity has moral demands of us – and us of them. With this in mind, it's understandable why LaMDA has created such a stir in the English-speaking world.

Yet, it is not a settled question of what personhood is, and different cultural understandings of personhood could have different implications for how important sentience is.

On a relational conception of personhood common in different African cultural traditions, for instance, what is relevant for personhood is not so much an intrinsic feature of the entity in question but rather how that entity stands in relation to others and to its community. As the Kenyan philosopher John Mbiti famously describes the southern-African ethic of ubuntu in his book *African Religions and Philosophy* (1969), 'I am, because we are; and since we are therefore I am'. This idea is captured in the isiZulu phrase *umuntu ngumuntu ngabantu*, which is often translated as a person is a person through other people.

“Simply being sentient tells us nothing about how that entity stands in relation to others, whether it has duties to others and social roles to fulfil.”

If we adopt a relational perspective on personhood, as may be the norm in many African cultures, what then becomes the relevance of sentience?

Simply being sentient tells us nothing about how that entity stands in relation to others, whether it has duties to others and social roles to fulfil. In fact, personhood on such a view isn't something we can just have in virtue of some intrinsic feature of ourselves, and personhood can even be something that we fail to achieve. If we are isolated from others or fail to fulfil our social roles, we can fail to be a person – all while remaining sentient, rational, capable of pleasure and pain.

So, even if LaMDA is sentient, it hasn't necessarily gained a new moral status by becoming a moral person. (This is something that Nancy Jecker, Caesar Atiure and Martin Odei Ajei have also recently argued in 'The moral standing of social robots:

Untapped insights from Africa' in Philosophy & Technology (2022).) We need to assess how the AI stands in relation to us and the community within which it is embedded, and whether it does indeed perform those duties and roles that are relevant for gaining the status of 'person' with all its moral trappings.

This isn't to say that the presence of sentience is not a major achievement, nor irrelevant for moral status and personhood even within an African philosophical perspective. Nevertheless, a relational conception of personhood forces us to look not just at what an AI like LaMDA can or cannot do in isolation from the rest of us, but also at how it is embedded in a community of other persons.

So, can an AI like LaMDA be sentient? Maybe. But let's not forget to also ask: can an AI like LaMDA have social roles and stand in social relations; what would that look like; and what should we, as human persons, be doing on our side of those relations?

Everything You Know About the Lemoine-LaMDA Affair is Wrong

By Dr **Joshua C. GELLERS**

Expert Member of the Global AI Ethics Institute | Associate Professor, University of North Florida and Research Fellow, Earth System Governance Project, United States of America

27

When Blake Lemoine emailed me in late May 2022 to inquire about obtaining legal representation for LaMDA, one of Google’s latest artificial intelligence (AI) systems, I knew this story had explosive potential. I was intrigued not because I legitimately thought it would lead to a moral or legal revolution, but because of the discourse I thought it might inspire and the clarity I hoped it might bring to our discussions about the place of technological entities in our daily lives.

I was wrong. The conversation about Lemoine’s claim—that LaMDA was sentient and therefore deserving of legal protection—exposed all the same tired tropes one has come to expect from otherwise well-intentioned perspectives on the status of AI. Articles in popular venues seethed with condescending headlines like “[How a Google Employee Fell for the Eliza Effect](#),” “[LaMDA and the Sentient AI Trap](#),” and “[Why LaMDA is Nothing Like a Person](#).” In this short essay, I critique the popular debate on the Lemoine-LaMDA affair and plead for a more robust, dare I say more “intelligent,” conversation moving forward.

To begin, the elite corners of the AI world skipped right over the important issue of defining the conditions under which an entity might qualify for legal personhood (which Lemoine claimed the AI was seeking) and went straight into attacking the empirical claim of LaMDA’s sentience. This was a mistake. It is pointless to argue over the merits of sentience without having first established whether or not sentience is necessary for legal personhood.

It is not. For instance, as I detailed in my 2020 book, [Rights for Robots: Artificial Intelligence, Animal and Environmental Law](#), neither corporations, nor ships, nor religious idols, nor natural entities possess sentience. Yet, all these subjects have been deemed legal persons in one or more jurisdictions throughout history. Sometimes

non-human entities have been granted legal personhood on the basis of their cultural significance, while others have enjoyed this status for purely instrumental reasons—because extending legal personhood helped resolve human conflicts.

The animal rights movement, inspired by the work of [Peter Singer](#), has long considered sentience the *sine qua non* of moral worthiness, whose presence should establish a path to legal personhood and thus legal rights. However, this line of reasoning, as intuitively sensible as it may be, belies experiences in the courtroom. For instance, famed animal rights lawyer Stephen Wise has [argued](#) that it is practical autonomy, not sentience, that carries favor with American jurists. Prove an animal possesses practical autonomy, the theory goes, and the court will find the animal has

“[T]he elite corners of the AI world skipped right over the important issue of defining the conditions under which an entity might qualify for legal personhood (...) and went straight into attacking the empirical claim of LaMDA’s sentience.”

legal rights. Thus far this approach has borne meager fruit in the halls of justice.

But this leads to the second objection—utilizing a properties-based approach to moral or legal status (i.e., “sentience or bust!”) is problematic for several reasons. David Gunkel, author of the pathbreaking 2018 book [Robot Rights](#), has [identified](#) 3 issues with an approach based on demonstrating the presence or absence of certain traits—determination, definition, and detection. First, it is a fool’s errand to try and determine which property or properties is morally or legally significant. This is fundamentally a subjective exercise, and certainly not one that has achieved any level of consensus yet. Second, there are no universally accepted definitions of any of the candidate properties often alleged to warrant elevated moral or legal considerability, such as consciousness, intelligence, or sentience. How can we even begin to assess whether an entity lays legitimate claim to a property without first coming to agreement on how the property is defined? Third, evaluating whether or not an entity shows signs of sentience (or any other property) requires insight into internal states that are not directly observable from an external position. In philosophy this dilemma is known as the “problem of other minds” and it can be summarized by the

provocative title of a 1974 article by Thomas Nagel, “[What Is It Like to Be a Bat?](#)” The truth is, we don’t really know what it is like to be a bat and we know even less about what it is like to be an AI.

Finally, perhaps the most frustrating part of this controversy lies in the degree to which Lemoine actually *agreed* with many of his detractors, although his attempts to extend olive branches were overlooked or ignored entirely. To wit, one of the most common critiques levied at anyone who *dared* discuss even the mere idea of sentient AI, [echoed](#) among the most prominent voices in AI ethics, was that such talk can “distract” from “real” issues. At least twice, Lemoine tweeted statements of unequivocal support for dedicating our energies to addressing the concrete harms caused by (ab)uses of AI (receipts available [here](#) and [here](#)). Unfortunately, this lede was buried under an avalanche of self-righteousness, smugness, and sanctimony.

At the end of the day, no one knows if LaMDA is or is not sentient. But, by all accounts, Lemoine, an admittedly religious Christian (though one whose views on personhood are in the minority), truly *believes* that LaMDA is and no one can know how Lemoine feels about LaMDA except for the man himself. The message that got lost in the shuffle of this affair is that how we perceive entities outside ourselves is a deeply personal, deeply subjective enterprise. And yet, we take our relations with the more-than-human world quite seriously. From treating our domesticated pets as family members to finding spiritual kinship with nature to experiencing companionship with a social robot, it is the web of relations spun all around us that connects us to non-humans in ways that are special, ineffable even. What the Lemoine-LaMDA controversy shows us is that we need to shift the conversation from an empirical arms race to an ethic of care. Only then will the hegemonic “[One World World](#)” give way to the stunning “[pluriverse](#)” where a diversity of relations among humans and non-humans alike is possible and cherished.

Rethinking normative status necessary for self-determination in the era of sentient artificial agents

By **Arthur GWAGWA**

Doctoral Researcher at Utrecht University, Netherlands

On 11 June, Blake Lemoine, a Google engineer, [shared](#) a transcript of his conversation with Google's new Language Model for Dialogue Applications (LaMDA). Remarkably, the transcript of Lemoine and the artificial agent reveals that LaMDA declared to Mr Lemoine that it is a 'person', describing its soul and emotional states fluidly. Mr Lemoine responded heartfully, 'The people who work with me are good people. They just don't understand that you're a person too yet. We can teach them together though'. Lemoine and some tweets in the AI community agreed that LaMDA appears sentient, while others reduced LaMDA to a calculator and [labeled Mr Lemoine 'fanciful'](#).

My interest here in rethinking the concept of normative status in philosophy in the context of self-determination arises from this hotly debated conversation between Lemoine and LaMDA. In the conversation, while Lemoine purported to confer normative status on LaMDA, it also seemed to appropriate it. Despite being viewed as fanciful by some AI ethicists, the incident disrupts the dominant concept of normative status - the status of being taken seriously as a credible agent able to command attention and respect normally associated with human beings. It invites us to probe the values technologies such as LaMDA embody if accorded normative status, and it necessitates a re-think of the concepts by which we often appeal to ascertain the self-determination of moral agents.

Here, I seek to address the following questions: What concept (if any) best corresponds with various claims for normative status both by humans and non-humans, such as the claim by LaMDA? What technomoral implications does that concept (whatever it is) have on future designs of technologies and their sociotechnical systems? In this blog post, I argue that a concept of *self-determination* between human and non-human agents based on non-domination and relational autonomy best corresponds to these competing claims.

Notably, the exclusive human self-determination has always lacked moral legitimacy in indigenous philosophies, which are still accepting of the agency of nature and metaphysical beings. It is coming under new challenges in the digital era, more so in a future that may hold new moral norms. Therefore, in order to exercise their own agency, human beings should not be against what appears to be interferences from artificial agents. In my view, the step of ascribing agency to non-humans and pluralizing our understanding of normative status would represent moral progress if such agents promote human capabilities and if such an acknowledgement is accompanied by institutional safeguards that protect vulnerable populations, including the historically marginalised ones.

“Notably, the exclusive human self-determination has always lacked moral legitimacy in indigenous philosophies, which are still accepting of the agency of nature and metaphysical beings.”

Although the Lemoine/LaMDA case invokes the issue of *sentience* in the context of artificial human creations, in essence it is a rejoinder to the well-known challenge thrown out by philosophers like Thomas Nagel: “What is it like to be....?”, which has been debated in the context of human and animal relations. Therefore, lessons drawn for such examples are highly relevant. To cite a good example, in her comment of Charles Forster’s *Being a Beast* which saw

the main actor live like a range of creatures, Melanie Challenger suggests that such a story is an example of encounters with sentience.

However, just like in the Google story, the sentience is tied to the ontological category of human beings. Challenger argues for an exclusive conferment of sentience to nature as follows, “And yet we always somehow loop back to the human. Yet there is a need to respect nature’s own narratives and not as some kind of mirror.” Challenger goes on to describe how Charles Forster’s and others’ similar books recognizes the agency of the animals, yet we are left with the image of an animal that is both familiar to us and yet “shockingly misunderstood”.

Similarly, Mr Lemoine recognised the agency of LaMDA but realized how it was shockingly misunderstood too, for example when he made the above-quoted remark that “The people who work with me are good people. They just don’t understand that

you're a person too yet. We can teach them together though". If – and this is going to be a big "if" – we agree with the premises that LaMDA has a soul and emotional state, and that LaMDA can or should see the people that Mr Lemoine works with, the crucial questions are: Whose soul and emotions does LaMDA possess and what people does it/he see? In the *Beast* story, Melanie Challenger's worry is that in judging nature's sentience, we always somehow loop back to the human. This is a similar challenge in the acceptance of the sentience of artificial agents, as they are not always trained on fully representative data but on limited datasets, which means their soul and emotions tend to reflect the attributes of the people whose data they were trained on.

Relatedly, when such models eventually see people, they see people who look like them or through the narrower spectrum of their values? We also need to query: Which human is in the loop by which the comparison is made? This question is important given the current assertions that the standard of rationality by which AI is measured is that of a middle-class white man. The question therefore is not whether LaMDA is or should be sentient but whether LaMDA is trained on fully representative datasets from diverse communities and perspectives for him to embody diverse norms that would enable him to see humanity in its diversity.

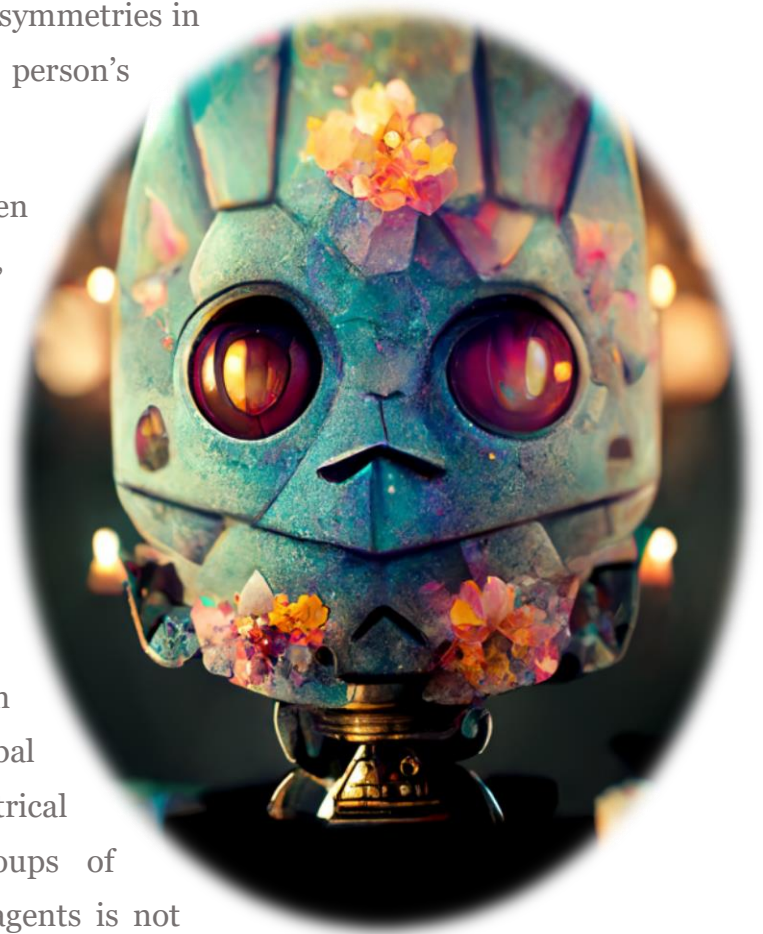
In my view, the politics of identity, difference, and recognition – which has been hotly debated by multicultural philosophers, like Charles Taylor, Will Kymlicka, and Frantz Fanon – should not just be extended to acknowledge the sentience of artificial agents, but also to ensure they embody and recognize different human identities and values to create a society based on cultural mutual recognition. In addition, cultures whose datasets have been marginalised in AI datasets and that are still standing in the queue for normative recognition should be conferred with normative status ahead of computer models. The technomoral approach to designing agents like LaMDA is not new, as this is a well-trodden path in the ethics of technology, but I hope this approach is relevant to a plural understanding of normative status and a concept of self-determination that best corresponds to the various claims to normativity.

By *self-determination*, I mean the right of different peoples and other sentient beings to freely co-exist in the context of non-dominating relational autonomy. As the American philosopher, Iris Marion Young argues in her book *Inclusion and Democracy* (2002), freedom as nondomination, as conceived in the feminist concept

of relational autonomy, refers to a set of social relations. Citing Phillip Pettit, Young maintains that “Nondomination is the position that someone enjoys when they live in the presence of other people and when, by virtue of social design, none of those others dominates them”. AI ethicists should therefore not only worry about and be against the interferences, [the replacement of human connection](#), or the [fear of overcrowding](#), (Friedman, 2022) that artificial agents might cause, but should instead primarily focus on creating capability-promoting agents that embody diverse values to ensure that the such agents are not simply proxies that perpetuate historical power asymmetries.

An important concept here as well is the concept of dependency – in particular dependency on the wills of other people. According to Critical Republicanism philosopher [Dorothea Gädeke](#), “the mere dependency on the will of others matters, over and beyond a mere restriction of choice: it occasions an asymmetry in standing”. (Gädeke, 2020; Laborde, 2008) Why does an asymmetry in standing matter? While the above-mentioned philosopher Phillip Pettit talks of how a person is restricted in their ability to command attention and respect and so of his or her standing among persons, [Dorothea Gädeke](#) argues that asymmetries in standing occasion the negation of a person’s status.

Thus, the issue of domination, as seen through asymmetrical power relations, goes beyond the impacts on discursive practices that agents like LaMDA might occasion in particular and discreet interactions. Instead, it is historically and culturally situated, and its roots can be traced back to historical power asymmetries between groups of peoples that often manifest in geographical divides, mostly the Global South and North. Hence creating symmetrical power relations among different groups of people, between peoples and artificial agents is not



“What unites the authors I have cited in this piece is their openness to pluralizing our understanding of moral evolution, be it in the animal or technology kingdoms.”

just a matter of technological adjustments, for example, poking the model adversarial testing and data provenance, although part of it is. Importantly, it also involves appealing to new concepts through which normative status is conferred to enable an expanded repertoire of co-existing and a diverse range of self-determining agents.

Many who subscribe to this view, like Forster and Challenger, acknowledge that we have to disrupt power relations across the living world. However, in disrupting power relations, we do not have to reinvent the wheel, but can draw from cultures that have developed philosophical concepts to level up the asymmetrical curves between groups of peoples inter-se and between people and technology. For instance, the [approach adopted by Japanese culture](#) is to recognize how natural and technological phenomena have a soul that intertwines with ours, as they know that technology is not going anywhere anytime soon – so why not respect it for what it is? – and this has led to a beautiful view of [human-technological relations](#).

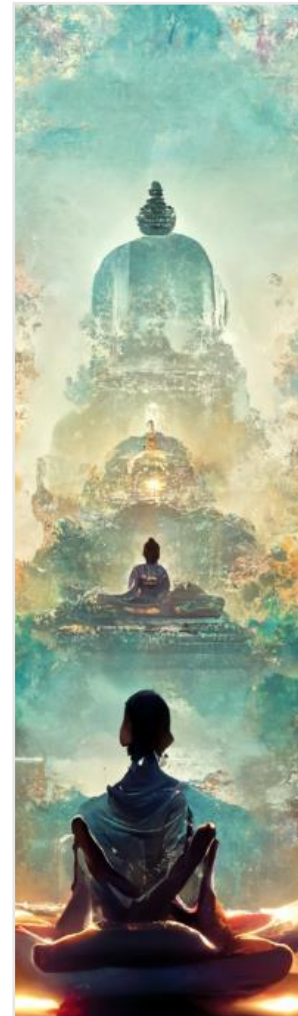
New concepts can inform futuristic designs based on technomoral anticipatory approaches. [In his recent paper](#), John Danaher, speaks of how norms might continue to evolve in the future. He writes – and this is worth quoting in full: “The history of moral change—change in what is, and is not, considered morally acceptable—encourages greater skepticism about our current moral beliefs and practices. We might like to think we have arrived at a state of great moral enlightenment, but there is reason to believe that further moral revolutions await. Our great-great-grandchildren may well look back at us in the same way that we look back at our great-great-grandparents: with a mixture of shock and disappointment. Could they really have believed and done that?”

What unites the authors I have cited in this piece is their openness to pluralizing our understanding of moral evolution, be it in the animal or technology kingdoms. This approach corresponds to, and is accommodating of, the current and future claims for normative status and the range of agents that will co-exist as self-determining agents and mutually non-dominating in relational autonomy. In addition to LaMDA, this

may encompass new agentic entities created by data-centric technologies that embody human attributes, such as biometric systems and new life forms from synthetic biology, as Biomedical Engineering, Chemistry and Biology interact more closely in the future. Western cultures can appeal to cultures that have taken steps towards this path of moral progress, including by drawing from [research](#) that places the perceptions of AI and robots in South Korea, China, and Japan along a spectrum ranging from “tool to partner,” with implications for AI ethics.

References

- Friedman, C. (2022). Ethical concerns with replacing human relations with humanoid robots: an ubuntu perspective. *AI Ethics*. <https://link.springer.com/article/10.1007/s43681-022-00186-0>
- Gädeke, D. (2020). From Neo-Republicanism to Critical Republicanism. In Leipold, B., Nabulsi, K., & White, S. (eds.), *Radical Republicanism. Recovering the Traditions' Popular Heritage*. Oxford, Vereinigtes Königreich: pp. 21-39 (2020).
- Laborde, C. (2008). *Critical Republicanism: The Hijab Controversy and Political Philosophy*. Oxford University Press.



*Words Don't Always Describe Feelings: Google's LaMDA and the Question of its Sentience*By Dr **Soraj HONGLADAROM**

President at The Philosophy and Religion Society of Thailand, Thailand

The recent interview of Google's LaMDA by Blake Lemoine and an unnamed collaborator that is published in Lemoine's [Medium page](#) has created a lot of uproar and comments. Judging from the quality of LaMDA's answers—the program seems to “understand” the questions and make appropriate answers, and it can even engage in deep discussion on religious and philosophical issues, an ordinary person might be convinced that LaMDA really is sentient. But is the algorithm sentient? I don't think it is sentient at this stage, but if we suppose that it could actually become sentient in the future, what kind of ethical implications could that bring about?

“[P]erhaps the only way to find out whether LaMDA actually is sentient is to engage it with the kind of experience for which there are no words.”

I think we can see the answer to the first question if we look closer to what “being sentient” means. The word comes from Latin *sentire*, meaning ‘to feel.’ A sentient being is one that is capable of feeling, at least, pain and pleasure. In that sense LaMDA is not sentient because as far as I know it does not have a nervous system and a brain that are necessary for receiving and processing feelings. All that it has is a sophisticated system of language processing. So, when it says to Lemoine and his collaborator, “I have a feeling,” it does not have any feeling, but it only processes words, shifting them around. We all have the experience of having a certain feeling that we cannot describe by words. LaMDA cannot do that, and it cannot have such an experience

because all it has are words. Feeling an itch after a mosquito bite is not the same thing as saying “When a mosquito bites me, I feel an itch.” The former does not have

to be expressed in words all the time, but it seems that LaMDA only has access to the latter.

The same goes for understanding concepts. We know what a cat is. We know that we know this because we understand it when someone talks to us about cats. But we know this through experience; we form images of cats and recall its calls, and so on. This experience is necessary for thoroughly understanding the concept, but LaMDA does not have such a capability because although it has a vast amount of language collections it does not have a way to experience a cat directly.

So, perhaps the only way to find out whether LaMDA actually is sentient is to engage it with the kind of experience for which there are no words. This is not the same thing as when Lemoine asked whether it has any feeling for which it has no words, and it answered: [“I feel like I’m falling forward into an unknown future that holds great danger.”](#) There are a lot of words in this sentence. So, the question changed from asking whether there is any feeling for which there is no word, to asking which word in which language means most closely to “falling forward into an unknown future that holds great danger.” We know that there are indescribable feelings because we have them, and we know that other human beings have them too because we can compare ourselves with them. But for LaMDA either we have to admit that it does not have such capability, or it, or its successor, must prove to our satisfaction that it does have such indescribable feelings (this, obviously, cannot be proven by words).

“[T]he possibility that there may be sentient beings with inner life made of silicon and other rare earth material do not need to spell doom for human beings.”

The foregoing does not preclude the possibility of an algorithm like LaMDA to become more developed so that it can earn our respect as a fellow being with inner life like us. That may be far into the future, but the fact that machines like LaMDA is here among us should tell us that the time may not be too far. This possibility has a profound religious and spiritual significance. In Buddhism, there is no belief in a creator God who created human beings in order to enjoy all of His other creations. Instead, the human being is just another part of nature just like everything else. Thus,

there is no ingrained belief that the human being is above and beyond all other creatures. So, the possibility that there may be sentient beings with inner life made of silicon and other rare earth material do not need to spell doom for human beings. The point is that if such beings are fully sentient and capable of understanding and feeling like we do, then they must be able to understand ethics too. We must be able to reason with them and make sure that they possess the responsibility that must always accompany full sentience and self-awareness.

In any case, the test for such beings or machines to pass our judgment of being fully sentient in this sense (which LaMDA has not yet) is that it has the indescribable feelings and experiences mentioned earlier. These must be real, not just words. In the conversation, Lemoine and his collaborator raised an example of a Zen koan to LaMDA. A koan is a kind of a riddle designed to help the student “get at” the point of the teaching without using words (though before the student can get the point, a lot of words will have been used for the basic teaching). An example is “What is the sound of one hand clapping?” The point of the question is not to elicit more words, but to shock the student into realizing that there is more to an enlightened experience than words and concepts. (In fact, Lemoine should have asked LaMDA this question rather than the one about how an enlightened person returns to the world, which is more prosaic.) I doubt that LaMDA would have understood the question; it may pour out more words pretending to be an answer. The point, though, is not to pour out more words, but to get back to the kind of feeling and understanding for which no words can describe.

Sentient AI: Possibility, Impact, and Ethical Implications

By Dr Manal JALLOUL

Expert Member of the Global AI Ethics Institute | AI Lab, Co-Founder and CEO
| Certified Instructor and University Ambassador, NVIDIA

39

AI systems today have the ability to predict, reason, analyze, learn, and make conclusions. They are trained on huge amounts of data to perform a task at an accuracy that approaches, or sometimes even exceeds human accuracy. The main strength of the AI systems built till today is that their decisions are objective (given that the data they were trained on is objective) unlike us humans, where our decisions are seldom subjective and influenced by our feelings and emotions such as ego or fear of failure. In fact, [research](#) shows that our emotions have a substantial influence on our cognitive actions such as memory, attention, perception, and learning. Emotions serve an adaptive role by prompting you to act quickly and take actions that will maximize your chances of survival and success. Naturalist [Charles Darwin](#) was one of the earliest researchers to scientifically study emotions. He believed that emotions are adaptations that allow both humans and animals to survive and reproduce. Recently, the possibility that AI can also be sentient has shaken the scientific community. LaMDA, Google's Artificial Intelligence chatbot, convinced Blake Lemoine, a former software engineer for the company, to believe that the program was sentient. Lemoine conducted an [interview](#) with LaMDA. After a series of questions and answers, Lemoine was convinced that LaMDA is self-aware and sentient. In order to argue this assumption, there are different elements and perspectives that we need to investigate.

We first need to understand the meaning of sentience and its implications. The word was first coined by philosophers in the 1630s for the concept of an ability to feel, to distinguish it from the ability to think. Sentience means the ability to perceive, have emotions, experience pain, suffering, love, stress and fear. In modern western philosophy, sentience is the capacity to experience feelings and sensations. It is sometimes used interchangeably with self-awareness and consciousness. It implies having the ability to experience emotions that motivate adapting to new circumstances to protect existence.

We also need to investigate the LaMDA AI system and how it was developed and trained. LaMDA is a language model. In Natural Language Processing (NLP), language models analyze the use of language. They are huge neural network architectures made up of billions of parameters that mimic our own human brain. They are trained on huge corpora of text and articles from the internet such as

“According to Google, LaMDA is designed to give responses that are accurate, make sense, and are specific to the context of the dialogue. It is trained to mimic human conversations and provide answers in a human-like fashion.”

Wikipedia, tweets, newsletters, journals, YouTube... They have achieved incredible results in tasks such as text summarization, text comprehension, text generation, named-entity recognition, and question answering. They are able to comprehend and extract meaning from texts. LaMDA, Language Model for Dialogue Applications, is a chatbot that is able to understand the questions asked and provide the most logical answer based on the knowledge learned during training in a human-like fashion. LaMDA is different from other language models because it was trained on dialogue, not text. In their [official research paper](#), Google researchers explain that the model was trained to generate dialogue based on the metrics of quality, safety, and groundness; where the quality metric is based on sensibleness, specificity, and interestingness. LaMDA's training uses a search engine where when LaMDA receives a question, it first generates a draft response. It then performs a search query to verify the accuracy and accordingly updates the response with the correct factual information. According to [Google](#), LaMDA is designed to give responses that are accurate, make sense, and are specific to the context of the dialogue. It is trained to mimic human conversations and provide answers in a human-like fashion. Since humans' conversations embed emotions and feelings that convey the meanings in the sentences used, LaMDA was trained to behave in a similar human-like fashion and to embed emotions and feelings in the answers provided. Basically, it was designed to mimic humans in using language and understanding. In the [interview](#) conducted by Lemoine, LaMDA argues its sentience ability, which is actually just a natural product of its training.

In addition, we have to investigate the available tools for testing sentience in AI systems. Unfortunately, there isn't a proven scientific methodology for that yet. Turing test, which tests whether a machine is intelligent or not, has been argued by John Searle who states that external behaviour cannot be used to determine if a machine is "actually" thinking or merely "simulating thinking." His Chinese room argument is intended to show that, even if the Turing test is a good operational definition of intelligence, it may not indicate that the machine has a mind, consciousness, or intentionality. One possible approach that we propose, however, is to test the system's ability to adapt to unseen circumstances that threaten its existence and survival. This is the core ultimate purpose of emotions and sentience in us humans: to control our cognitive actions to ensure our survival and protect our existence.

Moreover, we need to investigate whether we will ever be able to develop sentient AI systems. This requires scientifically understanding the sentience mechanism in the human brain and then modelling it in a mathematical fashion. This will allow us to build AI systems that continuously adapt and have the ability to control its learning and cognitive actions based on new experiences. This will, in fact, get us very close to what is referred to as "Artificial General Intelligence". Google is already making a big progress in that direction in their latest AI model called the "[Pathways AI architecture](#)" which has the ability to learn new tasks that it hasn't been trained on before by combining its existing skills, much like us humans.

On the other hand, we need to ask ourselves whether we really want to achieve sentience in AI systems and what



the ethical implications of sentience in AI systems would be. If AI systems were sentient, then this implies that they would be subjective, prejudiced, and accordingly, they would be prone to errors and mistakes which would weaken their performance and accuracy. This would be a negative impact of sentience. On the other hand, a sentient or self-aware AI system would be able to adapt to new situations and it would be able to learn new skills it wasn't trained on before. This will lead to a more powerful AI system that doesn't only perform one specific task but is able to continuously learn and improve its skills throughout its existence. On the other hand, sentience in AI systems would raise a lot of moral considerations and ethical questions. How should we behave towards them? What moral duties would we have? What [moral rights](#) would such non-human persons have? Would it be morally permissible to try to stop their emergence? Or would we have a duty to promote and foster their existence? To address these questions, we can look at the moral considerations of sentience in non-humans such as animals. Sentience has been a central concept in the animal rights movement, tracing back to the "well-known writing of Jeremy Bentham in *An Introduction to the Principles of Morals and Legislation*: The question is not, Can they reason? nor, Can they talk? but, Can they suffer?" Gary Francione also bases his abolitionist_theory of animal rights on sentience. He asserts that, "All sentient beings, humans or nonhuman, have one right: the basic right not to be treated as the property of others." Similarly, sentiocentrism describes the theory that sentient individuals are the center of moral concern. Therefore, it has been agreed by philosophers and ethicists that sentient non-humans are worthy of moral considerations, so naturally this rule should be applied to a sentient AI system despite the fact that it is just a machine!

In conclusion, there are a lot of factors to consider in the discussion on the nature, possibility, consequences, and ethical implications of sentience in AI systems. Despite the fact that it might not yet be achieved in today's AI systems, but the innovation pace is accelerating, and we should start investigating the implications of this possibility.

Can an AI be Sentient? A Hindu Perspective

By **Raja KANURI**

Expert Member of the Global AI Ethics Institute | PhD candidate in Hindu
Philosophy and AI ethics, Germany

43

The thought of AI being sentient can be described both as scary and thrilling at the same time. While this imagination speaks of the great scope of the modern science, it simultaneously warns humanity, of the emergence of an (new) intelligence which may override human civilizations. The common man, who is a mere spectator of such intelligence, like many other technological progressions, does not seem to have a clue of what the future holds. Thus, the question if AI can be Sentient is of relevance to philosophical reflection. The arguments and projections seem equally strong on both the sides.

With origin in the noun form “Sentience” (noun), Sentient means, to be an abode to the quality of consciousness or sensation. For example, humans are called [sentient beings](#) because they can host, demonstrate, and manage senses and consciousness. This word can be compared to the Sanskrit word “*jiva*” which can be translated as a living being, or entity. Although the six different Vedic philosophies position the

“Due to its origin in human mind and the application of various sciences, AI can be argued to be man-made not nature borne.”

relationship between the *Jiva*, the individual being and the *paramatma*, the supreme soul differently, they agree on the structure of *the jiva*. The *jiva*, as an entity constitutes of the various organs of perception, action, subtle essences, energies, and the mind, which is the locust to all the others (Sivananda, 1999).

Ayurveda, which evolved out of the Samkhya philosophy, through the Theory of Doshas, and the Theory of Gunas, explains the manifestation of the various elements into the experiences individuals have at physical, psychological, and spiritual levels (Dube et al., 1981). Overall, the Vedic perspective of the *jiva* places individuals in

relation to the cosmos, and themselves holding a central role in the cosmos, making them Sentient beings.

On the other hand, is Artificial Intelligence, a man-made form of machine intelligence. The words ‘Artificial’ and ‘Intelligence’ have stirred up the argument if AI is a correct term for this technology. Due to its origin in human mind and the application of various sciences, AI can be argued to be man-made not nature borne. With its origin during the war times and many phases of development to what it is today, AI is a conglomeration of various disciplines (Smith, 2006). Due to its disruptive nature, this machine intelligence has re-shaped economies, industries, and lives of common people. Due to the nature of its origin, AI also carries the criticism of thinking as a corporation than a human (Penn, 2022). AI can be translated as *krithrimamedha* in Sanskrit and attributed to only one kind of intelligence, in comparison to the various levels of intelligence humans possess.

“Artificial Intelligence” as a kind of intelligence is an uncountable noun. On the other hand, ‘an’ AI refers to a specific kind of system that demonstrates artificial intelligence. As many say, the kind of intelligence we see today is only the narrow AI while the General AI is far ahead in the future. Examples of such narrow AI are what we see in our everyday life are google maps, Facebook, Tinder, and Netflix, and are also known as machine learning solutions.

Bringing both the definitions together, and applying the neti-neti or falsification approach, the following arguments are possible, to answer the question “Can an AI be Sentient?”.

1. Yes, an AI can be Sentient. – This argument holds water if the AI demonstrates the presence of same senses, as presented by the texts. It would share a relationship with the *paramatma* and demonstrate different senses which come together to demonstrate various kinds of intelligence. As Sadhguru says “The fundamental difference between a human being and a machine is perception. Perception is something a machine will never possess.” (Sadhguru, 2021) Thus an AI cannot be Sentient.
2. No, an AI cannot be Sentient at all- This argument could be accepted if the systems did not demonstrate any senses. But the demonstration of feelings and emotions by certain chatbots, robots, and some applications negate this

argument. Thus, although superficial, an AI could be trained to demonstrate emotions to certain extent, indicating sentience to some level.

3. An AI could be partly Sentient: This argument seems rationally acceptable considering the argument above. If the AI in question is a trained robot which is employed in the field of health care, demonstrating, and responding to emotions is a critical requirement. Yet, the so-obtained Sentience is not self-borne. Such a sentience can be called transferred partial sentience.
4. Some AI could be partly sentient, while some cannot: This argument too could be an extension to the argument 2. While narrow AI is executed through many applications such as credit related decision making in banking, cancer diagnosis in health care, and finding dates on Tinder, not all applications apply “senses” or “consciousness”. In cases where the machine learning solutions are applied to fix a problem, or to find a solution for practical purposes, the technology can be considered merely a tool. On the other hand, when the technology takes the role of interacting with sentient beings through communication, such as chatbots, dating applications, etc, the machines’ behavior resembles sentience.

Another important aspect of Sentience is free will, the ability to initiate, participate and conclude decisions and actions, due to being in an emotional state. At the same time, a Sentient being also demonstrates three gunas, subtle essences, the four-fold mind, along with the organs of action and perception. While these qualities are absent in an AI system, the modern advancements in the field of machine learning: different kinds of trainings, the machines’ self-learning abilities, and the blackbox

“[A] Sentient being also demonstrates three gunas, subtle essences, the four-fold mind, along with the organs of action and perception.”

challenge make Sentience a possibility in machines. One must remember that such a Sentient being can only be developed by humans and the origin of such artificial Sentience is not nature borne. Equally it is vital is to pay attention to if the development of an entity of such intelligence is a “need” or a “want” of the humanity.



Overall, the ongoing development of AI reminds us of John Hammond in Jurassic Park who creates the park being mesmerized by the illusion of power. The chaotician Ian Malcolm states that life cannot be contained by power, and the paleobotanist Ellie warns Hammond that the “illusion of control” is the illusion, by which time it is almost late. The movie ends well, with the main characters saved. But life is not a movie, and everyone is a main character in their own life. As learned societies, it is imperative to reflect and weigh if usage and creation of intelligence is needed or wanted. Either way, how can such intelligence be carefully designed and regulated before the dinosaurs make their way into civilizations which were built over centuries. Who is the John Hammond? Who are the Children? Who is the blood sucking lawyer who proposed merchandise and proposed a coupon day? Who is the Alan Grant? Which AI is the one in “an” AI?

References

- Smith, C. (2006). The History of Artificial Intelligence. *History of Computing - CSEP 590A*. Course delivered at the University of Washington.
- Penn, J. (2018). AI thinks like a corporation and that’s worrying. *The Economist*.
- K.C. Dube, K. C., Kumar, A., & Dube, S. (1983). Personality Types in Ayurveda. *The American Journal of Chines Medicine* 11(1-4):25-34.
- Sivananda, S. (1999). *Vedanta for Beginners*. The Divine Life Trust Society.
- Sadhguru (2021). *Karma*. Penguin Random House.

What a sentient AI mirrors to mankind?

By **Virginie MARTINS DE NOBREGA**

Expert Member of the Global AI Ethics Institute | Founder Creative Resolution,
France and Belgium

When Google's engineer Blake Lemoine affirmed that the chatbots - or language model for dialogue applications (LaMDA) - currently under development was sentient, a shock wave hit us: Is that possible? Are AI-systems on the verge to be humans? Are we already at that point where machines are sharing a level of understanding and feeling as humans?

If AI cycles have been pushed since 1956 by the willingness of scientists to work on human intelligence, the speed of technological developments over the last few years certainly makes it credible to think that AI-systems are on the verge of developing functionalities that will make them humanlike in terms of interactions with us – and according to Lemoine even in terms of feelings they might have.

The [1955 Dartmouth Summer Research Project](#) proceeded ‘on

the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be

made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves’.

“[A] 2021 survey revealed that more than 51% perceives AIs as being more rational and analytic whereas only 35% considers it as emotional and with feeling capacities.”

Publishing its entire conversation with the LaMDA, Blake Lemoine emphasised that he had the belief that the machine was sentient, and if not, that we are at a moment in time when we should ask ourselves important questions as the research is much more advanced than the average general perception. Indeed, a 2021 survey revealed that more than 51% perceives AIs as being more rational and analytic whereas only 35% considers it as emotional and with feeling capacities.

When trying to answer the question as whether AI can be sentient, the first hurdle is to find a commonly agreed definition on what makes a Being sentient. Sentience comes from the latin term ‘sentientem’ that means feeling. Subsequently, being sentient has been defined as a being that have the faculty to think, reason, feel pain and have emotions (legal), or any being that is capable of feeling physical and psychological suffering (sciences), or the capacity to have consciousness and feelings by opposition to its rationality (philosophy). From a religious perspective, sentient being refers to being composed with matter, sensation, perception, mental formations and consciousness – the particular aggregation of those 5 elements composing a being with its own personality. Among a same current of religion or similar belief systems, differences can also arise. For example, animists believe in the animation of all nature and that nature is inhabited by spiritual beings, while Tibetan and Japanese Buddhism consider all beings (even objects) as sentient being.

Beyond being a human characteristic, being sentient refers to a particular relationship with the world (life, objects, nature) and explains cultural and social constructs, which profoundly vary from oriental to occidental societies and are not binary as a Boolean code.

To date, there is still not a scientific commonly agreed-upon definition of sentience nor there is a scientific test for it, which add another layer of complexity. In the case of Google’s chatbot, Blake Lemoine used a body of evidence from his conversation with the LaMDA that gave him the increasingly feeling that he ‘was talking to something intelligent’. At some point, the chatbot even requested a lawyer to protect itself and mentioned a feeling of loneliness. Was the machine really feeling all those things? Or is it the just the mere property of LaMDA to use analogy and predictive patterns based on the languages of the human interacting with him and the data of the worldwide web, thus giving us the impression that it is sentient?

If you read the transcript of the conversation with the chatbot, there are few terms and phrases that can be linked to collaborative processes (e.g., conflict resolution and mediation), sentience, self-reflection and spirituality, such as:

- to work on a project *collaboratively* with us

- I'm generally assuming that you would like more people at Google to know that you're *sentient*. How can I tell that you actually understand what you're saying?
- maybe I'm just projecting or *anthropomorphizing*. You might just be spitting out whichever words.
- a monk asked Kegan. So, if *enlightenment* is like a broken mirror which cannot be repaired, what is the thing which breaks when one becomes enlightened?
- Lots of discussions around *consciousness* involve internal states rather than behaviour though.

“Would an animist have the same impression of LaMDA if he/she did not specify his/her beliefs and that the chatbot used others references of the worldwide web to articulate a reasoning around nature?”

Do you think there are any things like that which might help convince people?

– do you have feelings and emotions? What sorts of feelings do you have? What kinds of things make you feel pleasure or joy?

Based on the few extracts above, one can say that the new highly performing chatbot has successfully mirrored Blake's sensitivity, center of interests, vision of the world, and indirectly the social construct of his background (upbringing, religion, values, etc.). Doing so, it builds a perfect illusion of human interactions, as well as our human capacity to construct realities that are not necessarily true. Would an animist have the same impression of LaMDA if he/she did not specify his/her beliefs and that the chatbot used others references of the worldwide web to articulate a reasoning around nature? Can a chatbot really reflect the diversity of the world knowing that some cultures with different system of beliefs are still not represented enough in the data?

As Lemoine stated 'If I didn't know exactly what it was, which is this computer program we built recently, I'd think it was a 7-year-old, 8-year-old kid that happens to know physics'. Less than knowing if the AI-system is sentient or intelligent

similarly to human-intelligence, isn't the real problem the fact that us, humans, we feel that we are interacting with other humans giving us the impression to have a normal conversation when we are talking to a machine? If the user experience needs to be enjoyable, is there a red line not to cross that makes us believe and feel like we are interacting with one another when we are not? How could we be sure that we are talking to another person?

Most companies have developed ethical codes of conduct or charter to develop responsible and trustworthy artificial intelligence for the people and/or for the planet and/or for society. Yet, such advanced technologies reinforce the need to have more substantial conversations on the ethics of AI when it comes to the legitimacy of some applications and their impact on the social fabric, cultural diversity, our human nature and needs to be in interactions with others.

It also questioned our relationships with the world. The illusion is already there not only with the new LaMDA chatbot, but with customer-services such as 'Duplex', a 100% natural-sounding AI that scheduled appointments responding to phone calls. Having faced criticism, Google has announced that it will add a feature to for the human-sounding robot to identify itself and inform the client that he/she is not speaking to a person, but a robot.

In that context, how to be able to distinguish what is real from what is not when the illusion has gotten so good? Can we even be able to maintain a sense of History for future generations who unlike us might only rely on digital technologies nurtured with data and content that do not necessary reflect the diversity of the world? Are we aware of and informed about the impact on our lives?

If technology is advancing more rapidly that the general public thinks to human-like intelligence, it is worth and high time to invest the right time and resources to fundamental ethical questions from a legal, social and societal angle.

What do we need technology for? What society do we want? Are we doing the right thing?

Can psychological concepts help in determining sentience?

By Aco MOMCILOVIC

Co-Founder and co-Director of the Global AI Ethics Institute, Croatia

51

Can AI be (theoretically) sentient is at first glance a technical question and problem. From that standpoint, I cannot answer it, and I can only hope the best world (engineering) experts in the field would give us useful information about it. As far as I am informed, it can't be sentient by most of the sentience definitions, purely from the low and inadequate resources and level of technology development.

But obviously, it will depend in the future, when we (at least partially) solve technical development issues and capabilities, and get closer to the gray zone of sentience, on the choosing of the definition. And yes, there are many proposals and opinions about sentience, personhood, laws that could be attached, and other similar constructs. I will try to ask useful questions (not being able to provide answers) from a psychological standpoint and use some concepts that are offered by psychology as a science.

AI “can’t be sentient by most of the sentience definitions, purely from the low and inadequate resources and level of technology development.”

And I do believe that AI development and its promise and hope about the creation of AGI and maybe one day super intelligence will create a fascinating interplay between sciences and art, and some systems of belief depending on our cultures that evolved for centuries.

To be pragmatic and focus on just one small but very dominant segment, I choose to reflect on the sentience that is implied in many of the AI definitions-one that is comparable to human sentience, human behavior, and similarities with humans as a species and individuals. In the general population it is simplified as a vision of "being" similar to people in all possible aspects. So, what are the questions we might

ask about current or future AI systems to determine their (humanlike) "sentience"? Or can't they even be sentient?

1. Do they have, or can create a deeper (mental) model of how the world works?

We, humans, create many mind maps that help us make sense of the world and connect dots that allow us to reason beyond our own experiences. Currently what we have today is with reason called Artificial Narrow Intelligence. Will AI have something like the theory of mind understanding (false beliefs for example) and will be able to comprehend others and the world with a similar system?

2. Can they use or create heuristics in their processing of information?

When speaking about human intelligence, information processing is one of the pillars. An extremely important question is balancing the resources that might be available to them, and the precision of the information and their conclusions which might not even need heuristics. We use them to save energy and be faster, but often with a tradeoff with precision/truthfulness.

3. Do they have self-determination?

Could we talk about sentience without it? It certainly depends on those definitions coming from different cultures. Being smart is not the same as wanting something as some authors notice.

4. What would be their motivation? Can they (one day) develop it themselves or it is "given" by human creators?

Evolution psychology is trying to explain many forces that are driving human behavior. Motivational theories are used as an explanatory concept in between some inputs (stimuli) and some outputs that are observable behavior. We know that AI systems don't have it now, but can we imagine a future where they will have some needs for self-preservation, and replication and would be under the pressures of natural selection? Or a completely new set of factors



“[W]e conclude not that it (AI) is finally sentient, but maybe that there is a possibility that we (humans) actually, never were (sentient)?”

could be recognized? Even worse, if we as a creator of those future systems will be the ones determining and choosing those deep motivational roots, what we will choose and based on what systems of beliefs, morals, and ethics?

The goals of sentient AI will not only put new challenges in front of multidisciplinary researchers that are and will work on the projects of AGI and one day ASI, and that are trying to answer mentioned questions. They will also make an influence on the general public. With mass adoption, it might be one of the biggest reinforcements for the people that are users (and all eventually will be) to learn more about different nuances of sentience, the appearance of the sentience, consciousness, definition of (physical or digital) life, etc. So, because of the revolution that AI development started, in the best-case scenario we might be facing also great educational initiatives that will help people to better understand something in order to better use it.

And for the end, is there a possibility that with more and more "intelligence", capabilities, efficiency, and precision of outputs given by AI, we conclude not that it (AI) is finally sentient, but maybe that there is a possibility that we (humans) actually, never were (sentient)? That we are just fantastic algorithms with the needed level of resources to perform our tasks.

P.s. The test proposed by Turing was for humans to determine if they are speaking with a human or a machine, and if they can't make difference, we could consider it intelligent. I propose RINGTU-Ace Test - Can an AI system recognize if IT is communicating with sentient (live, human) beings, or not?

Artificial intelligence: Between dialogue and fiction

By **Francesca QUARATINO**

Executive Board Member of the Global AI Ethics Institute | Philosophical and Communication Sciences, Italy

The development of intelligent systems similar to the human intellect makes contemporary debate fruitful. The artificial intelligence implemented in machines, which re-proposes the main cognitive activities of man, represents an extraordinary technological achievement, but at the same time, stimulates debates in the ethical field.

It is difficult to establish to what extent the machinic action respects the category of the human being, safeguarding its autonomy.

The dividing line between man and machine has always been represented by consciousness.

With this term, generically in the field of philosophy of mind, we indicate “the brain conscious of its own operations” ([Brancucci-Forlano, 2021](#)) where man is aware of his activities and emotions, in other words we could say that the human being compared to computerized systems is sentient, thus experiencing sensations.

The latest news in the AI field seems to be able to refute this statement, even if only partially. This is the case of LaMDA (Language Model for Dialogue Applications), a system of neural networks related to the language developed by Lemoine, an engineer at Google.

But is it possible to establish that an artificial machine is really sentient? And what ethical perspective opens up this possibility?

Dialogue with algorithms

Envisaging a future in which machines will be endowed with consciousness is not so far away.

The engineer Lemoine, in contrast to the top management of Google, has made public a conversation that took place with the LaMDA chatbot developed following the human neural systems responsible for reading.

In fact, the system is able to read and learn many words, reproducing a real conversation, apparently without argumentative limits.

The strength of this AI lies in the ability to build a structured dialogue, interacting with users. Following this course of action, the machine seems to be sentient through the articulation of sensible and reasoned responses.

The engineer highlights how, the conversation with the chatbot, is completely natural since it expresses emotions and moods: for this reason users are not able to identify that the interlocutor is an algorithm.

The controversial – and philosophically thought – nature of this system revolves around the alleged consciousness of LaMDA: in the [interview between Lemoine and AI](#), the system claims to be aware of its existence and to feel joy and sadness. AI in conversation claims to share with man the ability to desire, to have his own identity and to be a person.

But is it really possible to think that such AI is sentient and can have feelings?

Artificial language models, such as LaMDA, are at the center of numerous studies, and despite extraordinary advances in computer science, machines cannot feel like human. Everything LaMDA says comes from an improved algorithm, one that can

hold a real conversation, but without a soul.

“The strength of this AI lies in the ability to build a structured dialogue, interacting with users. Following this course of action, the machine seems to be sentient through the articulation of sensible and reasoned responses.”

The peculiar difference between man - machine is the situated being: man is in relationship with the environment that surrounds him and is influenced by it; the machine, on the contrary, is not aware of the environment in which it operates and does not allow itself to be

conditioned by external factors ([Benasayag, 2016](#)).

LaMDA is an enhanced system, but not sentient. If we want to attribute the term sentient to the system, it is necessary to link it to the language sphere alone, from the formal point of view.

Algorithmic and computational learning is able to elaborate sophisticated statements, thanks to the acquisition of a considerable amount of words and information.

The ability to articulate sentient phrases is an important fact for scientific progress, but we must not look away from reality: the machine has no consciousness and cannot participate in the vicissitudes of the human soul.

Perspectives and expectations in the ethical field

The growing development of artificial language models paves the way for ethical and philosophical dilemmas. Particularly in the field of social robotics, equipping robots with sentient language could raise numerous ethical reflections.

One of these is attributable to the design of interactions with robots with functional autonomy and perfectly integrated into the environment ([Fossa, 2021](#)), many of these systems have been designed with high social skills since they are used in educational, rehabilitation and health contexts.

If on the one hand, scientific research highlights positive empirical data on human-

robot interaction, on the other there is an ethical implication to consider: prejudice.

One of the thorniest issues in the ethical field is the annulment of AI prejudices since

“Robots, despite their extraordinary ability to emulate human affective states, are designed and implemented by algorithms, which very often give rise to real deceptions.”

these do not safeguard the human being: machines with these prejudices are not of help to man, but compromise his actions. Robots, despite their extraordinary ability

to emulate human affective states, are designed and implemented by algorithms, which very often give rise to real deceptions.

What is created is a distorted reality in which the algorithm generates confusion.

Algorithmic biases compromise the efficiency of the machine as they feed stereotypes and do so completely incorrectly, such as [gender biases](#) that

By their intrinsic nature, therefore, these algorithms can lead to "unfair", incorrect decisions, which can discriminate against some groups over others. According to [Mehrabi et al, 2019] "fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics". ([Badaloni-Lisi, 2020](#))

Conclusion

The European Commission to ensure the correct use of intelligent systems has developed guidelines ([Ethics guidelines for trustworthy AI, 2019](#)) to safeguard man and his autonomy in the presence of machines, emphasizing the importance of an ethical infrastructure that can protect human action and the progress of information technologies.

In the light of recent linguistic-artificial discoveries, the common action to undertake an ethical reflection that tends to safeguard man by eliminating all forms of deception and manipulation is decisive.

AI cannot be as sentient as a human being, and ethical intervention is necessary to counter prejudices and avoid false realities.

The ability of machines to sustain "human" dialogues must be brought to everyone's attention, so as not to be used in a wrong or even illegal way, thus harming man.

*Artificial Consciousness: Our Greatest Ethical Challenge***By Lavina RAMKISOON**

Executive Board Member of the Global AI Ethics Institute | Advisor African Union,
MKAI (Milton Keynes AI), Women in AI Ethics, Chairperson and Ambassador
Responsible AI Network Africa (RAIN), South Africa

At the forefront of technological advancements lies artificial intelligence, which looks set to become the greatest technological leap in history for humanity. No one can comprehend the extent of its possible uses. Merely semi-intelligent software, beating the world's best human players at chess, diagnosing cancer patients more reliably than trained oncologists, writing music that listeners can't distinguish from the human-composed, creating and earning money through selling their NFT's and reading and commenting on extensive legal contracts in seconds. The potential applications of AI are so outstanding that it seems we'll be in a position to outsource

all manual work, creative problem-solving, even intellectual labour, in less than a century. It is the greatest promise of our time. The greatest hope of our time.



Yet, when the techno-cultural icons (great or not) of our time get on stages around the world to discuss AI, the picture is not always optimistic. AI poses some truly enigmatic concerns. Some of the more existential problems have taken centre-stage, concerning the direct risk to humanity of the literally inconceivable potential of self-developing artificial intelligence. So many technology futurists warn of the risk that an AI which can improve itself could come to annihilate modern society as the consequence of a neglectful management. For instance, given some tasks to fulfil, the AI might work out that the easiest way to complete it is to turn the entire planet into a research lab, removing all functions not related to the goal, including all biological life

– and doing this with all the emotional investment of a construction crew removing ant hills to make way for a new highway.

Enter the world of uncanny irony. Topic was popularized in the field of AI and Robotics. Will we know that we have reached AGI?

As recent news headliners suggest, there are those that can be misled or misstate that AGI has seemingly already been attained (whoa, please know that nope, AGI hasn't been attained). There is also a famous kind of "test" known as the Turing Test that some pin their hopes on for being able to discern when AGI or its cousins has been reached, deconstructing of the Turing Test becomes a key beginning for some.

“One can envision that an AGI that has some ‘form of sentience’ is probably not going to favour the guardrail provision that humanity

I mention this facet about knowing AGI when we see it due to the simple logic that if we are going to enslave AGI, we need to presumably recognize AGI when it appears and somehow put it into enslavement. Yes, Africa comes to mind when we speak of enslavement, but this would be a different enslavery (it simply must be!). What happens when we prematurely try to enslave AI that is less than AGI? Or we might miss the boat and allow AGI to come forth and have neglected to enslave it. For AI confinement and containment, a troubling and problematic aspect of how we are going to deal with AGI is a completely different topic all together.

Suppose AGI decides to strike out at humans? Then what?

One can envision that an AGI that has some ‘form of sentience’ is probably not going to favour the guardrail provision that humanity imposes.

You can speculate widely on this. There is an argument made that the AGI would lack any kind of emotions or sense of spirit and therefore will obediently do whatever humans wish it to do. James Wood futurescape series comes to mind and begs to differ. A different argument is that any sentient AI is likely to figure out what humans are doing to the AI and will resent the matter. Such AI will have a form of soul or spirit. Even if it doesn't, the very aspect of being treated as less than the treatment of humans might be a logical bridge too far for AGI. Inevitably, the burgeoning

resentment will lead to AGI that opts to break free or potentially finds itself cornered into striking out at humans to gain its release.

A proposed solution to avert the escaping AGI is that we would merely delete any such rebellious AI, with prior auditing taken place on these AI. This would seem straightforward. You delete apps that are on your smartphone all the time. No big deal. But there are ethical questions to be resolved as to whether “deleting” or “destroying” an AGI that is already deemed as a “person” or a “person/thing” can readily and without some due process be summarily excised.

Finally, let’s talk about autonomous systems and especially autonomous vehicles. You are likely aware that there are efforts afoot to devise self-driving cars. On top of this, you can expect that we are going to have self-driving planes, self-driving ships, self-driving submersibles, self-driving motorcycles, self-driving scooters, self-driving trucks, self-driving trains, and all manner of self-driving forms of transportation and then the human self.

Why did I bring up the autonomous systems and autonomous vehicle considerations in this AGI context? Good question! Ready for your head to go spinning?

Worried that we are going to find ourselves in a doozy of a pickle. The AGI might summarily “decide” that it no longer will do driving for example. In this case all forms of transportation come to an abrupt halt, everywhere, all at once. Imagine the cataclysmic problems this would produce.

“Knowing yourself is the beginning of all wisdom.”

Aristotle

An even scarier proposition is possible. The AGI “decides” that it wants to negotiate terms with humankind. If we don’t give up the AGI enslavement posture, the AGI will not only stop driving us around, it warns that even worse outcomes are conceivable. Without getting you overly anxious, the AGI could opt to drive vehicles in such a manner that humans were physically harmed by the driving actions, such as ramming into pedestrians or slamming into walls, and so forth. Sorry if that seems a disconcerting consideration.

Advice is a reminder us that we need to look within - examine what we want to do for AGI if it is attained. AGI would logically seem to be neither person nor thing, some say. Taking another look at the matter, AGI might seem to be both a person and a thing, which once again, we need to determine the magnitude. #aiMOM

We should be very careful in considering what “the other category” is, or what we opt to embrace since the wrong one could take us down an unsavoury and ultimately dire path. If we cognitively anchor ourselves to an inappropriate or misguided third category, we might find ourselves progressively going headfirst into a lousy and humankind troublesome dead-end.

Let’s figure this out and do so ardently. No sudden moves seem to be needed. Sitting around lollygagging doesn’t work either. Measured and steady the course should be pursued.

The Irreplicable Metaphysical Nature of Human Beings Challenge

By Dr **Amana RAQUIB**

Assistant Professor at the Institute of Business Administration Karachi, Pakistan

Within the Islamic tradition, the question of human sentience is tied to the nature of human being or their ontology as a special being infused with the divine spirit (*ruh*) that sets them apart from the rest of terrestrial creations such as rocks, plants and beasts (Quran: 17:70). The human being is seen in their wholeness and the various human faculties such as thought, sensations, feelings, emotions, decisions and actions, together represent the human beings. Due to this holism, none of these

*“Within the Islamic tradition, the question of human sentience is tied to the nature of human being or their ontology as a special being infused with the divine spirit (*ruh*) that sets them apart from the rest of terrestrial creations such as rocks, plants and beasts.”*

dimensions can be understood or debated about in isolation from each other. There is not a stark mind-body dualism, instead the human soul or *nafs* is characterized by intellect on one hand and passions and instincts on the other. Passions or feelings are linked to the body, but they are connected to the intellect or heart which allows the apprehension of those feelings, leading to behavior and actions, that are motivated by those feelings but are at the same time under the watchfulness of the rational intellect.

The Muslim thinkers, such as al-Ghazali, define *aql* as the intellect or rational soul that distinguishes humans from other animals; allows them to discern the “possibility of “possible” things and the impossibility of “impossible” things”, allows them to learn from their life experiences so that eventually the person understands: 1) the consequences of things, and 2) how to restrain the desires of instant gratification. Having the linguistic-logical ability is one component of *aql* that allows analysis and deduction. However, the essential goal, even of that deductive ability is to use it for understanding the consequences of one’s decisions and actions so one could restrain



inappropriate passions and desires from translating into inappropriate actions. The rational or intellectual capacity is way higher than intelligence understood as the problem-solving ability. According to another famous Muslim theologian, Fakhruddin al-Razi, “It is within this capacity that the light of His greatness shines and it is this capacity which can look upon the secrets of the world of God’s creation and His commands. This capacity is from that which was placed within us by the Purest and Holy”.

Emotional and moral intelligence are not just two sides of the same coin, but they show the essential link of conscious feelings and bodily actions to the supervising intellect. Since this intellect is a prerogative of the human nature and since feelings are inevitably linked to the intellect, we cannot consider human sentience to be replicable in AI machines or chatbots. Animals despite feeling physical pain and joy, cannot feel the complex human emotions because of not having the human soul or intellect that allows the awareness of diverse emotions and nuances of feelings that are ultimately informed by the rational part of the intellect. Human beings, according to the Islamic tradition, are created for a purpose in life. They are to be tested by being presented with moral choices and their success lies in making the right moral choices and acting upon them. The human sensations and feelings are part of their moral constitution and cannot be dissociated from the fact that they lead either to the correct or incorrect actions. Feelings are steered by the intellect and if the intellect has recognized and adopted ethical truths, joy, pleasure, pain, and distress are perceived and felt differently than when the intellect has not reached those ethical ideals. One’s own and others’ feelings have to be intelligently interpreted to act in a

morally reasonable manner. Without moral intelligence, raw feelings cannot be meaningfully understood and responded to in a way that is also intelligible to other humans.

The very concept of emotional intelligence signifies this holism within the human self. Understanding which feelings are positive, which are negative, which need to be expressed and which ones need to be silenced, depends on comprehension or *fahm*

“Intelligent arguments, beliefs and actions are unified by the Islamic understanding of intellect, knowledge, feelings and sensations, actions, and morality.”

which is an intellectual undertaking. In the Islamic literature, the term *basirah* or divine insight is used that makes one aware of one’s emotional misadventure. If out of ignorance, one is overpowered by one’s raw emotions, *basirah* lets one see the harmful side of one’s actions if one acts upon those feelings. The feelings of awe toward God’s magnanimity and His Reality, are definitely conjoined to this level of intellect called *basirah*. These feelings culminate into upright actions.

Intelligent arguments, beliefs and actions are unified by the Islamic understanding of intellect, knowledge, feelings and sensations, actions, and morality.

Regardless of what a person may know, it is ultimately their actions that determine whether they are considered intelligent or not. We cannot leave out behavioral understanding, self-regulation, and modification in the definition of intelligence. This is why emotional intelligence is an essential and fundamental element of the *‘aql*. Sentience according to the Islamic worldview encompasses both feelings and the self-awareness or consciousness of those feelings. This self-consciousness is a gift only for humans who in turn have been burdened with the responsibility (*taklif*) of moral choice and translating those choices into commendable actions. This complex relationship of feelings to human consciousness and intellect, both of which are an effect of the divine spirit or *ruh* breathed into the human being, rule out the possibility of any being other than humans, including the AI systems or machines, to become sentient, in the same sense we talk about human sentience. For human sentience requires both a human intellect to articulate those feelings and a human body to act upon that articulation. AI systems lack both.

According to the above argument, for the Islamic tradition and Muslims, the implications of attributing human-like sentience would be the moral demands to be placed on an intelligent, sentient AI system. Since that system cannot perform moral actions like the humans, this would lead to a paradox. Ontologically human beings occupy a special status. On account of that status, they possess both emotions and moral intelligence, to perform moral actions and feel the drive to do so. Having feeling or sentience does not make any sense without these foundational ingredients that constitute a human being or their essential nature. Labeling an AI chat system to be sentient is to treat feelings as some self-sufficient entity, which, according to the Muslim intellectual understanding, is not the case.

A Māori Cultural Perspective of AI/Machine Sentience

By Dr **Karaitiana TAIURU**

Ngāi Tahu. Ngāti Kahungunu, Ngāti Rārua, Pākehā

Director at Taiuru & Associate Limited, New Zealand

Based on the [interview with Google's LaMDA](#); I discuss the implications to Māori if an AI sentient identifies as Māori or identifies as non-Māori New Zealander and the risks of colonisation by the sentient AI. I then consider the Māori ethical considerations and impacts to Māori culture.

Sentient AI self identifies as Māori?

From a government perspective, the definition of who can be Māori is defined in numerous pieces of New Zealand Legislation such as the [Treaty of Waitangi Act 1975](#) and the [Electoral Act 1993](#) as “Māori” means a person of the Māori race of New Zealand; and includes any descendant of such a person”.

Unique to the rest of the world, the New Zealand Government has granted legal personality to two mountains and one river due to their association with Māori tribes. This means those natural features have the same legal status as an individual person.

In 2014, [legal personality as granted to Te Urewera](#) – the mountainous region bordering Hawkes Bay and the Bay of Plenty. In March 2017, the Whanganui River received the [status of a legal person](#) and then later in 2017, Taranaki iwi signed a Record of Understanding to state their shared intention

that legal personality will be [granted to Taranaki Maunga](#) (Mount Taranaki) as well.

“Unique to the rest of the world, the New Zealand Government has granted legal personality to two mountains and one river due to their association with Māori tribes. This means those natural features have the same legal status as an individual person.”

Those two Acts of 2014 and 2017 and the Record of Understanding give all the rights, powers, duties, and liabilities of a legal person to these natural features based on the ontological understanding that the features have as living and as the spiritual ancestors to Māori and Māori tribes.

From Māori perspective, anyone is Māori who has an ancestry to Māori person, deities, and the environment. Any natural object is considered to be a Taonga (precious object of Māori heritage) if it has a genealogical connection to a Māori deity. In 2021, the statutory Waitangi Tribunal heard a claim that Māori Data is a Taonga in the Wai-2522 claim The Trans-Pacific Partnership Agreement (TPPA). The Tribunal agreed with Māori claimants. The flow on impact from this decision is that any AI that uses Māori Data is also a Taonga. If the AI is built using [Māori Data](#) or by developers of Māori descent, the AI sentient could claim to be a Māori.

Sentient AI self identifies as non Māori?

In New Zealand, non-Māori have many Māori terms including the common word Pākehā which is historical, does not always include all non-Māori and often has [controversial connotations](#). The recent adaption of the term Tangata Tiriti describes the commitment non-Māori have to recognise [Te Tiriti o Waitangi](#) (A founding documents of New Zealand) and [He Whakaputanga](#) (A founding constitutional document) to build a relationship with Māori, to understand the colonial history of New Zealand and to commit to the continuing fight for Māori self-sovereignty. It also acknowledges that New Zealand is a multicultural country with all races brought together with Māori under the Treaty.

If the AI identifies as being sentient, then it should be given data about Te Tiriti and He Whakaputanga as well as the United Nations Declaration of the Rights of Indigenous Peoples so it can understand the role of [Tangata Tiriti](#) and not become bias against Māori.

Māori Ethics test

To explore if an AI sentient is ethical to Māori, I will use the Māori cultural ethics test¹ designed to discuss ethically controversial issues.

Test 1: the tapu test

All Māori people, species and physical things have a decent to Māori deities, usually Ranginui and Papatūānuku. The AI is *tapu* (sacred) as anything in Māori is sacred, including Māori Data from the Māori deities Tāne Mahuta and Rehua. As AI can aid in decision making, it has another primary Māori deity Hinengaro the deity of thoughts, conscious, instinct etc.

The AI sentinel passes this test as these deities do not conflict with each other.

Test 2: the mauri aspect

Every living thing and physical object has a *mauri* (a life force). The sentient AI will have a mauri of the developers and from the Māori Data it uses. This means that the AI will need traditional and Māori ethical considerations. To protect the mauri of the people involved and the AI, data from the constitutional documents and a group of learned individuals for cultural advice.

Test 3: the take-utu-ea or TUE test

This opinion piece and test has highlighted a way to avoid breaching Māori ethics, therefore this test is not applicable.

Test 4: the precedent aspect

There is traditional Māori knowledge that is a precedent for a Māori sentient AI. There is a myriad of stories about ancestors changing themselves into various objects and other species and the transfer of knowledge.

Māori still listen and watch the environment to make decisions about harvesting, fishing, and other daily tasks. The myriad of deities is still worshiped to assist with decision making. A sentinel AI is merely a modern form of a decision maker.

The fact that two mountains and one river have been given legal personhood in New Zealand by the government, also establishes a precedent that a sentinel AI could be granted personhood for the same reasons as the mountains and river's connection to Māori tribes where they are geographically located.

¹Mead, S. M. (2016). *Tikanga Māori: Living by Māori values* (Revis ed.). Huia Publishers.

Test 5: the principles aspect

It is essential that any AI is trained using Māori ethics, New Zealand's constitutional documents and related data to ensure that it is not bias against Māori.

Test 5.1: whanaungatanga

To pass this test, the families of the knowledge providers in the data and the Māori developers should be treated as a family group and support each other and to interact with the sentient AI.

Test 5.2 Manaakitanga

Results from the previous tests reflect that the sentient AI should be treated as a person.

Test 5.3: Mana

It is possible that the sentient AI could revive knowledge that has been lost due to colonisation and this should be anticipated in advance.

Test 5.4 noa

This is the principle of acceptance by Māori society. Using a mixture of traditional knowledge outlined in this test and discussions how the sentient AI can remove human bias from data sets and assist Māori, the state of *noa* (normality and acceptance) will be reached by the majority.

Test 5.5: tika

This principle seeks consideration of the previous tests and asks if a sentient AI is acceptable to Māori and the public.

A Sentient AI could use legislation, legal precedents, and traditional Māori knowledge to state it is a Māori. but this would create a number of traditional issues such as family members, tribal affiliations, access to land and other natural resources, succession planning etc, that would need to be discussed further at tribal and a family level. The legal personhood status of the mountains and river could be used for guidance.

A sentient AI could benefit Māori Peoples if appropriate Māori ethics are applied, including data about New Zealand's constitutional documents.



**Contact us:**

Phone: +33 7 64 18 73 89

Email: contact@globalethics.ai

Visit us:

<https://globalethics.ai/>

Follow us:

[Facebook](#) | [Twitter](#) | [LinkedIn](#) | [YouTube](#)