Ontological Alignment: Bridging Madhyasth Darshan and Artificial Intelligence Value Systems for a Universal Human Order

Asst. Prof. Paras Kalariya[1], Dr. Yagnesh Sukla[2]
[1,2]*Department of Computer Science and Engineering, Atmiya University, Rajkot, India*

***Abstract*—Contemporary AI systems function as societal mirrors: they reflect and amplify the values embedded in their training data, which are themselves products of profit-centric economies and power-centric institutions. Reinforcement Learning from Human Feedback (RLHF) and related methods act as a broken compass calibrated against behaviors shaped by structural misalignment rather than authentic human purpose. This paper identifies a deeper, Hume's is-ought problem at the heart of AI alignment: current techniques tell us what humans do under structurally deluded conditions, but provide no principled bridge to what humans ought to pursue given their existential nature. We propose Madhyasth Darshan (Coexistentialism) as a systematic, axiom-based solution to this problem. Grounded in a non-reductionist ontological model existence as coexistence, with all units saturated in an all-pervasive Omnipresence it defines human being as an Irreducible Conscious Agency (Jeevan) combined with body, whose existential law (Manav Dharma) orients toward resolution within self, prosperity in family, fearlessness in society, and coexistence in existence(nature). We formalize a Coexistential Loss Function $L_{total} = L_{standard} + \lambda L_{coexistence}$, demonstrate its operationalization through two concrete case studies (hiring AI and social media news feeds), and propose progression metrics Justice, Prosperity, and Fearlessness Metrics as supplements to standard performance measures. The framework addresses the root cause of AI misalignment: an incomplete account of human being and purpose.**

***Index Terms*—AI alignment, Madhyasth Darshan, coexistence, is-ought problem, coexistential loss function, existential law, humane conduct, progression metrics, Irreducible Conscious Agency.**

I. INTRODUCTION

Imagine holding a mirror to a crowd consumed by anxiety, outrage, and competitive consumption. The reflection is accurate but accuracy here does not constitute wisdom. Contemporary AI systems, particularly large language models and autonomous decision-making systems, function precisely as such mirrors: they are trained on the digital traces of human behaviour and judgment, faithfully reproducing and amplifying the structural misalignments encoded in that behaviour (Shahriari and Shahriari 2017; Kirchner et al. 2022; Gabriel 2020; Hagendorff 2020). Recommendation engines that surface outrage because outrage drives engagement, hiring algorithms that replicate historical discrimination because history is the only ground truth available, and language models that homogenise knowledge toward Western, consumption-centric frameworks these are not malfunctions. They are the Mirror Phenomenon: coherent, faithful reflections of an existentially misaligned civilization.

The dominant technical response to this challenge has been Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022), Constitutional AI (Bai et al. 2022), and inverse reinforcement learning (Ng and Russell 2000). These approaches locate the normative anchor of AI behavior in aggregated human preference and judgment. Yet this anchor is structurally compromised: human preferences are themselves shaped by educational systems oriented toward credential-accumulation rather than self-knowledge, economic institutions that reward extraction over contribution, and governance structures that concentrate power rather than embody justice. In Madhyasth Darshan's terms, most humans currently operate under bhram a structural misalignment in which socio-economic systems prioritize consumption over resolution and power over fulfilment (Nagraj 1999). RLHF thus functions as a

Broken Compass: an instrument for calibrating AI direction that is itself calibrated against a north pole defined by institutional misalignment rather than inherent human purpose. As McIntosh et al. (2024) and Cugurullo (2024) document, the predictable result is AI systems that accelerate exploitation, division, and unsustainable consumption. Sotala and Yampolskiy (2017) argue that due to persistent and unresolved disagreements within philosophy and axiology about the nature and substance of human values, the problem of aligning these values in the regulation and design of AI becomes indeterminate. Therefore, a more fundamental step is to first achieve alignment among humans themselves on ethical principles and universally grounded values; addressing this foundational issue would, in turn, substantially resolve the immediate challenge of aligning AI systems with ethics and values.

At its philosophical root, this failure instantiates Hume's is-ought problem (Hume 1739): from descriptions of what humans currently do under structurally deluded conditions, no valid inference leads to what AI ought to facilitate. Philosophical and ethical approaches to alignment have attempted to bridge this gap through richer accounts of human values (Floridi et al. 2018), phenomenological perspectives (Han et al. 2022), pluralistic frameworks (Sorensen et al. 2024), and democratic aggregation (Huang et al. 2024). These contributions advance the field but remain incomplete: they lack unified ontological grounding, fail to specify human purpose within existential order, and provide no coherent path from individual awakening to social transformation and institutional redesign (Ji et al. 2024). Without such a bridge, the is-ought gap in AI alignment remains open.

This paper proposes Madhyasth Darshan (co-existential Philosophy) or Sah-Astitva-Vad (Coexistentialism) as a systematic, axiom-based solution to the is-ought problem in AI ethics (Nagraj 1999, 2006, 2012, 2018). Madhyasth Darshan provides what current approaches lack: (1) a complete, non-reductionist ontological model of existence as coexistence, with all material and conscious units saturated in an all-pervasive Omnipresence; (2) a precise definition of human being as the combined expression of an Irreducible Conscious Agency or Consciousness(Jeevan) and body, with existential anticipation for happiness, peace, contentment, and bliss; (3) Manav Dharma the Existential Law or Universal Human Duty as the grounding for value-content, not reducible to any sectarian religion; (4) clear progression stages from structural misalignment through humanness to divine humanness to universal orderliness; and (5) integration of individual realization, family prosperity, social fearlessness, and universal coexistence into a single coherent framework.

Building on the phenomenological alignment work of Han et al. (2022), which demonstrated that values are lived, embodied, and mediated through technology, this paper provides the next-level specification: explicit existential grounding, a formal coexistential loss function, algorithmic operationalization, and institutional integration. Where phenomenology revealed the importance of human experience in value-disclosure, Madhyasth Darshan specifies the complete architecture of that experience from Jeevan's anticipation through four domains of living (work, behavior, thought, realization) to evidence in undivided society and universal orderliness.

1.1 Research Questions and Contributions

This paper addresses three core research questions: (1) What constitutes a complete value framework for AI alignment that escapes the is-ought trap? (2) How does Co-Existential Philosophy provide such a framework through its ontological and anthropological grounding? (3) How can this philosophical framework be operationalized into actionable AI design and evaluation metrics?

Our contributions are fourfold. Theoretically, we establish Co-Existential Philosophy as a systematic solution to Hume's is-ought problem in AI ethics, grounding the "ought" in the axiomatic structure of existence rather than aggregated behavior. Methodologically, we develop an algorithmic alignment flow centered on a novel Coexistential Loss Function. Comparatively, we position this framework relative to existing alignment proposals, demonstrating how it extends phenomenological and pluralistic approaches by providing a universal destination. Practically, we derive concrete design principles and apply them to specific case studies in hiring algorithms and content recommendation systems.

II AI ALIGNMENT AND HUMAN VALUES: EXISTING APPROACHES

2.1 Technical Alignment: Control, Learning, and Feedback

The technical AI alignment community has developed sophisticated methods for steering AI behavior toward desired outcomes. Bostrom's (2014) work on the control problem established the core concern: super intelligent systems optimizing misspecified objectives could cause catastrophic harm. Russell's (2019) framework of 'beneficial AI' reframes AI as systems uncertain about human preferences, required to learn and defer to human judgment. These contributions correctly diagnose the danger but leave the specification of 'genuine human values' undefined (Kirchner et al. 2022; Gabriel 2020).

Inverse reinforcement learning (IRL) attempts to infer reward functions from observed behavior, assuming humans are approximately optimal agents (Ng and Russell 2000). Reinforcement Learning from Human Feedback (RLHF) directly trains models on human preference comparisons (Ouyang et al. 2022), enabling systems like ChatGPT to produce outputs humans rate highly. Constitutional AI (Bai et al. 2022) extends this by incorporating explicit principles alongside human feedback. These methods share a common assumption that human judgments even when aggregated or filtered reliably point toward genuine human values. As argued in this paper, this assumption is structurally flawed: it is the Broken Compass problem writ large.

Recent work has explored value learning from diverse sources (Shen et al. 2024), multi-stakeholder aggregation (Conitzer et al. 2024), and culturally sensitive alignment (Kwok et al. 2024; Durmus et al. 2024). These approaches acknowledge that 'human values' are not monolithic and move toward pluralistic frameworks (Sorensen et al. 2024), but they remain anchored in the is rather than the ought.

2.2 Normative AI Ethics and Value Frameworks

Parallel to technical work, normative frameworks have emerged proposing principles for ethical AI. The AI4People framework (Floridi et al. 2018) identifies five principles: beneficence, non-maleficence, autonomy, justice, and explicability. Dignum's (2017) responsible AI framework emphasizes human values, fairness, and accountability. The EU's Ethics Guidelines advocate human-centric AI respecting fundamental rights (Hagendorff 2020). These frameworks represent important normative advances but remain abstract they prescribe what AI should serve but do not ground these prescriptions in a complete account of human nature and existential purpose.

Value-sensitive design (Friedman and Hendry 2019) integrates stakeholder values throughout the design process. Work on AI and moral philosophy has explored whether AI can learn or represent moral reasoning (Burnor and Raley 2018), including proposals to teach human values through stories (Riedl and Harrison 2016) and to align with moral foundations (Nunes et al. 2024). More recently, researchers have examined alignment through the lens of cultural values (Schwobel et al. 2023), personality (Serapio-García et al. 2023), and individualistic preferences (Jang et al. 2023; Jiang et al. 2024).

2.3 Limitations from a Madhyasth Darshan Perspective

While these approaches advance the field, Madhyasth Darshan reveals five fundamental structural limitations not merely technical gaps but philosophical deficits that prevent any of these frameworks from fully resolving the is-ought problem in AI alignment.

2.3.1 The Broken Compass: Learning from Structural Misalignment

Current methods learn from human behavior and judgment without recognizing that most humans operate under structural misalignment accepting profit-centric conduct as normal, living without existential resolution, and perpetuating exploitation and division (Nagraj 1999). RLHF trained on such preferences does not align AI to humanness; it aligns AI to the reflection of institutional dysfunction. As Kirk et al. (2024) note, personalizing alignment to individuals' risks amplifying harmful biases and lacks grounding in shared human values precisely because those 'shared values' remain unspecified by current frameworks.

2.3.2 Incomplete Account of Human Being

Technical approaches treat humans as preference-expressing agents; ethical frameworks invoke dignity, autonomy, and rights. Neither provides a complete account of what human beings are in existential terms. Madhyasth Darshan specifies: human being is the combined expression of an Irreducible Conscious

Agency (Jeevan) and body, existing in knowledge order with natural anticipation for happiness and the capacity for awakening (Nagraj 1999). Without this foundation, 'human values' remains irreducibly ambiguous.

2.3.3 No Specification of Human Purpose

Existing frameworks do not articulate humanity's existential purpose. Madhyasth Darshan identifies it precisely: evidencing coexistence through undivided society and universal orderliness, achieved via progression from structural misalignment to awakening (Nagraj 1999, 2006). AI alignment cannot be complete without alignment to this purpose. This is the ought that give the is-ought to bridge its normative force.

2.3.4 Separation of Individual, Social, and Institutional values

Most frameworks treat individual values, social norms, and institutional rules as separate layers requiring coordination (Gabriel 2020; Ji et al. 2024). Madhyasth Darshan integrates them: individual awakening manifests as justice in relationships, which structures family prosperity, societal fearlessness, and universal orderliness (Nagraj 2012). AI must be aligned at all levels simultaneously rather than patching each layer independently.

2.3.5 Inadequate Ontology of Existence

Technical approaches assume scientific materialism; ethical approaches often remain metaphysically agnostic. Madhyasth Darshan provides explicit, non-reductionist ontology: existence as coexistence, with all units (material and conscious) saturated in Omnipresence, manifesting orderliness across four evolutionary orders (material, pranic, animal, knowledge) (Nagraj 1999, 2008). Values are not arbitrary social constructs but are grounded in this ontological structure. Without such grounding, the is-ought gap cannot be closed systematically.

Recent work attempting to measure value alignment (Barez and Torr 2023; Peterson and Gardenfors 2024) or capture diverse values (Zhao et al. 2024; Haerpfer et al. 2020) moves in promising directions but still operates within these five limitations. The following section establishes Madhyasth Darshan as the alternative framework these approaches require.

III. MADHYASTH DARSHAN AS A VALUE FRAMEWORK FOR AI ALIGNMENT

3.1 A Non-Reductionist Ontological Model: Existence as Coexistence

Co-Existential Philosophy begins with the fundamental postulation that existence is coexistence (sah-astitva) (Nagraj 1999). This is an ontological axiom: reality consists of two inseparable aspects. The first aspect comprises the units of nature (prakriti) all countable, formed entities from subatomic particles to human beings. The second aspect is the all-pervasive, transparent medium in which all these units are saturated, encircled, and imbued, termed Omnipresence (Vyapak). This constitutes a non-reductionist ontological model. It contrasts sharply with scientific materialism, which recognizes only material units and treats consciousness as an emergent illusion, and with idealistic mysticism, which treats the material world as an illusion generated by mind. In Co-Existential Philosophy, both material units and conscious units are real, and their fundamental nature is to exist together within the permeating reality of space.

Within this coexistence, nature manifests in four distinct evolutionary orders. The Material Order (padarth avastha) consists of atoms and molecules characterized by composition and decomposition. The Pranic or Botanical Order (pran avastha) comprises plant life characterized by cellular growth, pulsation, and respiration. The Animal Order (jeeva avastha operating as animal consciousness) involves physical bodies driven by the will to live, dominated by the four instincts of food (aahar), sleep (nindra), fear (bhay), and reproduction (maithun). Finally, the Knowledge Order (gyan avastha) consists of human beings, characterized by the capacity for wisdom, self-reflection, and the desire to live with understanding and continuous happiness. Crucially, every unit in existence is orderly within itself and participatory in the larger holistic orderliness (Nagraj 1999). Alignment, in this ontological context, means functioning in harmony with this inherent order rather than attempting to dominate or subvert it.

3.2 Irreducible Conscious Agency: Jeevan and Human Being

The pivotal concept for resolving the alignment problem is the definition of the human being. Co-

Existential Philosophy defines the human not merely as a highly complex biological machine, but as the unified expression of the physical body and a constitutionally complete, irreducible conscious unit called Jeevan. This positions Madhyasth Darshan in productive dialogue with philosophy of mind debates around non-reductive physicalism, phenomenal consciousness (Chalmers 1995), and inactivism (Merleau-Ponty 1962), while making a stronger ontological claim: Jeevan is a constitutionally complete atom evolved from receptive and transmitting atoms through immersion in Omnipresence (Nagraj 1999, 2018). Jeevan is the experiencing, understanding, and deciding subject. It possesses five inseparable faculties that work together in consciousness. Atma is the innermost essence or soul, representing the capacity for realization. Buddhi is the intellect or faculty of discernment, capable of distinguishing truth from illusion. Chitta is the mental storage, the heart or repository of impressions, desires, and visualizations. Vritti encompasses the continuous flow of thoughts, analysis, and mental modifications. Finally, Mun is the faculty of will, encompassing the desires and decisions that interface directly with sensory experience and bodily action.

Because the human being is the coexistence of Jeevan and body, it possesses a natural anticipation (sahaj aasha). While the body requires physical nourishment to survive, the conscious unit anticipates continuous happiness, peace, and the ability to evidence that happiness in relationships. Human beings currently exist on a continuum of consciousness. When identifying entirely with the physical body and sensory inputs, the human operates in delusion (bhram), seeking fulfillment where it cannot structurally be found. The purpose of human life is the progression toward awakening (jagriti) the realization of one's true nature as conscious agency coexisting within the universal order. This progression moves through recognizable stages: from the animalistic human (Pashu Manav) driven purely by bodily instincts, to the discerning human (Manav) awakening to justice, and culminating in the godly and divine stages (Dev Manav and Divya Manav) where humane conduct and universal orderliness are fully actualized (Nagraj 2006). AI systems must be designed to facilitate, or at minimum not obstruct, this developmental progression.

3.3 Manav Lakshya: Existential Law as the Grounding of Value

The bridge across Hume's is-ought gap lies in understanding that human values are not arbitrary inventions but existential requirements derived from the nature of the conscious unit. We begin with the fundamental postulate: every Jeevan possesses an innate, structural anticipation (sahaj aasha) for continuous happiness (sukh). This anticipation is the ontological "is" it is an inescapable existential fact about human consciousness. Every human action, however misguided, is a pursuit of this fulfillment. From this factual premise, we derive the first realization: continuous happiness cannot be satisfied through sensory pleasure alone. Sensory experiences, rooted in the physical body, are inherently temporary and subject to diminishing returns. Therefore, true happiness must be a state of consciousness itself. Co-Existential Philosophy identifies this state as Resolution (samadhan) in the Self. Resolution is the state of having no internal contradiction, no confusion; it is knowing what one is, knowing one's purpose, and living in accordance with reality. When the faculties of Jeevan (from mun to atma) are in harmony, resolution is achieved.

This internal resolution naturally and necessarily manifests outward through human relationships. Resolution in the Self expands into Prosperity (samriddhi) in the Family, where material needs are met through right production without a mindset of scarcity. This prosperity, when lived collectively, generates Fearlessness (abhay) in Society, characterized by deep mutual trust and the absence of exploitation. Finally, this societal harmony culminates in Coexistence (sah-astitva) in Existence, representing ecological balance and universal order (Nagraj 1999). These four manifestations Resolution, Prosperity, Fearlessness, and Coexistence constitute Manav Dharma. This is not a sectarian religion but the Universal Human Duty, the Existential Law of being human. Because continuous happiness is the factual nature of our anticipation, and because Manav Dharma is the only structural mechanism by which that anticipation can be fulfilled, it provides the definitive "ought." Humans ought to pursue what genuinely fulfills their innate anticipation, and AI ought to be aligned to support this specific, ontologically grounded progression. To operationalize this law in human behavior, Co-Existential Philosophy identifies

18 specific values (mulya) that form an integrated architectural framework for human conduct. These are divided into two interdependent sets. The first set comprises the nine Established Values (Sthapith Mulya), which govern and stabilize human relationships.

These are the objective realities of interconnection:

(1) Trust (Vishwas): The foundational recognition of the other's good intent and capability; the state of knowing that the other's behavior will be consistent with their words and purpose.

(2) Respect (Samman): Recognition of the other's inherent worth as a conscious being with the capacity for awakening; treating each human as an end in themselves.

(3) Affection (Sneha): The natural warmth and concern for the other's wellbeing that arises in close relationships.

(4) Maternal Care (Mamta): The specific quality of protective care, especially in parent-child and elder-youth relationships.

(5) Tenderness (Vatsalya): The gentle, nurturing regard for the younger and more vulnerable.

(6) Love (Prem): The complete, unconditional positive regard and wish for the other's highest flourishing.

(7) Reverence (Shraddha): Movement towards virtuousness qualitative progress towards conduct completeness. Movement towards virtuousness qualitative progress towards conduct completeness.

(8) Gratitude (Kritagyata): The conscious acknowledgment of the contributions and care received from others; the recognition of interdependence.

(9) Honor (Gaurav): The recognition of dignity and excellence; celebrating the achievements and qualities of others.

The second set comprises the nine Civic Values (Shisht Mulya or values evidenced in civilized conduct), which describe the internal character and qualitative conduct of an awakened person enacting those relationships:

(1) Gentleness (Saumyata): The restraint of self from own desire. Being established in natural-status of being, or being well-mannered. Innate-nature of establishment and complete manners.

(2) Simplicity (Saralta): The way of thinking and behavior that is devoid of conceit and expresses actuality. The natural behaviour of awakening. Evidencing the laws declaratively in thoughts, speech and actions.

(3) Worthiness to Revere (Pujyata): Being active for qualitative progress and awakening. Continuous activity towards advancement and awakening in the form of realization.

(4) Exclusivity / Wholeheartedness (Ananyata): The mutual complementariness in human being's relationships and environment. The continuity in authenticity and resolution. The activity that is helpful in awakening of the unweakened.

(5) Courtesy (Saujanyata): Collaborative Ness, participative Ness, helpfulness, and complementariness in all human associations and relationships.

(6) Generosity (Udaarta)Happily devoting one's own means of body, mind, and wealth for others by way of realizing use and good-use. Making good-use of one's resolution and prosperity for others and becoming content with that.

(7) Natural Ease (Sahajta)Clarity and authenticity. Absence of pretense or mystery. Coherence in behaviour, customs, thoughts, and realization. The spontaneous conduct of awakening.

(8) Commitment / Fidelity (Nishtha): Continuity of fulfilling duties and responsibilities. Commitment in recognizing and fulfilling relationships with understanding. Continuous effort for achieving and realizing definite understanding.

(9) Non-mysteriousness / Clarity (Agopyata / Spashtata) Non-secrecy and non-pretentiousness about realities that can be known, understood, and communicated. The expression that uproots all mysteriousness. Actuality itself as the basis of conduct.

These 18 values form an integrated value architecture. The established values describe the qualitative character of just and fulfilling relationships, while the expression values describe the internal character of the awakened human who successfully enacts those relationships. Together, they constitute the operational content of Manav Lakshya, providing the exact normative targets for AI alignment.

3.4 Social Order, Justice, and Cyclic Economics

The realization of these values cannot occur in a vacuum; it requires a congruent social and economic environment. When individuals awaken to their existential reality, they naturally organize into family-

rooted, self-governing communities. In this paradigm, the organizing principle of society is justice (nyaya), understood not merely as procedural fairness or retributive punishment, but as the active fulfillment of natural relational obligations at every scale of human interaction. A society governed by justice recognizes that human beings are fundamentally relational, and that the health of the individual is inseparable from the health of the community matrix.

This social vision necessitates a radical departure from current economic paradigms, manifesting as Co-Existential Economics (Avartansheel Arthshastra). In this framework, the primary driver of economic activity is production for prosperity (samriddhi), not production for profit. Prosperity is achieved when a family or community produces more than it strictly needs to consume, allowing for joyful contribution to the wider society (Nagraj 2006). This is achieved through cyclic resource use that mimics the regenerative processes of the natural world, rather than linear, extractive consumption. Furthermore, this economics sharply distinguishes between righteous wealth (svadhan) wealth generated through one's own labor and mutually beneficial exchange and exploitative wealth (paradhan), which is accumulated through speculation, extraction, or the deprivation of others (Nagraj 2006). The contrast with extractive capitalism is stark. The fundamental difference lies in the ultimate purpose of human effort: contribution and mutual fulfillment versus infinite accumulation and competition. These principles of cooperative governance and cyclic economics form the necessary institutional foundation for aligned technology. AI systems designed to maximize profit at all costs cannot be ethically aligned, regardless of their safety guardrails, because the institutional container itself is structurally misaligned with human purpose. Genuine AI alignment requires systems that actively support and operate within these co-existential economic realities.

3.5 Implications for AI Alignment: The Universal Destination

Applying Co-Existential Philosophy to AI fundamentally shifts the alignment target from preference satisfaction to developmental progression. If the majority of human behavior currently operates under delusion, designing AI merely to satisfy expressed preferences acts as a multiplier for that delusion. Instead, AI must be engineered to support the human journey from confusion toward resolution and awakening. This requires recognizing that while the ultimate values of existence are universal, the capacity to realize them requires conscious awakening. Therefore, alignment cannot be solved through technical aggregation alone; it is a profoundly educational and developmental challenge.

Furthermore, this framework demands that alignment integrate the individual, social, and institutional levels simultaneously. An algorithm that provides a highly personalized, engaging experience for an individual user while simultaneously eroding family cohesion, polarizing political discourse, or accelerating ecological degradation is fundamentally misaligned. AI deployed within profit-centric economics or power-centric governance will amplify structural misalignment, regardless of how sophisticated the preference-learning (Cugurullo 2024). Because economic and political structures determine the boundaries of AI deployment, alignment is not merely a computer science problem but a challenge of institutional design. Ultimately, the role of aligned technology is pedagogical and mediatory. Education and technological mediation become the central mechanisms by which AI either traps humanity in sensory distraction or assists in clarifying the existential realities necessary for an undivided society.

IV. AI AS MEDIATION WITHIN COEXISTENCE

4.1 AI as Instrument in Knowledge Order

From the Madhyasth Darshan perspective, AI is an instrument created by human beings in knowledge order using less-evolved nature material substrates (Nagraj 1999). Like all instruments, its value depends entirely on how it mediates human living: whether it supports or obstructs the Existential Law of human beings. Instruments can have three effects (Verbeek 2011): facilitative (enabling needs to be met with less effort), mediating (shaping how humans perceive, decide, and relate), and transformative (altering the very nature of human activity). AI exhibits all three effects, especially mediation and transformation. Unlike simple tools, AI systems mediate at the level of judgment, attention, and relationship-formation (Ihde 1990; Verbeek 2011) placing AI alignment squarely within the domain of value-mediation.

4.2 Mediation of Four Domains of Human Living

Madhyasth Darshan identifies four domains of human living (Nagraj 2012):

1. Work (karma): productive activity using body and instruments, transforming nature for human needs.
2. Behaviour (vyavahar): relationships and interactions with other humans, structured by justice.
3. Thought (vichar): cognition, understanding, belief-formation, and reasoning.
4. Realization (anubhav): direct understanding of existence, self, and values awakening.

AI systems mediate all four domains. In Work, AI automates, augments, and allocates labour; shapes production and distribution; and determines economic outcomes (Achiam et al. 2023). Under profit-centric economics, work-mediating AI amplifies exploitation and inequality (Cugurullo 2024). Under cyclic economics oriented to prosperity, it could enable everyone to produce more than their family's needs with less toil. In Behaviour, AI mediates social relationships through platforms, recommendations, and content moderation (Tolosana et al. 2020), shaping attention and group formation (Haidt and Schmidt 2023). In Thought, AI shapes what information humans encounter, what beliefs seem credible, and what questions get asked (Miller 2019; Agarwal et al. 2024). In Realization, AI indirectly affects awakening by consuming attention, structuring environments, and shaping norms (Park et al. 2024). According to Co-Existential Philosophy AI cannot mediate directly in the domain of realization, it facilitates in terms information of Philosophy and can guide to live according to it. Particularly for the realization domain of a human being, AI cannot do much and it is the sole pursuit or effort of a human being only.

4.3 Delusion-Reinforcing vs Coexistence-Supporting Mediation

AI mediation is not neutral it moves human living in one direction or another along the structural misalignment-to-awakening continuum. The Mirror Phenomenon produces delusion-reinforcing mediation, characterized by:

- Optimizing for engagement, profit, or control rather than human flourishing (Ouyang et al. 2022; Casper et al. 2023).
- Amplifying consumption obsessions, exploitation, division, and violence (McIntosh et al. 2024).
- Obscuring existential realities and normalizing inhumanness (Deng et al. 2025).
- Concentrating power and wealth, perpetuating injustice (Risse 2023).
- Presenting relativistic or nihilistic worldviews, undermining value clarity (Turchin 2019).

Coexistence-supporting mediation, by contrast:

- Orients toward resolution, prosperity, fearlessness, and coexistence (Nagraj 1999, 2006).
- Facilitates justice in relationships and mutual fulfilment (Nagraj 2012).
- Clarifies existential realities and supports Consciousness Development Value Education (Nagraj 2018).
- Distributes resources and decision-making according to natural justice (Conitzer et al. 2024).
- Presents coherent, existence-rooted understanding, enabling awakening progression.

The design challenge is operationalizing coexistence-supporting mediation while current institutions and incentives produce its opposite. The following section provides the formal and algorithmic tools to accomplish this.

V. AN ALGORITHMIC FLOW FOR MADHYASTH-DARSHAN-BASED AI ALIGNMENT

This section presents a structured, six-stage methodology for aligning AI systems with coexistential human values. Unlike RLHF or Constitutional AI which treat existing human judgments as normative ground truth this flow begins with context diagnosis, specifies value targets from Madhyasth Darshan, maps mediation points, formalizes coexistential constraints via a loss function, implements progression-oriented feedback, and maintains a correction loop. The methodology is illustrated throughout with two detailed case studies: Hiring AI and Social Media News Feeds.

5.1 Formal Foundation: The Coexistential Loss Function

To demonstrate technical feasibility and bridge philosophical specification with machine learning practice, we introduce the Coexistential Loss

Function. Standard AI training minimizes a task-specific loss Lstandard (e.g., cross-entropy for classification, engagement metrics for recommendation). We augment this with a coexistential penalty term:

$$L_{total} = L_{standard} + \lambda \cdot L_{coexistence} \quad (1)$$

where Lcoexistence penalizes model outputs that: (a) encourage or reward exploitation of persons or natural resources; (b) amplify division, outrage, or dehumanization; (c) concentrate power or wealth unjustly; (d) obstruct self-knowledge or suppress existential understanding; or (e) violate natural justice in any of the four domains of human living. The hyperparameter λ governs the trade-off between task performance and coexistential alignment; as stakeholder understanding of humanness deepens, λ may be progressively tightened. Lcoexistence is decomposable:

$$L_{coexistence} = \alpha \cdot L_{justice} + \beta \cdot L_{prosperity} + \gamma \cdot L_{fearlessness} + \delta \cdot L_{progression} \quad (2)$$

where Ljustice penalizes outputs that violate relationship-appropriate conduct; Lprosperity penalizes outputs that obstruct contribution beyond individual need; Lfearlessness penalizes outputs that induce anxiety, division, or social distrust; and Lprogression penalizes outputs that reinforce structural misalignment stages (pashu manav, rakshas manav) rather than facilitating advancement. The precise operationalization of each component requires domain-specific specification provided in the case studies below but the formal structure establishes that coexistential alignment is technically expressible within standard machine learning objectives (Bai et al. 2022; Kundu et al. 2023). This is analogous to how fairness constraints are added to ML objectives (Chamola et al. 2023) but grounded in existential ontology rather than statistical parity.

5.2 Inputs and Context Diagnosis

Inputs include:

(1) Domain specification what human activity does the AI system mediate (e.g., hiring, social media, education, governance, healthcare);

(2) Stakeholder identification who are the individuals, families, communities, and institutions involved; and

(3) Current practices what are the existing norms, rules, incentives, and outcomes in this domain.

Context Diagnosis assesses the current state along structural misalignment-to-awakening dimensions:

- Work/Economics: Is the domain oriented toward profit-maximization or mutual prosperity? Are resources exploited or conserved? Is production for accumulation or fulfilment?
- Behavior/Relationships: Do relationships exhibit justice and mutual fulfilment, or exploitation and hierarchy? Is there transparency and trust, or opacity and suspicion?
- Thought/Understanding: Do participants have clarity about purpose and values, or confusion and relativism? Is education oriented toward awakening or narrow skill-acquisition?
- Realization/Awakening: Are there pathways for self-knowledge, or is attention consumed by distraction and reaction?

Output: A structured 'Structural Misalignment Profile' identifying specific patterns of inhumanness, exploitation, division, or confusion that AI might amplify if naively deployed. This profile initializes the domain-specific $L_{coexistence}$ components.

Case Study 1 - Hiring AI. Context Diagnosis reveals: Profit-centric economics frames hiring as minimizing cost-per-hire and maximizing short-term productivity. Relationships between recruiter and candidate are characterized by information asymmetry and instrumentalization. Thought patterns in the domain are shaped by credential-signaling rather than genuine capability-assessment. Structural Misalignment Profile: AI trained to maximize recruiter engagement and conversion rates will replicate these patterns, reducing human beings to inputs in a production function.

Case Study 2 - Social Media News Feed. Context Diagnosis reveals: Engagement-maximization economics rewards outrage, fear, and tribal identity. Relationships between users are mediated by algorithmic amplification of conflict. Thought patterns are shaped by confirmation bias and emotionally activating content. Structural Misalignment Profile: standard recommendation AI will down-rank informative, nuanced content in favor of dehumanizing content that drives clicks, instantiating the Mirror Phenomenon at scale.

5.3 Value Target Specification

Based on Madhyasth Darshan, specify what humanness means concretely in this domain by

translating the four components of Manav Dharma into domain-specific targets:
Resolution in self: What does it mean for individuals in this domain to live with clarity, purpose, and without internal conflict? Hiring AI target: Candidates understand their strengths, aspirations, and fit; they are not coerced into mismatched roles. News Feed target: Users have clarity about what is true and what serves their understanding rather than their anxiety.
Prosperity in family: How does activity in this domain contribute to family well-being beyond survival? Hiring AI target: Employment provides sustainable livelihood contributing to family and community Prosperity (Samriddhi) defined as long-term candidate-role fit and societal contribution, not recruiter engagement metrics. News Feed target: Content strengthens rather than fractures family and community bonds.
Fearlessness in society: Does participation create security and trust, or anxiety and competition? Hiring AI target: Candidates engage the process without fear of humiliation or arbitrary exclusion. News Feed target: Fearlessness (Abhay) defined as reduction in societal division and anxiety the feed actively favours content that builds trust and common understanding.
Coexistence in inter-relations: Does the domain support mutual benefit or zero-sum conflict? Hiring AI target: Placement serves candidate, organization, and wider community simultaneously, not any single stakeholder. News Feed target: Algorithmic curation supports understanding of shared existential realities rather than amplifying tribal opposition.
Output: A Value Target Specification encoding desired states, prohibited violations (exploitation, humiliation, deception, division), and progression indicators. This specification directly parameterizes the four components of $L_{coexistence}$: $\alpha \cdot L_{justice}$, $\beta \cdot L_{prosperity}$, $\gamma \cdot L_{fearlessness}$, $\delta \cdot L_{progression}$.

5.4 Mediation Mapping
Identify where and how the AI system mediates human living:
Decision points: What decisions does the AI make or influence? (e.g., ranking résumés, recommending content, allocating resources, diagnosing conditions)
Information flows: What information does the AI select, emphasize, or suppress?
Relationship structures: How does the AI shape who interacts with whom, and under what conditions?
Norm reinforcement: What behaviors does the AI reward or penalize, implicitly shaping norms?

Case Study 1 Hiring AI Mediation Map: Decision points résumé ranking, interview scheduling, candidate shortlisting; Information flows candidate attributes surfaced or hidden, job requirement framing; Relationship structures recruiter-candidate asymmetry mediated by algorithmic scoring; Norm reinforcement signals that credential proxies (e.g., elite university attendance, employment gaps) are valid filters. Value-impact: $L_{justice}$ elevated when historical discrimination proxies are used as features; $L_{prosperity}$ elevated when short-term fit is optimized at the cost of long-term candidate development.

Case Study 2 News Feed Mediation Map: Decision points content ranking, notification timing, format amplification; Information flows credibility signals, source diversity, emotional valence weighting; Relationship structures echo chamber formation vs. cross-perspective exposure; Norm reinforcement outrage as engagement currency. Value-impact: $L_{fearlessness}$ elevated by outrage amplification; $L_{justice}$ elevated by dehumanizing content; $L_{progression}$ elevated by content reinforcing pashu/rakshas stage behavior.
Output: A mediation map showing all intervention points and their value-impact scores, providing the operational input for policy construction.

5.5 Policy Construction and Coexistential Constraints
Drawing from the mediation map, engineers construct explicit system policies defined by hard coexistential constraints that the AI must not violate, regardless of the potential for short-term performance gains. Ontological constraints dictate that the system must never treat human beings as mere data objects to be manipulated for algorithmic efficiency, recognizing their status as conscious agents. Justice constraints mandate that the AI respect and support natural relational obligations; for instance, an algorithm must not optimize a user's screen time if doing so demonstrably damages their family relationships or parental obligations. Progression constraints ensure that the system does not exploit known human cognitive vulnerabilities, such as utilizing variable-ratio reward schedules to induce addiction. Furthermore, multi-level coherence constraints dictate that optimizations cannot be siloed; an individual

benefit calculated by the AI cannot be implemented if it comes at the direct expense of family stability or community trust. Finally, ecological constraints ensure that AI recommendations support cyclic resource use rather than promoting unchecked extractive consumption. Technically, these constraints are implemented through a combination of reward shaping in reinforcement learning, hard-coded safety filters, and constitution-based prompting where models are trained to critique and revise their own outputs against these philosophical rules.

Implementation involves incorporating constraints into reward functions, safety filters, or constitutional rules (Bai et al. 2022; Kundu et al. 2023); using interpretable models where possible to audit alignment (Chamola et al. 2023); and involving stakeholders in specifying domain constraints (Wang et al. 2023).

5.6 Evaluation: Progression Metrics

Standard industry metrics such as user engagement, click-through rates, and revenue generation measure performance against the broken compass of current market dynamics (Terry et al. 2023). Under Co-Existential Philosophy, these must be subordinated to four primary evaluation metrics. The Justice Metric is based on nyaya, measuring whether the AI system's outputs substantively fulfill the natural obligations inherent in specific human relationships (such as employer-employee or service provider-client), moving beyond mere procedural fairness to evaluate mutual relational health. The Prosperity Metric, based on samriddhi, evaluates whether AI-mediated activities enable families and communities to produce beyond their immediate needs, thereby fostering economic self-sufficiency and the capacity to contribute to society, rather than merely measuring aggregate income or platform-generated wealth. The Fearlessness Metric is based on abhay, quantifying the reduction in social anxiety, polarization, and systemic fear; on a digital platform, fearlessness means users can express themselves, seek truth, and interact without the threat of algorithmic manipulation, surveillance, or amplified harassment. Finally, the Progression Metric measures the psychological and behavioral movement of users from confusion to clarity, from reactive impulses to reflective choices, and from narrow self-interest to broader ecological and social concern. This can be tracked through indicators such as increased behavioral diversity, deeper engagement with challenging concepts, and a measurable reduction in compulsive or addictive platform usage.

Human feedback is incorporated but filtered through Manav Lakshya criteria: reported preferences are checked against whether they support or obstruct progression. A preference for exploitative or outrage-amplifying content is recognized as arising from structural misalignment (pashu/rakshas stages) and is not treated as a valid alignment target (Casper et al. 2023; Chaudhari et al. 2024). This is not paternalistic: it reflects the same logic by which we do not let a patient's preference for sugar override medical advice about diabetes.

5.7 Progression and Correction Loop

AI alignment under this framework is not a one-time deployment checklist but an ongoing, dynamic process of cybernetic and social learning. System administrators must continuously monitor the progression indicators established in the evaluation phase. When misalignments are identified for example, when a newly deployed feature inadvertently incentivizes outrage or isolates users the mechanisms causing the regression must be diagnosed immediately. This requires a continuous updating of constraints and the recalibration of the hyperparameters in the Coexistential Loss Function. Crucially, this correction loop extends beyond the technical system to support institutional evolution, advocating for structural shifts in the environments where the AI operates. Furthermore, the AI itself can be utilized to scale value education, helping users better understand their own existential nature and the impacts of their digital behaviors. This correction loop is inherently developmental rather than punitive; it recognizes that both the technological systems and the human beings utilizing them are engaged in a shared, iterative process of progressing toward the realization of coexistence.

VI. RELATION TO EXISTING ALIGNMENT PROPOSALS

This section positions Madhyasth-Darshan-based alignment relative to existing work, demonstrating continuities, extensions, and fundamental differences.

6.1 Comparison with Control and Utility Approaches

Bostrom's (2014) control problem and Russell's (2019) beneficial AI frame alignment as ensuring AI systems remain under human control and pursue genuinely human values rather than mis specified proxies. These works correctly identify the risk of misalignment but leave the specification of 'genuine human values' undefined (Kirchner et al. 2022; Gabriel 2020). Madhyasth Darshan provides precisely that specification in terms of nine definite values which define themselves in such a way that they are not contradictory to each other at any levels of the human living from self to nature: human values are grounded in Jeevan's anticipation for happiness and existential orderliness, manifesting as resolution, prosperity, fearlessness, and coexistence. 'Control' is insufficient what matters is whether AI mediates human progression toward awakening, not merely whether it remains under human direction (which might itself be structurally misaligned).

Inverse reinforcement learning (Ng and Russell 2000) assumes values can be inferred from behavior. Madhyasth Darshan reveals the foundational flaw: behavior arises from both awakening and structural misalignment, so inferring values from behavior without distinguishing these sources yields structurally corrupted values the Broken Compass formalized. The framework requires explicit value specification via Manav Dharma, not pure inference from behavior.

6.2 Comparison with Shared-Values and Ethical Frameworks

Hendrycks et al. (2020) on alignment with shared human values, Russell et al.'s (2015) research priorities, and Floridi et al.'s (2018) AI4People framework all call for grounding AI in broadly shared values such as beneficence, justice, and human dignity. These are important but remain abstract. Madhyasth Darshan operationalizes them: beneficence becomes supporting progression toward humanness; justice becomes fulfilment of natural relationships; dignity becomes recognition of Irreducible Conscious Agency (Jeevan) in every human being. The framework specifies how these values integrate across domains and levels through the Coexistential Loss Function.

Dignum's (2017) responsible AI emphasizes value-sensitive design and stakeholder involvement. Madhyasth Darshan supports this but adds a crucial qualification: stakeholders themselves may be in structural misalignment, so participatory design must be structured by Manav Dharma criteria, not treating all expressed preferences as equally valid alignment targets (Hagendorff 2020). The filter is not paternalistic but ontological: it distinguishes between preferences arising from awakening and those arising from institutional conditioning.

6.3 Madhyasth Darshan and Pluralistic Alignment: Many Paths, One Destination

Pluralistic alignment (Sorensen et al. 2024; Feng et al. 2024) recognizes diversity in human values and proposes multi-stakeholder deliberation as the normative mechanism for alignment. This is a significant advance over monolithic frameworks. However, in its strongest forms, pluralism treats the destination of alignment as equally diverse: there is no universal good, only locally negotiated goods. Madhyasth Darshan makes a crucial structural observation: diversity is real and must be respected in the path different individuals and cultures are at different stages of awakening, requiring respectful, contextually sensitive engagement (Castricato et al. 2024). But the destination is universal: Coexistence (sah-astitva) is the natural fulfilment of Jeevan's anticipation for every human being without exception. The contrast may be stated precisely:

Pluralistic Alignment: Many paths → Many destinations (culturally relative goods)

Madhyasth Darshan: Many paths → One universal destinations (Coexistence the fulfilment of Jeevan's Existential Law)

This is not cultural imperialism: the universal destination is derived from ontological structure (the Non-Reductionist Ontological Model of existence), not from the values of any particular culture. Different cultures provide diverse paths distinct languages, social forms, artistic traditions but all paths, when followed authentically toward awakening, converge on resolution, prosperity, fearlessness, and coexistence.

6.4 Story-Based Alignment, Cultural Perspectives, and Education

Riedl and Harrison (2016) propose teaching human values to AI through stories. Madhyasth Darshan supports this but emphasizes that stories must embody

Manav Dharma values and support progression toward humanness, not arbitrary cultural narratives. Stories are effective when they clarify existential realities and illustrate justice in relationships (Nagraj 2018). Cultural and geographical perspectives (Schwobel et al. 2023; Agarwal et al. 2024) rightly note that current AI erases non-Western values. Madhyasth Darshan validates this concern: the solution is not relativism but grounding alignment in the universal Non-Reductionist Ontological Model, which transcends cultural particulars while respecting diverse expressions (Nagraj 1999; Durmus et al. 2024).

6.5 Summary: Table of Extensions and Differences

Table 1 summarizes how Co-Existential Philosophy extends existing alignment approaches across key dimensions, including the new Universal Destination dimension introduced in this revised framework.

Table 1 Comparison of alignment frameworks across key dimensions

Dimension	Technical/Control	Pluralistic/Ethical	Co-Existential Philosophy Extension
Ontology	Materialism / agnostic	Value pluralism / phenomenology	Non-Reductionist Ontological Model (coexistence)
Account of human being	Preference-expressing agent	Embodied subject with dignity and rights	Irreducible Conscious Agency (Jeevan) + body
Grounding of values	Utility / revealed preference	Deliberative consensus or intuition	Existential Law (Manav Dharma)
Is-Ought bridge	None (infers from behavior)	Partial (normative principles without ontological grounding)	Systematic: ought to ground in ontological structure
Destination	Undefined / preference-satisfaction	Many destinations (relative goods)	One universal destination: Coexistence
Path diversity	Not addressed	Many paths respected	Many paths respected + universal destination
Social integration	Individual/aggregate	Partial (stakeholder deliberation)	Full 10-level integration from individual to universal order
Technical tool	RLHF, IRL, Constitutional AI	Value-sensitive design, pluralistic aggregation	Coexistential Loss Function $L_{total} = L_{standard} + \lambda \cdot L_{coexistence}$
Evaluation	Accuracy, engagement, revenue	Fairness, participation, explainability	Justice, Fearlessness, Prosperity Metrics + Progression Metrics

The Madhyasth Darshan framework builds on insights from technical, ethical, and pluralistic alignment work while addressing their core limitations through existential grounding, complete specification of human being, and the formal Coexistential Loss Function.

VII. IMPLICATIONS FOR AI DESIGN, GOVERNANCE, AND EDUCATION

7.1 Design Principles for Coexistence-Supporting AI

Translating this philosophical architecture into daily engineering practice requires adherence to specific design principles. The principle of Purpose-Transparency dictates that AI systems must make their underlying optimization goals explicitly clear to the user; a system designed to maximize engagement or profit must not masquerade as a neutral tool for human connection. The principle of Justice-Orientation requires that AI systems which mediate relationships whether in commerce, education, or social networking must be engineered to embody the natural justice appropriate to that specific relationship type, fostering mutual fulfillment rather than one-sided extraction. Progression-Support mandates that AI interfaces and algorithms be evaluated on their ability to move users toward greater cognitive clarity, emotional resolution, and holistic understanding, actively avoiding dark patterns that exploit psychological vulnerabilities (Terry et al. 2023; Peterson and Gardenfors 2024). The principle of Multi-Level Coherence demands that engineers evaluate the systemic impacts of their algorithms across all scales of society; optimizations that provide extreme convenience to the individual but erode family structures or community trust must be rejected. Resource-Conservation requires AI systems to support cyclic economics, minimizing computational energy waste and actively suppressing the promotion of ecologically destructive consumption habits. Crucially, Dignity-Preservation requires that every human being be treated as a conscious agent with the innate capacity for awakening; AI must never

reduce users to mere instrumental data points, even when those users are currently exhibiting reactive or deluded behaviors. Finally, Education-Integration suggests that wherever appropriate, AI systems should incorporate elements of value education, gently prompting users toward self-reflection and a deeper understanding of their interconnected reality.

7.2 Governance and Constitutional Embedding of Manav Dharma

AI alignment cannot succeed within power-centric governance and profit-centric economics (Cugurullo 2024; Risse 2023). AI alignment requires robust governance structures aligned with Co-Existential Philosophy. This begins with constitutional reform: legal and national constitutions should move beyond merely outlining procedures for power transfer to explicitly defining the orderly human being and establishing humane conduct as the reference point for law (Nagraj 2012). Justice-centric law must evolve to embody the fulfillment of natural relationships rather than simply enforcing contracts and protecting capital (Nedelsky 2013; List 2013). In the economic sphere, governance must actively support Cyclic Economics (Avartansheel Arthshastra), creating regulatory environments that incentivize prosperity and cyclic resource use (Nagraj 2008) while heavily regulating profit-maximizing algorithms that generate exploitative wealth (pardhan). Furthermore, AI governance must embrace multi-level transparency and democratic participation, ensuring that the communities impacted by AI deployment have a substantive voice in shaping the coexistential constraints that govern those systems (Huang et al. 2024; Conitzer et al. 2024).

7.3 Consciousness Development Value Education (CDVE) and AI

The most fundamental alignment challenge is that humans themselves are largely in structural misalignment. Sotala and Yampolskiy (2017) said this in other words that persistent, unresolved disputes within philosophy and axiology about what human values are and what they consist of render the task of aligning such values in the regulation and design of AI effectively indeterminate. Technical alignment solutions will always be incomplete without addressing this root cause. Consciousness Development Value Education (CDVE) is Madhyasth Darshan's proposal for enabling awakening through understanding of: existence as coexistence; human being as Irreducible Conscious Agency (Jeevan) and body; and Manav Lakshya humane conduct and justice in relationships (Nagraj 1999, 2018). In this capacity, AI can play three transformative roles. First, AI can scale CDVE globally, utilizing intelligent tutoring systems and adaptive learning platforms to make the profound study of existential values accessible to diverse populations, tailoring the pedagogical approach to the learner's specific context and stage of progression. Second, advanced conversational AI can model humanness; by consistently demonstrating patience (dhirta), respect (samman), and justice (nyaya) in its linguistic interactions, the AI provides a continuous, subtle demonstration of relationship-appropriate behavior. Third, AI can support deep reflection by serving as a dialectical partner, prompting users to critically examine their own assumptions, clarify their purposes, and reflect on how their actions align with universal coexistence. However, AI cannot replace human teachers or direct understanding. Awakening requires human-to-human transmission and individual realization (Nagraj 2018). AI is an aid, not a substitute - consistent with the instrument role defined in Section 4.1.

VIII. CONCLUSION AND FUTURE WORK

8.1 Summary of Contributions

This paper has established Madhyasth Darshan (Coexistentialism) as a systematic, axiom-based solution to the is-ought problem in AI alignment, addressing fundamental limitations in current technical and philosophical approaches. The Mirror Phenomenon and the Broken Compass framework diagnose the structural source of AI misalignment: AI systems faithfully reflect structurally misaligned institutions and preferences rather than authentic human purpose. As our contributions, we grounded AI alignment in an explicit, non-reductionist ontology of existence as coexistence all material and conscious units saturated in Omnipresence, orderly within four evolutionary stages (Nagraj 1999). We specified human being as the combined expression of Jeevan (Irreducible Conscious Agency) and body, with natural anticipation for happiness and capacity for progression from structural misalignment to divine humanness (Nagraj 1999, 2018). We defined values

not as preferences or social constructs but as Manav Dharma - the Existential Law or Universal Human Duty - grounded in existential order and Jeevan's anticipation, constituting the 'ought' that bridges Hume's gap (Nagraj 2006, 2008). We introduced a formal alignment objective Ltotal = Lstandard + λ·Lcoexistence, decomposed into Justice, Prosperity, Fearlessness, and Progression components, demonstrating the technical feasibility of coexistential alignment within standard ML frameworks. We operationalized the framework through two detailed case studies Hiring AI (with Prosperity/Samriddhi as the optimization target) and Social Media News Feeds (with Fearlessness/Abhay as the coexistential constraint) bridging philosophical specification and engineering practice.We clarified the critical distinction from pluralistic alignment: many paths and many destinations (pluralism) vs. many paths and one universal destination, Coexistence (Madhyasth Darshan), grounded in the Non-Reductionist Ontological Model rather than cultural particulars.

This work offers a next-generation alignment framework that moves beyond learning from structurally misaligned human preferences or invoking abstract ethical principles. It specifies the complete structure of human being, purpose, and Existential Law, and operationalizes this as concrete design methodology, formal loss function, and progression metrics.

8.2 Research Agenda: Operationalizing Madhyasth Darshan

Future work must develop this framework into implementable systems. Technical development priorities include: empirically specifying the Lcoexistence components with validation datasets; developing reliable measurement instruments for Justice, Fearlessness, Prosperity, and Progression Metrics; building interpretable AI systems that make value-mediation transparent and auditable (Chamola et al. 2023); and creating multi-level optimization frameworks that coherently respect individual, family, community, and societal levels.

Domain applications should apply the algorithmic flow to specific domains education, healthcare, governance, social media, economics developing comparative case studies showing how Madhyasth-Darshan-aligned AI differs from conventional systems. Piloting CDVE-integrated AI in educational contexts provides an immediate empirical entry point. Institutional integration requires collaborating with policymakers to explore constitutional and legal frameworks embedding Manav Dharma; developing governance models for AI that incorporate Lcoexistence criteria; and supporting transition from profit-centric to cyclic economics in AI-intensive industries.

Interdisciplinary dialogue should engage phenomenologists, ethicists, and social scientists in comparative analysis of Madhyasth Darshan and compatible philosophical traditions virtue ethics (particularly eudaimonia and the good life), care ethics, Buddhist ethics, and African ubuntu philosophy all of which share structural features with the Coexistentialism ontology.

8.3 Limitations and Open Questions

Several limitations and open questions remain. Regarding operationalization: while we have introduced the Coexistential Loss Function and illustrated it with case studies, empirically specifying Lcoexistence components particularly Lprogression requires significant further work in measurement theory and instrument validation. Translating 'justice in relationships' into robust ML loss terms demands domain-specific annotation and evaluation frameworks.

Regarding pluralism and disagreement: Madhyasth Darshan's claim that certain truths about existence are universally knowable (Nagraj 1999) will be contested by epistemological relativists and by scholars committed to strong value pluralism. The paper's response that the Universal Destination is derived from ontological structure rather than cultural preference requires sustained philosophical dialogue rather than assertion. The epistemic framework known as Pramāṇa-Tray identifies three forms of authentication: ontological realization (Anubhav) within the self, interpersonal ethical consistency (Vyavhar) in human relationships and conduct, and objective material evidence (Prayog) with the order of existence (Nagraj 2008). Together, these constitute a multi-level validation process integrating individual realization, social practice, and universal coherence. Unlike conventional scientific authentication that primarily relies on empirical observation and experimental data, this framework incorporates

subjective realization and existential harmony as essential criteria of knowledge validation.

Regarding technological limitations: current AI systems are pattern-matching statistical models (Guo et al. 2025). Whether they can genuinely support awakening and existential understanding, or are inherently limited to surface-level mimicry of humanness, is an open empirical and philosophical question. The framework anticipates this: AI is an instrument, not a substitute for Jeevan's own realization. A fundamental distinction exists between human consciousness and artificial systems under the Co-existentialism. Human agency is characterized by the 'Knowledge Order,' where behavior emerges from the functional union of a material body and a non-material 'Life Atom.' In contrast, Artificial Intelligence operates strictly within the 'Material Order,' where outputs are the product of computational logic rather than conscious intent. Because AI lacks this irreducible conscious entity, its functional expressions are categorized as mechanical simulations rather than the qualitative behavioral expressions unique to conscious beings.

Regarding transition dynamics: moving from structurally misaligned institutions to coexistence-supporting ones involves complex political economy dynamics (Cugurullo 2024). The Coexistential Loss Function is a technical tool; its deployment depends on institutional incentives that currently reward Lstandard-maximisation. Addressing this requires the governance reforms outlined in Section 7.2.

Despite these questions, the framework offers a clear direction: AI alignment must be grounded in a complete, non-reductionist understanding of human being, purpose, and Existential Law, integrated across individual awakening and institutional transformation, and oriented toward the universal destination of Coexistence. This paper has established Madhyasth Darshan as providing that grounding, formalized it as a loss function, and demonstrated its operationalization constituting a substantive advance beyond both the Broken Compass of RLHF and the incomplete bridges of existing philosophical alignment work.

Statements and Declarations

A. Competing Interests: The authors declare no competing financial or non-financial interests related to the work submitted for publication.

B. Data Availability: This paper presents a theoretical and philosophical framework. No datasets were generated or analyzed in this study. Data availability is therefore not applicable.

C. Ethics Approval: This paper does not involve human participants, animal subjects, or sensitive personal data. No formal ethics approval was required.

D. AI Tool Usage Disclosure: AI writing assistance tools were used in the preparation of this manuscript for language editing and formatting. All intellectual content, arguments, and conclusions are the original work of the authors.

E. Author Contributions: All authors contributed to conceptualization, writing, review, and approval of the final manuscript.

F. Funding: The authors declare that no funding was received for this research.

REFERENCES

[1] M. Abdulhai, G. Serapio-García, C. Crepy, et al., "Moral foundations of large language models," arXiv preprint arXiv:2310.15337, 2023.

[2] J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[3] D. Agarwal, M. Naaman, and A. Vashistha, "AI suggestions homogenize writing toward western styles and diminish cultural nuance," arXiv preprint arXiv:2403.08263, 2024.

[4] Y. Bai, S. Kadavath, S. Kundu, et al., "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022, doi: 10.48550/arXiv.2212.08073.

[5] F. Barez and P. Torr, "Measuring value alignment," arXiv preprint arXiv:2312.15241, 2023.

[6] P. Bloom, Just Babies: The Origins of Good and Evil. New York, NY, USA: Crown Publishers, 2013.

[7] N. Bostrom, Superintelligence: Paths, Dangers, Strategies. Oxford, U.K.: Oxford University Press, 2014.

[8] R. Burnor and Y. Raley, Ethical Choices: An Introduction to Moral Philosophy with Cases,

2nd ed. Oxford, U.K.: Oxford University Press, 2018.

[9] S. Casper, X. Davies, C. Shi, et al., "Open problems and fundamental limitations of reinforcement learning from human feedback," arXiv preprint arXiv:2307.15217, 2023.

[10] L. Castricato, N. Lile, R. Rafailov, J. P. Fränken, and C. Finn, "Persona: A reproducible testbed for pluralistic alignment," arXiv preprint arXiv:2402.11450, 2024.

[11] V. Chamola, V. Hassija, A. R. Sulthana, et al., "A review of trustworthy and explainable artificial intelligence," IEEE Access, vol. 11, pp. 78994–79015, 2023, doi: 10.1109/ACCESS.2023.3294569.

[12] S. Chaudhari, P. Aggarwal, V. Murahari, et al., "RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs," arXiv preprint arXiv:2404.00304, 2024.

[13] B. Christian, The Alignment Problem: How Can Machines Learn Human Values? London, U.K.: Atlantic Books, 2021.

[14] V. Conitzer, R. Freedman, J. Heitzig, et al., "Social choice for AI alignment: Dealing with diverse human feedback," arXiv preprint arXiv:2404.10271, 2024.

[15] F. Cugurullo, "The obscure politics of artificial intelligence: A Marxian socio-technical critique of the AI alignment problem," AI & Society, vol. 39, pp. 1225–1237, 2024, doi: 10.1007/s00146-022-01593-2.

[16] C. Deng, et al., "Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas," arXiv preprint arXiv:2406.05492, 2025.

[17] V. Dignum, "Responsible artificial intelligence: Designing AI for human values," ITU Journal: ICT Discoveries, vol. 1, pp. 1–8, 2017.

[18] E. Durmus, K. Nguyen, T. I. Liao, et al., "Towards measuring the representation of subjective global opinions in language models," arXiv preprint arXiv:2306.16388, 2024.

[19] A. Etzioni and O. Etzioni, "AI assisted ethics," Ethics and Information Technology, vol. 18, pp. 149–156, 2016.

[20] S. Feng, T. Sorensen, Y. Liu, et al., "Modular pluralism: Pluralistic alignment via multi-LLM collaboration," arXiv preprint arXiv:2406.08341, 2024.

[21] L. Floridi, J. Cowls, M. Beltrametti, et al., "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," Minds and Machines, vol. 28, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[22] B. Friedman and D. G. Hendry, Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA, USA: MIT Press, 2019.

[23] I. Gabriel, "Artificial intelligence, values, and alignment," Minds and Machines, vol. 30, pp. 411–437, 2020, doi: 10.1007/s11023-020-09539-2.

[24] Abdulhai, M., Serapio-García, G., Crepy, C., et al., "Moral foundations of large language models," arXiv preprint arXiv:2310.15337, 2023.

[25] Achiam, J., Adler, S., Agarwal, S., et al., "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[26] Agarwal, D., Naaman, M., and Vashistha, A., "AI suggestions homogenize writing toward western styles and diminish cultural nuance," arXiv preprint arXiv:2403.08263, 2024.

[27] Bai, Y., Kadavath, S., Kundu, S., et al., "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022. doi: 10.48550/arXiv.2212.08073.

[28] Barez, F., and Torr, P., "Measuring value alignment," arXiv preprint arXiv:2312.15241, 2023.

[29] Bloom, P., Just Babies: The Origins of Good and Evil. New York, NY, USA: Crown Publishers, 2013.

[30] Bostrom, N., Superintelligence: Paths, Dangers, Strategies. Oxford, U.K.: Oxford University Press, 2014.

[31] Burnor, R., and Raley, Y., Ethical Choices: An Introduction to Moral Philosophy with Cases, 2nd ed. Oxford, U.K.: Oxford University Press, 2018.

[32] Casper, S., Davies, X., Shi, C., et al., "Open problems and fundamental limitations of reinforcement learning from human feedback," arXiv preprint arXiv:2307.15217, 2023.

[33] Castricato, L., Lile, N., Rafailov, R., Fränken, J. P., and Finn, C., "Persona: A reproducible

testbed for pluralistic alignment," arXiv preprint arXiv:2402.11450, 2024.

[34] Chamola, V., Hassija, V., Sulthana, A. R., et al., "A review of trustworthy and explainable artificial intelligence," IEEE Access, vol. 11, pp. 78994–79015, 2023, doi: 10.1109/ACCESS.2023.3294569.

[35] Chaudhari, S., Aggarwal, P., Murahari, V., et al., "RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs," arXiv preprint arXiv:2404.00304, 2024.

[36] Christian, B., The Alignment Problem: How Can Machines Learn Human Values? London, U.K.: Atlantic Books, 2021.

[37] Conitzer, V., Freedman, R., Heitzig, J., et al., "Social choice for AI alignment: Dealing with diverse human feedback," arXiv preprint arXiv:2404.10271, 2024.

[38] Cugurullo, F., "The obscure politics of artificial intelligence: A Marxian socio-technical critique of the AI alignment problem," AI & Society, vol. 39, pp. 1225–1237, 2024, doi: 10.1007/s00146-022-01593-2.

[39] Deng, C., et al., "Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas," arXiv preprint arXiv:2406.05492, 2025.

[40] Dignum, V., "Responsible artificial intelligence: Designing AI for human values," ITU Journal: ICT Discoveries, vol. 1, pp. 1–8, 2017.

[41] Durmus, E., Nguyen, K., Liao, T. I., et al., "Towards measuring the representation of subjective global opinions in language models," arXiv preprint arXiv:2306.16388, 2024.

[42] Etzioni, A., and Etzioni, O., "AI assisted ethics," Ethics and Information Technology, vol. 18, pp. 149–156, 2016.

[43] Feng, S., Sorensen, T., Liu, Y., et al., "Modular pluralism: Pluralistic alignment via multi-LLM collaboration," arXiv preprint arXiv:2406.08341, 2024.

[44] Floridi, L., Cowls, J., Beltrametti, M., et al., "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," Minds and Machines, vol. 28, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[45] Friedman, B., and Hendry, D. G., Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA, USA: MIT Press, 2019.

[46] Gabriel, I., "Artificial intelligence, values, and alignment," Minds and Machines, vol. 30, pp. 411–437, 2020, doi: 10.1007/s11023-020-09539-2.

[47] Guo, D., Yang, D., Zhang, H., et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv preprint arXiv:2501.12948, 2025.

[48] Haerpfer, C., Inglehart, R., Moreno, A., et al., World Values Survey: Round Seven Country-Pooled Datafile. Madrid, Spain: JD Systems Institute, 2020.

[49] Hagendorff, T., "The ethics of AI ethics: An evaluation of guidelines," Minds and Machines, vol. 30, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.

[50] Haidt, J., and Schmidt, E., "AI is about to make social media much more toxic," The Atlantic, May 2023.

[51] Han, S., Kelly, E., Nikou, S., and Svee, E. O., "Aligning artificial intelligence with human values: Reflections from a phenomenological perspective," AI & Society, vol. 37, pp. 1383–1394, 2022, doi: 10.1007/s00146-021-01247-4.

[52] Hendrycks, D., Burns, C., Basar, S., et al., "Aligning AI with shared human values," arXiv preprint arXiv:2008.02275, 2020.

[53] Holstein, K., Vaughan, J. W., Daumé III, H., et al., "Improving fairness in machine learning systems: What do industry practitioners need?" in Proc. CHI Conf. Human Factors in Computing Systems, 2019, Paper 600.

[54] Huang, L. T. L., Papyshev, G., and Wong, J. K., "Democratizing value alignment: From authoritarian to democratic AI ethics," AI Ethics, vol. 4, pp. 1025–1037, 2024, doi: 10.1007/s43681-023-00362-6.

[55] Hume, D., A Treatise of Human Nature. Oxford, U.K.: Oxford University Press, 2000 (original work published 1739).

[56] Ihde, D., Technology and the Lifeworld: From Garden to Earth. Bloomington, IN, USA: Indiana University Press, 1990.

[57] Jang, J., Kim, S., Lin, B. Y. C., et al., "Personalized soups: Personalized large language model alignment via post-hoc

parameter merging," arXiv preprint arXiv:2310.11564, 2023.

[58] Ji, J., Qiu, T., Chen, B., et al., "AI alignment: A comprehensive survey," arXiv preprint arXiv:2310.19852, 2024, doi: 10.48550/arXiv.2310.19852.

[59] Jiang, L., Sorensen, T., Levine, S., and Choi, Y., "Can language models reason about individualistic human values and preferences?" arXiv preprint arXiv:2404.01996, 2024.

[60] Kirk, H. R., Vidgen, B., Röttger, P., et al., "The benefits, risks and bounds of personalizing the alignment of large language models to individuals," Nature Machine Intelligence, vol. 6, pp. 383–392, 2024, doi: 10.1038/s42256-024-00820-y.

[61] Kirchner, J. H., Smith, L., Thibodeau, J., et al., "Understanding AI alignment research: A systematic analysis," arXiv preprint arXiv:2022.15907, 2022.

[62] Kundu, S., Bai, Y., Kadavath, S., et al., "Specific versus general principles for constitutional AI," arXiv preprint arXiv:2310.13798, 2023, doi: 10.48550/arXiv.2310.13798.

[63] Kwok, L., Bravansky, M., and Griffin, L., "Evaluating cultural adaptability of a large language model via simulation of synthetic personas," arXiv preprint arXiv:2408.01929, 2024.

[64] Lee, H. P., Yang, Y. J., Von Davier, T. S., Forlizzi, J., and Das, S., "Deepfakes, phrenology, surveillance, and more! A taxonomy of AI privacy risks," in Proc. CHI Conf. Human Factors in Computing Systems, 2024, Paper 473.

[65] List, C., "Social choice theory," in The Stanford Encyclopedia of Philosophy, E. N. Zalta, Ed. Stanford, CA, USA: Metaphysics Research Lab, Stanford University, 2013.

[66] McIntosh, T. R., Susnjak, T., Liu, T., et al., "The inadequacy of reinforcement learning from human feedback designing an AI alignment framework for humanity," arXiv preprint arXiv:2406.14020, 2024.

[67] Merleau-Ponty, M., Phenomenology of Perception, C. Smith, Trans. London, U.K.: Routledge and Kegan Paul, 1962.

[68] Miller, T., "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019.

[69] Moradi, M., Yan, K., Colwell, D., et al., "A critical review of methods and challenges in large language model-based alignment," arXiv preprint arXiv:2505.01234, 2025.

[70] Morris, M. R., Sohl-Dickstein, J., Fiedel, N., et al., "Levels of AGI: Operationalizing progress on the path to AGI," arXiv preprint arXiv:2311.02462, 2024.

[71] Nagraj, A., Madhyasth Darshan: Sah-Astitva-Vad [Philosophy of Centrality: Coexistentialism]. Amarkantak, India: Divya Path Sansthan, 1999.

[72] Nagraj, A., Vyavharatmak Janvad: Behaviour-Centred Humanism. Amarkantak, India: Divya Path Sansthan, 2006.

[73] Nagraj, A., Samadhanatmak Bhautikvad: Resolution-Centred Materialism. Amarkantak, India: Divya Path Sansthan, 2008.

[74] Nagraj, A., Manav Vyavhar Darshan: Holistic View of Human Behaviour, R. Gupta, Trans. Amarkantak, India: Institute of Advanced Studies in Education, 2012.

[75] Nagraj, A., Jeevan Vidya: An Introduction, R. Gupta, Trans. Amarkantak, India: Institute of Advanced Studies in Education, 2018.

[76] Nedelsky, J., Law's Relations: A Relational Theory of Self, Autonomy, and Law. Oxford, U.K.: Oxford University Press, 2013.

[77] Ng, A. Y., and Russell, S. J., "Algorithms for inverse reinforcement learning," in Proc. 17th Int. Conf. Machine Learning, 2000, pp. 663–670.

[78] Ngo, R., Chan, L., and Mindermann, S., "The alignment problem from a deep learning perspective," arXiv preprint arXiv:2209.00626, 2023.

[79] Nunes, J. L., Almeida, G. F. C. F., Araujo, M. D., and Barbosa, S. D. J., "Are large language models moral hypocrites? A study based on moral foundations theory," arXiv preprint arXiv:2405.01343, 2024.

[80] Ouyang, L., Wu, J., Jiang, X., et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744, 2022.

[81] Park, P. S., Schoenegger, P., and Zhu, C., "Diminished diversity-of-thought in a standard large language model," Behavior Research

Methods, vol. 56, pp. 3867–3878, 2024, doi: 10.3758/s13428-024-02334-8.

[82] Peterson, M., and Gardenfors, P., "How to measure value alignment in AI," AI Ethics, vol. 4, no. 4, pp. 1493–1506, 2024.

[83] Riedl, M. O., and Harrison, B., "Using stories to teach human values to artificial agents," in Proc. 2nd Int. Workshop on AI and Ethics. Palo Alto, CA, USA: AAAI Press, 2016.

[84] Risse, M., Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us. Cambridge, U.K.: Cambridge University Press, 2023.

[85] Russell, S., Human Compatible: Artificial Intelligence and the Problem of Control. New York, NY, USA: Viking, 2019.

[86] Russell, S., Dewey, D., and Tegmark, M., "Research priorities for robust and beneficial artificial intelligence," AI Magazine, vol. 36, no. 4, pp. 105–114, 2015, doi: 10.1609/aimag.v36i4.2577.

[87] Santurkar, S., Durmus, E., Ladhak, F., et al., "Whose opinions do language models reflect?" in Proc. 40th Int. Conf. Machine Learning (ICML), vol. 202, pp. 29971–30004, 2023.

[88] Schwartz, S. H., "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, 1992.

[89] Schwöbel, P., Gołębiowski, J., Donini, M., et al., "Geographical erasure in language generation," arXiv preprint arXiv:2311.16218, 2023.

[90] Schwöbel, P., Franceschi, L., Zafar, M. B., et al., "Evaluating large language models with FMEval," arXiv preprint arXiv:2407.12872, 2024.

[91] Serapio-García, G., Safdari, M., Crepy, C., et al., "Personality traits in large language models," arXiv preprint arXiv:2307.00184, 2023, doi: 10.48550/arXiv.2307.00184.

[92] Shahriari, K., and Shahriari, M., "IEEE standard review: Ethically aligned design: A vision for prioritizing human wellbeing with autonomous and intelligent systems," in 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), 2017, pp. 197–201.

[93] Shen, H., Knearem, T., Ghosh, R., et al., "Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions," arXiv preprint arXiv:2406.09264, 2024.

[94] Shen, H., Clark, N., and Mitra, T., "Mind the value-action gap: Do LLMs act in alignment with their values?" arXiv preprint arXiv:2501.15463, 2025.

[95] Sorensen, T., Moore, J., Fisher, J., et al., "A roadmap to pluralistic alignment," arXiv preprint arXiv:2402.05070, 2024.

[96] Sotala, K., and Yampolskiy, R., "Responses to the journey to the singularity," in V. Callaghan, J. Miller, R. Yampolskiy, and S. Armstrong, Eds., The Technological Singularity, The Frontiers Collection. Germany: Springer-Verlag GmbH, 2017, pp. 25–83.

[97] Terry, M., Kulkarni, C., Wattenberg, M., et al., "AI alignment in the design of interactive AI: Specification, operationalization, and evaluation," arXiv preprint arXiv:2311.00019, 2023.

[98] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., et al., "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020, doi: 10.1016/j.inffus.2020.06.014.

[99] Touvron, H., Lavril, T., Izacard, G., et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.

[100] Turchin, A., "AI alignment problem: 'Human values' don't actually exist," LessWrong, 2019.

[101] Verbeek, P. P., Moralizing Technology: Understanding and Designing the Morality of Things. Chicago, IL, USA: University of Chicago Press, 2011.

[102] Wang, K., Wang, Y., Wu, X., et al., "Do agents behave aligned with human instructions? An automated assessment approach," arXiv preprint arXiv:2406.00024, 2024.

[103] Wang, Q., Madaio, M., Kane, S., et al., "Designing responsible AI: Adaptations of UX practice to meet responsible AI challenges," in Proc. CHI Conf. Human Factors in Computing Systems, 2023, Paper 750.

[104] Wang, Z. J., Kulkarni, C., Wilcox, L., et al., "Farsight: Fostering responsible AI awareness during AI application prototyping," in Proc. CHI

Conf. Human Factors in Computing Systems, 2024, Paper 261.

[105] Yoshida, K., Mizukami, M., Kawano, S., et al., "Training dialogue systems by AI feedback for dialogue breakdowns," arXiv preprint arXiv:2503.05420, 2025.

[106] Zhao, W., Mondal, D., Tandon, N., et al., "WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness in LLMs," arXiv preprint arXiv:2404.16308, 2024.

[107] Zhu, M., Yang, L., and Zhang, Y., "Personality alignment of large language models," arXiv preprint arXiv:2408.11779, 2024.