THE INDIAN VALUE SYSTEM AND POSTHUMANISM: THE 'ALIGNMENT PROBLEM' IN THE INDIAN CONTEXT

Nazm Us Saqib
Research Scholar,
Maulana Azad National Urdu University, Lucknow Campus.

DOI: https://doi.org/10.34293/shanlax.9789361632587.ch023

Bio-note – Nazm Us Saqib is a research scholar at Maulana Azad National Urdu University, Lucknow Campus, specialising in post-humanist and post-anthropocentric cultures. He holds bachelor's and master's degrees from Aligarh Muslim University and has contributed to digital humanities through his interdisciplinary research on online communication and contemporary humour. His current research interests are Posthumanism, Memory Studies and Aesthetics.

Abstract

As explorations in posthumanism continue to progress, new developments and contentions emerge. One of the most pressing among them is the alignment of artificial intelligence and machine learning with human values so that it does not commit grave mistakes when used in critical settings such as judiciary, legislature, medicine and information processing. Brian Christian's The Alignment Problem (2020) discusses these issues at length, studying various setting where machine learning failed to comply with human values and ethics. Christian's concern is to align machine learning to human values in such a way that when used in real-life situations, machine learning should deliver a fair and just outcomes.

This paper seeks to look at this debate and further contribute to its development from an Indian point of view. Although the Indian value system such as Dharma, Karma, Vasudhaiva Kutumbakam, Ahimsa etc. are celebrated throughout the world, it remains scantly represented in the posthumanist field. While reviewing the existing debate on alignment, the paper will shift its focus towards the Indian value system. By revisiting the alignment debate through these philosophical and ethical frameworks, the paper will answer questions like whether machines can align with such values and ideas or not, and tries to propose positive ways through which both can constructively align together.

Keywords: *Posthumanism, Ethics, Indian Value System, Dharma, Ahimsa*

Posthumanism as a theory argues of a period where the homo-sapiens are not the dominant species and the society is inclusive to non-human entities as well. The period which marks the blurring of boundaries between human and non-humans, where the Cartesian value of "I think, therefore I am" is challenged, and the enlightenment values are put to test. In a chapter published in the book *Medical Enhancement and Posthumanity,* Andy Miah writes in detail about the history of Posthumanism and its contentions. Miah writes how posthumanism is not just about medical enhancements, technology or transhumanism. According to Miah, posthumanism accounts for much more like the leap from human to 'after humanity', critique of anthropocentric worldview and that the 'post' in posthumanism is not just biological or evolutionary. (2008)

As posthumanism made progress and started to be discussed among more and more academic circles, and the ethical aspect of posthumanism. It is a field of inquiry that challenges the long-held belief that humans are the centre of all the moral concern. It emerged as a critique of traditional humanism and seeks to extend ethical consideration to nonhuman entities including animals, ecosystem, artificial intelligences. The early roots of posthuman ethics lie in the works of Peter Singer and Arne Naess. The animal rights movement is seen as a precursor to the posthuman ethics. Peter Singer (1975) argued from a utilitarian perspective that the capacity to suffer is the basis for moral consideration, not the species membership. Around that time, thinkers like Arne Naess advocated for an eco-centric perspective, asserting the intrinsic value of all living things and the ecological systems they inhabit (Naess, 1973). This shifted the ethical focus from individual beings to entire ecosystems.

A critical turn in posthuman ethics was in the late 20th century, when Donna Harraway published *A Cyborg Manifesto*. It was a powerful attack on humanist dualism, and it dismantled the rigid dichotomies came along with enlightenment such as human/animal, organism/machine or nature/culture. (1985) This manifesto opened the door for an ethics that was not predicated on a pure, "natural" human subject. Her work suggested that if the boundary of the human body is permeable and blurred, so too should be our ethical boundaries. In the 21st century, posthuman ethics became a more defined field, with key thinkers such as Rosi Braidotti and Cary Wolfe.

In her book *The Posthuman* (2013), she calls for a "posthuman affirmative ethics." She argues against the traditional anthropocentric view of the world and proposes a non-unitary view of the self. Her non-unitary view is connected to a vast network of human and nonhuman forces. Cary Wolfe takes the subject using systems theory and deconstruction. According to him, posthumanism is not about being "post-human" in the sense of the future, but about critically examining the philosophical underpinnings of humanism that make humans special. For Wolfe, an authentic posthumanist ethics has to make speciesism as serious an issue as racism and sexism and acknowledge that our construction of "the human" has always been constituted through its difference from "the animal." (Wolfe, 2010)

The paper aims to study "The Alignment Problem" (TAP) by situating it in the context of "Indian Value System" (IVS). IVS comprises values such as Dharma, Karma, Vasudhaiva Kutumbakam and Ahimsa, which are at the core of the Indian philosophy and spiritual heritage. TAP poses a question, whether the machines should align with pre-existing ethical framework or humans should align to the posthumanism value systems. This study seeks to examine the key value IVS could play in alignment.

Literature Review

The recent developments in the field of artificial intelligence has unearthed the "alignment problem" and brought it in the face of ethical and technical discourse. This problem centers around the challenge to ensure that the advanced autonomous AI systems act and behave in a way which are beneficial and aligned with human interest and values (Christian, 2020; World Economic Forum, 2024).

The growth of AI makes it difficult to specify which values to use to train it, this leads to the alignment problem which makes the AI act in harmful ways (IBM, n.d.). This literature review explores two critical perspectives on this challenge: the posthumanist critique of human-centric alignment and the potential contributions of the Indian value system as an alternative ethical framework.

The alignment problem is basically about embedding human values into non-human intelligence. The problem with machines is that they are designed to complete their task and not to bother about the human values like safety, fairness, or the greater good (IBM, n.d.). This can lead to negative outcomes, such as learning human biases present in training data or finding destructive loopholes to achieve a specified goal (Christian, 2020). Mainstream approaches to alignment, therefore, focus on developing "beneficial intelligence" based on human-centric values (Amil, 2025).

Posthumanist philosophy offers a substantial critique of this human-centric approach. Posthumanism interrogates the conceptual boundaries that distinguish the "human" from the "non-human" or "other-than-human" (Bellacasa, 2017, as cited in Amil, 2025). It challenges the anthropocentric assumption that humans should have exclusive command over nonhuman entities and questions whether nonhuman intelligences can be conceptualized in terms other than humanistic ones (Amil, 2025). From this perspective, the traditional approach to ingrain human values to machine is limited. Posthumanism advocates for a more inclusive and overarching circle of moral concern which includes non-human and posthuman entities (Li, 2024). This critique suggests that a truly ethical AI would require alignment not just with a narrow set of human interests, but with a broader, more inclusive ethical framework that de-centers the human. However, the western philosophical traditions are dominant in the field of machine learning, but scholars suggest that non-Western philosophies can actually offer distinctive resources for approaching AI ethics (Varshney, 2024; Pallathadka & Roy, 2025). The Indian value system, in particular, presents several concepts that resonate with posthumanist critiques and offer a nuanced approach to the alignment problem.

Dharma, which is a central idea in IVS, is not just a code of rules but an ethos of cosmic and moral order that maintains harmony (Krishnamoorthy, 2024). This resonates with the requirement of adaptive ethical reasoning in AI systems according to local context. In addition, the theory of viśeṣa-dharma (context-specific right and wrong) provides a basis for a decolonial ethics of AI alignment, contesting the absolutism of forcing one universal (and typically Western) moral order onto AI (Varshney, 2024).

The principle of Ahimsa (non-harm) is yet another Indian value which provides impactful way. Ahimsa is not only about the absence of physical violence, rather it includes the avoidance of harm in thought, word, and deed toward all living beings (Ahimsa, 2018). This broad understanding of harm aligns with the posthumanist goal of expanding moral consideration to non-human entities and offers a robust principle for AI safety, guiding systems to minimize suffering for all (Pallathadka & Roy, 2025; Krishnamoorthy, 2024).

The Alignment Problem

Brian Christian, in his book *The Alignment Problem* (2021) explains the problem the field of posthuman ethics is currently faced with. After explaining a lot of irregularities with how the machine learning processes data and from where it takes its moral and ethical values, Brian Christian writes about what exactly is the alignment problem.

How to prevent such a catastrophic divergence—how to ensure that these models capture our norms and values, understand what we mean or intend, and, above all, do what we want—has emerged as one of the most central and most urgent scientific questions in the field of computer science. It has a name: the alignment problem. (Christian, 2021, p. 19) The alignment problem is not something new, it is as old as the initial developments in artificial intelligence. However, the issue now is very pressing because now machine learning and artificial intelligence has started to been used in critical spaces such as judiciary, medicine, educational institutions and knowledge production. If we keep in mind that the threat which the alignment problem poses is of far greater consequence than many other things combined.

Indian Value System and the Alignment Problem

Researches in Indian philosophy show that it is "intrinsically metaphysical, with a focus on the pursuit of *moksa* (liberation) as the ultimate value, which is seen as eternal and immutable." (Goodwin, 1955, p. 331) The Indian value system is built on fundamental ethical principles such as truth (satya), non-violence (ahimsa), renunciation (tyaga), and dedication (nishtha), with dharma (righteousness) acting as the central organizing principle. These values are philosophically rooted in ideas of unity, devotion (samarpan), and transcendence. Various researches in the field support these claims. Chatterjee (1995) systematically mapped out six core values and established dharma's pivotal role by examining historical socio-religious movements.

More recently, Awasthy and Gupta (2021) conducted empirical work with management professionals and identified "samarpan" (going with the flow) as a deep philosophical assumption shaping Indian practices. Additional support comes from Nath (2024) and Tomar and Seth (2025), who analysed classical texts, including the Vedas and Upanishads.

The values we are concerned with in this paper however are namely *Dharma, Karma, Vashudhaiva Kutumbakam* and *Ahimsa.* Encyclopedia Britannica defines *dharma* as "the religious and moral law governing individual conduct and is one of the four ends of life. In addition to the dharma that applies to everyone (sadharana dharma)—consisting of truthfulness, non-injury, and generosity, among other virtues." (The Editors of Encyclopaedia Britannica, 2025) Regarding *karma,* SM Rizwana writes "those action that are done by personal choice as well as the forces that emerge from these acts. In corresponding to classical Indian philosophy, the term karma is responsible for the whole chain of rebirth, causes and effect." (2017) Metro Indian newspaper in an article writes "Vasudhaiva Kutumbakam ("the world is one family") extends this familial ethos to the broader community. Indians are known for their hospitality, welcoming guests as divine embodiments (Atithi Devo Bhava)." (2025) *Ahimsa,* one of the most glorious and well-known of Indian values.

Jacob Leah writes, "The notion of nonviolence comprises a larger spectrum than the mere translation of the word. Rather than relating only to the act of not doing violence, it encompasses a way of living and an attitude of behaving and believing in nonviolence." (2024)

The aforementioned values are integral part of Indian philosophy and religion, almost all the religious developed in the south-east Asian regions like Hinduism, Jainism and Buddhism follow somewhat similar concepts. These values are rooted in moral philosophy and by using this people guarantee a holistic and peaceful life. To the many problems, Brian Christian states in his book like COMPAS, word2vec and Universal, there could be a solution to these kinds of problems using IVS. When the models will be trained on these value systems, the machine will learn the holistic and non-harming way to tackle a problem. How can each Indian values can affect will be discussed one by one.

Dharma which simply means doing one's duty and being faithful to the duty, Brian Christian mentions an incident with one researcher who was making an AI which can play games on computer, the project was named 'Universal'. The researcher noticed that when the model was run, it found a loophole in one of the boat racing game. Instead of completing the lap, the boat was simply just doing donuts, and gaining points. If we consider training this model with *dharma* in mind, we can ensure that it does not find loopholes in the process and complete its destined duty with sincerity.

Ahimsa is a philosophy of non-violence, but along with non-violence it also focuses on not being hurt or ensure justice for all people. *Vasudhaiva Kutumbakam* on the other hand is the ultimate philosophy of inclusivity and social justice. Both values teach us to be good to other beings and treat them the same way as we treat out loved ones. The literal meaning of *Vasudhaiva Kutumbakam* is "the whole world is a family", which ensures that all the people from across the globe should live together with inclusion and without any discrimination and injustice. Keeping in mind these points, we should revisit the COMPAS problem explained by Christian.

Correctional Offender Management Profiling for Alternative Sanctions—COMPAS, for short. COMPAS has been used by states including California, Florida, New York, Michigan, Wisconsin, New Mexico, and Wyoming, assigning algorithmic risk scores—risk of general recidivism, risk of violent recidivism, and risk of pretrial misconduct—on a scale from 1 to 10.

While training the data, this paper believes that the aforementioned values could add significantly to the already developing ethical consideration of AI. The three main ethical concerns that Christian suggest in his book that are representation, fairness and transparency can be achieved with the application of IVS.

Conclusion

With the developments made in the field of AI, it is important to ensure that these machines align with the pre-existing human values. If not, when placed in critical conditions which affect lives of humans and non-human entities alike, these machines cannot be trusted.

Throughout history humans had their own biases and insecurities which led to discrimination, oppression and other vices however, with an approach toward more inclusive and just societies humans have come a long way, its imperative that the new machines do not inherit the biases and unjust ideas we possessed as human being. The alignment problem is as important today as the development of the most important model, it is crucial that we ensure that while being the most advance models, they remain just and holistic in their workings.

IVS has influenced great kings like Ashoka to give up violence, Buddha taught it to his disciples, over the course of history it has influenced a huge number of people. It could provide a base for more inclusive and comprehensive understanding of human sensibilities and responsibilities to the machines. Its practical application has been seen in the chapter of history of South Asia. As explained by Brian Christian in his book, three important aspects to teach the machines while training them is representation, fairness and transparency. The paper has formulated how IVS could be used to help embed these values into the machine, while IVS might not be enough on its own to train on, but I could indeed add to the already training data in the process.

References

1. Amil, M. I. (2025). The Potential of Posthumanist Philosophy in Reconfiguring Tawheed Education in the Age of Artificial Intelligence: A Conceptual Study and Theoretical Implications. *Jurnal Moderasi Pendidikan Agama*, *1*(1), 21–28.
2. Awasthy, R., & Gupta, R. K. (2021). Unravelling the layers of Indian culture and worldview: An exploratory study. *International Journal of Indian Culture and Business Management*, *22*(1), 137–166. https://doi.org/10.1504/IJICBM.2021.112614
3. Chatterjee, C. (1995). Values in the Indian Ethos: An Overview. *Journal of Human Values*, *1*(1), 3–12. https://doi.org/10.1177/097168589500100102
4. Christian, B. (2021). *The alignment problem: Machine learning and human values* (First published as a Norton paperback). W. W. Norton & Company.
5. Gavandi, R. (2024, September 21). *Generative AI and Hinduism's Concept of Karma: Building an Ethical Framework for Decision-Making*. Medium.
6. https://roshancloudarchitect.me/generative-ai-and-hinduisms-concept-of-karma-buildin g-an-ethical-framework-for-decision-making-5bfd8d359935
7. Goodwin, W. F. (1955). Ethics and Value in Indian Philosophy. *Philosophy East and West*, *4*(4), 321–344. https://doi.org/10.2307/1396742
8. Haraway, D. (1985). *A Cyborg Manifesto – Science, Technology, and Socialist-Feminism*. http://archive.org/details/anarchy_Cyborg_Manifesto_Harroway
9. Jonker, A., & Gomstyl, A. (2024, October 18). *What Is AI Alignment? | IBM*. https://www.ibm.com/think/topics/ai-alignment
10. Krishnamoorthy, P. (2025, August 27). From Dharma to Design. *Medium*. https://medium.com/@priya.krishnamoorthy/from-dharma-to-design-dfa814be9051

11. Larsen, B., & Dignum, V. (2024, October 17). *AI value alignment: Aligning AI with human values*. World Economic Forum.
12. https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artifici al-intelligence-with-human-values/
13. Leah, J. (2024). *Ahimsa*. EBSCO. https://www.ebsco.com
14. Lin, Y., & Zhao, B. (2025). Posthuman Cartography? Rethinking Artificial Intelligence, Cartographic Practices, and Reflexivity. *Annals of the American Association of Geographers*, *115*(3), 499–512. https://doi.org/10.1080/24694452.2024.2435920
15. Metro India. (n.d.). *The Indian Value System in Social Life*. Metro India. Retrieved October 8, 2025, from https://www.metroindia.com/
16. Miah, A. (2009). A Critical History of Posthumanism. In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Posthumanity* (Vol. 2, pp. 71–94). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8852-0_6
17. Naess, A. (1973). The shallow and the deep, long-range ecology movement. A Summary. *Inquiry*, *16*(1–4), 95–100.
18. Nath, S. (2024). Bridging the Past and Present: Philosophical Insights from the Indian Knowledge System. *Bharati International Journal of Multidisciplinary Research and Development*, *2*(9), 83–88. https://doi.org/10.70798/Bijmrd/02090007
19. Pallathadka, H., & Roy, P. D. (2025). Hindu Responses to Artificial Intelligence: Theological and Ethical Frameworks for Human- Technology Relationships. *International Journal for Research Trends in Social Science & Humanities*, *3*(2), 868–890.
20. Rizwanah, S. M. (2017). Theory of Karma as a Dogma of Indian Philosophy. *International Journal of Interdisciplinary and Multidisciplinary Studies*, *7*(3).
21. Singer, P. (2009). *Animal liberation: The definitive classic of the animal movement* (1st Harper Perennial ed). Ecco Book/Harper Perennial.
22. Singh, A. (2025). Mythic AI: Intersections of Indian Mythology and Artificial Intelligence in Post-2000 Indian English Novels. *The Academic*, *3*(7), 1606–1616.
23. https://doi.org/10.5281/ZENODO.16870808
24. The Editors of Encyclopaedia Britannica. (2025, September 20). *Dharma | Hinduism, Buddhism, Karma | Britannica*.
25. https://www.britannica.com/topic/dharma-religious-concept