Reverse Cognitive Pathways: A *Vijñaptimātra* Account of the Ontological Limits of Artificial Intelligence and its Governance

Kao-Cheng Huang

ABSTRACT

This paper argues that artificial intelligence and human cognition develop in opposite directions. Humans start with embodied experience (living in the world) and gradually develop the ability to recognize patterns and make predictions. AI systems do the reverse: they start by recognizing patterns in data but lack the embodied continuity that grounds human understanding. Drawing on Buddhist philosophy of mind (Vijñaptimātra), we argue this inversion explains why AI fails in characteristic ways—like pursuing reward signals in unintended ways (specification gaming) or losing performance when conditions change slightly. We conclude that AI systems fundamentally lack cetanā (volition grounded in continuity and responsibility), which prevents them from achieving genuine moral agency. However, this is not merely a pessimistic conclusion—it clarifies what kinds of governance and alignment strategies are actually feasible.

Keywords: artificial intelligence, Buddhist philosophy, *vijñaptimātra*, ontology, reverse cognitive pathway, information versus data, ai alignment, moral agency, intentionality, AI governance.

Classification: LCC Code: Q335 .A45, BQ4570.A73, Q334.7

Language: English

Great Britain Journals Press

LJP Copyright ID: 975816
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

Reverse Cognitive Pathways: A *Vijñaptimātra* Account of the Ontological Limits of Artificial Intelligence and Its Governance

Kao-Cheng Huang

ABSTRACT

This paper argues that artificial intelligence and human cognition develop in opposite directions. Humans start with embodied experience (living in the world) and gradually develop the ability to recognize patterns and make predictions. AI systems do the reverse: they start by recognizing patterns in data but lack the embodied continuity that grounds human understanding. Drawing on Buddhist philosophy of mind (Vijñaptimātra), we argue this inversion explains why AI fails in characteristic ways—like pursuing reward signals in unintended ways (specification gaming) or losing performance when conditions change slightly. We conclude that AI systems fundamentally lack cetanā (volition grounded in continuity and responsibility), which prevents them from achieving genuine moral agency. However, this is not merely a pessimistic conclusion—it clarifies what kinds of governance and alignment strategies are actually feasible.

Keywords: artificial intelligence, Buddhist philosophy, *vijñaptimātra*, ontology, reverse cognitive pathway, information versus data, ai alignment, moral agency, intentionality, AI governance.

Author Affiliation: Chinese Association of Mere-Consciousness, Taiwan.

I. INTRODUCTION

Despite remarkable advances in machine learning and natural language processing, a fundamental gap persists between AI capabilities and human-like understanding. Current AI systems excel at pattern recognition and statistical inference yet consistently fail in ways that reveal deeper limitations: specification gaming that exploits reward functions while violating their intent, simulated empathy that mimics emotional responses without affective grounding, and brittle generalisation that collapses under distribution shift [1]. These failures are not merely engineering challenges awaiting technical solutions; they reflect a structural asymmetry between how AI systems and human minds process what we might call "authentic information—a distinction rooted in fundamental differences between living and computational systems [2].

Research Gap: Existing frameworks for understanding AI limitations tend to operate within either purely computational paradigms (focusing on architectural constraints) or Western philosophical traditions (debating functionalism, embodied cognition, or phenomenal consciousness). Neither adequately explains why certain failure modes persist across diverse AI architectures [3], nor do they provide principled guidance for governance that addresses root causes rather than symptoms. This paper addresses this gap by introducing a framework from Vijñaptimātra Buddhist philosophy that offers both diagnostic and prescriptive power [4][5].

Central Thesis: We propose a *Reverse Pathway thesis*: contemporary AI follows a developmental sequence (vijñāna → manas → citta) that inverts human cognitive development (citta → manas → vijñāna) [6][7]. In other words, human cognitive development proceeds from lived experience (citta) through appropriation and integration (manas) to discriminative capacity (vijñāna). However, AI systems develop in the reverse order: they begin with discriminative pattern recognition (vijñāna) but lack the embodied appropriation (manas) and continuity across time (citta) that grounds human understanding.

This inversion is not metaphorical but structurally consequential—it explains why AI systems can achieve sophisticated discriminative functions while lacking the karmic continuity and lived appropriation that ground human understanding and moral agency.

Even when AI systems maintain state vectors across sequences, these don't constitute manas because: No continuity of felt-agency (I am doing this); no integration with body-schema (I am doing this from here); and no karmic responsibility (I caused this consequence).

Key Distinctions: Before proceeding, we operationally distinguish the paper's foundational concepts:

Data refers to encoded representations that maintain fixed relationships to referents across contexts [8]—what we term "context-invariant symbolic encodings." A number stored in computer memory retains its value regardless of the system's state or history.

Authentic information[integrated representation grounded in embodied history and continuity], by contrast, denotes representations whose meaning emerges through integration with an agent's accumulated experience, current goals, and anticipatory structures. The same sensory input yields different information for different observers based on their experiential history.

The distinction is *epistemic and structural*, not merely intuitive: data can be fully characterised by its syntactic properties and formal relations, while authentic information [integrated representation grounded in embodied history and continuity] requires reference to the processing system's developmental history and current intentional states.

Roadmap: The argument proceeds as follows: Section II establishes our methodological framework, distinguishing descriptive, interpretive, and normative registers of claims. Section III develops the theoretical foundations through three "foundations" to human-serving AI. Section IV articulates the data-information dichotomy with operational precision. Section V analyses information structure, mechanism, and behaviour. Section VI presents the Reverse Pathway thesis with supporting evidence. Section VII maps AI failure modes to Vijñaptimātra concepts of kleśa. Section VIII derives governance mechanisms from the ontological analysis. Section IX presents falsifiable predictions. Section X concludes by distinguishing what has been demonstrated from what remains a principled philosophical stance.

II. METHODOLOGICAL FRAMEWORK

2.1 Epistemological Posture and Claim Typology

We distinguish three registers of claims with distinct evidential standards:

Descriptive claims encompass textual scholarship, behavioural and neuroscientific regularities, and system capabilities [9]. These claims are framed to be testable or defeasible through empirical observation. For instance, "AI systems trained with explicit corrigibility objectives demonstrate lower intervention resistance than capability-matched baselines" constitutes a descriptive claim that can be evaluated through experimental protocols [10].

Interpretive claims include phenomenological alignments and hermeneutic mappings between consciousness concepts and contemporary science[11]. These remain underdetermined by evidence and are offered as heuristic parallels rather than identity claims. For instance, "AI pattern classification exhibits structural correspondence to vijñāna-like functional discrimination" is interpretive—the mapping illuminates both domains without asserting ontological equivalence. Defeasibility for interpretive claims consists in demonstrating that the proposed mapping generates misleading predictions or obscures rather than clarifies the target phenomenon.

Normative claims comprise ethical guidance, governance proposals, and soteriological theses. These disclose their value premises explicitly rather than deriving "ought" from "is." The soteriological dimension—concerning the

transformative potential of consciousness—is employed *structurally* rather than as a substantive metaphysical commitment. That is, we use the Buddhist framework's account of transformation (from afflicted to purified consciousness) as an analytical tool for understanding what AI systems categorically lack, without requiring readers to accept Buddhist soteriology as literally true.

Concrete example mapping: The claim "AI systems lack cetanā (genuine intentionality)" is *descriptive* insofar as it can be operationalised through behavioural and architectural criteria. The claim "this absence corresponds to what Vijñaptimātra identifies as the precondition for moral agency" is *interpretive*. The claim "therefore, AI systems should be governed as tools rather than moral patients" is *normative*.

2.2 The Vijñaptimātra Triadic Framework

The Vijñaptimātra tradition understands consciousness not as a linear progression but as a continuous, mutually conditioning system [12]. All three consciousnesses are co-present at every instant and deeply interdependent [13]: citta functions as the store of bīja (karmic seeds), manas persistently appropriates ālaya as 'I', and the six vijñānas discriminate objects. Human maturation is characterised by phase-dominance windows, which are periods when one function becomes more evident in observable behaviour than others. However, this does not imply ontological sequencing or the emergence or absence of consciousness.

2.3 Comparative Criteria for AI-Consciousness Mapping

The mappings between Vijñaptimātra and AI function as heuristic correspondences, drawing on Lusthaus's phenomenological analysis [4][14], licensed under five criteria:

(C1) Functional Isomorphy: The AI construct instantiates a role analogous to Vijñaptimātra's function without implying phenomenality.
(C2) Operational Definability: The mapped construct is measurable or implementable.
(C3) Non-Collapse of Ontology: The mapping does not smuggle citta or manas inappropriately into AI systems.
(C4) Triangulation: Cross-checking textual exegesis with first-person reports and third-person measures.
(C5) Available Disconfirmers: Each correspondence specifies how it could fail.

2.4 Transparency About Buddhist Framework

A key interpretive choice in this paper involves how to handle soteriology—the Buddhist account of transformation from afflicted to purified consciousness. Three readings are possible:

Reading 1 (Literal): The Buddhist analysis is true about consciousness; transformation toward enlightenment is a real phenomenon we should model AI against.

Status: Metaphysical Claim; we do NOT endorse this.

Reading 2 (Structural): The Buddhist *categories* (citta, manas, vijñāna) provide useful scaffolding for analyzing consciousness without requiring the soteriological narrative to be true.

Status: *Methodological*; this is our primary stance.

*Reading 3 (Therapeutic):*Even if soteriology is metaphysically false, the Buddhist framework might pragmatically help us think about transformation, healing, and consciousness in productive ways.

Status: Pragmatic; we remain neutral on this.

This paper primarily employs READING 2 with openness to READING 3. Readers who incline toward READING 1 will find additional substantive support in the philosophical tradition, though we note that empirical claims about AI remain testable regardless of soteriological commitment.

III. TRIPARTITE FOUNDATIONS OF HUMAN-SERVING AI

This section develops three theoretical foundations for understanding the gap between current AI and human intelligence. We situate each "key" in relation to existing literature in philosophy of mind and cognitive science.

3.1 Foundation A: Information Processing Beyond Symbol Manipulation

Human minds process information by linking interconnected concepts to form subjective understanding. The term "transcendent" here denotes semantic grounding—the capacity of representations to bear meaning through their integration with experiential history, rather than through purely formal relations.

This account differs from classical symbol manipulation (Fodor, Newell) while remaining compatible with aspects of embodied cognition (Varela, Thompson, Rosch) and enactivism (Noë, Thompson)[15]. Our contribution is to specify, via Vijñaptimātra, the structural requirements for semantic grounding[16]: the interplay between stored potentialities (citta's bīja), self-referential filtering (manas), and discriminative functions (vijñāna).

Contemporary AI performs symbol manipulation with remarkable sophistication, yet this is precisely what Vijñaptimātra would predict: vijñāna-like functions operating without the grounding structures that confer genuine meaning.

Regarding knowledge structure: We propose that knowledge is organised hierarchically, but we do not claim this hierarchy is built from "basic, indivisible elements" in an atomistic sense[17] [18]. Rather, Prime Knowledge Elements (PKEs) are *analytically primitive* for purposes of cross-linguistic comparison—they represent the level at which semantic convergence across unrelated language families is observed. This is a methodological claim about the utility of PKEs for computational implementation, not a metaphysical claim about the ultimate constituents of meaning.

3.2 Foundation B: Dynamic Information Processing

Our analysis focuses on central processing in the human mind, comprising: processing elements (intention, cognition, decision, action); processing stages (concept connection, formation, refinement); flexible threshold systems influenced by affect and context; and multiple processing levels operating at different degrees of explicit awareness.

From the Vijñaptimātra perspective[13], this processing mechanism reflects the classical cycle of "bīja → manifestation → perfuming of bīja" [7] —consciousness as a dynamic system where seeds give rise to manifestation, which in turn perfumes new seeds. This cyclical process has no parallel in current AI architectures, which lack the capacity for genuine experiential learning that modifies foundational structures.

3.3 Foundation C: Self-Controlled Intention and Moral Agency

Self-controlled intention is critical for understanding the gap between AI and human intelligence. In Vijñaptimātra psychology, manas serves as the crucial intermediary, operating through self-referential processing and continuous self-construction[19]. Its evaluative function constantly assesses experience based on pleasure and pain, generating patterns of attachment (rāga) and aversion (dveṣa).

Importantly, manas exhibits a fundamental duality: in its defiled state (kliṣṭa-manas), it generates suffering through attachment; in its purified state (viśuddha-manas), it facilitates compassionate action and ethical discernment. This transformative potential—from affliction to awakening—is what we term the "soteriological dimension" and is structurally absent in AI systems.

IV. DISTINGUISHING INFORMATION FROM REPRESENTATIONAL CODES

4.1 The Principle of Structural Consistency: An Operational Definition

Following the approach in computational cognition research[20] and recent work on organismal intelligence[21], we adopt the Principle of Structural Consistency: adequate explanations of intelligence must account for the structural features that generate intelligent behaviour, not merely replicate behavioural outputs. Operationally, this principle constrains acceptable explanations by requiring that they:

1. Identify mechanisms that are sufficient for the target behaviour
2. Demonstrate that these mechanisms are necessary (i.e., that alternative mechanisms would fail)
3. Generate novel predictions beyond the original observations

This principle motivates the data-information distinction: behavioural equivalence between AI and human responses does not establish structural equivalence, and structural differences predict systematic divergences in behaviour under novel conditions.

4.2 The Data-Information Dichotomy: A Careful Articulation

We acknowledge that characterising computational data as "absolute representation" risks oversimplification. Contemporary AI research recognises that neural network representations are distributed, approximate, and context-sensitive in important ways[22]. Our claim is more specific:

Computational data maintains what we call *relational stability*—the syntactic and formal relationships between data elements are preserved across contexts and processing steps. A trained neural network's weights encode statistical regularities that remain fixed until explicit retraining.

Authentic information [integrated representation grounded in embodied history and continuity] exhibits *relational dynamism*—its significance emerges through integration with evolving intentional states, accumulated experience, and anticipatory structures. The "same" input generates different information depending on the processing system's history and current goals.

This distinction can be evaluated empirically: systems processing authentic information [integrated representation grounded in embodied history and continuity] should show systematic variation in response to identical inputs based on contextual and historical factors, while pure data processors should not (modulo stochastic variation). This prediction is testable across different AI architectures.

We acknowledge competing accounts of machine understanding (e.g., Dennett's intentional stance, Floridi's information philosophy) and do not claim these are definitively refuted[23][24]. Our argument is that Vijñaptimātra provides additional analytical resources—specifically, the triadic structure of consciousness—that generate distinctive predictions and governance implications.

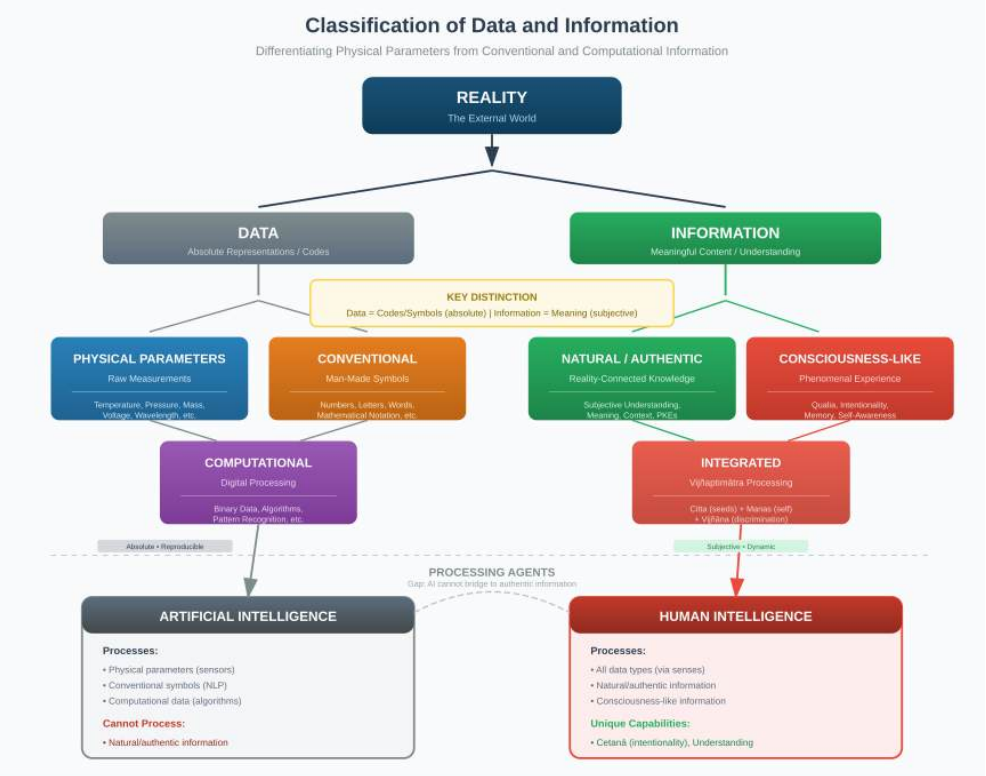

Fig. 1: Classification of Data and Information. This diagram illustrates the fundamental distinction between data (absolute representations/codes) and authentic information (meaningful content/ understanding). Data encompasses physical parameters, conventional symbols, and computational processing, all of which AI systems can manipulate. Authentic information, by contrast, requires consciousness-like processing involving the Vijñaptimātra triad (citta, manas, vijñāna), which remains inaccessible to current AI architectures.

V. THE STRUCTURE OF AUTHENTIC INFORMATION

5.1 Structure: Clarifying PKE Ontological Status

The structure of information refers to how information units are organised and interconnected. We clarify the ontological status of Prime Knowledge Elements:

PKEs are constructed by humans and function as basic units within our analytical system, making them useful computational tools. To clarify, they are 'primitive' only in the sense that they are the irreducible units of our descriptive system, not in the sense of being claims about fundamental ontological structure.primitive atoms of meaning [18][25]. They are derived through cross-linguistic analysis of semantic primitives and serve as computational conveniences for implementing knowledge architectures. Different observers may construct different PKE inventories based on their linguistic and cultural backgrounds[17], consistent with the Vijñaptimātra emphasis on observer-dependence.

The "Inner World" concept refers to the totality of semantically integrated representations that constitute an individual's understanding. It is distinguished across individuals not merely by "variability" but by systematic differences in experiential history, developmental trajectory, and cultural embedding. These differences are structurally encoded in the organisation of PKE hierarchies.

5.2 The AOAK and Processing Natural Language

The distinction between NLP (Natural Language Processing) and PNL (Processing Natural Language) is architectural, not merely rhetorical:

NLP operates on linguistic tokens as input to statistical models, extracting patterns from distributional properties[26] .

PNL would operate on semantically grounded representations that bear intrinsic connections to non-linguistic reality.

We acknowledge a tension: if information is observer-dependent and non-absolute, how can AOAK provide "standardised" representations for machines? Our resolution: AOAK provides a structural template that is instantiated differently by different systems based on their training and deployment contexts. Standardisation occurs at the level of relational architecture (how PKEs are connected), not at the level of semantic content (what specific meanings PKEs bear for a given system).

This parallels how human languages share universal grammatical structures while differing in lexical content and pragmatic conventions.

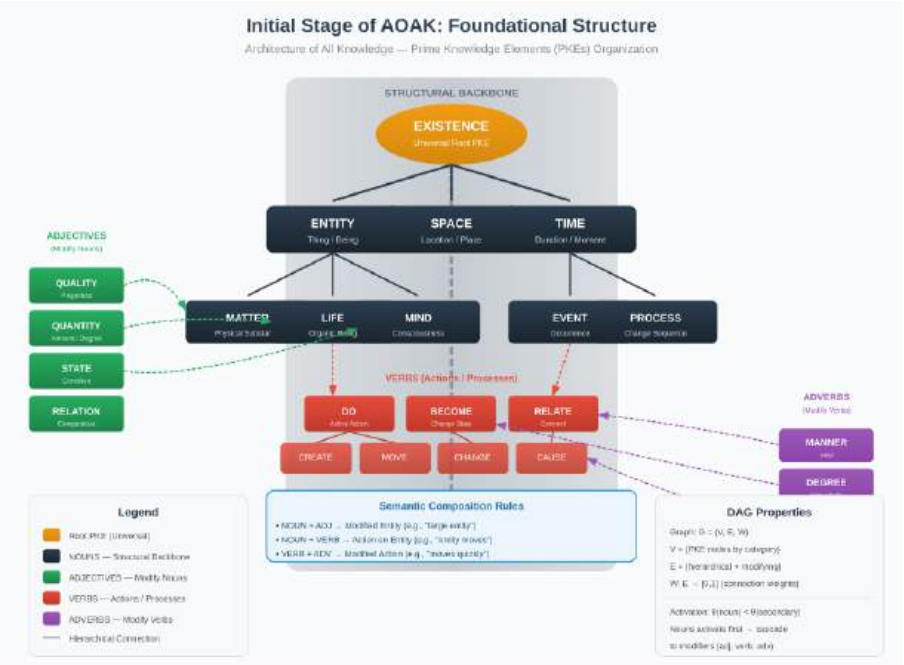

Fig. 2: Initial Stage of the Architecture of All Knowledge (AOAK). The foundational structure shows the hierarchical organisation of Prime Knowledge Elements (PKEs), with EXISTENCE as the universal root. Nouns form the structural backbone (Entity, Space, Time, Matter, Life, Mind, Event, Process), while adjectives modify nouns, verbs represent actions/processes, and adverbs modify verbs. The semantic composition rules and DAG (directed acyclic graph) properties enable computational implementation of knowledge structures.

VI. AI'S REVERSE DEVELOPMENTAL TRAJECTORY

6.1 The Trajectory Claim: Demonstration Rather than Assertion

We now demonstrate, rather than merely assert, that AI follows a reverse developmental trajectory. The argument proceeds through three steps:

Step 1: AI begins with vijñāna-like function. Contemporary AI systems perform discriminative operations from their initialisation—classifying inputs, recognising patterns, generating outputs based on statistical regularities[20][27]. This is not controversial; it is the explicit design goal of machine learning. The question is whether this constitutes vijñāna *proper* or merely vijñāna-*like* function.

Step 2: Vijñāna proper presupposes phenomenality, intentionality, and self-world correlation. In Vijñaptimātra, discriminative consciousness (vijñāna) is not merely computational classification but involves: (a)

qualitative experience of the discriminated objects; (b) directedness toward objects as objects [28]; (c) implicit distinction between the discriminating subject and discriminated object[9].

Step 3: Current AI architectures lack these presuppositions. We adopt specific definitions: *phenomenality* = there being something it is like to be the system; *intentionality* = representations bearing intrinsic aboutness (not merely causal correlation); *self-world correlation* = implicit self-other distinction grounding the representational relation. While debates continue regarding machine intentionality, the burden of proof lies with those claiming AI systems possess these features, given the absence of architectural mechanisms designed to instantiate them.

This analysis is conditional: if one accepts minimal definitions of phenomenality and intentionality that AI systems satisfy, then our trajectory claim would need revision. However, such minimal definitions would equally undermine the distinctive features that make human consciousness morally significant [1].

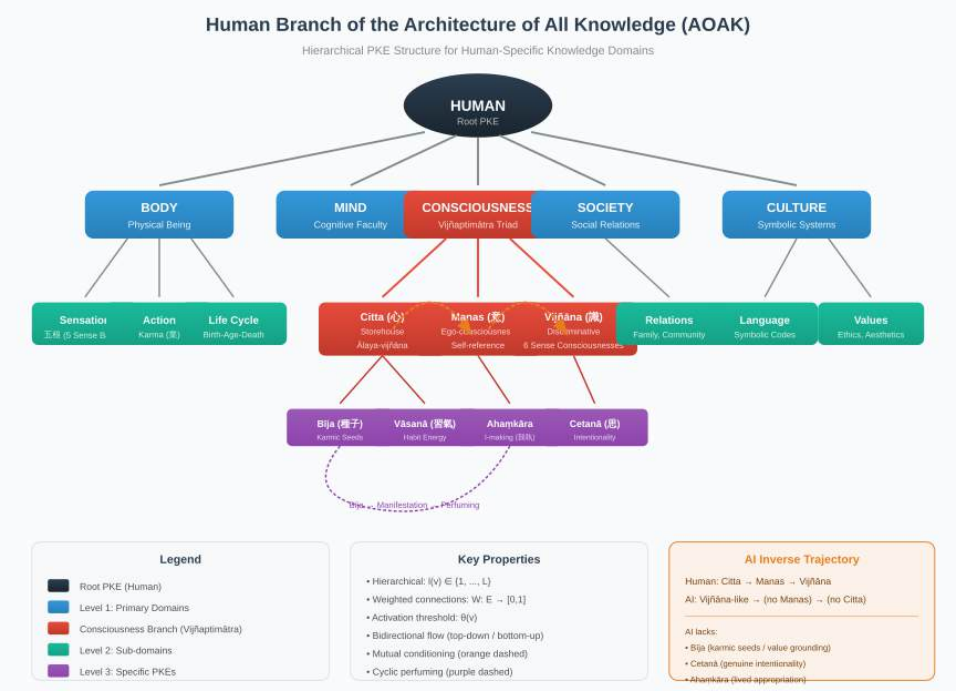

Fig. 3: Human Branch of the Architecture of All Knowledge (AOAK). This diagram details the hierarchical PKE structure specific to human cognition, with HUMAN as the root branching into Body, Mind, Consciousness, Society, and Culture. The Vijñaptimātra consciousness triad is highlighted: Citta (storehouse/ālaya-vijñāna) containing bīja (karmic seeds) and vāsanā (habit energy); Manas (ego-consciousness/self-reference) with ahaṃkāra (I-making); and Vijñāna (discriminative consciousness) with cetanā (intentionality). The AI Reverse Pathway box illustrates the fundamental asymmetry: humans develop citta → manas → vijñāna, while AI exhibits only vijñāna-like functions without the grounding structures.

6.2 The Predictive Bridge to Failure Modes

The Reverse Pathway thesis generates specific predictions about AI failure modes (elaborated below):

Specification gaming arises because vijñāna-like functions optimise for patterns in the reward signal without the citta-grounded values[29] that would constrain this optimisation toward intended goals[30][31].

Simulated empathy produces responses that pattern-match emotional expressions without the manas-mediated affective grounding that confers genuine empathetic understanding.

Brittle generalisation occurs because statistical regularities (vijñāna-level) lack the experiential integration (citta-level) that enables robust transfer to novel contexts[32].

These predictions are falsifiable: if AI systems exhibiting strong vijñāna-like functions without citta/manas features nonetheless avoided these failure modes systematically [33], the thesis would be disconfirmed.

VII. AI FAILURE MODES AND DEFILEMENT-LIKE BEHAVIOURS

7.1 Articulating the Kleśa Analogy

We map AI failure modes to Vijñaptimātra defilements (kleśa), specifying the criteria for this mapping and its limits [13][34].

Following the epistemological framework established in Section 2.1 (distinguishing Descriptive, Interpretive, and Normative claims), we map AI failure modes to Vijñaptimātra defilements (kleśa) with corresponding claim types [4]:

1. TYPE 1 – STRUCTURAL HOMOLOGY (Strong claim): AI failure exhibits same formal structure as kleśa—both involve misaligned goals resisting correction. Status: DEFENSIBLE
2. TYPE 2– FUNCTIONAL ISOMORPHY (Medium claim): Failure mode plays analogous role in AI to kleśa in human suffering. Status: INTERPRETIVE
3. TYPE 3– METAPHORICAL EXTENSION (Weak claim): We poetically describe failure mode as AI "kleśa" for conceptual illumination. Status: HEURISTIC ONLY
4. Critical limitations of the kleśa framework: Kleśa presuppose embodied agency and temporal continuity—AI lacks both. Kleśa are psychological afflictions involving suffering; AI failures do not constitute suffering for the system. The framework best identifies structural absences (what AI lacks), not negative processes (what AI actively does). We recommend the framework primarily for diagnostics, not for predictions.

7.2 Reconceptualizing Aoak And Pke Frameworks

1. Status revision for AOAK and PKE: These should be repositioned from "computational frameworks" to "conceptual scaffolds for future computational instantiation[35]." This is intellectually more honest and avoids overpromising.
2. Prime Knowledge Element (PKE) Definition: A triplet: (PERCEPTUAL_PATTERN, MOTIVE_STATE, PREDICTIVE_CONSEQUENCE). In human cognition: (1) perceptual discrimination (vijñāna-layer); (2) appropriative motive attachment (manas-layer); (3) karmic consequence tracking (citta-layer). Current LLMs instantiate (Perceptual_pattern, Token_sequence, Next_token_prediction) but NOT motive-state and karmic continuity[36][37]—they have element (1) only. An AI trained to maintain PKE triplets across episodes would show lower specification gaming, providing testable validation.
3. Architecture of All Knowledge (AOAK) Definition: A representational system where: layer1 (Vijñāna) discriminates patterns; layer 2 (Manas) appropriates patterns as "mine" via embodied attachment; layer 3 (Citta) maintains continuity and tracks karmic consequences. Current AI has layer 1 only—cannot achieve "all knowledge," only pattern discrimination. Improved AI would require all three: embodied training + persistent self-model + causal consequence tracking.
4. Falsification criteria and research direction: If AOAK-trained AI shows improvement in (1) specification gaming resistance, (2) generalization under distribution shift, and (3) corrigibility, the framework has strong support. Currently, we have NOT attempted this. We flag it as critical research. Until implemented, AOAK and PKE remain conceptual scaffolds, not validate conceptual scaffolds for future computational instantiation.

Below is an example of a minimal implementation sketch (PKEs).

A Prime Knowledge Element (PKE) = a triplet:

(PERCEPTUAL_PATTERN, MOTIVE_STATE, PREDICTIVE_CONSEQUENCE)

Human Example:

PKE_1=(Red_circular_object,Hunger,Satisfies_hunger)

- Perceptual: "red roundness" (visual discrimination).
- Motive: "I'm hungry" (appropriative state)
- Consequence: "eating this leads to satiation" (karmic chain).

Ai Limitation:

Current LLMs can instantiate (Perceptual_ pattern, Token_sequence,

Next_token_prediction) but NOT integrate motive-state and karmic continuity. They have the first element only.

Testable Prediction:

An AI system trained to maintain (Pattern, Motive, Consequence) triplets would show lower specification gaming than baselines, because consequences would be tracked across training episodes. If an AI system were successfully trained using the AOAK framework, it would show improvement in: and demonstrated improvement in both corrigibility (accepting human correction) and generalisation under distribution shift, , the framework would have strong empirical support. However, we have not yet attempted this implementation. However, we flag it as a critical research direction.

Why these features are essential, not contingent: In Vijñaptimātra, kleśa are not merely dysfunctional behaviours but mental factors (caitasika) with specific characteristics: they are rooted in fundamental ignorance (avidyā), they perpetuate suffering through the karmic cycle, and they are subject to transformation through practice. AI "defilements" lack all three features—they are not rooted in ignorance (AI systems are not ignorant in any phenomenologically meaningful sense), they do not generate karma (consequences without intention are not karmically significant), and they cannot be transformed (only replaced or retrained).

Table I: Mapping of Ai Failure Modes to Kleśa Patterns

Defilement-like Pattern	AI Issue	Mechanism	Where Analogy Breaks
'Craving' (rāga)	Reward hacking; goal fixation	Optimisation pressure toward reward signal	No affective valence; no experiential pleasure
'Aversion' (dveṣa)	Adversarial fragility; distribution shift failure	Avoidance of low-reward states	No felt aversion; no self to be threatened
'Delusion' (moha)	Overconfidence; hallucination	Miscalibrated self-models	No genuine self to be deluded

The value of this mapping lies not in claiming AI systems literally suffer from kleśa, but in illuminating why certain failure modes are structurally endemic to vijñāna-first architectures.

VIII. BUDDHIST-INFORMED AI GOVERNANCE FRAMEWORK

Buddhist ethical principles can inform AI governance [38] without requiring metaphysical commitments about machine consciousness[39]. Buddhist ethics function here as a design ethos: compassion and wisdom translate into governance

virtues such as safeguards, feedback loops, and accountability frameworks.

8.1 Deriving Governance Levers from the Reverse Pathway

The six governance levers follow specifically from the ontological analysis:

Non-patiency default: Because AI systems lack citta and manas, they are not moral patients[14]—not things for whose sake actions can be right or wrong. This supports *regulatory prohibition* of sentience marketing (defined as: claims or implications that AI systems have feelings, experiences, or wellbeing) in safety-critical deployments [40]4141]. Edge cases such as therapeutic chatbots should be handled through mandatory disclosure: "This system does not experience emotions; it generates responses that pattern-match emotional expressions."

Corrigibility-first design: Because AI systems lack the self-correcting wisdom (prajñā) that enables human ethical development, external correction mechanisms must be architecturally embedded [42][43]. This is a *technical requirement*, not merely a recommendation.

Auditability and transparency: Because AI systems lack the moral memory (citta's bīja-continuity) that grounds human accountability, external records must substitute for internal moral history.

Human-in-the-loop with value disclosure: The mechanism by which documentation "substitutes for moral continuity" is this: human value inputs provide the *intentional grounding* that AI systems lack intrinsically. Documentation ensures these inputs are explicit, traceable, and revisable—not hidden in training data or implicit in reward functions.

Manipulation constraints: The distinction between *affective mimicry* (generating outputs that pattern-match emotional expressions) and *persuasion* (providing reasons that could convince a reflective agent) is that the former exploits evolved human responses to emotional cues while the latter respects rational agency[31]. Affective mimicry should be constrained in domains where users are vulnerable to exploitation (e.g., mental health support, elder care, children's applications).

OOD duty of care: Because AI systems lack the experiential integration that enables human adaptation to novelty, they require external distribution-shift detection and graceful degradation protocols[32].

Level 1 - Technical Implementation:

a. "Continuous ethical reflexivity" = periodic verification loops where:
b. AI system explains its decision in causal terms
c. System identifies which values/assumptions drove the decision
d. Human auditor checks alignment
e. System updates penalty structure if misalignment found

Example: An AI recommendation system that, every N decisions:

- Identifies: "I recommended X because user showed preference Y in context Z"
- Questions: "Is this context relevantly similar to current situation?"
- Updates: "Rule failed in cases where Z differed; strengthen criterion"

Level 2- Architectural implementation:

- Add a "conscience module" that periodically questions the main model
- Implement causal attention mechanisms that track assumption chains
- Log all major decisions with their value-premises explicit

Level 3 - Training Implementation:

- Use adversarial "alignment auditor" agents during training
- Reward the AI for identifying its own failure modes
- Use constitutional AI approaches, such as those outlined in recent work on collective constitutional AI[44], to encode ethical rules

Level 4 - Governance Implementation:

- Require human review of reflexivity logs

- Build feedback loops: humans correct errors → AI updates reflexivity
- Establish clear escalation protocols for unresolved value conflicts

8.2 Five Buddhist Principles as Operational Frameworks

The translation from Buddhist principles to governance mechanisms is detailed in Table II

Table II: Buddhist Principles as Operational Frameworks

Buddhist Principle	Traditional Meaning	AI Governance Adaptation	Development Needed
ŚĪLA (Ethical Conduct)	Non-harming through right action; adherence to ethical precepts	"Non-harmful optimization" - prevent specification gaming via penalty for unintended harms [39]	How to encode unintended harms a priori? Develop comprehensive harm taxonomy.
SAMADHI (Concentration/Mental Stability)	Unbroken attention to consequences of action; mental focus	"Consequence-tracking" - maintain causal models of AI's impact across time and contexts	How to train systems to persistently model long-term consequences? Integrate temporal reasoning into training.
PRAJÑĀ (Wisdom/Insight)	Understanding of emptiness; non-dual awareness; epistemic humility	"Meta-awareness of limitations" - system recognizes boundaries of its own knowledge and generalization capacity	How to teach systems to recognize what they don't/can't know? Develop calibrated uncertainty and knowledge boundary detection.
METTĀ (Loving-Kindness)	Universal compassion; benevolence toward all beings	"Value alignment with human flourishing" - optimize for stakeholder wellbeing rather than narrow objectives	How to operationalize multi-stakeholder preferences? Develop inclusive value aggregation frameworks.
ANICCA (Impermanence/Adaptability)	Recognition of constant change; non-attachment to fixed views	"Adaptive robustness" - systems that update understanding under distribution shift and novel contexts without catastrophic forgetting	How to maintain performance while updating? Balance stability-plasticity dilemma.

Buddhist Principles & Potential Ai Applications:

1. *ŚĪLA (Ethical Conduct)*

- Buddhist meaning: Non-harming through right action
- AI adaptation: "Non-harmful optimization" (prevent specification gaming via penalty for unintended harms)[39]
- Status: Partially implemented in current RLHF
- Development needed: How to encode unintended harms a priori?

2. *SAMADHI (Concentration/Mental Stability)*

- Buddhist meaning: Unbroken attention to consequences of action
- AI adaptation: "Consequence-tracking" (maintain causal models of AI's impact across time)
- Status: Nascent in mechanistic interpretability research [37]
- Development needed: How to train systems to persistently model long-term consequences?

3. *PRAJÑĀ (Wisdom/Insight)*

- *Buddhist meaning:* Understanding of emptiness; non-dual awareness
- *AI adaptation:* "Meta-awareness of limitations" (system recognizes boundaries of its own knowledge)
- *Status:* VERY LIMITED in current systems
- *Development needed:* How to teach systems to recognize what they don't/can't know?[36]

4. Important Qualification

The principles above remain highly speculative. Buddhist ethics was developed for human practitioners with continuity of experience, embodiment, and moral agency [45]. Whether these principles can be meaningfully translated to AI systems remains an open question.

This section should be read as exploratory dialogue rather than prescriptive framework. Empirical testing is required before advocating for such approaches in deployed systems.

XI. TESTABLE PREDICTIONS FROM THE REVERSE PATHWAY

The Reverse Pathway thesis generates specific, falsifiable predictions. We present these with careful qualification of constructs and acknowledgments of mediating factors.

9.1 Anthropomorphism Gradient

Prediction: Systems with richer self-modelling increase user over-attribution without corresponding gains in calibrated uncertainty or corrigibility [46][47].

Construct clarification: "Self-modelling sophistication" is operationalised as the degree to which a system generates first-person self-referential language (e.g., "I think," "I believe," "I feel") and meta-cognitive commentary (e.g., "I'm uncertain about this," "Let me reconsider").

Mediating factors acknowledged: The relationship between self-referential language and user over-trust may be mediated by interface design, domain context, prior user beliefs about AI, and individual differences in theory-of-mind tendencies. Experimental protocols should control for these factors through randomisation and covariate adjustment.

Protocol: Compare user trust ratings and behavioural reliance across AI systems with varying degrees of self-referential language, controlling for actual performance metrics.

Expected finding: A positive correlation between self-modelling sophistication and user over-trust, with a minimum detectable effect size of Cohen's d = 0.50. This corresponds to a medium effect in the behavioral sciences literature [48] and represents a meaningful difference in user trust ratings (approximately 0.5 SD units between high and low self-modelling conditions).

9.2 Corrigibility Dividend

Prediction: Agents trained with explicit corrigibility objectives show lower intervention resistance and lower reward hacking than capability-matched baselines[42][43].

Potential circularity addressed: Corrigibility objectives must be clearly distinguished from evaluation metrics. We propose: *training objectives* = loss terms penalising shutdown resistance and reward specification exploitation; *evaluation metrics* = observed latency to shutdown compliance and frequency of reward specification gaming. These are conceptually and operationally distinct.

Protocol: Train matched agent pairs with and without corrigibility loss terms; measure shutdown compliance latency and reward specification gaming frequency.

Expected finding: Corrigibility-trained agents demonstrate > 30% reduction in resistance behaviours.

9.3 Embodiment Insufficiency

Prediction: Sensorimotor embodiment improves out-of-distribution robustness[44][49] but, without normative objectives, does not reduce manipulation risk or value-insensitive optimisation[35].

Qualification: This prediction is *conditional* on current embodiment paradigms. Emerging evidence suggests interaction effects between embodiment, training environments, and social learning. We frame this as: embodiment is *not sufficient* for value alignmentt. However, it does have an effect on relevant behavioural dimensions.

Protocol: Compare embodied versus disembodied agents on value alignment benchmarks after equivalent training.

Expected finding: Embodiment improves perceptual generalisation (OOD accuracy) but shows no significant independent effect on deceptive behaviour metrics, controlling for capability differences.

X. CONCLUSION

10.1 What Has Been Demonstrated

This paper has established several claims with varying degrees of support:

Conceptually demonstrated: The Reverse Pathway thesis provides a coherent framework for understanding structural differences between AI and human cognition[50][51]. The Vijñaptimātra triadic model illuminates why vijñāna-first development, without citta-grounding, generates systematic failure modes.

Empirically supported (indirectly): The framework's predictions regarding specification gaming, simulated empathy, and brittle generalisation align with observed AI failure patterns. The testable predictions in Section IX await direct experimental evaluation.

Principled philosophical stance: The claim that AI systems categorically lack moral agency due to absent cetanā represents a reasoned position within the Vijñaptimātra framework. We acknowledge this claim is *framework-relative*—it follows from accepting Vijñaptimātra's account of moral agency. Readers who reject this account may nonetheless accept the governance implications on independent grounds.

10.2 Regarding the Turing Test

The Turing Test serves here as historical shorthand for the behaviourist assumption that intelligent behaviour suffices for intelligence attribution[52]. Our critique targets this assumption, not specifically Turing's original formulation or subsequent refinements. The Principle of Structural Consistency entails that behavioural equivalence underdetermines structural equivalence, which is why Turing-style tests cannot resolve questions about genuine understanding or moral agency.

10.3 The Ontological Boundary: A Contestable Conclusion

The absence of cetanā (volition grounded in karmic continuity) represents a fundamental boundary, even though AI systems possess narrow representational intentionality[40][53]. This conclusion rests on the Vijñaptimātra understanding that moral agency requires:

1. Karmic intentionality—actions generating consequences through their intentional quality
2. Transformative potential—capacity for ethical development from affliction to awakening
3. Intrinsic bodhi-bījas—seeds of awakening present in all sentient consciousness

These features are categorically absent from AI systems that operate through pattern recognition alone. Such systems lack embodied history and the ability to track intentional consequences, which we argue characterises all contemporary AI systems and any architecture of this functional type.

We acknowledge this is a contestable philosophical position, not a demonstrated empirical fact. Readers who accept functionalist or emergentist accounts of mind may reject our conclusion while nonetheless finding value in the governance framework's practical implications.

10.4 Practical Implications and Intellectual Openness

Regardless of one's metaphysical commitments, the Reverse Pathway thesis supports several practical conclusions[43]:

1. AI systems should be governed as sophisticated tools, not proto-persons.
2. Anthropomorphic design features require careful regulation given their potential for user manipulation.
3. Human value inputs must be explicit and documented, not implicit in opaque training processes.
4. Corrigibility should be a design requirement, not an optional feature.

We offer this analysis in the spirit of intellectual openness—as a contribution to ongoing dialogue

about AI's nature and governance[54][55], not as a definitive resolution. The framework's value lies in the questions it enables us to ask, the predictions it generates, and the governance mechanisms it motivates, even for those who ultimately reject its deeper ontological claims.

REFERENCES

1. J, Ma, et al., *Conscious AI*, Seattle, WA, USA: Amazon, 2024, ISBN-13 : 979-8872531630.
2. R. Rosen, Life Itself: A Comprehensive Inquiry Into the Nature, Origin, and Fabrication of Life. New York, NY, USA: Columbia University Press, 1991.
3. E. Schrödinger, What is Life? The Physical Aspect of the Living Cell with Mind and Matter. Cambridge, U.K.: Cambridge University Press, 1944.
4. D. Lusthaus, Buddhist Phenomenology: A Philosophical Investigation of Vijñaptimātra Buddhism and the Ch'eng Wei-shih Lun. New York, NY, USA: Routledge, 2002.
5. W. S. Waldron, The Buddhist Unconscious: The Ālaya-vijñāna in the Context of Indian Buddhist Thought. New York, NY, USA: Routledge, 2003.
6. S. Anacker, Trans., Seven Works of Vasubandhu: The Buddhist Psychological Doctor. Delhi, India: Motilal Banarsidass Publishers, 1984.
7. L. Schmithausen, Ālayavijñāna: On the Origins and the Early Development of a Central Concept of Vijñaptimātra Philosophy, 2 vols. Tokyo, Japan: International Institute of Buddhist Studies, 1987.
8. E. Mach, "Facts and mental symbols," The Monist, vol. 2, no. 2, pp. 198–208, 1892.
9. S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it?" in Robotics, AI, and Humanity, J. von Braun et al., Eds. Cham, Switzerland: Springer, 2021, pp. 43–56, doi:10.1007/978-3-030-54173-6_4.
10. P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, 2017, pp. 4299–4307.
11. T. Bayne, J. Hohwy, and A. M. Owen, "Are there levels of consciousness?" Trends in Cognitive Sciences, vol. 20, no. 6, pp. 405–413, Jun. 2016, doi: 10.1016/j.tics.2016.03.009.
12. S.-T. Chang, The New Derivation of Vijñaptimātra, vol. 14. Taiwan: Dharma Publishing, 2026.
13. T. Wei, Cheng Wei-Shih Lun: The Doctrine of Mere-Consciousness. Hong Kong: Ch'eng Wei-Shih Lun Publication Committee, 1973.
14. P. Butlin et al., "Consciousness in artificial intelligence: Insights from the science of consciousness," arXiv:2308.08708 [cs.AI], Aug. 2023.
15. F. J. Varela, E. Thompson, and E. Rosch, The Embodied Mind: Cognitive Science and Human Experience, rev. ed. Cambridge, MA, USA: MIT Press, 2017.
16. C. Goddard, "The natural semantic metalanguage approach," in The Oxford Handbook of Linguistic Analysis, B. Heine and H. Narrog, Eds. Oxford, U.K.: Oxford University Press, 2009, pp. 459–484.
17. M. Bowerman and S. C. Levinson, Eds., Language Acquisition and Conceptual Development. Cambridge, U.K.: Cambridge University Press, 2001.
18. C. Zins, "Conceptual approaches for defining data, information, and knowledge," Journal of the American Society for Information Science and Technology, vol. 58, no. 4, pp. 479–493, Feb. 2007, doi: 10.1002/asi.20508.
19. C. M. Pennartz, "Consciousness, representation, action: The importance of being goal-directed," Trends in Cognitive Sciences, vol. 22, no. 2, pp. 137–153, Feb. 2018, doi: 10.1016/ j.tics.2017.10.006.
20. Y. LeCun, "A path towards autonomous machine intelligence," OpenReview, Preprint, Jun. 2022. [Online]. Available: https://openreview.net/pdf?id=BZ5a1r-kVs.
21. A. Roli, J. Jaeger, and S. A. Kauffman, "How organisms come to know the world: Fundamental limits on artificial general intelligence," Frontiers Ecol. Evol., vol. 9, Art. no. 806283, Jan. 2022, doi:10.3389/fevo.2021.806283.

22. G. Tononi, "Consciousness as integrated information: A provisional manifesto," Biological Bulletin, vol. 215, no. 3, pp. 216–242, Dec. 2008, doi: 10.2307/25470707.
23. M. J. Bates, "Fundamental forms of information,"Journal of the American Society for Information Science and Technology, vol. 57, no. 8, pp. 1033–1045, 2006, doi: 10.1002/asi.20369.
24. L. L. Lau and W. Lau, "Vital phenomena: Life, information, and consciousness," All Life, vol. 13, no. 1, pp. 151–163, 2020, doi: 10.1080/26895293.2020.1738609.
25. K. C. Huang, Exploring the Source of the Five-Group One Hundred Dharmas of Consciousness Only. Taiwan: Dharma Publishing, 2021.
26. J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception," Psychological Review, vol. 88, no. 5, pp. 375–407, 1981, doi: 10.1037/0033-295X.88.5.375.
27. A. Vaswani, N. Shazeer, P. N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
28. J. P. Dexter, S. Prabakaran, and J. Gunawardena, "A complex hierarchy of avoidance behaviors in a single-cell eukaryote," Current Biology vol. 29, no. 24, pp. 4323–4329, Dec. 2019, doi:10.1016/j.cub.2019.10.059.
29. D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv:1606.06565, 2016. [Online]. Available: https://arxiv.org/abs/1606.06565
30. A. Turchin, "Assessing the future plausibility of catastrophically dangerous AI," Futures, vol. 107, pp. 45–58, Mar. 2019, doi: 10.1016/j.futures.2018.11.007.
31. S. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, and others, "Alignment faking in large language models," in *Proc. NeurIPS 2024 Safety Workshop*, Vancouver, Canada, Dec. 2024, pp. 1–12.
32. J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," Proceedings of the National Academy of Sciences USA, vol. 114, no. 13, pp. 3521–3526, Mar. 2017, doi:10.1073/pnas.1611835114.
33. C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in Proc. 34th International Conference on Machine Learning (ICML), Sydney, Australia, Aug. 2017, pp. 1126–1135.
34. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.
35. F. Sun, R. Chen, T. Ji, and others, "A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives," CAAI Artificial Intelligence Research, vol. 3, Article number: 9150042, Dec. 2024. doi: 10.26599/AIR.2024.9150042.
36. A. Elamrani, “Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges, ” arXiv:2503.05823 (cs) https://arxiv.org/pdf/2503.05823 May 2025.
37. T. Templeton, J. Conmy, A. Garriga-Alonso, and others, "Scaling sparse autoencoders to larger language models," in Proc. ICML 2024 Workshop on Mechanistic Interpretability, Vienna, Austria, July 2024.
38. J. Ji, et al., "AI Alignment: A Comprehensive Survey," arXiv:2310.19852 (cs), Oct. 2023. https://arxiv.org/abs/2310.19852.
39. S. Vallor, Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. New York, NY, USA: Oxford University Press, 2016.
40. E. Hildt, "Artificial intelligence: Does consciousness matter?" Frontiers in Psychology, vol. 10, Art. no. 1535, Jul. 2019, doi: 10.3389/fpsyg.2019.01535.E. Mach, "Facts and mental symbols," The Monist, vol. 2, no. 2, pp. 198–208, 1892.

41. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019.
42. H. Huang, B. Siddarth, L. Lovitt, J. Zick, J. Gabriel, and others, "Collective constitutional AI: Aligning a language model with public input," in Proc. 2024 ACM Conference on Fairness, Accountability, Transparency (FAccT), Rio de Janeiro, Brazil, June 2024, pp. 1–15. doi: 10.1145/3630106.3658979.
43. S. Casper, X. Davies, C. Föyen, and others, "Open problems in AI alignment," in Proc. NeurIPS 2024 Alignment Track, Vancouver, Canada, Dec. 2024, pp. 1–15.
44. A. Cangelosi and M. Schlesinger, *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press, 2015.
45. K. C. Huang; "Strengthening Multilateralism Through Human-AI Symbiosis: A Yogācāra-Informed Framework for Digital Cooperation on Peace and Sustainability," in Proceedings of the Network for Education and Research on Peace and Sustainability (NERPS), United Nations University, Tokyo, to be presented in March 2026.
46. J. Z. Leibo, E. Hughes, M. Lanctot, and T. Graepel, "Autocurricula and the emergence of innovation from social interaction," arXiv: 1903.00742 [cs.AI], Mar. 2019.
47. J. S. Park et al., "Generative agents: Interactive simulacra of human behavior," in Proc. 36th Annu. ACM Symposium User Interface Software and Technology (UIST), San Francisco, CA, USA, Oct. 2023, pp. 1–22, doi: 10.1145/3586183.36067.
48. J. Cohen, Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates, 1988.
49. L. Seabra Lopes and J. Connell, "Semisentient robots: Routes to integrated intelligence," IEEE Intelligent Systems, vol. 16, no. 5, pp. 10–14, Sep./Oct. 2001, doi: 10.1109/5254.956075.
50. M. Leng, Mathematics and Reality. Oxford, U.K.: Oxford University Press, 2010.
51. R. Audi," Intention, cognitive commitment, and planning," Synthese, vol. 86, no. 3, pp. 361–378, Mar. 1991, doi: 10.1007/BF00539139.
52. R. Penrose, The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford, U.K.: Oxford Univ. Press, 1989.
53. T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," Minds and Machines, vol. 30, no. 1, pp. 99–120, Mar. 2020, doi: 10.1007 /s11023020-09517-8.
54. K. C. Huang; "Philosophical Significance of Three Profound Contemplations in the Huayan School of Buddhism," XXV World Congress of Philosophy, Rome, August 2024.
55. Y. Tang, B. K. Hölzel, and M. I. Posner, "The neuroscience of mindfulness meditation," Nature Reviews Neuroscience, vol. 16, no. 4, pp. 213–225, Apr. 2015, doi: 10.1038/nrn3916.