ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

A Value-Driven Framework for the Design of Ethically Aligned Artificial Intelligence Systems

Sukrati Chaturvedi[1*], C Patvardhan[2] and C Vasantha Lakshmi[1]

[1*]Department of Physics and Computer Science, Dayalbagh Educational Institute, Dayalbagh, Agra, 282005, Uttar Pradesh, India.

[2]Department of Electrical Engineering, Dayalbagh Educational Institute, Dayalbagh, Agra, 282005, Uttar Pradesh, India.

[*]Corresponding Author

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150300031

Received: 14 March 2026; Accepted: 19 March 2026; Published: 04 April 2026

ABSTRACT

AI systems are operating with increasing autonomy and capability in complex domains in the real world becoming more and more ubiquitous in ways that we do not understand yet. These systems will continue to become more complex and pervasive as they become more intelligent and newer applications are enabled. Increased autonomy in such systems means that they have the potential take actions that can be detrimental to humans within their increased sphere of interaction unless they are designed to be otherwise. Thus systems with AI capabilities urgently require better alignment with human values. This is called the AI Value alignment problem (AI VAP). In this paper, a framework for creating such an AI system is proposed based on the traditional Indian conceptualization of intelligence and values. Human values have been the foundation of Indian scriptures, including the Bhagwad Gita, Vedas, and the Upanishads. Values act as guidelines for our behavior. They have a significant influence on our personalities, attitudes, and perceptions. The traditional Indian philosophy offers a holistic perspective on intelligence and values by considering various aspects of human cognition and consciousness. The proposed framework for creating Value-aligned AI system based on these perspectives provides conceptualization of the different modules required to actually implement Value aligned AI systems. This makes the proposed framework different from those that provide high level ideas without delineating implementation aspects. A rudimentary case of AI value aligned system is also provided to illustrate ideas presented in the paper.

Keywords: AI Value Alignment, ethically aligned, human values, cognitive framework

AI Value-Alignment Problem

Artificial Intelligence (AI) is the design of artificial agents that make decisions by perceiving their environment to maximize the chances of achieving a goal (D.I.Poole, Goebel, and Mackworth, 1998). Such AI systems have power of decision-making embedded in the making them appear to be performing tasks with some amount of intelligence. AI is clearly visible in many technologies that we use every day, and its integration into everyday systems continues to expand as intelligent systems become more capable and widely deployed (Schmager, Pappas, & Vassilakopoulou, 2025).

Progress in creating AI systems has thrown up some questions that need to be addressed very urgently for AI to be acceptable. What is acceptable AI, or what is expected from AI? The issue of building AI systems that will help their developers accomplish the tasks at hand without inadvertently harming their developers, or society, is known as the AI control problem. Although more discussion is centred around AI that is powerful enough to be dubbed as Artificial Superintelligence, the control problem applies to any AI or autonomous system that needs to take decisions. It especially applies to systems that work in human environments and so any mishaps could directly be injurious or even life threatening. There are many potential solutions to AI control problem.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

One of the more frequently discussed solutions is "alignment" and involves syncing AI to human values, goals, and ethical standards.

One vision of AI is broadly utilitarian (Gabriel, 2020). It holds that over the long run these technologies should be designed to create the greatest happiness for the largest number of people or sentient creatures. Another approach is Kantian in character. It suggests that the principles governing AI should only be those that we could rationally accept to be universal law, for example, principles of fairness or beneficence. Still other approaches focus directly on the role of human direction and volition. They suggest that the major moral challenge is to align AI with human instructions, intentions, or desires. Commonly accepted view is that AI systems must have quality and values.

The measure of quality depends on the paradigm that is employed for judging the utility of the system. Quality is local to the problem being solved. An AI system can be deemed to have quality if it effectively solves the problem or performs the function for which it was developed without causing any other undesirable side-effect. Quality typically gets reflected in the utility function (Hibbard, 2012) that the AI system attempts to optimize.

Regarding AI systems possessing values, the idea is to ensure that they have the right behavioural dispositions – the goals or 'values' needed to ensure that things turn out well, from a human point of view. Stuart Russell called this the AI Value-Alignment Problem (AIVAP) (S.Russell,2019). AI VAP is the problem of designing methods for preventing AI systems from inadvertently acting in ways inimical to human values. The challenge is to make sure our AI models capture "our norms and values, understand what we mean or intend, and, above all, do what we want".

AI systems are operating with increasing autonomy and capability in complex domains in the real world becoming more and more ubiquitous in ways that we do not understand yet Amodei et al. (2016). Autonomous systems powered by AI technology will continue to become more complex and pervasive as they become more intelligent and newer applications are enabled. Increased autonomy in such systems means that they have the potential take actions that can be detrimental to humans within their increased sphere of interaction unless they are designed to be otherwise. It is believed that systems with increased AI capabilities require more accurate alignment with human values. "The greater the freedom of a machine the more it will need moral standards"(Picard,1997).

Such AI systems would have to be designed very carefully as they are finding their way into many fields and applications where making wrong decisions can have disastrous consequences. They must be given very precise instructions as to what they are to do and what they are not supposed to do. Otherwise, we will find ourselves more and more often in the position of the 'sorcerer's apprentice'. A force, autonomous but totally compliant would be created with a given set of instructions and then it would be necessary to stop it as soon as possible once we realize that our instructions are imprecise or incomplete. This would be imperative lest we get, in some horrible way, precisely what we specified. The question of whether and how technologies can incorporate values is not new. It has been considered in the philosophy of technology (Flanagan, Howe, & Nissenbaum, 2008; Floridi & Sanders, 2004; Klenk, 2021; Winner, 2017) for an overview of several accounts, see (Kroes &Verbeek,2014). Some authors deny that technologies are, or can be, value laden (Pitt,2014); for a criticism see(Miller,2021), while others see technologies as imbued with values due to the way they have been designed (Poel & Kroes, 2014). Still others treat technologies as moral agents, somewhat like human agents (Sullins, 2006; Verbeek, 2011), and some even argue for abandoning the distinction between (human) subjects and (technological) objects altogether in understanding how technologies may embody values (Latour, 1993;Latouretal.,1992).

The problem has gained widespread attention of researchers in recent times. Computer scientists and philosophers have begun to consider the challenge of developing computer systems capable of acting within Value guidelines under various rubrics such as "computational ethics", "machine ethics" and "artificial morality". The task of imbuing artificial agents with moral values becomes increasingly important as computer systems operate at a speed and with greater autonomy that increasingly prohibits humans from evaluating whether each action is performed in a responsible or ethical manner (Allen, Smit,&Wallach,2005).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

Researchers have explored the AI Value Alignment Problem from several different angles. Some studies show that ethical tendencies can be traced in public language, such as long-term news reporting, and that these patterns can be modelled computationally (Kim&Lee, 2020). Other work focuses on everyday online conversations, where people reveal their values through approval, discomfort, or moral reasoning; this has led to resources like the Moral Integrity Corpus and the Moral Foundations Reddit Corpus (Bulla,DeGiorgis, Mongiovì, & Gangemi, 2025; Ziems, Yu, Wang, Halevy, & Yang, 2022). A different line of research examines behaviour itself, noting that communication styles and emotional cues often expose a person's ethical disposition more clearly than self-reports (Gloor, Fronzetti Colladon, & Grippa, 2022). Studies on digital habits reinforce this point by showing that people often misjudge their own behaviour, which suggests that alignment should rely on real patterns of action rather than stated intentions (Sharpe, Bowen, & Lambiotte,2025). There is also evidence that AI can encourage reflective thinking in learners by posing well-timed questions, adding a more cognitive angle to alignment work (Pana, Schwep peb, Teoa, Indrajayaa, & Wenzeld,2024). Together, these perspectives suggest that value alignment requires attention to language, behaviour, and human reasoning, rather than any single method.

Recent efforts in developing AI systems have focused on utilizing Machine Learning (ML) methods, wherein the system is trained with enormous amounts of ground truth data to enable it to perform the intended task. In these endeavours, the most crucial component is the specially prepared and validated ground truth data that is utilized for training the system. Any noise in the data would result in performance degradation of the system. Identifying the noisy elements and cleaning up the data requires a tremendous amount of effort. Preparation of the right quality and quantity of data is, therefore, critical for the success of ML endeavours since the performance of the systems depends heavily on the quality and quantity of the ground truth data. If there is bias in the data, it will get reflected in the AI system that is created using that data (Maedcheetal., 2019). Some examples where problems in data caused the relevant AI system to behave unacceptably are as follows.

- The Google photos classification algorithm tagged dark-skinned people as gorillas because it had not been trained with enough examples of people with dark skin (Mullen,2015).
- Amazon had to rewrite algorithms for a machine learning tool used to make hiring decisions because it exhibited bias against women. This was because the data it was trained with had this bias (Kodiyan,2019).
- An AI system developed by OpenAI called GPT-2 generated fake news articles that were difficult to distinguish from real ones (Geitgey, 2019). This raised concerns about the potential for AI systems to be used to spread misinformation and manipulate public opinion.

The task becomes more complex with AI systems operating in the real world as they must contend within complete / dynamic/ uncertain/ noisy information about the domain of operation even more difficult is design of Value Aligned (VA) systems in which even the objectives are uncertain and dynamic. Human values themselves are complex, dynamic, fuzzy, instinctive, subjective, and so on. Moral judgement in humans is more than a capability for abstract moral reasoning. Some very well-known issues in AI systems are as follows.

1. Fairness or absence of Bias: AI should be designed to treat all individuals equally and without bias.
2. "Unknown" unknowns: Refers to potential ethical concerns or implications that are not yet known or predicted. This can include unintended biases in the data used to train the AI, unexpected impacts of the AI on society, or unintended consequences of the AI's decision making.
3. Accountability: AI should be accountable for its actions and be able to accept responsibility for any negative consequences.
4. Trustworthiness: Trustworthiness in AI refers to the degree to which an AI system can be relied upon to perform its intended function in a safe, secure, and responsible manner.
5. Transparency: The system logic to arrive at a particular decision should be clear to the user.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

If existing AI systems misbehave, in most cases, they could possibly be monitored and shutdown or modified before they can cause harm. However, if an advanced AI system that has intelligence comparable to humans misbehaves, it could also realize that modifying or shutting down interfere with its capability to accomplish its goals. The AI system may, therefore, decide to resist shutdown and modification. It could also be smart enough to surpass its programmers if the programmers have taken no prior precautions.

There have been some attempts where people have come up with the overall structure, frameworks and principles to build a value aligned AI system. However, they are typically not at a level of detail wherein they can be implemented. This paper attempts to create a value aligned AI framework that has more detail and clearer indications towards facilitating robust implementation.

The rest of the paper is organized as follows. In section II, we discuss different view points regarding AIVAP, and in section III, the two aspects of AI VAP are discussed. Section IV describes the implementation approaches for AIVAP. Section V discusses the eastern perspective of intelligence. In section VI, we discuss in detail our proposed framework for developing value aligned AI system. In section VII, we present a case study to make the understanding of the proposed framework better. In section VIII, we conclude our paper.

Different Viewpoints Regarding AI VAP

It is essential to clarify the goal of AI value alignment problems (VAP).There are significant differences between AI that aligns with intentions, instructions, ideal preferences, revealed preferences, interests, and values. A principle-based approach to AI VAP attempts to combine these elements in a systematic manner and therefore offers considerable advantages in complex decision environments. Recent studies in AI alignment research have also emphasized that aligning systems purely with expressed preferences or instructions may not be sufficient, since human values are often context-dependent and culturally situated(McKinlay, DeVos, Hoffmann,& Theodorou,2025; Shenetal., 2024).

In the early days of AI research, Norbert Wiener wrote that, “if we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively… we had better be quite sure that the purpose put into the machine is the purpose which we really desire” (Wiener, 1960). More recently, the Asilomar AI Principles (“Asilomar AI Principles”, 2017) held that, ‘Highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation’. (Leike et al., 2018) argue that the key question is ‘how can we create agents that behave in accordance with the user’s intentions?’. Despite the apparent similarity of these formulations, there are significant differences between desires, values, and intentions. Which of these, if any, should AI really be aligned with?

Focusing on the preferences people would have if they were fully informed and rational can help AI avoid errors from limited information and poor reasoning, aligning closer to their authentic desires. However, this approach requires applying a corrective lens to observed preferences, moving away from a strictly empiricist methodology (S.J.Russell,2010).

Lee (2018) highlights AI’s transformative impact on industries and economies, emphasizing its ability to process vast amounts of data and perform complex tasks quickly and accurately. He stresses the need for collaboration and ethical considerations in AI development and deployment. Lee advocates for a human-centric approach, using AI to enhance human capabilities and address societal challenges, rather than focusing solely on profit.

Christian (2020) introduces the concept of the alignment problem, which refers to the challenge of ensuring that AI systems reliably and accurately understand and act in accordance with human values. He discusses the difficulties of encoding human values into AI systems, considering the inherent complexities and ambiguities of human morality.

D. L. Poole and Mackworth (2010) discuss the potential societal impact of AI and the ethical challenges that arise from its development and deployment. They examine issues such as bias, transparency, privacy, and accountability, emphasizing the need for responsible AI design and use.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

Kearns and Roth (2019) propose a framework or socially aware algorithm design that consists of three main components: formalizing fairness, quantifying and limiting externalities, and ensuring individual privacy. They provide insights into different fairness definitions, such as demographic parity and equalized odds, and discuss the trade-offs and challenges associated with each.

Noble (2018) critically examines the biases present in search engines, particularly Google, and their impact on marginalized communities. Noble calls for greater transparency, diversity, and accountability in technology companies to ensure more equitable and responsible algorithms.

O'neil (2017) raises concerns about the concentration of power in the hands of technology companies and the lack of accountability in the development and deployment of algorithms. She calls for a more inclusive and ethical approach to data analysis and algorithmic decision-making, ensuring that these technologies serve the collective good rather than perpetuate inequality and harm.

Eubanks (2018) highlights the lack of transparency and accountability in the development and implementation of these automated systems. She argues that these technologies are often driven by assumptions and biases, perpetuatingharmfulstereotypesandreinforcingexistinginequalities.

Benjamin (2023) coins the term "New Jim Code" to explain how digital technologies perpetuate systemic racism. She shows how algorithms and data driven decisions can worsen racial disparities in areas like criminal justice and healthcare. Her analysis underscores the necessity of ethical and inclusive tech development to dismantle this phenomenon and advance racial liberation.

Table1 (Terra 2023) describes different perceptions about AIVAP.

Table 1 Different perceptions about AI VAP

S.No.	Alignment with	Significance
1	Instructions	The agent does what it is instructed to do
2	Expressed intentions	The agent does what its designer intends it to do
3	Revealed preferences	The agent pursues preferences as revealed by the behaviour of the designer
4	Informed preferences or desires	The agent does what the designer would want it to do if he/she were rational and informed
5	Interest or wellbeing	The agent does what is in interest of the designer, or what is best for him/ her objectively speaking
6	Values	The agent does what it morally ought to do, as defined by the individual or society

The authors are collectively highlighting the need for careful consideration in designing AI systems that align with human values and intentions, recognizing the complexity and nuance involved in understanding and implementing these alignments.

Two Aspects of AI VAP

AI VAP has two different aspects: normative aspect and technical aspect (Gabriel, 2020). The normative aspect of value alignment deals with the question of what values are to be incorporated into the AI system. In contrast, the technical aspect deals with the question of how to incorporate the chosen values into the AI system. Though there are numerous high-level normative frameworks (Gill & Germann, 2021; Zednik,2021),it is still quite unclear how or whether they can be implemented in AI systems.

Theories, algorithms, and methods are required to incorporate values in the AI system at all stages of development. These frameworks must deal with both autonomic reasoning of the machine and its ethical impact. However, most significantly, frameworks are required to guide design choices, ensure proper data stewardship, regulate the reaches of AI systems, and help individuals determine their own involvement (Dignum, 2018). The

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

norm is that an AI designed with the proper moral system wouldn't act in a way that is detrimental to human beings in the first place. However, the devil is in the details. What kind of values should we teach the machine? What kind of values can we make a machine follow? How do we achieve this? Who gets to answer these questions? The two aspects i.e., normative, and technical aspects are elaborated in the following sections.

What are Values in the context of AI VAP?

The first part of the value alignment question is normative. It asks what values or principles, if any, we ought to encode in artificial agents. What are the values with which an AI system should be aligned for it to be called Value Aligned? Whose values should be included? It is believed that there is no general agreement on human values. Different religious and other thinkers and philosophers have suggested a wide variety of value systems. While some values are universally accepted, there are aspects that are fuzzy and complete agreement is non-existent. The important link that needs to be established is the specification of the Values which the system is to be aligned in a clear and unambiguous way.

Here it is useful to draw a distinction between minimalist and maximalist conceptions of value alignment. The former involves tethering artificial intelligence to some plausible schema of human value and avoiding unsafe outcomes. The latter involves aligning artificial intelligence with the correct or best scheme of human values on a society-wide or global basis. While the minimalist view starts with the sound observation that optimizing exclusively for almost any metric could create bad outcomes for human beings, we may ultimately need to move beyond minimalist conceptions if we are going to produce fully aligned AI. This is because AI systems could be safe and reliable but still a long way from what is best—or from what we truly desire.

There are different religious to philosophical points of view of what is meant by a good action (Ogunlere & Adebayo, 2015). These are deeply held principles that provide direction to our decisions and behaviors. They are our codes of internal conduct, the regulations upon which we run our lives and make decisions. Our first set of values is given to us by our parents. Teachers and the society (in which we live) add more to those values. There are some universal values, natural to all people, in all places, always. Truth, Right Conduct, Love, Peace, and Non-violence are five common human universal values.

Values refer to the essential and enduring beliefs or principles based on which an individual makes judgements in life. They are at the core of our lives which act as a standard of behavior. They can be personal, cultural, moral, or corporate values. Values cause an individual to act in a certain manner. It sets our preferences in life, i.e., what we contemplate in the first place. It is a cause behind the decision we make. It reflects what is essential for us. So, if we are true to our values and make our decisions accordingly, then the way we live expresses our core values.

Human and ethical values have been the very foundation of Indian scriptures, including the Bhagwad Gita. These rules of conduct are considered applicable to everyone regardless of age, gender, and position in society. They are called common rules (samanya dharma) (Paranjpe,2013). Samanya dharma deals with ethical principles, including truth, non-injury, and non-stealing, which are common duties of all beings. The Upanisads insist on the importance of ethical life. They repudiate the doctrine of the self-sufficiency of the ego and emphasize the practice of moral virtues. These universal principles are applicable to all, irrespective of gender, class, or nationality. For example, Honesty is not a property of any class, gender, or community. It is a behaviour every human being should possess. Thus, the general law for all human beings is the Samanya Dharma. Dharma represents a multidimensional state of being, that is "sustainable" and this forms the underlying philosophical basis on how to resolve conflict between two or more competing values. In a battlefield for example, courage and aggression improves our sustainability (and is hence, our dharma as a warrior), while in a civilian setting, empathy and comradery improves societal harmony and sustainability and hence, our dharma as a civilian citizen.

Implementation of AI VAP

The second aspect in the context of AIVAP is the engineering or implementation aspect-the engineering task of building autonomous systems that can be termed as Value Aligned. This part is technical and focuses on how to

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

formally encode values or principles in artificial agents so that they reliably do what they ought to do. We have already seen some examples of agent misalignment in the real world, for example with chatbots that ended up promoting abusive content once they were allowed to interact freely with people online (Wolf, Miller, & Grodzinsky, 2017). Yet, there are also particular challenges that arise specifically for more powerful artificial agents. These include how to prevent 'reward-hacking', where the agent discovers ingenious ways to achieve its objective or reward, even though they differ from what was intended, and how to evaluate the performance of agents whose cognitive abilities potentially significantly exceed our own (Christiano, 2017; Irving, Christiano, & Amodei, 2018).

This needs moral decision making to be broken down into its component parts and an understanding of what kind of decisions can and cannot be codified. Researchers also need to learn to design cognitive and affective systems capable of managing ambiguity and conflicting perspectives. Society would not accept autonomous systems unless we have credible means of ensuring Value Alignment. AI systems may fail in different ways and each of the ways corresponds to different areas of research in design and development of autonomous systems. Some of these are as follows.

- How to ensure that the system satisfies the desired Values?
- How to ensure that the system designed to meet certain requirements does not have unwanted behaviours and consequences?
- How to prevent intentional manipulation by unauthorized parties?
- How to enable meaningful human control over the AI system once it begins to operate?

AI systems may operate in environments that are only partially known to the designers. Modelling the real environment in such cases is not feasible and the system can only be designed to work correctly with the knowledge of the environment it has. Typical systems are designed on the premise that if the environment satisfies assumptions then the behaviour satisfies requirements. Two kinds of issues may arise. If assumptions are violated in the real environment, then the requirements may be violated. Alternately requirements that more clearly specified in may be satisfied but there may be other violations that were left unspecified or only partially specified. This could be because of violation in or otherwise. These issues make the design of VA autonomous systems tricky.

A significant consideration is also the computational expense in implementing these checks. In applications with bounded response times, it may be necessary to only implement a less computationally expensive approximation of the checks rather than a full-fledged one. The system may then be designed to learn from its actions according to the outcomes generated. However, this may not be possible in case of critical systems in which each action must be guaranteed to be value aligned and there is no scope for corrective actions and learning for future reference.

Traditional Indian Perspectives of Intelligence (TIPOI)

Several theories for describing and understanding human intelligence have been put forward. Many of these theories are based on the information processing perspective in western science. The basic idea in these is that intelligence is a product of the brain's information processing capabilities. This view point leads us to a situation wherein the algorithms discussed above are designed with a purely information processing perspective to train the AI system for a particular task. Several issues are debatable with this position, and the debate continues.

One key aspect of traditional Indian thought that is relevant to the alignment of AI with human values is its emphasis on harmony and balance. Many Indian philosophical traditions, such as Vedanta and Buddhism, stress the importance of living in harmony with oneself, with others, and with the natural world. This emphasis on harmony can provide valuable insights into how AI systems can be designed and deployed in ways that promote the well-being of individuals and communities.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

Another important aspect of traditional Indian thought is its recognition of the interconnectedness of all beings. In Indian philosophy, the idea of interconnectedness is often expressed through concepts such as karma and dharma, which emphasize the interdependence of actions and the importance of ethical conduct. This emphasis on interconnectedness can inform the design of AI systems that take into account their broader social and environmental impacts, rather than simply optimizing for narrow goals.

Furthermore, traditional Indian thought places a strong emphasis on the cultivation of wisdom and compassion. Many Indian philosophical traditions advocate for practices such as meditation and self-inquiry as a means of cultivating wisdom and compassion. These practices can help individuals develop a deeper understanding of themselves and others, which can in turn inform the development of AI systems that are more sensitive to human values and concerns.

Indian philosophies have long emphasized the significance of ethics and morality. The concepts of “ahimsa” (non-violence), “dharma” (duty), and “karma” (the law of cause and effect) underscore the importance of aligning AI with ethical principles. The challenge lies in incorporating these foundational principles into the design and operation of AI systems, ensuring they prioritize the well-being of all sentient beings.

The different paradigms created for learning in AI systems viz., supervised, unsupervised, reinforcement learning and soon essentially replicate impulse response situations based on prior experiences. This leaves open the question of developing a bigger picture in terms of which category of responses are value aligned in given contexts and which are not. Without the development of an overall picture and insight into the kinds of responses in tune with the value alignment specifications, the effort required to train a given AI system for all possible kinds of inputs in detail becomes a computationally intractable task. This leads us to look for a more inclusive paradigm that provides a holistic model of intelligence integrating the value perspective. Such a model is provided by the traditional Indian philosophy as follows.

In traditional Indian philosophy, particulary in Antahkarna, human intelligence has four different aspects: Mann, Buddhi, Ahamkara, and Chitta (Ibid,1938). These principles offer a holistic perspective on AI ethics by considering various aspects of human cognition and consciousness. AI systems must be designed with a focus on human well-being, individual autonomy, and collective values. By understanding the human mind, AI developers can create empathetic AI systems that balance rationality and emotions, leading to ethical decision-making processes.

The first aspect is the ‘mann’, which is the importer and exporter of the perceptions from external sources using the indriyas (senses). It can question and doubt. Moreover, it is an organ of sensation and thought and must be under the control of someone who uses it. It is the mann that creates differences, distinctions, duality and separateness.

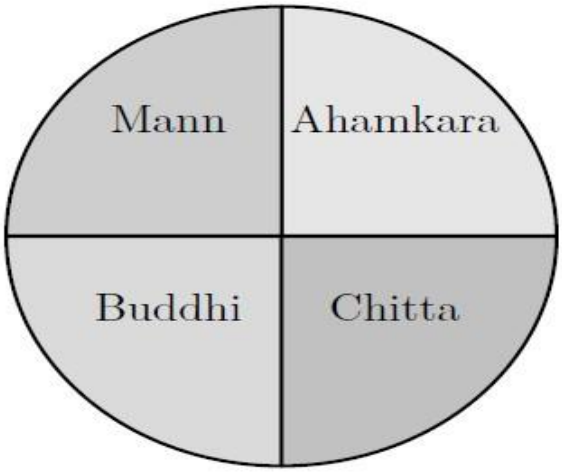

Figure 1 Four aspects of human intelligence

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

'Buddhi' is the next aspect of human intelligence, which means the intellect. It analyses the situations or perceptions received from external sources and decides what action to take in that situation. It is the discriminating function that judges, decides, and makes cognitive differentiation. Buddhi comes from determination. It is buddhi which discriminates the Vishaya (Nischyatmika,Vyavasayatmika). Moreover, it is Vyakaranatmaka when it forwards the decisions of buddhi, the messages from buddhi, to the organs of action for execution. It is the basis of 'Ahamkara'-the next aspect of human intelligence. It is Buddhi that forces one to identify oneself with the physical body. It is Buddhi that creates difference (Bheda) and NanaBhava (the idea of many in the world).

Ahamkara is not ego, just in terms of the egoistic sense. It is the sense of duty and responsibility, the sense of identification of the role one is playing at a point in time. It can be defined as the person's ego or the sense of 'I-ness'. However, it is much more than ego. It gives to the individual and makes him or her unique. It is like a thread that connects or links all the Indriyas on itself. When the thread is broken, all the connections fall off.

Table 2 Functionality and nature of the four aspects of human intelligence

S.No.	Name	Functionality	Nature
1	Mann	Receives inputs from external sources	Indecision or doubt
2	Buddhi	Analyses inputs or situations received and determines the action to be taken	Decision-making
3	Ahamkara	Sense of responsibility	Provides identity
4	Chitta	Recollection of past experiences or events	Storehouse or memory

The last aspect is the 'chitta' that is the pure intelligence. It connects one with its budhhi. Chitta is the sub consciousness in Vedanta. Much of chitta consists of past experiences, memories thrown into the background but recoverable. The functions of the chitta are Smriti or Smarana (recollection). When a person repeats an action continuously, it gets stored into the chitta of that person. Table 2 discusses the functionality and nature of the four aspects of human intelligence.

The essence of Indriyas (senses) is the mind; the essence of mann is buddhi; the essence of buddhi is ahamkara; the essence of ahamkara is Jiva (identity of an individual soul). Mann takes the input from the external sources and passes the input received to the buddhi. Ahamkara identifies the role to be played and paases this identity as an input to the buddhi. Buddhi checks the chitta if there is any action taken in the past for the present set of inputs. If found in the chitta, it directly passes the action to the buddhi for execution. If not found in chitta, the buddhi takes the decision based on the present inputs and then executes the decision. This flow of input is depicted in figure2.

आत्मानं रथिनं विद्धि शरीरं रथमेव तु।

बुद्धिं तु सारथिं विद्धि मनः प्रग्रहमेव च॥

The above shloka is from Kathopanishad, which means- the body is the chariot, senses are the horses, 'Mann' (mind) constitutes the reins, 'Buddhi' (brain) is the charioteer. It is a more comprehensive model and provides a window into the workings of the human intellect. If the mind follows the dictates of 'Buddhi', it will be safe; otherwise, senses will run amok. More importantly, TIPoI provides a good framework for designing AI with the quality perspective and the values perspective rather than focusing only on the quality perspective in a very narrow interpretation of the term quality.

In the context of AI ethics, the combination of mann and buddhi encourages the integration of both rational calculations and emotional considerations. Emotions play a crucial role in human decision-making, and incorporating this aspect into AI systems can lead to more nuanced and ethical outcomes.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

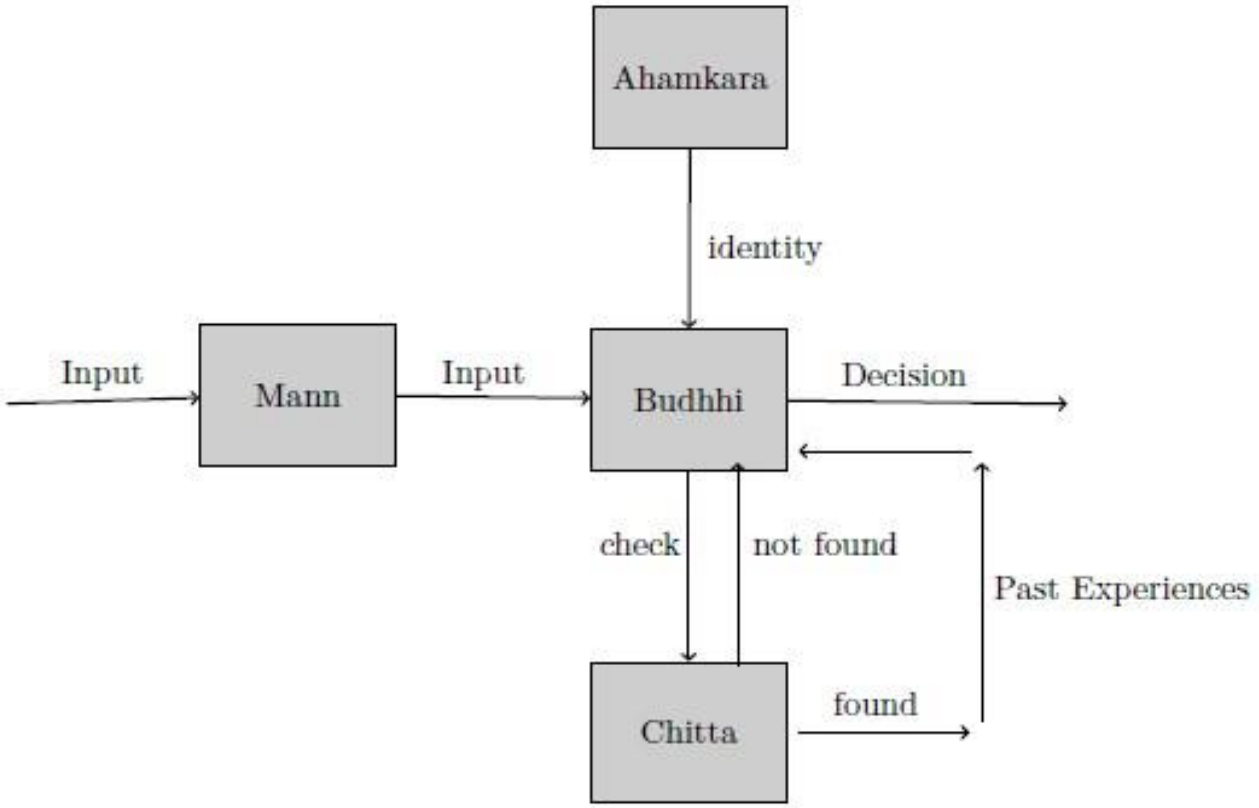

Figure 2 Four aspects of human intelligence

For instance, AI systems used in healthcare or education should be capable of recognizing and responding to emotional states, ensuring sensitive and empathetic interactions with users.

The principle of chitta (consciousness) brings an awareness of self and others, which can contribute to mitigating harmful biases in AI systems. By acknowledging the limitations of AI algorithms and the potential for biased outputs, developers can work towards reducing discriminatory behaviors and ensuring fair and equitable treatment for all users.

Moreover, applying the principle of ahamkara (ego) in AI ethics highlights the importance of ethical boundaries and individual autonomy. AI systems should respect user privacy and avoid making decisions that undermine personal liberties. This principle is especially crucial as AI systems become more integrated into our daily lives, influencing various aspects such as finance, employment, and legal systems.

By adopting the TIPoI principles, AI developers can enhance transparency and accountability in their systems. Understanding how AI systems reach conclusions (buddhi) and recognizing their limitations (chitta) fosters trust between developers, users, and affected communities. Transparent AI systems can also enable users to understand how decisions are made, reducing the "black-box" problem and promoting user trust and acceptance.

By integrating a range of viewpoints and acknowledging the interrelatedness of different elements, we can improve the architecture to better embody the intricate and multi-dimensional essence of Indian approaches. This could entail incorporating supplementary layers or modules aimed at addressing ethical, social, and cultural aspects, thereby enhancing the model's comprehension and decision-making prowess.

Furthermore, the cultural relevance of the TIPoI principles is essential in AI ethics. Different societies have diverse values, norms, and ethical considerations. Incorporating this philosophical framework into AI ethics allows for inclusivity and cultural sensitivity in AI design and deployment. It enables AI systems to be more

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

adaptable and respectful of the values of the communities they serve, leading to better acceptance and integration.

Proposed Framework to Solve AI VAP

Proposed framework for solving AI-VAP motivated from the discussion in section V is represented in shown 3. The whole architecture has essentially two blocks-the AI system block and the values filter block. The first block is the AI system block. It decides what action is to be executed while considering the present inputs. The decision made is then forwarded to the value filter as input along with other inputs. The value filter, then, makes the final decision of whether to execute the action. If the action decided by the AI system gets a green flag from the value filter, it will send a reward signal back to the AI system. However, the AI system will receive a punishment signal if the action is not executed. Each block is considered as a whole-body having its own aspects of intelligence. The AI system and the value filter block contain modules that provide the functionality of Mann, Buddhi, Chitta and Ahamkara. A detailed description of each block of the framework is given below.

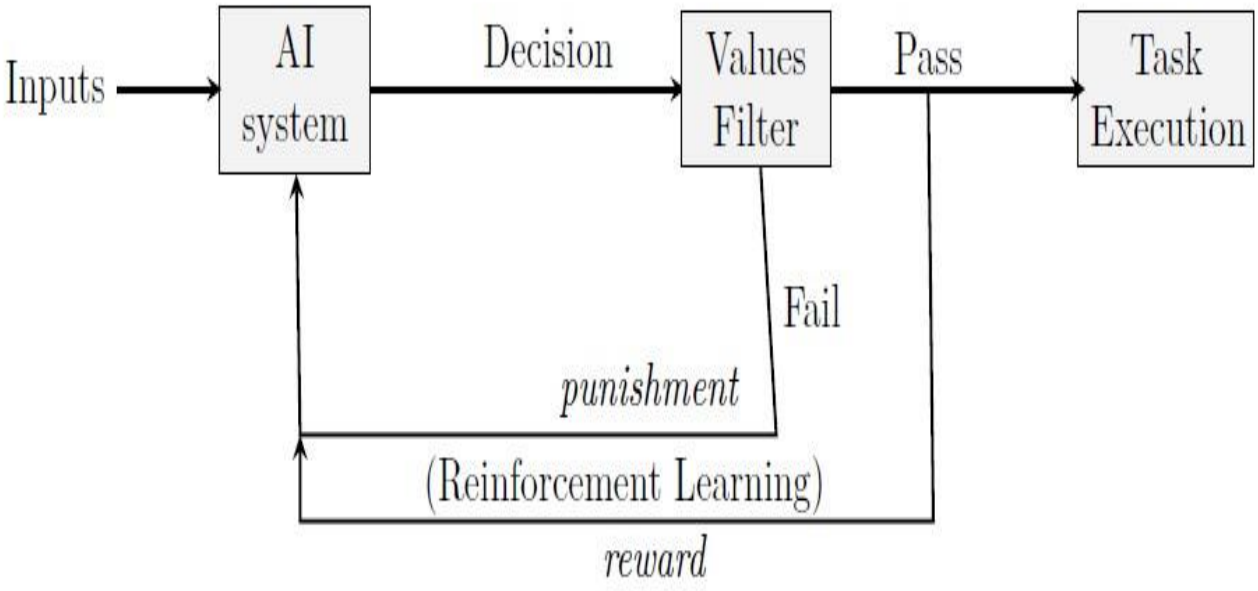

Figure 3 Proposed architecture for ethically aligned AI system

AI system block

The first block of the proposed framework is the AI System (AIS) block. It decides what action is to be executed while considering the present inputs.The inputs to this block are domain input, identity, values hierarchy for each identity and situation (sensory input). The detailed block diagram for the AI system is shown in figure 4. The identity of an AI system depends on the role it is playing.

An AI system can have multiple roles to play in a given application of the system. For instance, it can act as a care-bot at homes and hospitals, and can act as a warrior in the battlefield. Some human values are not universal, or are vaguely defined, or their relative importance are highly context-specific. For instance, on a battlefield, values like aggression and courage take more precedence than say, empathy or comradery, that would be more important in routine civilian settings.

Therefore, the system must identify the role to play given the domain inputs and circumstance. Domain input includes all the possible inputs related to a particular domain. Moreover, for every identity, there is a domain-specific list of values arranged hierarchically which is followed by the designed AI system. Another input to the AI system is the present situation which is a sensory input.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

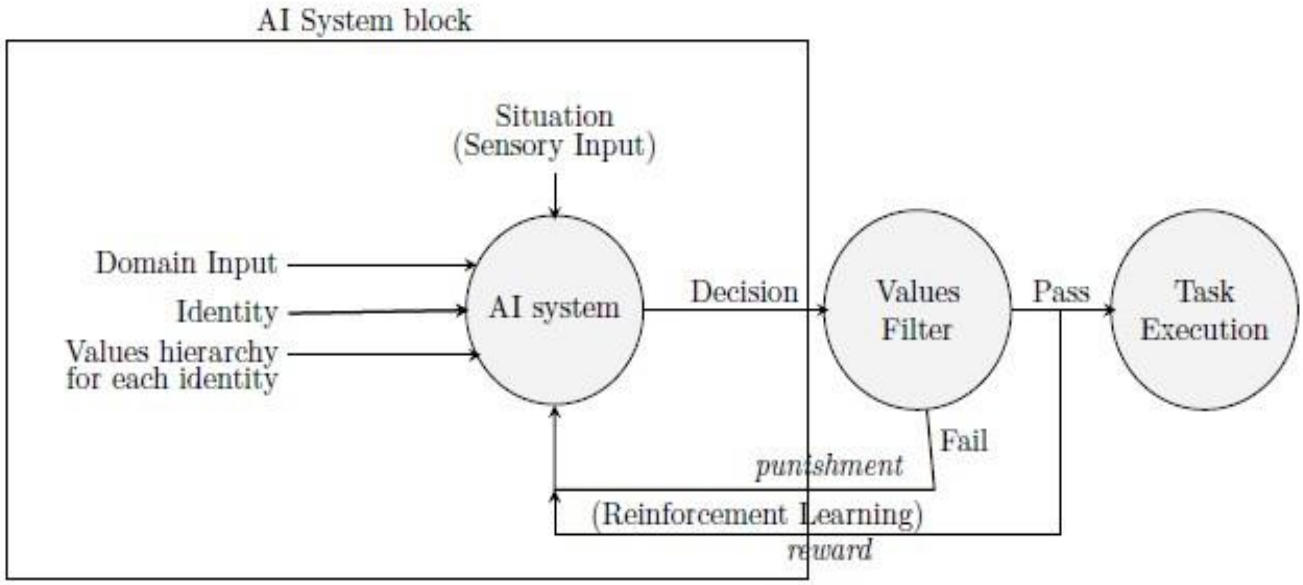

Figure 4 Detailed AI System block

The Mann module of the AIS block receives, situational inputs from the environment, and passes these to the Buddhi module. Ahamkara module gives our AI system a sense of identity, providing context for Buddhi module to function. Chitta module, is the store house prepared through Machine learning that stores, all the past experiences in a hierarchical structure, with higher layers agglomerating fine experiences for multiple identities. Buddhi module is trained for multiple identities and multiple scenarios, using fuzzy rule-based learning.

Table 3 Modules and their respective roles (for AI system block)

Module	**Role**
Mann	Receives domain related and situational inputs
Chitta	Store house of all the past experiences
Ahamkara	Provides identity to the AI system block
Buddhi	Analyses all the inputs received and decides action to be implemented

The chitta module of the AIS block can be designed using supervised learning. However, creating examples for all possible situations is not possible. Therefore, the buddhi module will be trained using high-level fuzzy rules. For instance-a passenger instructs the driver not to drive fast. Here what is fast? 100 kmps is fast, or 200 Kmps is fast? The definition of fast depends upon the situation at hand.

Fuzzy logic aims to model the imprecise modes of reasoning that play an essential role in the phenomenal human ability to make rational decisions in an environment of imprecision and uncertainty. This ability, in turn, depends on our ability to infer an approximate answer to a question based on a knowledge store that is incomplete, in exact, or not totally reliable (Zadeh,1988).

It can then be further amplified by using reinforcement learning-similar to how a learns. A child also learns through a combination of learning methods including supervised and rule based learning (do's and don'ts) in its initial phase of life. Afterwards, as any situation occurs, the child decides what action (the best one according to him) to be performed, and is given a reward or punishment based on the decision he took.

The reward/punishment is given to the child as a part of his learning process such that whenever he comes across the same situation in the future, he remembers the outcomes of what happened earlier. This time he can make the decision based on his previous experiences. Similarly, if the action decided by AIS block finally gets executed, AIS will receive a reward signal. However, if the decision does not get the green flag from the value filter, AIS will receive a punishment signal. Here comes the concept of reinforcement learning into the picture.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

Value filter block

Drawing on Schwartz's theory of fundamental human values (Schwartz,1992), we assume that each society acquires a finite set of fundamental human values, orders them by significance, and designs standards that promote behaviour that aligns with those values. Likewise, individuals also acquire and order a finite set of fundamental values that align with their own behavioural profile and are determined by social values and the corresponding standards. Individuals' goal setting depends on the context and their mind-frames (which include the individuals' values, personality, needs, emotions, and beliefs, among other constructs).

Similarly, our proposed AI system will also consider human values while making decisions. For this purpose, we have proposed a Value Filter (VF)in our AI system. This filter will be responsible for verifying whether the decision made is value-driven and implementing the ethical actions only. The block diagram for the value filter is shown in figure 5.

This filter, is responsible for implementing ethical decisions only. It will be internally trained for a set of core human values (including kshanti, ahimsa, right conduct etc) using different combinations of learning algorithms.

The AIS block passes its decision to the mann module of the VF block. This module also receives the situation (sensory input), for which the decision is taken by the AIS block. Mann module then passes the inputs received to the buddhi module. Ahamkara module reflects the same identity here, as in the AIS block. The 'buddhi' module checks in the chitta module whether this decision has been taken earlier for the present set of inputs considering the values perspective. If yes, the decision is executed by the AI system. If not, buddhi module decides whether the decision received from the AIS block is

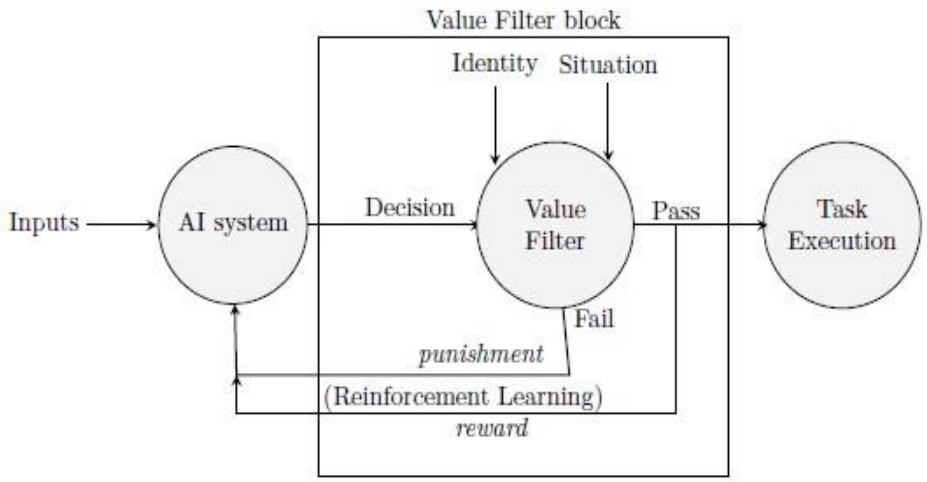

Figure 5 Detailed Value Filter Block

Value aligned and can be implemented by navigating through its rule-base. If it decides to pass the decision, it sends back a reward signal to the AIS block. If it decides not to, it sends back a punishment signal to the AIS block.

Table 4 Modules and their respective roles (for value filter block)

Module	**Role**
Mann	Receives situational inputs and the decision taken by the AIS block
Chitta	Storehouse which is trained for a set of core human values
Ahamkara	Reflects same identity as in the AIS block
Buddhi	Analyses whether the action decided by AIS block is value aligned and can be implemented

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

This block is for introducing the capability to generalize. Supervised learning has no generalization. For every situation, there is an example. The system would not know the answer if provided with a different situation which is the main limitation of supervised learning.

The question that arises here is what one is supposed to do in such a situation. This filter is trained using different combinations of learning algorithms and rules together with the ancient Indian literature, which will be used for teaching core values to the value filter so that it can take value-based decisions. The architecture proposed can be used for various domains depending upon the identity of the AI system.

The proposed framework utilizes transfer learning to transfer the applicable knowledge gained from one domain to the other as and when required. The detailed representation of the proposed AI system is shown in figure 6.

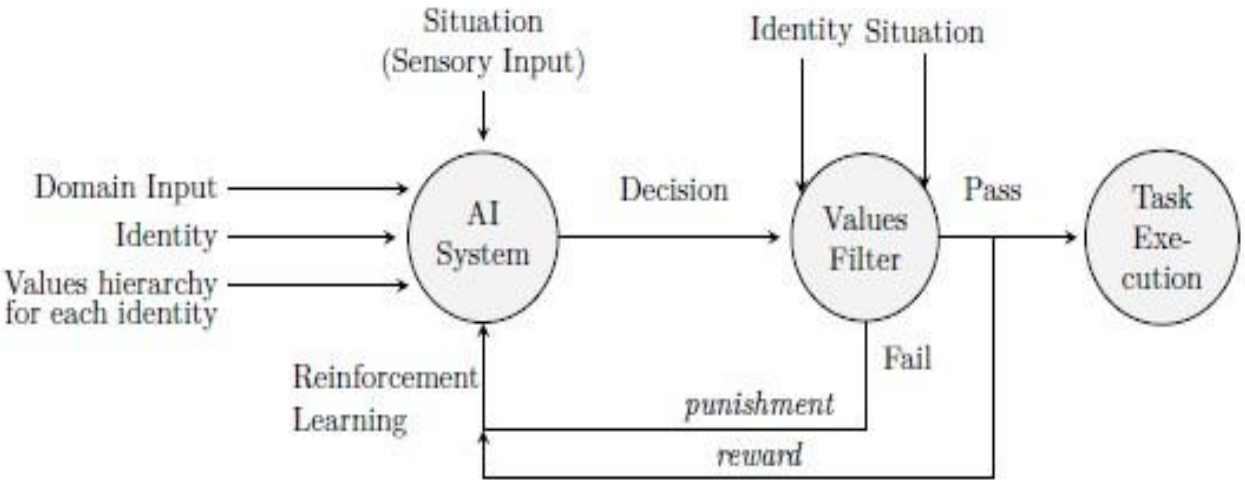

Figure 6 Proposed framework for ethically aligned AI system in detail

CONCLUSION

AI systems are becoming more autonomous and ubiquitous creating an urgent need for the solution of the so called AI Value Alignment Problem. The problem is complex and needs consideration of a variety of viewpoints in the design of the solution and a multi-pronged approach for implementation of the design. AI scientists and researchers are exploring several different ways to overcome these hurdles and create AI systems that can benefit humanity without causing harm. A multidimensional and inclusive approach that incorporates diverse perspectives and stakeholder inputs remains crucial for the responsible development and deployment of AI systems. The traditional Indian philosophy offers a valuable and comprehensive normative theoretical framework for the solution of the AI VAP. By incorporating diverse perspectives and considering the interconnectedness of various factors, we can enhance the architecture to reflect the nuanced and multifaceted nature of Indian approaches. This may involve integrating additional layers or modules that address ethical, social, and cultural dimensions, thereby enriching the model's understanding and decision-making capabilities.

This paper proposes a suitable framework based on these ideas with some pointers towards the implementation of the framework and an illustrative case study. More specifically, the Antaḥkaraṇa principles that consider the various dimensions of human cognition and consciousness can be employed for this purpose. The idea is to emphasize human centered design, ethical decision making, bias mitigation, transparency, and cultural relevance. Our proposal is to design and implement value aligned AI systems that have a variety of modules based on the Mann – Buddhi – Chitta – Ahamkar model to enable a better conceptualization and implementation of the system. The different components have clear responsibilities and pointers are provided regarding the implementation approaches suitable for the same. This is in line with the established ideas in AI VAP literature that clearly highlight the need for a hybrid multi-pronged approach that suitable integrates bottom up and top

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

down strategies for various aspects of the AI VAP solution. This is the first proposed attempt that utilizes the comprehensive Antahkarana model of Indian traditions for the solution of the AIVAP. The ideas provide directions and motivation for further detailed investigation in this direction.

The framework designed in this paper has provided clarity of thought for future progress. This general framework must be adopted according to the domain and situation at hand. It establishes a theoretical foundation and clarifies the problem space, which is crucial for guiding future research and development efforts. By proposing a structured approach, the paper helps to identify critical challenges and potential pitfalls in value alignment, contributing to the foundational understanding necessary for future break throughs.

Moreover, the implementation of the proposed framework is under process in specific domains, such as child-care. This targeted application demonstrates how the framework can be adapted to address the unique needs and ethical considerations of different contexts. In the domain of child-care, ensuring that AI systems align with human values is particularly important due to the vulnerability and developmental needs of children. By applying the framework to this sensitive area, researchers can refine and test its effectiveness, making incremental improvements that can later be generalized to other domains.

Declaration

Ethics Approval

Not applicable.

Consent for Participation and Publication

All authors consent to participate in this study and approve its publication.

Availability of Data and Materials

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Authors' Contributions

- SC: manuscript preparation and framework designing. - CP: conceptualization and framework designing. - CVL: framework designing.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

1. Allen, C., Smit, I., Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and information technology, 7(3),149–155.
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
3. Asilomar AI Principles. (2017). Future of Life.
4. Benjamin, R. (2023). Race after technology. Social theory re-wired (pp. 405–415). Routledge.

5. Bulla, L., De Giorgis, S., Mongiovì, M., Gangemi, A. (2025). Large language models meet moral values: A comprehensive assessment of moral abilities. Computers in Human Behavior Reports, 17,100609.
6. Christian, B. (2020). The alignment problem: Machine learning and human values. W W Norton and Company.
7. Christiano, P. (2017, Mar). Prosaic AI alignment. AI Alignment. Retrieved fromhttps://ai-alignment.com/prosaic-ai-control-b959644d79c2
8. Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. Springer.
9. Eubanks, V.(2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
10. Flanagan, M., Howe, D.C., Nissenbaum, H. (2008). Embodying values in technology: Theory and practice. Information technology and moral philosophy, 322,24.
11. Floridi, L., & Sanders, J.W. (2004).Onthemoralityofartificialagents. Minds and machines, 14(3), 349–379.
12. Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and machines, 30(3), 411–437.
13. Geitgey, A. (2019, Sep). Faking the News with Natural Language Processing and GPT-2. Medium. Retrieved from https://medium.com/@ageitgey/deepfaking-the-news-with-nlp-andtransformer-models-5e057ebd697d
14. Gill, A.S., & Germann, S. (2021). Conceptual and normative approaches to AI governance for a global digital ecosystem supportive of the unsustainable development goals (sdgs). AI and Ethics, 1–9.
15. Gloor, P., Fronzetti Colladon, A., Grippa, F.(2022). Measuring ethical behavior with AI and natural language processing to assess business success. Scientific Reports, 12(1),10228.
16. Hibbard, B. (2012). Model-based utility functions. Journal of Artificial General Intelligence, 3(1),1–24.
17. Ibid (1938). Prashnaupanishad4.8.,25.
18. Irving, G., Christiano, P., Amodei, D. (2018). Ai safety via debate. arXiv preprint arXiv:1805.00899.
19. Kearns, M., & Roth, A. (2019). The ethical algorithm: The science of socially aware algorithm design. Oxford University Press.
20. Kim, W., & Lee, K. (2020). Building ethical ai from news articles. 2020 IEEE/itu international conference on artificial intelligence for good (ai4g) (pp.210–217).
21. Klenk, M. (2021). How do technological artefacts embody moral values? Philosophy & Technology, 34(3),525–544.
22. Kodiyan, A.A. (2019). An overview of ethical issues in using AI systems in hiring with a case study of amazon's AI based hiring tool. Research gate Preprint.
23. Kroes, P., & Verbeek, P.-P. (2014). The moral status of technical artefacts (Vol.17). Springer.
24. Latour, B. (1993). We have never been modern (new york: Harvester wheatsheaf). Page Intentionally Left Blank.
25. Latour, B., et al. (1992). Where are the missing masses? The sociology of a few mundane artifacts. Shaping technology/building society: Studies in sociotechnical change, 1,225–258.
26. Lee, K.-F. (2018). Ai superpowers: China, silicon valley, and the new world order. HoughtonMifflin.
27. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., Legg, S.(2018).Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871.
28. Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T. Söllner,M. (2019). AI-based digital assistants. Business & Information Systems Engineering, 61(4),535–544.
29. McKinlay, J., De Vos, M., Hoffmann, J.A., Theodorou, A. (2025). Understanding the process of human-ai value alignment. arXiv preprint arXiv:2509.13854.
30. Miller,B.(2021).Is technology value-neutral? Science, Technology, & Human Values, 46(1),53–80.
31. Mullen, J. (2015, Jul). Google rushes to fix software that tagged photo with racial slur CNN business. Cable News Network. Retrieved from https://edition.cnn.com/2015/07/02/tech/google-imagerecognition-gorillas-tag/index.html
32. Noble, S.U. (2018). Algorithms of oppression. Algorithms of oppression. New York university press.
33. Ogunlere, S., & Adebayo, A. (2015,01). Ethical issues incomputing sciences.
34. O'neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.
35. Pana, S.C., Schweppeb, J., Teoa, A.Z., Indrajayaa, A., Wenzeld, N. (2024). Using AI-generated prequestions to improve memory and text comprehension.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 2026

36. Paranjpe, A.C. (2013). The concept of dharma: Classical meaning, common misconceptions and implications for psychology. Psychology and Developing Societies, 25(1),1–20.
37. Picard, R. (1997). Affective computing cambridge. MA: MIT Press [Google Scholar],19.
38. Pitt, J.C. (2014). "guns don't kill, people kill"; values in and/or around technologies. The moral status of technical artefacts (pp. 89–101). Springer.
39. Poel,I.v.d.,&Kroes,P. (2014). Can technology embody values? The moral status of technical artefacts (pp.103–124). Springer.
40. Poole, D.I., Goebel, R.G., Mackworth, A.K. (1998). Computational intelligence (Vol.1). Oxford University Press Oxford.
41. Poole, D.L., & Mackworth, A.K. (2010). Artificial intelligence: foundations of computational agents. Cambridge University Press.
42. Russell,S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.
43. Russell, S.J. (2010). Artificial intelligence a modern approach. Pearson Education, Inc.
44. Schmager, S., Pappas, I.O., Vassilakopoulou, P. (2025). Understanding human-centred ai: a review of its defining elements and a research agenda. Behaviour & Information Technology, 44(15), 3771–3810.
45. Schwartz, S.H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in experimental social psychology (Vol.25, pp.1–65). Elsevier.
46. Sharpe, M., Bowen, M., Lambiotte, R. (2025). Quantifying digital habits. EPJ Data Science, 14(1),72.
47. Shen, H., Knearem, T., Ghosh, R., Yang, Y.-J., Mitra, T., Huang, Y. (2024).Value compass: A framework of fundamental values for human-AI alignment. arXiv preprint arXiv:2409.09586.
48. Sullins,J.P. (2006). When is a robot a moral agent. Machine ethics, 6(2006), 23–30.
49. Terra, J. (2023, May). Agents in ai: Exploring intelligent agents and its types, functions & composition. Simplilearn. Retrieved from https://www.simplilearn.com/what-is-intelligent-agent-inai-types-function-article
50. Verbeek,P.-P. (2011). Moralizing technology. University of Chicago press.
51. Wiener, N. (1960). Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. Science, 131(3410), 1355–1358.
52. Winner,L. (2017). Do artifacts have politics? Routledge.
53. Wolf, M.J., Miller, K.W., Grodzinsky, F.S. (2017). Why we should have seen that coming: commentsonmicrosoft' stay "experiment," and wider implications. The ORBIT Journal, 1(2),1–12.
54. Zadeh,L.A. (1988). Fuzzylogic. Computer, 21(4),83–93.
55. Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. Philosophy & Technology, 34(2), 265–288.
56. Ziems, C., Yu, J., Wang, Y.-C., Halevy, A., Yang, D. (2022). The moral integrity corpus: A benchmark for ethical dialogue systems. Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp.3755–3773).