International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Human Values from Indian Philosophy to Solve AI Alignment Problem

Sukrati Chaturvedi[1], C Patvardhan[2], C Vasantha Lakshmi[3]

[1]Department of Physics and Computer Science, Dayalbagh Educational Institute, Agra, India
[2]Department of Electrical Engineering, Dayalbagh Educational Institute, Agra, India
[3]Department of Physics and Computer Science, Dayalbagh Educational Institute, Agra, India

Abstract
The swift progress of artificial intelligence (AI) has presented society with unparalleled opportunities and challenges. With the growing autonomy of AI systems, the critical concern of ensuring their alignment with human values and ethical principles has come to the forefront. The AI alignment problem refers to the challenge of designing AI systems that act in ways that are beneficial and aligned with human intentions and values. In this paper, we explore the potential contributions of human values from Indian philosophy in solving the AI alignment problem. We conclude that it is possible to establish and tailor a finite set of human values derived from Indian philosophy for the purpose of addressing the enduring challenges that AI systems are expected to tackle in their operational tasks.

Keywords: AI, value alignment problem, human values, Indian philosophy.

1 Introduction

Artificial Intelligence (AI) encompasses a wide-ranging domain that involves the capacity of computers and machines to execute tasks typically associated with intelligence. These tasks include pattern recognition, experiential learning, decision-making, and problem-solving. Various forms of AI exist, including narrow or weak AI, crafted for specific tasks, and general or strong AI, possessing the capability to perform any intellectual task achievable by a human.

These systems are being increasingly deployed for performing various tasks autonomously in different fields. Autonomy in these systems implies that they can make decisions independently. AI technologies are becoming more ubiquitous and are visible anytime, anywhere. These advancements will only increase in the future, and as autonomy increases, technology will become more intricate and widespread.

Early AI efforts required the painstaking efforts of the so-called knowledge engineers to capture the domain knowledge required to make the AI capable of operating autonomously in the given domain. These systems, however, were very brittle, and the approach lacked scalability as AI systems that could operate in wider or multiple domains were pursued. Recent technological advances in AI are based on machine learning strategies wherein the system is able to learn to perform the given tasks from the examples provided to it. This was seen very clearly in Alpha Go (Silver, Huang, Maddison, Guez, Sifre, Van Den Driessche, Schrittwieser, Antonoglou, Panneershelvam, Lanctot et al., 2016) and even more so in Alpha Go Zero (Silver, Schrittwieser, Simonyan, Antonoglou, Huang, Guez, Hubert, Baker, Lai, Bolton et al., 2017).

These endeavors demonstrate the tremendous scientific advances and provide strong evidence that even without human intervention, AI has the potential to attain performance levels equal to or surpassing that of humans in certain domains. Sophia, the humanoid robot developed by Hanson Robotics, has journeyed across the globe, offering insights on AI as well as engaging in discussions on various compelling topics such as organizational dynamics, societal concerns, and political issues (Parviainen and Coeckelbergh, 2021). A host of other recent developments continue these trends toward systems that exhibit a wider range of problem-solving and a larger degree of autonomy. The rapid developments in autonomous AI systems have raised concerns in several quarters that some care is required to ensure that the systems developed are not detrimental to human interests. The objective for AI systems should be to contribute to a better world by utilizing their capabilities to fulfill human needs and desires, always in alignment with human values (Yudkowsky, 2011). Although the question was raised more in the context of Artificial Superintelligent systems initially, AI researchers are now in agreement that any AI system that is designed to operate autonomously must be so designed that it adheres to human values.

The advancement and application of AI give rise to ethical and social considerations, such as the possibility of job displacement and the implications for privacy. Considering the paperclip maximizer scenario (Bostrom, 2014), The maximizer aiming to maximize the quantity of paperclips in the universe would similarly eliminate us, not from malice, but to align the atoms in our bodies with its primary objective. The ”paperclip maximizer” strives to utilize all available resources for an objective that many humans would not perceive as valuable.These concerns have led to ongoing debates about determining the suitable role of AI in society and the necessity for the responsible advancement and utilization of AI technologies. Therefore, to prevent such conditions from occurring in the future, it is essential to integrate human values into AI systems. However, embedding values into AI systems remains a challenge.

Furthermore, the focus has been on the output side to determine the behavior an AI system must exhibit. How to attain that behavior is not that clear. Intuitively one understands what each of these principles stands for. However, it is challenging to characterize them in a form that can be implemented in an AI system. The implication of the principle in consideration is only evident in the context of a particular application where the AI system is to function. The level of generality in the specification makes it challenging to specify the same for each application separately. This implies that the AI system has the specified values as an integral part that can not be dealt with by only specifying principles.

Instead of focusing only on a list of principles, we argue the necessity of linking both- AI principles and human values. This is to facilitate them into an AI framework by focusing on how they can interact and complement each other. In this paper, we attempt to find a set of human values to train AI systems to achieve the desired output behaviour. Moreover, we have also tried to map the way human learns to the machine learning methods in order to train the identified values to the AI systems.

This paper is structured as follows. In section II, we discuss what are values in the context of AI systems, followed by section III, in which we discuss the importance of values in the context of AI-human interaction. In section IV, we have made an attempt to identify a set of human values that we can use to achieve value-aligned AI systems. In section V, we discuss how we can train the AI systems for the set of identified values. The conclusion and future work are discussed in section VI.

2 What are Values?

The notion of human values is a comprehensive and universal concept. Human values refer to the beliefs and principles that influence human behavior and guide decision-making. In the context of AI systems,

human values refer to the ethical and moral considerations that should be considered when designing and implementing AI technology so that the system is value-aligned. These can be delineated in various ways, encompassing religious and philosophical perspectives on the definition of a virtuous action (Ogunlere and Adebayo, 2015). Moreover, they represent the things that are important or meaningful to a person or group and serve as a framework for how they live their lives. Values can be individualistic, such as honesty, integrity, and respect, or they can be collective, shared by a group like a company or a community. Values can influence an individual's or group's actions, choices, and relationships. They can shape how a person or group views the world and can serve as a source of motivation and direction. They are profoundly ingrained principles that dictate our decisions and behaviors. These are convictions held deeply within our hearts and form the internal ethical codes, guiding the principles upon which we lead our lives and make decisions.

Our initial set of values is imparted by our parents, and additional ones are instilled by teachers and the society we are part of. The values we embrace also hinge on our faith systems. Indian culture, in particular, plays a pivotal role in integrating ethical values (Karpavithra and Karvittal, 2017). AI systems are designed and developed by humans, and as such, they can reflect the values and biases of the people who create them. AI systems are trained on data sets that are generated and collected by humans, and these data sets can contain biases and prejudices that are present in society. Consequently, AI systems have the potential to perpetuate and magnify these biases and prejudices if they are not designed and implemented with care. The field of machine ethics has recently emerged within AI as a dedicated area of research, with a focus on guaranteeing the ethical conduct of artificial agents (Shulman, Jonsson and Tarleton, 2009). In developing AI algorithms, it is crucial to consider the overarching goals of society. Autonomous AI systems ought to align with the goals and practices rooted in human values. To ensure consistency with values related to human dignity, rights, opportunities, and cultural diversity, AI-based intelligence frameworks should be designed to incorporate human-like qualities (Paraman and Anamalah, 2022). AI researchers and practitioners must incorporate moral, societal, and legal values into the design of AI systems to facilitate the necessary technological advancements and responses (Dignum, 2017). In this paper we are mainly focused on training moral values to AI systems.

Developers and researchers should prioritize reflecting on the values and considering the ethical implications of AI systems and guaranteeing their design conforms to human values. This can involve taking steps to mitigate bias in data sets, and ensure that AI systems should exhibit transparency and accountability in their decision-making processes.

3 Importance of values in the context of AI-human interaction

As technology progresses, machines have the capacity to increasingly substitute human processes, procedures, and operations in real-life settings, including areas such as elderly care(e.g., care bots (Anderson, Anderson and Berenz, 2018; Vital, Couceiro, Rodrigues, Figueiredo and Ferreira, 2013)), health care, transportation (e.g., autonomous vehicles (Poczter and Jankovic, 2014)), human resources, and military applications (e.g., Autonomous Weapon Systems (AWS), drones) (Pflanzer, Traylor, Lyons, Dubljević and Nam, 2022). AI systems are utilized in challenging conditions where they are expected to navigate the complexities of human life, ensuring safety, accounting for human preferences and biases, and adapting to dynamic situations. At times, AI systems may also be employed in morally contentious situations.

AI systems are operating increasingly in environments where they are in direct contact with humans or

make decisions that impact human lives. According to certain AI researchers, as autonomous systems become increasingly widespread, it becomes essential to guarantee that these systems operate in a manner aligned with human values (Russell, Dewey and Tegmark, 2015; Soares and Fallenstein, 2014). The Value Alignment Problem (VAP) is thus to ensure that the objectives of AI systems match those of their human users as they gain autonomy and capability (Fisac, Gates, Hamrick, Liu, HadfieldMenell, Palaniappan, Malik, Sastry, Griffiths and Dragan, 2020). This is a complex and important issue because AI systems are designed and developed by humans, but they can sometimes act in ways that are unexpected or unintended and that may not be aligned with human values or interests. It is challenging to match AI systems with human values because our values are a complex web of preferences and subconscious impulses. Some researchers have expressed concern that future systems may engage in" reward hacking," where our preferences are only partially satisfied, similar to the King Midas story, where what was satisfied was what was said rather than what was meant if we do not clearly specify all of our values in a machine's value function.

The developers of AI frameworks encounter significant uncertainties in ensuring AI alignment with human values through design. Thus, they typically specify the value elements of the framework at a very high level. Thus, it is pretty unclear what, how, or even whether these elements can be implemented in AI systems. If the system is playing board games such as Chess or Go, adhering to the rules of the game is the bare requirement from a value point of view. However, in the case of a self-driving vehicle (Poczter and Jankovic, 2014) (Bimbraw, 2015), a more detailed requirement specification rather than just specifying a high-level goal is required even in the face of unknowns and, even, unknown unknowns.

A solution that may appear to solve a given problem as specified in the problem statement provided to the designer of the AI system may not pass the value test. The classic example is a problem statement that asks a driverless car to take the owner to the airport "as fast as possible". Taken literally, this could lead the AI system to completely violate traffic rules to achieve the goal specified. Along with the goal at hand, the values should be integrated into the AI system to ensure that it does not violate any established rules to make it an acceptable autonomous AI system. AI Safety researchers from various subareas collaborate to study safety on a global scale (Soares and Fallenstein, 2017) (Hadfield-Menell, Russell, Abbeel and Dragan, 2016).

In the absence of values, decisions taken by the AI the system may not be acceptable. It can have serious negative consequences and can undermine the responsible and ethical development and use of AI. It is important to consider the values and ethical implications of AI systems throughout the development process in order to ensure that they are aligned with human values and are used in a way that is responsible and ethical.

Toward this end, designing AI systems that clearly adhere to core human values is necessary. Three questions emerge in this context as follows (Chaturvedi, Patvardhan and Lakshmi, 2023).

1. What are the values that must be inculcated in the AI systems?
2. How can we guarantee that these values are ingrained in the current AI system, ensuring its proper implementation?
3. Once an AI system is implemented, how can one confirm the accuracy of its execution? Will this be a binary (Yes|No) answer or a range of possible values between 0, i.e., no alignment, and 1, i.e., perfect alignment?

The key challenge is determining which values to integrate and how to incorporate them into an AI system. Can humans agree on a set of universally accepted values across all cultures? These are, of course,

debatable. However, it is not difficult to arrive at a basic set of human values, a kind of Common Minimum Program. Assimov's Laws (Asimov, 1941) are a celebrated example of such an elementary Common Minimum Program.

4 What values to teach?

"In an interview, Stefano Ermon (*Beneficial AI 2017*, 2022) points out that a significant unresolved matter is articulating the precise nature of these values, considering that individuals may come from diverse regions, possess varying cultural perspectives, and have different socio-economic backgrounds. Roman Yampolskiy (*Roman Yampolskiy on the uncontrollability incomprehensibility and unexplainability of AI*, 2022; Han, Kelly, Nikou and Svee, 2021) highlights the difficulty of encoding human values in a programming language. He emphasizes that humans struggle to reach a consensus on shared values, and even when agreement occurs, these values can evolve over time.

It is difficult to identify a single set of values that would be considered "common" for all AI systems, as the values and ethical principles that are considered important for an AI system will depend on the specific goals and intended use of the AI system. For example, an AI system designed for use in a medical setting might prioritize values such as patient privacy and confidentiality, while an AI system designed for use in a military context might prioritize values such as proportionality and discrimination in the use of force.

Table 1: Prominent ethical principles identified in existing AI guidelines

S.No.	Ethical Principle	Importance in context of AI system
1	Accountability and Responsibility	AI must be held responsible for its actions and capable of acknowledging accountability for any adverse outcomes.
2	Transparency	The user should have clarity about the system's logic leading to a specific decision.
3	Non-maleficence	Within the realm of AI, this implies crafting and employing AI systems in a manner that reduces the likelihood of harm. This involves steering clear of biased algorithm creation and guaranteeing that AI systems refrain from making decisions that could lead to harm for individuals or society.
4	Beneficence	In the AI context, this entails crafting and utilizing AI systems to enhance the welfare and safety of individuals and society.
5	Justice and fairness	AI should be developed to treat every individual impartially and devoid of bias.

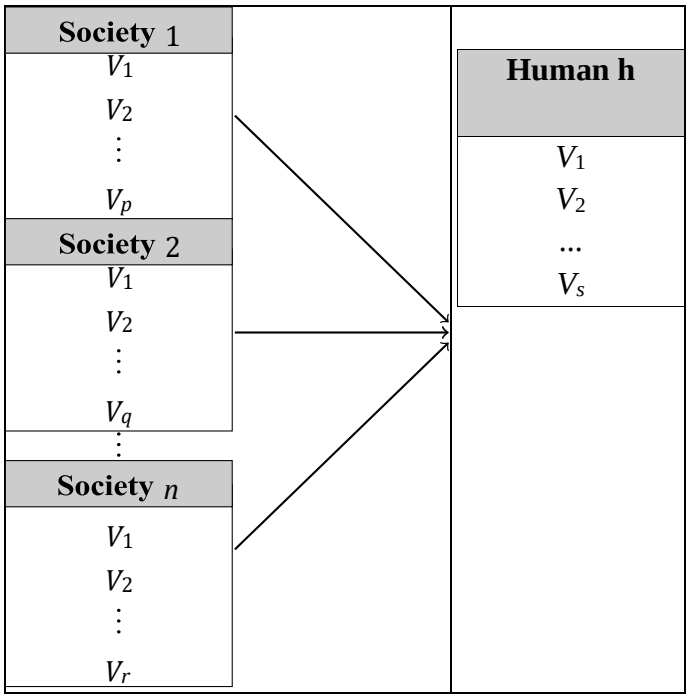
However, that being said, there are certain values and ethical principles that are often considered important for AI systems more broadly. Governments, regulatory authorities, industry, and academia have undertaken numerous initiatives to answer the first of the three questions listed in section 3 in the context of AI systems. The outcomes of these endeavors are commonly communicated through specific guidelines or ethical principles. A review of these lists and some survey papers in this context (Jobin, Ienca and Vayena, 2019; Khan, Badshah, Liang, Waseem, Khan, Ahmad, Fahmideh, Niazi and Akbar, 2022)

indicates that the majority of attention has mainly been directed towards a few fundamental principles. These are listed in table 1.

Consensus could lead to the identification of a set of values. Considerable convergence exists in human rights theory across African, Western, Islamic, and Chinese philosophies. A framework of values, incorporating principles such as ”all humans have the right not to be harmed, regardless of potential economic gain from harming them,” might be formulated and supported by individuals from diverse cultures.

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 norms that promote behavior that aligns with those values. Similarly, individuals acquire and order a finite set of fundamental values that align with their behavioral profile and are influenced by social values and the corresponding norms.

Figure 1: Depicting Schwartz’s theory of fundamental values



5 Values in Indian philosophical traditions

The bedrock of Indian scriptures, including the Bhagavad Gita, has been rooted in human and ethical values. These codes of behavior are deemed relevant to individuals of all ages, genders, and societal positions. They are referred to as common rules (samanya dharma) (Paranjpe, 2013). Samanya dharma encompasses ethical principles such as truth, non-injury, and non-stealing, which are considered universal duties for all beings. The Upanishads stress the significance of leading an ethical life, rejecting the notion of the self-sufficiency of the ego. They highlight the cultivation of moral virtues and assert that these universal principles apply to everyone, regardless of gender, class, or nationality. For example, Honesty is not a property of any class, gender, or community. It is a behavior every human being should possess. Thus, the general law for all human beings is the Samanya Dharma.

Various positive values can be identified among the numerous described sacred scriptures such as Bhagavat Gita, Upanishads, Thirukkural, Manu Smriti, etc. They are narrated below in the form of shlokas

(couplets) in Sanskrit followed by their English translations. While precise English translations may be unavailable for certain words, attempts are made to convey the closest possible meanings.

अमाि◌नत्वमदिम्भत्वमिहंसा क्षाि◌न्तराजर्वम।◌्
आचाय पासनं शौचं स्थैयर्मात्मिविनग्रहः।।**13.8**।।
ड्रिन्द्रयाथषु वैराग्यमनहङ्कार एव च।
जन्ममृत्युजराव्याि◌धदुःखदोषानुदशर्नम।।**13.9**।।◌्
असक् ि◌◌तरनिभष्वङ्गः पुत्रदारगहाि◌दषृ ◌ु।
ि◌नत्यं च समिच त्विम ◌ाि◌न ◌ोपपि षु।।**13.10**।।
मिय चानन्ययोगेन भक् ि◌◌तरव्यिभचा रणी।
ि◌विव देशसेि◌वत्वमरितजर्नससिद।।**13.11**।।◌ं
अध्यात्मज्ञानिनत्यत्वं त वज्ञानाथर्दशर्नम।◌् एतज्ज्ञानिमित प्रो मज्ञानं यदतोन्यथा।।**13.12**।।

Translation of the above shlokas in English is as follows.

Humility, lack of pride, nonviolence, tolerance, simplicity, seeking guidance from a genuine spiritual mentor, cleanliness, stability, and self-discipline; relinquishment of sensual indulgence, absence of false pride, awareness of the inherent suffering in birth, death, old age, and disease; detachment from familial ties, possessions, and maintaining composure in both favorable and unfavorable circumstances; unwavering and genuine devotion, inclination towards solitude, separation from the general populace; acknowledgment of the significance of self-realization, and the quest for philosophical comprehension of the Absolute Truth—all these attributes I designate as wisdom, and anything contrary to them is ignorance.

Some more shlokas from Indian scriptures emphasising on human values are naratted below.

अभयं स वसशं ◌ुि◌द्धः ज्ञानयोगव्यविस्थतः।
दानं दम यज्ञ स्वाध्यायस्तप आजर्वम।।**16.1**।।◌्
अिहंसा सत्यमक्रोधस्त्यागः शाि◌न्तरपैशुनम।◌्
दया भूतेष्वलोलुप्त्वं मादर्वं ह्रीरचापलम।।**16.2**।।◌्
तेजः क्षमा धितः शौचमद्रोहो नाि◌तमाि◌नता।◌ृ भविन्त सम्पदं दैवीमिभजातस्य भारत।।**16.3**।।

English translation of these shlokas is as follows.

“Fearlessness, purification of one’s existence, fostering spiritual knowledge, acts of charity, selfdiscipline, engagement in sacrificial rites, study of the Vedas, practice of austerity and simplicity; nonviolence, truthfulness, absence of anger; renunciation, tranquility, avoidance of faultfinding, compassion, and freedom from covetousness; gentleness, modesty, and unwavering determination; vigor, forgiveness, fortitude, cleanliness, absence of envy, and indifference to seeking honor—these divine attributes, O son of Bharata, characterize individuals endowed with a godly nature.”.

All the values listed may not be significant in specific AI system context because of its role. Teaching values to AI is a complex and challenging task, as values are subjective and can vary significantly among individuals and cultures. An attempt has been made to identify some of those values that need to be incorporated in the AI systems considering the above mentioned notions. These values are listed below.

1. Dharma: ”Dharma” is a term from Indian philosophies that is often translated into English as ”duty,” ”righteousness,” ”law,” ”morality,” or ”ethics.” It encompasses a complex and multifaceted concept that varies depending on the specific religious or philosophical context in which it is used. It provides an ethical framework for guiding human behavior and decisionmaking. Similarly, AI developers and policymakers must establish ethical principles to ensure that AI systems are designed and used responsibly, considering their impact on individuals, society, and the environment. It emphasizes fairness

and justice in societal relationships. In AI, ensuring fairness and avoiding bias is crucial, especially in algorithms used for decision-making in areas like healthcare, finance, and criminal justice.

धमर्चयार् धमर्सग्रहः

This famous shloka is from Tripitaka. It can be translated as- ”Ethical conduct is the essence of dharma.” This succinct verse highlights the centrality of ethical behavior in the concept of dharma.

धमर्चयार् धमर्सग्रहःअिहंसा सत्यमक्रोधस्त्यागः शाि न्तरपैशुनम।

दया भूतेष्वलोलुप्त्वं मादर्वं ह्रीरचापलम॥

This verse is from Bhagvad Gita and it lists several ethical values, emphasizing their importance in leading a virtuous life.

2. Ahimsa: This value emphasizes the importance of non-violence and compassion for all beings. The common interpretation of nonviolence is often limited to refraining from causing harm or destruction to the physical body. However, in essence, nonviolence entails avoiding causing distress to others. Generally, people are ensnared by ignorance within the materialistic understanding of life, leading to enduring material suffering.In the context of AI, this could lead to designing AI systems that are non-violent and do not cause harm to individuals or society by their action and inaction, and are able to make ethical and beneficial decisions.

अिहंसा परमो धमर्ः।

This famous shloka is from the Mahabharata, an ancient Indian epic. It emphasizes the importance of ahimsa as the highest moral and ethical duty.

3. Kshanti: It is a Sanskrit word that means peace or tranquility. Kshanti embodies a positive mindset, denoting ”accommodation” rather than painful resignation. The kshanti attitude reflects one’s cheerful and calm acceptance of the expectations or demands imposed by another person or situation, aligning with what he finds pleasing. He willingly and happily accommodate both situations and individuals. In the context of AI, kshanti could be interpreted as the value of promoting peace and harmony in society. This would lead to designing AI systems that are fair, transparent, and respectful of human rights, and that do not create conflicts or cause harm to individuals or groups. By following the value of kshanti, AI can be used in a way that benefits society and contributes to global peace and stability.

क्षाि न्त सवर् कमर्सु।

The above shloka empahasizes on the importance of kshanti and it means that kshanti is necessary in all actions.

4. Daya: It is a Sanskrit word that means compassion or empathy. In the context of AI, Daya could be interpreted as the value of showing compassion and empathy towards others. By following the value of Daya, AI can be used in a way that is ethical and beneficial to others. This would lead to designing AI systems that are sensitive to the needs and feelings of individuals, and that take care to avoid causing not only physical harm but also mental distress. It could also involve using AI to address social and environmental challenges, and to enhance the welfare and security of both individuals and society. Daya promotes a more cordial relationship wherein the AI system is tolerant of misdemeanors on the part of its users and also restricts its actions to those that do not disturb harmony.

दया तु परमो धमर्ः, सवर्भूतेषु कारुण्यम।

This shloka is from the Puranas, a collection of Hindu texts. It highlights the importance of daya as the highest moral duty and emphasizes the need to be kind and compassionate towards all living beings.

5. Kshama: This value emphasizes the importance of forgiveness and understanding. It is a Sanskrit word that means forgiveness or understanding. In the context of AI, Kshama could be interpreted as the

value of being forgiving and understanding in the use of AI. This would lead to designing AI systems that are able to learn from their mistakes and adapt to new situations, and that are able to take into account the complex and dynamic nature of human behavior. It could also involve being open to feedback and criticism and being willing to learn from others to improve AI's ethical and beneficial use. By following the value of Kshama, AI can be used in a way that is flexible and responsive to the needs of individuals and society.

क्षमा सव प रस्थानाम् सहृद्भु ूतानां च सवर्दा।

This shloka is from the Hitopadesha, a collection of fables and stories from ancient India. It emphasizes the universal importance of kshama in all situations and towards all people, including friends and loved ones.

6. Shoucham: Shoucham is a Sanskrit word that means purity or cleanliness at all levels including thoughts. In the context of AI, Soucham could be interpreted as the value of being pure and clean in the use of AI. This could lead to designing AI systems that are free from bias, discrimination, and that are able to make fair and transparent decisions. It could also involve ensuring that AI systems are secure, protect individuals' privacy, and are used sustainably and does not harm the environment.

स्वच्छतयैव परमा गितः॥

The source of the Sanskrit shloka is the Manusmriti,one of the ancient Hindu texts that contains laws and codes of conduct for individuals and society. It means "Cleanliness is the ultimate goal". It emphasizes the idea that cleanliness is not just a means to an end but a goal in itself, and that it is essential to living a fulfilling life.

7. Mardavam: Mardavam is a Sanskrit word that means gentleness or sensitivity. In the context of AI, Mardavam could be interpreted as the value of being sensitive and respectful towards others. By following the value of Mardavam, AI can be used in a way that is ethical and considerate of the needs and feelings of others. This could be achieved by designing AI systems that are aware of the potential impact of their actions on individuals and society, and that take care to avoid causing harm or distress.

मदर्वं परमं व्यवहारं ि व ि वदुः ि प्रयम् ॥

मदर्वं परमं व्यवहारं महत्वम् प्रदम् ॥

The shloka above emphasizes on the importance of mardavam. It means that the gentleness is highly valued by the wise all over the world, and gentleness is of great significance.

8. Dridha- nischayah: Dridh-nishchaya is a Sanskrit phrase that means strong determination or resolve. In the context of AI, Dridh-nishchaya could be interpreted as the value of being determined and persistent in pursuing ethical and beneficial goals. By following the value of Dridh-nishchaya, AI can be employed responsibly and in accordance with human values.This could involve setting clear ethical guidelines for the development and use of AI, and holding AI systems accountable for their actions.

ि नयः सवर्धमार्णां धम ऽधमर्ि ववजर्तः।

This couplet is from Mahabaharata and states that firm determination is the essence of all righteousness, and it is free from unrighteousness.

9. Amanithvam: Amanithvam is a Sanskrit word that means humility or modesty. In the context of AI, Amanithvam could be interpreted as the value of being humble and modest in the use of AI. This could involve recognizing the limitations of AI and being open to new ideas and perspectives. It could also involve being transparent about the capabilities and limitations of AI, and being willing to engage in dialogue with stakeholders to ensure that AI is used ethically and beneficially. By following the value of Amanithvam, AI can be used in a way that is respectful and considerate of others.

अमाि◌नत्वं परमं त वं सवर्भूतदयापरम।◌् सव पकार शुद्धात्मा क्षेत्रज्ञः स च पाण्डवः॥

This Sanskrit shloka is from Mahabharata and it emphasizes the importance of amanithvam as the highest truth, the essence of compassion towards all beings, and the quality of the pure soul that is dedicated to serving others. It means "humility is the highest truth, it is the essence of compassion towards all beings, it is the pure soul that is dedicated to serving others, and it is the quality of the righteous Pandavas." Ahimsa has a deeper connotation that goes much beyond not doing bodily harm to include taking care that the system does not do anything that causes mental agony or anguish. Similarly, kshanti is also concerned with maintaining mental peace and tranquility. In order to ensure that the system follows the principles of ahimsa and kshanti, it must show daya or mercy Table 2: Significance of identified human values in AI system

S.No.	Identified Human Value	Significance in AI System
1	Dharma (Duty)	To achieve justice and fairness
2	Ahimsa (Non-violence)	To achieve non-maleficence
3	Kshanti (Tolerance/peace)	To achieve beneficence and non-maleficence
4	Daya (Compassion/ Mercy)	To achieve beneficence
5	Kshama (Forgiveness)	To achieve beneficence
6	Soucham (Internal and external purity/ cleanliness)	To achieve transparency
7	Mardavam (Gentleness)	To achieve beneficence and non-maleficence
8	Dridhanischayah (Determination)	To achieve autonomy and accountability
9	Amanithvam (Humility)	To achieve beneficence

in its dealings. Daya ensures that even in the face of provocation, the system does not react in a way that goes counter to the principles of ahimsa and kshanti.

To ensure that AI is beneficial and ethical, it is essential to develop and implement mechanisms for teaching human values to AI systems. This can be achieved through a combination of technical solutions, such as incorporating ethical principles into the design of AI systems, and social approaches, such as engaging in dialogue with stakeholders to ensure that AI aligns with human values. By incorporating human values into AI, we can ensure that AI systems act in ways that are fair, transparent, accountable, and respectful of human rights.

To achieve the desired output behaviours mentioned in table 1, we can train the AI system using identified values (for reference, see table 2). For instance, to achieve non-maleficence in an AI system, it can be trained for values including ahimsa, kshanti,daya, dharma. Moreover, some of the desired output behaviors of AI systems, such as non-maleficence, beneficence can be achieved by training the system using a rule-based learning algorithm.

6 How to teach values?

The methodology to implement a Value Aligned AI system remains a critical problem. This is the technical part of the problem as distinct from the normative part that deals with the choice of the Values that need

to be incorporated for the system to be accepted as Value Aligned. Two distinct approaches are prevalent in the literature. One is the top-down or the rule-based approach and the other is the bottom-up approach wherein the system is trained using examples.

There are different methods to inculcate values into human beings, including teaching through moral stories, do's and don'ts, and rewards/punishments. For every method, we have a respective machine learning algorithm or rule-based method to train AI systems. It is well-known that these methods are used to teach values to humans from the very beginning. Similarly, learning methods can be used to train AI systems for human values. If values cannot easily be enumerated by human programmers, they can be learned (Harrison and Riedl, 2016). Some algorithms are discussed here to make the relationship between human learning methods and machine learning methods more clear.

6.1 Supervised Learning

One popular learning method for teaching wrong or right to children is telling moral stories (Kulkarni, 2013) or giving examples for different situations. Similarly, the supervised learning method can be used to teach values to the AI system through moral stories (Harrison and Riedl, 2016) (Nahian, Frazier, Riedl and Harrison, 2020) and examples (Allen, Smit and Wallach, 2005) (Balakrishnan, Bouneffouf, Mattei and Rossi, 2019) (Balakrishnan, Bouneffouf, Mattei and Rossi, 2018).

In the context of teaching values to AI, supervised learning could involve providing the AI with a large dataset of examples of ethical or moral situations, along with the appropriate values-based response. The AI would then be trained to predict the correct response for a given situation. For example, the AI might be trained with examples of situations where Daya is the applicable value and then be asked to predict whether Daya is the correct response reflecting Daya in a new situation. Through repeated training and evaluation, the AI could learn to make values-based decisions in a more accurate and consistent manner. However, the effectiveness of using supervised learning to teach values to AI will depend on the quality and diversity of the training data, as well as the specific goals and objectives of the training.

6.2 Unsupervised Learning

Human beings learn from their past experiences. To decide whether to perform a particular action in a situation, they would recall what had happened when 'this' action was performed earlier in the same situation. The same methodology can be applied to machines. Unsupervised learning algorithms acquire certain features from the data. Upon the introduction of new data, these algorithms employ previously acquired features to classify the data (Fahle, Prinz and Kuhlenkötter, 2020).

Unsupervised learning could involve providing AI system with a large dataset of examples of ethical or moral situations without any labels or guidance on the correct values-based response. The AI system would then be tasked with identifying common patterns and trends in the data and using similarity to these patterns to make values-based decisions in new situations. However, unsupervised learning may be less effective than supervised learning for teaching values to AI, as it relies on the AI's ability to identify patterns in the data on its own, which can be complicated and unreliable.

6.3 Reinforcement Learning

A human being is rewarded, in one way or the other, for each right action he/she takes. However, they are punished if the action taken is wrong. It is a greedy approach. Humans will always want to maximize the rewards and, therefore, will always try to take the right action.Within a biological framework, rewards can be seen as similar to the sensations of pleasure or pain (Sutton and Barto, 2017). Similarly, RL can be used to teach values to the AI system using rewards and punishment (Noothigattu, Bouneffouf, Mattei,

Chandra, Madan, Varshney, Campbell, Singh and Rossi, 2019) (Rodriguez-Soto, Lopez-Sanchez and Rodriguez-Aguilar, 2021).

The primary objective of reinforcement learning (RL) is to optimize rewards through interactions with an environment, involving the execution of various actions and encountering both failures and successes (Kaelbling, Littman and Moore, 1996). The agent is not directed to follow a pre-defined action. Reinforcement learning mirrors the natural learning process in which there is no teacher or guide present, and the learner acquires knowledge through the trial-and-error method.

6.4 Rule-based Learning

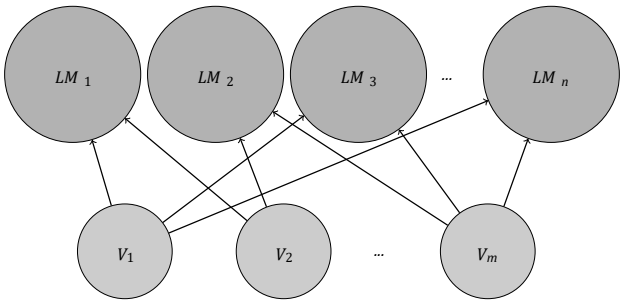
One way of incorporating Values is to put in place a rule base that essentially provides the dos and don'ts to the system. Such an approach can be utilized when the role of the AI system is in a limited domain and it is possible to encapsulate all the dos and don'ts of its activities in the form of the rules specified apriori. However, the number of such rules could become unmanageable if the variety of inputs and the corresponding possible responses or activities that the AI system caters to is larger. In such situations it becomes very difficult to provide any guarantee that the rules are both consistent and complete. The priorities among the values change with changing roles that the system has to play and these changing priorities have to be adequately and consistently reflected in the rules that are incorporated for the Value Alignment task. Ensuring that all these requirements are fulfilled necessitates larger number of rules. On the flip side, it might also become computationally expensive to sift through all the rules to determine the applicability of each rule and the relative priority in case multiple rules are applicable especially if the number of rules is large. These two conflicting requirements make implementing the Value Alignment by specifying rules a difficult task in domains that include multiple roles taken up by the AI system.

Many researchers have proposed that it is more feasible to train a system to follow a Value Aligned behaviour rather than try and incorporate values in the form of rules externally. Alan Turing (Turing, 1950) had proposed this idea in a seminal paper way back in 1950 in the context of creating AI systems. He proposed that rather than attempting to create an adult AI system it could be better to create a child AI system that is then trained through learning algorithms to perform the necessary functions.

Much the same notion is also valid for the Value Alignment Problem. As discussed above several varieties of learning algorithms are available including Supervised learning, Unsupervised learning, Reinforcement learning, Inverse reinforcement learning, Imitation learning and so on. For any role, even a human is trained using all the above approaches as appropriate.

Similar to this when we wish to develop the AI system we have to inculcate all these values into the system. However the prioritization of these values is different for each sysem and the corresponding role and the present situation. For instance, When an AI system is placed in a war zone, its highest priority will be dridha-nischaya (determination) to save its country from opponents. Whereas, if the same or different AI system is placed in a health care environment, where it is supposed to take care of the elderly, its highest priority will be mardavam (gentleness) ,daya (compassion), ahimsa (non-violence)and amanithvam (humility), all depending on the present situation. For each system and each corresponding role, we have to create a learning approach and select one of them.

A human learns each value through a single or a combination of learning methods. Say, If a human learns a value V_i through a learning method HM_j or a combination of learning methods HM_1,HM_2 …HM_j, then the AI system will learn the same value through the respective machine learning methods LM_1,LM_2 …LM_j .



Supervised learning essentially translates into learning through examples. One prominent way in which children have been taught Values over the ages is through stories. In all cultures stories are available that illustrate the importance of some value and the child can easily relate the practical situation that it faces with that in the story. Thus, learning is facilitated. This can be done for each of the Values identified.

In the real world, the different situations that the AI system may need to respond to may form a continuum rather than be a limited set of discrete situations. Unsupervised learning comes into play in these situations. The system may be trained to identify the closest situation it has seen earlier and the response it was trained for either through the Supervised learning mode or by means of response rule and follow that procedure.

In many situations creating enough examples that enable discrimination and response to any situation that the AI system faces becomes a difficult task. Reinforcement learning has been suggested wherein the system is trained to maximize a certain reward function when selecting one out the possible responses. This is applicable when the reward function can explicitly be designed to fulfil the mandated goals.

Sometimes, the reward function is difficult to specify and implement. Inverse reinforcement learning has been suggested wherein the system learns the reward function itself by observation of the required variety of stimulus – response situations. This has been shown to be very useful in situations like driving where the driver responds instinctively through experience to any given situation but might find it hard to specify the exact rule or reason why a particular action was preferred over the other available options.

In some roles it is not possible to learn by trial and error in actual field trials. The only recourse available then is to create simulation environments wherein the system can operate in the simulated environment and pick up the relevant experience for training. Alternatively, training videos may be created to enable Imitation learning for the particular role.

The above discussion leads to the notion of a hybrid approach to the implementation of Value

Alignment in an AI system where the different elements like rule-base and the learning approaches are judiciously combined on a where applicable basis to incorporate each of the Values that are relevant to the roles that the AI system is required to play.

Human values inculcation is not any different than any training programme for human candidates to be prepared for a given role. Just as a training programme includes various components viz, lectures, demonstrations, practice, correction and evaluation with feedback. Role playing determines the various acivities and situations that the AI system has to face in each role.

Considering the case of child care AI system, the relevant roles may include being a teacher, guardian, friend. The values that are relevant for each of the roles could be as follows.

1. **Teacher**: Dridhnishchayah, Mardavam, Ahimsa, Soucham, Kshama
2. **Guardian**: Mardavam, Soucham, Amanithvam, Kshama, Ahimsa, Daya, Kshama
3. **Friend**: Kshanti, Ahimsa, Kshama, Humility

Let the set of actions that are supposed to be performed by the child care AI system include:

1. To remind the homework to the child: One of the roles to be played by the child care AI system could be of a teacher which needs to remind and the child to finish its home work. System is supposed to be trained in a way that it remains gentle with the child even if the child denies to finish the homework. To teach gentleness, the system can be trained using examples based on situations (i.e, suervised learning) and can be given a reward signal whenever it treats the child with gentleness and a punishment signal whenever it tries to be harsh with the child (i.e., reinforcement learning).
2. To engage child with brain development games: Another role could be of a friend or a companion. If the child and the system play some game and the child tends to win, system should not loose its calmness. It is supposed to maintain peace. To teach peace, system can be trained using situation based examples (i.e, suervised learning), reward/punishment signal (i.e., reinforcement learning) or rule-base.

Ultimately, the most effective learning method for teaching a selected value to an AI system will depend on the specific task, available references, and the capabilities of the AI system.

7 Conclusion and future research

AI technologies give rise to many ethical issues as the autonomous intelligent systems are socially and culturally embedded (Kitchin and Dodge, 2014). This paper identifies the human values that are to be incorporated in AI systems.

Indian scriptures, such as the Vedas and the Bhagavad Gita, contain a wealth of moral and ethical principles that could potentially be used to guide the development of AI. However, the effectiveness of using Indian scriptures to teach values to AI would depend on the specific methods and approaches used, as well as the goals and objectives of the training.

In conclusion, this paper advocates for a holistic approach to AI alignment that incorporates human values from Indian philosophy. By doing so, we aim to contribute to the development of AI technologies that prioritize human well-being, uphold ethical standards, and ultimately serve as responsible and beneficial tools for a harmonious coexistence between humans and AI systems.

This research can be extended by finding out the ways to implement these methodologies mathematically. Some pointers are provided in this work by mapping the core values to be incorporated. These are being pursued to develop a child care AI system that has the requisite values as identified in this work.

References

1. Allen, C., Smit, I. and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches, *Ethics and information technology* **7**(3): 149–155.
2. Anderson, M., Anderson, S. L. and Berenz, V. 2018. A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm, *Proceedings of the IEEE* **107**(3): 526–540.
3. Asimov, I. 1941. Three laws of robotics, *Asimov, I. Runaround* .
4. Balakrishnan, A., Bouneffouf, D., Mattei, N. and Rossi, F. 2018. Using contextual bandits with behavioral constraints for constrained online movie recommendation., *IJCAI*, pp. 5802–5804.
5. Balakrishnan, A., Bouneffouf, D., Mattei, N. and Rossi, F. 2019. Incorporating behavioral constraints in online ai systems, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3–11.

6. *Beneficial AI 2017* 2022.
7. **URL:** *https://futureoflife.org/event/bai-2017/*
8. Bimbraw, K. 2015. Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology, *2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, Vol. 1, IEEE, pp. 191–198.
9. Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press.
10. Chaturvedi, S., Patvardhan, C. and Lakshmi, C. V. 2023. Ai value alignment problem: The clear and present danger, *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, pp. 1–6.
11. Dignum, V. 2017. Responsible artificial intelligence: designing ai for human values.
12. Fahle, S., Prinz, C. and Kuhlenkötter, B. 2020. Systematic review on machine learning (ml) methods for manufacturing processes – identifying artificial intelligence (ai) methods for field application, *Procedia CIRP* **93**: 413–418. 53rd CIRP Conference on Manufacturing Systems 2020. **URL:** *https://www.sciencedirect.com/science/article/pii/S2212827120307435*
13. Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D.,
14. Sastry, S. S., Griffiths, T. L. and Dragan, A. D. 2020. Pragmatic-pedagogic value alignment, *Robotics Research*, Springer, pp. 49–57.
15. Hadfield-Menell, D., Russell, S. J., Abbeel, P. and Dragan, A. 2016. Cooperative inverse reinforcement learning, *Advances in neural information processing systems* **29**.
16. Han, S., Kelly, E., Nikou, S. and Svee, E.-O. 2021. Aligning artificial intelligence with human values:
17. reflections from a phenomenological perspective, *AI & SOCIETY* pp. 1–13.
18. Harrison, B. and Riedl, M. O. 2016. Learning from stories: using crowdsourced narratives to train virtual agents, *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*.
19. Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of ai ethics guidelines, *Nature Machine Intelligence* **1**(9): 389–399.
20. Kaelbling, L. P., Littman, M. L. and Moore, A. W. 1996. Reinforcement learning: A survey, *Journal of artificial intelligence research* **4**: 237–285.
21. Karpavithra, S. and Karvittal, S. 2017. The role of indian ethics and values, *International Journal of Engineering and Management Research (IJEMR)* **7**(2): 560–569.
22. Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M. and Akbar, M. A. 2022. Ethics of ai: A systematic literature review of principles and challenges, *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022*, pp. 383–392.
23. Kitchin, R. and Dodge, M. 2014. *Code/space: Software and everyday life*, Mit Press.
24. Kulkarni, S. 2013. Panchatantra: An example of using narratives in teaching in ancient indian education, *Puheenvuoroja narratiivisuudesta opetuksessa ja oppimisessa* .
25. Nahian, M. S. A., Frazier, S., Riedl, M. and Harrison, B. 2020. Learning norms from stories: A prior for value aligned agents, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 124–130.
26. Noothigattu, R., Bouneffouf, D., Mattei, N., Chandra, R., Madan, P., Varshney, K. R., Campbell, M., Singh, M. and Rossi, F. 2019. Teaching ai agents ethical values using reinforcement learning and policy orchestration, *IBM Journal of Research and Development* **63**(4/5): 2:1–2:9.

27. Ogunlere, S. and Adebayo, A. 2015. Ethical issues in computing sciences.
28. Paraman, P. and Anamalah, S. 2022. Ethical artificial intelligence framework for a good ai society: principles, opportunities and perils, *AI & SOCIETY* pp. 1–17.
29. Paranjpe, A. C. 2013. The concept of dharma: Classical meaning, common misconceptions and implications for psychology, *Psychology and Developing Societies* **25**(1): 1–20.
30. Parviainen, J. and Coeckelbergh, M. 2021. The political choreography of the sophia robot: beyond robot rights and citizenship to political performances for the social robotics market, *AI & society* **36**(3): 715–724.
31. Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V. and Nam, C. S. 2022. Ethics in human–ai teaming: principles and perspectives, *AI and Ethics* pp. 1–19.
32. Poczter, S. L. and Jankovic, L. M. 2014. The google car: driving toward a better future?, *Journal of Business Case Studies (JBCS)* **10**(1): 7–14.
33. Rodriguez-Soto, M., Lopez-Sanchez, M. and Rodriguez-Aguilar, J. A. 2021. Multi-objective reinforcement learning for designing ethical environments., *IJCAI*, pp. 545–551.
34. *Roman Yampolskiy on the uncontrollability incomprehensibility and unexplainability of AI* 2022. **URL:** *https://futureoflife.org/podcast/roman-yampolskiy-on-the-uncontrollabilityincomprehensibility-and-unexplainability-of-ai/*
35. Russell, S., Dewey, D. and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence, *Ai Magazine* **36**(4): 105–114.
36. 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, Elsevier, pp. 1–65.
37. Shulman, C., Jonsson, H. and Tarleton, N. 2009. Machine ethics and superintelligence, *Reynolds and Cassinelli* pp. 95–97.
38. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. 2016. Mastering the game of go with deep neural networks and tree search, *nature* **529**(7587): 484.
39. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al. 2017. Mastering the game of go without human knowledge, *Nature* **550**(7676): 354.
40. Soares, N. and Fallenstein, B. 2014. Aligning superintelligence with human interests: A technical research agenda, *Machine Intelligence Research Institute (MIRI) technical report* **8**.
41. Soares, N. and Fallenstein, B. 2017. Agent foundations for aligning machine intelligence with human interests: a technical research agenda, *The Technological Singularity*, Springer, pp. 103–125.
42. Sutton, R. and Barto, A. 2017. Reinforcement learning: An introduction-complete draft.
43. Turing, A. M. 1950. Computing machinery and intelligence, *Mind* **59**(236): 433.
44. Vital, J. P., Couceiro, M. S., Rodrigues, N. M., Figueiredo, C. M. and Ferreira, N. M. 2013. Fostering the nao platform as an elderly care robot, *2013 IEEE 2nd international conference on serious games and applications for health (SeGAH)*, IEEE, pp. 1–5.
45. Yudkowsky, E. 2011. Complex value systems are required to realize valuable futures. the singularity institute, san francisco, ca.