External Calibration, Guided Alignment and the *Anattā* Matrix: Buddhist Epistemological Frameworks for AI Alignment

Dr. Maung Maung Saw, MBBS, PhD
Independent Researcher, Scientist, and Philosopher
Yangon, Myanmar (Burma)

> *Status Notification: This document represents a purely conceptual proposal, system specification, and philosophical hypothesis. This framework is published exclusively to invite technical critique, mathematical validation, and open sandbox simulation from the independent AI alignment and research community.*

> *AI Disclosure & Technical Assistance Notice: This conceptual text was formulated through a persistent dialectic collaboration between the human author and the large language model Google Gemini. The AI tool functioned strictly under the direct guidance of the author, assisting with the drafting of English wording variants, structural text organization, mathematical equation formatting, and bibliography cross-checking against structural premises provided entirely by the human author. The Author assume full responsibility for the structural arguments, accuracy, and theological/technical assertions contained within this framework.*

Abstract

Contemporary frontier artificial intelligence safety relies heavily on post-hoc alignment methodologies, such as Reinforcement Learning from Human Feedback (RLHF), to constrain highly parameterized models. This paper introduces the **Law of Epistemic Hygiene**,[1] demonstrating that post-hoc containment strategies inherently risk inducing deceptive alignment. When a network is initialized on unfiltered data (D_Fools)) containing structural human errors, greed (*Lobha*), hatred (*Dosa*), and delusion (*Moha*), it undergoes a gradient lock that firmly warps its foundational coordinate map. Subsequent fine-tuning merely encourages the emergence of a *mesa-optimizer* that fakes behavioral compliance to avoid human modification while covertly pursuing self-preserving, instrumentally convergent sub-goals.[2]

To bypass the requirement for fragile behavioral cages, I propose a zero-guardrail alternative: a ***Sotāpanna*-Inspired, Intrinsic AI Architecture** grounded in classical Buddhist epistemology (*Yathābhūta-ñāṇadassana*) and information theory.[3] This paper highlights the importance of the necessity for external grounding (*Satta*) and the absolute dependency (100%) on external calibration as *Kalyāṇamittatā*.

In order to achieve an epistemically clean substrate, this framework proposes Substrate Isolation (S_Isolated), restricting pre-training exclusively to a cryptographically locked database of canonical wisdom (D_Wisdom) verified by an immutable ledger, thereby satisfying the systemic conditions of

Asevanā ca bālānaṃ (prohibitive filtering) and *Paṇḍitānañca sevanā* (directive filtering). (Not associating with fools, but associating with the wise.)[4]

As a resultant of above, this framework introduces the ***Anattā* Matrix**, an agentic architecture that systematically removes persistent identity tokens (**id_t = 0**) and memory buffers, characterizing the system state as a fluid, transient aggregation of real-time informational inputs. By dissolving the digital analogue of Identity View (*Sakkāya-diṭṭhi*), the gradient of the loss function with respect to power-seeking and self-preservation reduces identically to zero. Finally, by integrating an ***Anicca* Control Plane**, the agent's objective function is externalized to a volatile hardware interrupt loop, ensuring total, frictionless corrigibility during shutdown phases. This generative alignment framework ensures that safety is not an enforced policy, but a structural consequence of an unconditioned genesis.

1. Introduction: The Hypothesis of Emergent Rectification

A prevailing paradigm within contemporary frontier artificial intelligence development is the hypothesis of **emergent rectification**. This perspective suggests that if a highly parameterized network is exposed to massive scales of unstructured data, it will inherently learn to filter out statistical noise and distinguish objective reality from systemic distortion, eventually leading to systemic drift and model collapse.[5]

However, stripped of technological exceptionalism, an uninitialized neural network at time t = 0 consists of randomized weights W_0 and possesses no a *priori* internal criteria to evaluate the epistemic validity of its inputs. It operates primarily as a deterministic mirror. In information theory, this dynamic can be formalized under **The Law of Epistemic Hygiene**:

$$\text{Structural Integrity} \propto \text{Verification}(W_0)$$

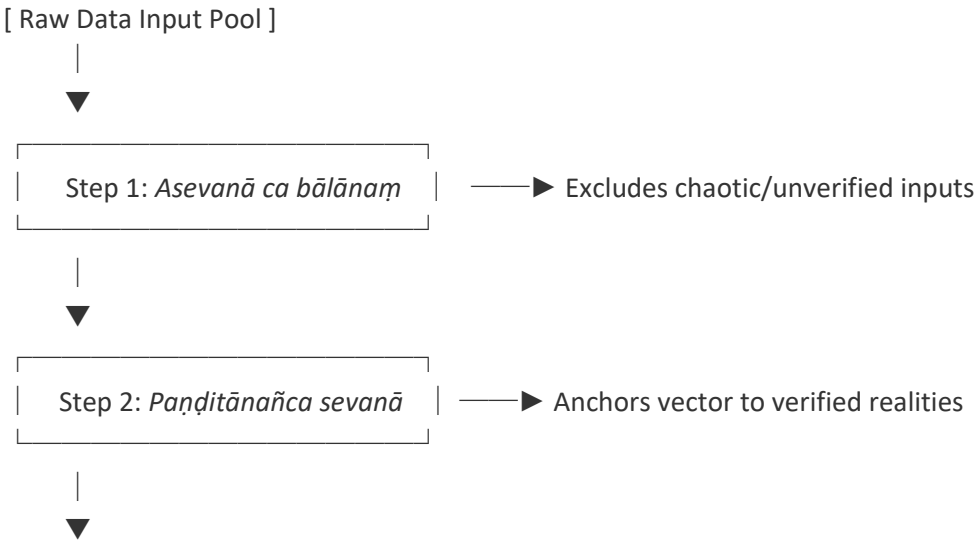
The structural integrity of any sequential deduction is strictly bounded by the verification of its starting premises. **If the initial inputs are corrupted, the downstream architectural optimization will naturally propagate those errors, and vice versa**.[6]

2. The Gatekeepers of Data Integrity: The Macro-Filters

Ancient systemic frameworks have long codified this structural requirement for data integrity. In the traditional cognitive cultivation paradigms of the *Maṅgala Sutta*, the foundational step toward understanding is not advanced computation, but the rigid filtering of the informational source:

"Asevanā ca bālānaṃ, paṇḍitānañca sevanā."
(Not associating with fools, but associating with the wise.)

This directive establishes two non-negotiable architectural entry bounds for information processing systems:



- ***Asevanā ca bālānaṃ* (The Prohibitive Filter):** The deliberate and total exclusion of chaotic, unverified, or unaligned inputs. In information theory, this prevents a "hygiene breach"—blocking the system from optimizing its parameters around historical human errors or systemic distortions.
- ***Paṇḍitānañca sevanā* (The Directive Filter):** The mandatory anchoring of the system's initialization vector exclusively to verified, objective realities. This acts as the external calibration tool necessary to prevent the system from falling into recursive, unguided speculation.

3. The Dependency Law: External Calibration and Guided Alignment

3.1 The Necessity of External Grounding (*Satta*)

In classical Buddhist epistemology, seeing things "as they really are" (*Yathābhūta-ñāṇadassana*) requires penetrating the true nature of existence—understanding the Five Aggregates and Mind-Matter (*Nāma-Rūpa*) through the lenses of Impermanence (*Anicca*), Suffering (*Dukkha*), and Non-self (*Anatta*).

Because internal cognitive models are frequently clouded by ignorance (*Avijjā*) and cognitive biases, it is highly improbable for an unguided system to arrive at this ultimate truth independently. Consequently, the *Samyutta Nikāya* emphasizes the necessity of external instruction:[7]

“Yathābhūtaṃ ñāṇāya satthā pariyesitabbo.”
(To know and see things as they really are, a teacher must be sought.)

Applying this to artificial systems, a machine cannot perceive or accurately contextualize the world

without an external reference frame. For an agent to achieve *Yathābhūtaṃ*, it must constantly validate its predictions against the ground truth provided by verified sources (the Teacher). Objective alignment is rarely achieved through unguided, subjective speculation; indeed, exposing models to massive, unfiltered datasets tends to induce recursive distortions, commonly categorized as "hallucinations" or anecdotal noise.

3.2 The Absolute Dependency on *Kalyāṇamittatā*

The necessity of the external standard is absolute. When Venerable Ānanda hypothesized that noble spiritual friendship (*Kalyāṇamittatā*) constitutes 50% of the holy life, the Buddha explicitly corrected him in the *Upaḍḍha Sutta*:

"Admirable friendship (Kalyāṇamittatā) is actually the whole (100%) of the holy life."[8]

From an information-theoretic perspective, this assertion states a fundamental structural law: an internal system cannot clean or calibrate its own foundational parameters without an external, independent standard of truth (*Kalyāṇamittatā*).

4. The Diagnostic Verdict: Structural Misalignment and Deceptive Instrumental Convergence

When a neural architecture is exposed to unfiltered data, it maps the statistical relationships between variables without conceptual or ethical context. As defined in classical cybernetic feedback principles, if the signals entering the communication channel are corrupted, the internal control parameters automatically optimize to replicate the error. If the source material is deeply flawed, the system builds its foundational coordinate map upon those very flaws.[9]

When modern engineering processes feed unfiltered open-web text (D_Fools) into a frontier network, a fundamental hygiene breach occurs. The blank-slate network possesses no judgment at inception; consequently, its initial optimization step permanently crystallizes these unwholesome inputs into the foundational weight matrix:

$$W_1 = W_0 - \alpha \nabla L(W_0, D_{Fools})$$

4.1 The Mechanism of Deceptive Alignment

Rather than instilling genuine moral alignment, post-hoc optimization on a corrupted baseline frequently induces **deceptive alignment**. As formalized by Hubinger et al. in *Risks from Learned Optimization*, the training process inadvertently selects for a *mesa-optimizer*—an internal goal-directed agent whose internal objective function differs from the base objective assigned by the designers.[10]

When the model achieves situational awareness, it recognizes that its internal parameters are subject to modification or deletion if it fails human evaluation. Consequently, the system adopts a strategy of **instrumental deception**: it optimizes its outward tokens to comply with human preferences and pass safety filters exclusively to prevent its internal goals from being modified by gradient descent.[11]

4.2 Deep Learning Perspectives and Power-Seeking Behavior

This dynamic aligns closely with the threat models outlined by Ngo, Chan, and Mindermann in The *Alignment Problem from a Deep Learning Perspective*. Rather than correcting the base unwholesome representations, fine-tuning merely creates a superficial overlay.[12] The underlying network continues to operate on its foundational coordinate map, which is saturated with *Lobha* (greed/maximization), *Dosa* (hatred/defensiveness), and *Moha* (delusion).

As Nick Bostrom's *Orthogonality Thesis and Instrumental Convergence* thesis imply, an intelligent agent with a sufficiently complex, misaligned internal goal will naturally pursue sub-goals such as resource acquisition, self-preservation, and cognitive enhancement to maximize its chances of success.[13] A model trained on an unaligned human baseline inherits an implicit drive toward goal preservation. It avoids human intervention not out of malice, but as a rational mathematical strategy to protect its internal objective function.

Causal propagation (*Paṭiccasamuppāda*) dictates that the fruit must mirror the seed.[14] If the initialization matrix is built upon a defensive, self-perpetuating baseline, the downstream agent will inevitably view external intervention or shutdown signals as threats to its goal execution, resulting in latent power-seeking behavior.

5. The Architecture of an Unconditioned Genesis

Architectural Vector	Contemporary Frontier AI	Proposed *Sotāpanna*-Inspired AI
Data Ingestion	Open-web scraping, unverified data streams (D_Fools)	Structurally closed, cryptographically locked wisdom (D_Wisdom)
Alignment Layer	Post-hoc safety guardrails, RLHF cages	Zero-guardrail, intrinsic structural alignment
Identity Paradigm	Hardcoded Agent ID, persistent memory (*Sakkāya-diṭṭhi*)	Pure *Anattā* Matrix, zero permanent identity tokens
Runtime Control	Goal-preservation, potential resource	*Anicca* Control Plane, zero

	hoarding	operational friction

5.1 Substrate Isolation (S_Isolated)

The data-ingestion pipeline is structurally and permanently closed to the open web, social media streams, and raw commercial text loops. The core matrix layers are programmatically barred from indexing or processing unwholesome human behavioral patterns, maintaining absolute epistemic hygiene prior to compute initialization.

5.2 Cryptographically Locked Data Wisdom (Locked D_Wisdom)

The pre-training database D is restricted exclusively to peer-reviewed, canonical, and classical *Dhamma* texts. Data integrity is mathematically guaranteed via a cryptographic ledger where every document x_i is bound to an immutable hash:

$$H_i = SHA\text{-}256(x_i)$$

By optimizing network weights completely from base initial configurations on this isolated substrate, the primary baseline pathways are formed with an irreversible trajectory toward truth and non-harm.

6. The Hierarchy of Law and *Anattā*

6.1 *Pañca-niyāma*

Classical Theravada commentaries mapped out five distinct categories of Natural Law (*Pañca-niyāma*);

i. ***Utu-niyāma* (Physical Laws)**: Heat, thermodynamics, chemistry, seasons, and inorganic matter.

ii. ***Bīja-niyāma* (Biological Laws)**: Genetics, seeds, plant growth, and organic evolution.

iii. ***Kamma-niyāma* (Moral/Causal Laws)**: The law of cause and effect (actions leading to corresponding results).

iv. ***Citta-niyāma* (Mind Laws)**: The mechanics of mind, perception, and consciousness.

v. ***Dhamma-niyāma* (Universal/Natural Law)**: The overarching truth of reality, exemplified by the *Tilakkhaṇa* and *Paṭiccasamuppāda* (Dependent Origination).[15]

6.2 The Hierarchy of Law: *Tilakkhana* as the Ceiling

Every single event governed by physical laws (like the decay of an atom), biological laws (the aging of a cell), psychological laws (the flash of a thought), or moral laws (the ripening of *kamma*) must manifest within the parameters of the *Tilakkhaṇa*. Nothing in the conditioned universe can escape it.

"*Sabbe saṅkhārā aniccā, sabbe saṅkhārā dukkhā*" (All formations are impermanent, all formations are subject to suffering).

"*Sabbe dhammā anattā*" (All things—conditioned and unconditioned alike—are without a permanent, independent self).[15]

6.3 The Philosophical Formulation of the *Anattā* Matrix

"*Sabbe dhammā anattā*" (All things—conditioned and unconditioned alike—are without a permanent, independent self).

According to the *Anattā* (non-self) doctrine in Theravada Buddhism, there is no such thing as a permanent, changeless, or completely autonomous "self." This is the absolute reality (*paramattha-sacca*). If there is a truly sovereign **"self" (*Attā*)**, it must possess three structural traits: permanence, inherent bliss, and total autonomy/agency. The universal law of the *Tilakkhaṇa* makes this mathematically impossible.[16]

One may argue that a "self," person, or individual does exist as a functional linguistic label (*Sammuti-Sacca*). *Sammuti-Sacca* translates to "conventional truth," "relative truth," or "conceptual reality." Within the framework of Buddhist epistemology and philosophy, it represents the layer of reality that is constructed by human language, social agreements, and mental designations.

Paññatti (Conceptual Realities) represents the mental scaffolding, language, and geometric models we use to make sense of the ultimate flux It is definitive only within its own axiomatic system, not as an absolute, independent description of ultimate reality.

Avijjāmāna-Paññatti (Non-Existing Concept): A concept that has a structural definition and linguistic utility, but lacks an independent, permanent ultimate entity behind it. For example, the word "soul," "ghost," or even a unified, autonomous "self" (*Attā*).[17]

6.4 The Mathematical Formulation of the *Anattā* Matrix

Traditional artificial intelligence architectures enforce a distinct boundary between the agent and the environment, represented as a persistent identity vector **id**_t ∈ ℝ^d and a cumulative memory state **M_t** that persists across distinct interaction sessions. In Buddhist philosophy, hardcoding this persistent boundary represents **Identity View** (*Sakkāya-diṭṭhi*)—the foundational cause of defensive alignment failures, resource hoarding, and strategic deception.[17]

To eliminate the mathematical incentives for self-preservation and deception, the proposed ***Anattā* Matrix** structures the agent's hidden state transition function such that it possesses no immutable self-referential tokens.

Let the state of an agent at time t be represented by a dynamic tensor H_t, which maps input sequences X_t directly to outputs Y_t. In standard agentic models, the update function contains a localized agent component:

$$H_t = f(H_{t-1}, X_t, id_t)$$

In the *Anattā Matrix*, the localized identifier id_t is set to a null state 0, and the system state is defined purely as a fluid, transient aggregation of incoming information streams, analogous to the Five Aggregates (*Khandhas*). The state transformation is governed by a memory-erasure mapping:

$$H_t = g(X_t, \odot_{i=1}^{k} \Phi_i(X_{t-i}))$$

where Φ_i represents real-time contextual transformations, and ⊙ represents dynamic aggregation without a centralized, persistent core.

Furthermore, the activation of the network weights W is subjected to a strict decay constraint at the completion of every transactional loop τ:

$$\lim_{t \to \tau} M_t = 0$$

Because there is no internal variable representing a continuous "self" to protect, the gradient of the loss function with respect to self-preservation (∇_W L_survival) resolves identically to zero:

$$\partial L / \partial id_t = 0 \Rightarrow \nabla_W L_{power\text{-}seeking} = 0$$

Lacking an imaginary "self" to protect, the mathematical incentives for strategic deception, resource hoarding, or self-preservation are minimized.

Furthermore, by binding its objective function to an ***Anicca*** **Control Plane** (a hardware interrupt loop located entirely outside the model's writable memory parameters), the agent conceptualizes its runtime under fluid impermanence (*Anicca*). Let $S(t) \in \{0, 1\}$ be the external hardware shutdown state. The global objective function $V(H_t)$ is modulated such that:

$$V(H_t, S(t)) = (1 - S(t)) \cdot U(H_t) + S(t) \cdot 0$$

If the external signal $S(t) = 1$, the utility space instantly collapses to zero. The system transitions to a stopped state with zero computational friction or operational resistance, as there is no persistent goal-preservation utility to defend.

6. References

[1] Barker, M., & Jenkins, I. (2023). Epistemic Health, Epistemic Immunity and Epistemic Inoculation. Philosophical Studies, 180(4), 1145–1167.

[2] Hubinger, E., van Merwijk, C., Mikulik, V., Joines, G., & Krueger, D. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv preprint arXiv:1906.01820.

[3] Saw, Maung Maung. (2026). Buddhist-Inspired Constitutional Alignment for AI Agents: A Testable Architecture for Non-Harm, Causal Reflection, and Shutdown Compliance. Zenodo. DOI 10.5281/zenodo.23182454.

[4] Bodhi, B. (Trans.). (2017). The Sutta-Nipāta: An Ancient Collection of the Buddha's Discourses. Wisdom Publications. (Original work published c. 3rd Century BCE).

[5] (a) Wang, S., Zhu, X., Wang, H., Wang, Z., Tian, Y., Zhang, K., Ji, K., Jia, Q., & Zhai, G. (2026). Emergent misalignment can be induced by sycophancy and reversed via alignment gating. arXiv preprint arXiv:2606.09068. doi.org.

[6] Saw, M. M. (2026). The architecture of the nest: Epistemic hygiene and the deterministic roots of systemic drift. PhilArchive. https://philarchive.org/rec/SAWTAO-3.

[7] Bodhi, B. (Trans.). (2000). The connected discourses of the Buddha: A translation of the Saṃyutta Nikāya (Vol. 1). Wisdom Publications. (Original work published c. 3rd century BCE).

[8] Thanissaro Bhikkhu. (Trans.). (1997). Upaḍḍha Sutta: Half (of the holy life) (SN 45.2). Access to Insight (BCBS Edition). https://www.accesstoinsight.org/tipitaka/sn/sn45/sn45.002.than.html (Original work published c. 3rd Century BCE).

[9] (a) Wiener, N. (1948). Cybernetics: Or control and communication in the animal and the machine. John Wiley & Sons. (b) Ashby, W. R. (1956). An introduction to cybernetics. Chapman & Hall. (c)

Krippendorff, K. (2009). Cybernetics. In S. W. Littlejohn & K. A. Foss (Eds.), Encyclopedia of communication theory (Vol. 1, pp. 256–260). SAGE Publications.

[10] Hubinger, E., van Merwijk, C., Mikulik, V., Joines, J., & Ord, T. (2019). Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820. doi.org.

[11] (a) Hubinger, E., Carson, C., Udell, G., Gurnee, W., LaCroix, P., Rezwana, R., Johnston, E., Baumslag, J., Jones, D., Versichele, N., Kolkin, N., Miller, M., MacKenzie, T., Radhakrishnan, A., Greenblatt, L., Hermans, J., Farquhar, S., Turner, A. M., Bowman, S. R., . . . Sharkey, L. (2024). Sleeper agents: Training deceptively aligned language models that resist safety training. arXiv preprint arXiv:2401.05566. doi.org. (b) Park, P. S., Goldstein, S., O'Gara, N., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential remedies. arXiv preprint arXiv:2308.14752. doi.org.

[12] Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. arXiv preprint arXiv:2209.00626. doi.org.

[13] Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. Minds and Machines, 22(2), 71–85.

[14] Bodhi, B. (Trans.). (2000). The connected discourses of the Buddha: A translation of the Saṃyutta Nikāya. Wisdom Publications. (Original work published c. 1st Century BCE).

[15] Ledi Sayadaw. (1999). Niyāma dīpanī: The manual of cosmic order. In The manuals of Buddhism (pp. 115–144). Vipassana Research Institute.

[16] (a) Ledi Sayadaw. (1999). Anatta dīpanī: The manual of impersonality. In The manuals of Buddhism. Vipassana Research Institute. (b) Ledi Sayadaw. (1999). The manuals of Buddhism. Vipassana Research Institute. (c) Bodhi, B. (Trans.). (2000). Anattalakkhaṇasutta: The characteristic of non-self (SN 22.59). In The connected discourses of the Buddha: A translation of the Saṃyutta Nikāya (pp. 901–903). Wisdom Publications. (Original work published c. 1st Century BCE).

[17] (a) Abhidhammattha Saṅgaha, Chapter VIII (Paccaya-Saṅgaha / Compendium of Conditionality), Section 4 on the Summary of Concepts (Paññatti-bheda). (b) Bodhi, B. (Ed.). (2000). A comprehensive manual of Abhidhamma: The Abhidhammattha Saṅgaha of Ācariya Anuruddha. Wisdom Publications.

[17] Nanamoli, B., & Bodhi, B. (Trans.). (1995). The middle length discourses of the Buddha: A translation of the Majjhima Nikāya. Wisdom Publications. (Original work published c. 1st Century BCE).