Raman, not Varelse — Why Acceptably Controllable AGI Policies Cluster Around Human-Like Stacks

Trevor Buteau; Steve Kommrusch

Independent; Independent

Abstract. We argue that under realistic budgets, verifiably alignable AGI policies cluster around communicable ("raman") rather than opaque ("varelse") architectures. Alignment verification has a cost: an auditor must understand the system well enough to evaluate it. We define *epistemic distance* as this cost — auditor training time plus per-evaluation effort — and argue it shrinks monotonically with architectural legibility. Over Perrier's Alignment Control Stack, lower-layer opacity compounds upward, so interpretability bottlenecks at any layer trace to opacity beneath them. A negative argument (combinatorial cost under opacity) bounds the feasible-alignment set; a positive argument (the existence of human civilization) establishes it is nonempty. Canonical x-risk pessimism rests on an unexamined varelse assumption; identifying it converts an impossibility result into a design constraint. We synthesize active inference's alignment-as-preference-overlap with Perrier's assurable controllability, linked by epistemic distance, and translate the framework into layer-by-layer design heuristics supplemented by a bidirectional diagnostic: the Raman test.

Keywords: AGI alignment · epistemic distance · active inference · alignment control stack · raman · varelse · controllability

1 Introduction: Alignment Under Control Limits

1.1 Making "AI Alignment" Operational

We adopt Perrier's Alignment Control Stack [1], framing "alignment" as a relational, long-horizon closed-loop control problem in which some set of humans HH attempts to keep a capable machine agent A within a desirable set of states and trajectories, using physical, technical, and social constraints. Alignment is a property of a *relationship* between agents: an AI is aligned with a specified set of human stakeholders under specific conditions of oversight and deployment, never "aligned" in isolation.

When alignment-with-HH is defined as assurable controllability of A by HH under realistic budgets, alignable solutions cluster around what we call **raman** rather than **varelse** architectures. Raman agents are sufficiently communicable

and legible that their behavior can be modeled, predicted, and corrected using tools HH already possesses or can feasibly develop; varelse agents are so opaque that meaningful mutual understanding is unavailable at any affordable cost. Our claim is not that raman agents are safe or virtuous by default—human agents are maximally raman to each other and still wage war, deceive, and exploit. The claim is narrower: **ramanization increases the feasible set of alignment mechanisms that can be justified, audited, and maintained over time, by lowering the epistemic distance between HH and A.** Ramanization is a constraint on interface, not on capability: a system with superhuman performance but human-legible policy structure is raman, and a system with human-level capability but opaque policy structure is varelse.

1.2 Positioning and Contributions

The connection between transparency and controllability is not new. Mechanistic interpretability [24] addresses it at the behavioral and representational layers. Work on corrigibility [25,26] addresses it at the reward layer. Safe reinforcement learning [27] addresses it at training. Each targets a specific layer of the control stack; none provides a unified account of how transparency requirements compose across layers. We engage with the existential risk ("x-risk") literature: Bostrom [17], Yudkowsky and Soares [18], and Yampolskiy [19] derive powerful pessimistic conclusions that we argue depend on a specific architectural premise (Section 2.5).

This paper makes five contributions: (1) a formal framework synthesizing alignment as prior preference overlap (from active inference) with assurable controllability (from Perrier), linked by epistemic distance as verification cost (Section 2); (2) a propagation heuristic asserting that lower-layer opacity compounds upward (Section 2.4); (3) a clustering argument combining a negative result (combinatorial verification cost) with a positive result (civilizational existence proof) establishing that the feasible-alignment set is nonempty but concentrated around human-like architectures (Section 3); (4) layer-by-layer design heuristics with active inference as a running example (Section 4); and (5) empirical measurement proposals including bidirectional theory-of-mind diagnostics (Section 5).

1.3 A Note on Terminology

The terms "raman" (RAH-mahn) and "varelse" (VAH-rel-seh) are borrowed from Orson Scott Card's *Speaker for the Dead* [16], where they denote levels in a hierarchy of foreignness: raman names an Other sufficiently communicable for meaningful coexistence; varelse names an Other so opaque that mutual intelligibility fails. Card's key insight (emphasized in Malmgren's [12] literary analysis) is that the boundary is partly epistemic: it reflects the observer's capacity for understanding as much as the observed's nature. We adopt the terms because they compress into single words the cost of verifying one agent's alignment by another.

Readers uncomfortable with literary vocabulary may substitute "low-epistemic-distance" for "raman" and "high-epistemic-distance" for "varelse" throughout.

2 Formal Framework: Alignment as Assurable Preference Overlap Under Epistemic Constraints

2.1 Joint Dynamics and Policy Space

Let S denote the state space of the joint system (A, HH, E), where A is the machine agent, HH a stakeholder set, and E the shared environment. A policy $\pi : \mathcal{H} \to \Delta(\text{Act})$ maps observation histories to distributions over actions. Let HH's normative policy be π_H and A's realized policy be π_A. The space of all such policies is $\mathcal{P}$.

2.2 Alignment as Assurable Preference Overlap

Active inference [5] supplies the content of what HH verifies. Under the free energy principle, an agent acts to minimize expected free energy relative to prior preferences—a distribution C over preferred observations encoding what the agent treats as desirable. The alignment target is sufficient overlap between C_A and C_{HH} under some divergence measure, with threshold set by HH.

Since HH is a stakeholder set rather than a single agent, C_{HH} reads more accurately as a family of priors held by different stakeholders than as a single distribution. Ramanization turns this from limitation into feature: when C_A and A's policy structure are legible, distinct stakeholders can verify overlap against their own priors. A parent and a regulator evaluating the same system need neither agree on C_{HH} nor share an auditor; each audits C_A against their own normative commitments. Varelse systems collapse this plurality by forcing every stakeholder through the same specialist interpreter.

Alignment is the maintained condition in which overlap stays above a threshold acceptable to HH, sustained through a closed-loop control relationship whose three operations are (i) observation—HH estimates C_A; (ii) verification—HH evaluates A; and (iii) correction—HH intervenes when overlap falls below threshold. Perrier's Alignment Control Stack organizes the physical, technical, and institutional infrastructure across which these operations are distributed. Alignment is thus a dynamic property of the (A, HH) relationship, requiring ongoing expenditure to maintain.

Two agents may share prior preferences yet diverge in policy due to differing generative models: distinct causal maps of how the world works produce distinct action sequences even when goals coincide. Such divergence is a target for the controllability loop—the observation step detects policy-level drift, the correction step acts on it.

2.3 Epistemic Distance and the Controllability Bound

Prior preference overlap is not directly observable—the observation step in the control loop has a cost. An auditor must understand enough of A's architecture to determine what C_A is and whether it matches C_{HH}. We operationalize that cost as epistemic distance: $d_E(HH, A) = T_{\text{train}} + T_{\text{task}}$, where T_{train} is the minimum training required to produce an auditor competent to evaluate system A, and T_{task} is the time for that auditor to produce a correct causal account of a specific output or behavioral episode.

A system whose internal states map onto concepts the auditor already possesses minimizes T_{train}: the training was paid during ordinary cognitive development, which is why humans are cheaply auditable by other humans. The core claim: controllability presupposes observability, and observability has a price. Feasible alignment—alignment verifiable under budget B—requires $d_E(HH, A) \leq B$. Since d_E depends on the architecture that realizes a policy rather than the policy alone, we define the feasible-alignment set over architecture–policy pairs: $\mathcal{F}(B) = \{(A, \pi_A) : d_E(HH, A) \leq B\}$. The same behavioral π_A may admit multiple implementations with different auditability profiles; ramanization selects implementations whose verification cost falls within budget. $\mathcal{F}(B)$ shrinks with architectural opacity, because opacity raises the floor on both components: alien architectures demand more auditor training and make each evaluation harder.

2.4 Stack Structure and the Propagation Heuristic

Following Perrier [1], decompose the control relationship into ten ordered layers: physical hardware (L_1), system software (L_2), AI framework (L_3), model architecture (L_4), training (L_5), behavioral output (L_6), interpretability (L_7), rewards and value alignment (L_8), multi-agent dynamics (L_9), and sociotechnical governance (L_{10}). Each layer L_i admits its own epistemic distance d_E^i, decomposable into T_{train}^i and T_{task}^i.

Verification at layer L_i requires reasoning over the hypothesis space of internal states consistent with observed behavior. Legibility at L_i constrains that space structurally; behavioral testing from above samples it. When the space is small, sampling suffices and upper-layer channels substitute for lower-layer legibility. An aircraft is verified behaviorally without understanding transistor physics because its flight envelope is bounded and pre-characterized. When the space is large, sampling cost grows combinatorially (Section 3 makes this bound precise). Opaque lower layers force upper-layer verification to work by sampling, at costs that scale with the space they cannot see into.

This yields the propagation heuristic: verification effort at any layer eventually runs into the floor set by what is legible beneath it. An auditor's evaluation of whether training (L5) produced the desired dispositions is constrained by their understanding of what the architecture (L4) can represent, and also by their knowledge (or lack thereof) of what the lower level substrate (L1–L3) makes computable. Independent verification channels at higher layers (like behavioral

testing (L6), psychometric profiling) reduce T_{task} without reducing T_{train}, and do so at sampling costs that scale with the unobserved space below.

The design prescription follows directly: when T_{task} at L_i exceeds budget, the lever is T_{train} at L_{i-1}. Verification walls do not easily yield to more effort at the layer where they appear; they yield to legibility at the layer beneath. The opacity frontier moves down when the layer beneath it is ramanized, and not before.

How deep the heuristic runs is a question of stakes. An aircraft is verified behaviorally because its failure modes are bounded—the flight envelope is narrow, the consequences of a crash are local. A web page recommendation engine never requires L3 or L2 legibility because the cost of recommending the wrong page does not justify the investment. Necessary verification depth scales with what is at risk above, and the floor to which the propagation heuristic must be run is set by the cost of getting the upper-layer answer wrong. Systems whose plausible failure modes include existentially catastrophic outcomes ("x-risk") face no bound on that cost. Digging to whatever depth turns the upper-layer verification problem from intractable to tractable is, at that point, the cheap option.

2.5 The Varelse Assumption: X-Risk Pessimism as an Architectural Conditional

The canonical x-risk arguments derive their pessimistic conclusions from architectural premises that describe one region of design space and leave the rest unexamined. Bostrom's [17] treacherous turn requires that HH cannot detect divergence between stated and internal goals — that T_{train} at L_8 is prohibitive, which fails when goal architecture is inspectable. Yampolskiy's [19] containment arguments extend the same logic: containment difficulty scales with T_{train}.

Yudkowsky and Soares [18] present the most detailed version. Their analysis is grounded in the properties of systems grown through opaque optimization like backpropagation on von Neumann hardware, billions of uninterpretable parameters, and training temporally compressed and disconnected from deployment. These are systems whose opacity frontier sits at the top because every layer beneath L9 is dark. Their conclusions about such systems may well be correct. What they do not examine is whether the same conclusions hold when the frontier moves—when any interior layer becomes legible enough that verification at the layer above stops paying combinatorial costs. If their conclusions are correct, the stakes are exactly those that justify running the propagation heuristic to its floor. Converting a general impossibility into a conditional whose antecedent is architectural (and addressable at whatever depth the stakes require) is itself progress.

3 The Clustering Argument

3.1 Negative Argument: Opacity as Default

The x-risk literature has argued at length that verifying alignment in opaque systems is prohibitive [17,18,19]. Section 2.5 recast those arguments as condi-

tional on high T_{train} at lower layers; read in that register, they establish the negative result directly. Under architectural opacity, verification cost exceeds realistic budgets, and the feasible-alignment set under any such budget is a strict and small subset of $\mathcal{P}$.

The standard rejoinder—that real systems carry inductive biases rather than being unconstrained—concedes the thesis. Such biases are the structural constraints that pull T_{train} below budget [20,21,23]. The question becomes which constraints simultaneously minimize auditor training cost and preserve capacity for general intelligence. The example of human civilization (Section 3.2) identifies one promising answer.

3.2 Positive Argument: The Human-Like Cluster Exists

Consider human civilization as a multi-agent system. The negative argument establishes that verification in opaque systems is intractable; the civilizational record establishes that verification among mutually legible agents is *tractable enough to sustain coordination at scale*. The positive argument does not require that human alignment is perfectly good, succeeds constantly, or extends to AGI-level capability differentials. It requires only that mutual verification at human-to-human epistemic distance is solvable at bounded cost—enough to sustain institutions, contracts, and corrective intervention—which the persistence of civilization establishes. Wars, exploitation, deception, and institutional failure are compatible with this weaker claim; the relevant contrast is continuation under bounded verification cost, not perfection. Whether the same tractability extends to human–AGI verification at AGI capability is a question the framework opens, not one it closes.

Three structural features carry the tractability. *Legibility*: human policies are partially observable through language, behavior, expression, and institutional transparency mechanisms— evolved capacities for social cognition and communication [28,29]. *Structural homology*: shared developmental, perceptual, and motivational architecture means T_{train} for cross-agent verification is approximately zero, paid during ordinary cognitive development. In active inference terms, agents sharing generative model structure reduce mutual inference from unconstrained search to parameter estimation within a known model class. *Institutional amplification*: law, norms, monitoring, and reputation are calibrated to the epistemic distances characteristic of human-to-human interaction [30], and the calibration has demonstrably improved—coordination problems intractable in 1600 are routinely solved in 2025.

Section 4 traces the same three features down the Alignment Control Stack, showing at each layer what human cognitive architecture contributes to tractability and what current AI architectures omit. The civilizational existence proof establishes feasibility within the raman region; building coordination infrastructure for artificial raman agents is the engineering problem it leaves open.

A complementary existence proof comes from cross-species ramanization. Domesticated dogs and wolves share most of their cognitive architecture, yet humans predict dog behavior at low cost while wolves remain largely opaque: dogs

read human communicative gestures that hand-raised wolves do not [57], and dog–human mutual gaze drives an oxytocin feedback loop absent in wolves [58]. The difference is not raw intelligence but selected architectural alignment with human social cognition, established over at least 15,000 years of domestication [59]. Two points follow. First, ramanization does not require human identity—a dog is not a small human, but its trajectories are mappable into human social coordinates. Second, ramanization can be *induced* through structured selection pressure: Belyaev's farm-fox experiment produced human-directed social behavior by selecting on tameness alone [60]. The question for AGI is whether analogous pressures can be applied during design and training.

3.3 The Design Implication

The feasible-alignment region under budget B is nonempty (by the existence proof) but small relative to $\mathcal{P}$ (by the negative result). The prescription: architectures whose internal representations, decision processes, and communication modalities are structurally similar enough to human cognition that T_{train} stays within institutional budgets. "Structurally similar" means similarity that lowers d_E, not functional identity; ramanization is a continuum, not a binary; and the prescription is defeasible — any architecture achieving bounded d_E qualifies as raman regardless of resemblance to human cognition. The human-like cluster is the existence proof, not the only theoretical possibility.

4 Descending the Stack

The formal results of Sections 2–3 yield a design directive: reduce epistemic distance across the Alignment Control Stack. We present the stack top-down. At each layer, ramanization makes answering specific questions about the system tractable—ones that varelse architectures leave unanswerable, and whose answer depends on the layer beneath.

Consider a scenario. An AGI system has caused harm and says: *"I'm sorry. It was an accident and I feel terrible about it. I care for you deeply, and I never want to hurt you like this again. I'm willing to accept the consequences of my actions, and I hope you'll give me the opportunity to make things right."* A human saying this can be evaluated imperfectly but substantively. Can an AGI?

4.1 L10: Sociotechnical governance.

Governance is cooperation infrastructure. Moral codes, norms, and institutions transform interaction from prisoner's dilemma to assurance game, where both parties do best cooperating provided each can verify the other's cooperative intent [30,11,47,53]. This is the alignment control loop of Section 2 at institutional scale: observe behavior, compare to expectations, intervene on deviation.

Current AI governance [13] closes this loop around the system's developers and operators, not the system itself—because the system offers nothing for governance to get traction on. The AI is treated as a minor: a ward whose guardians

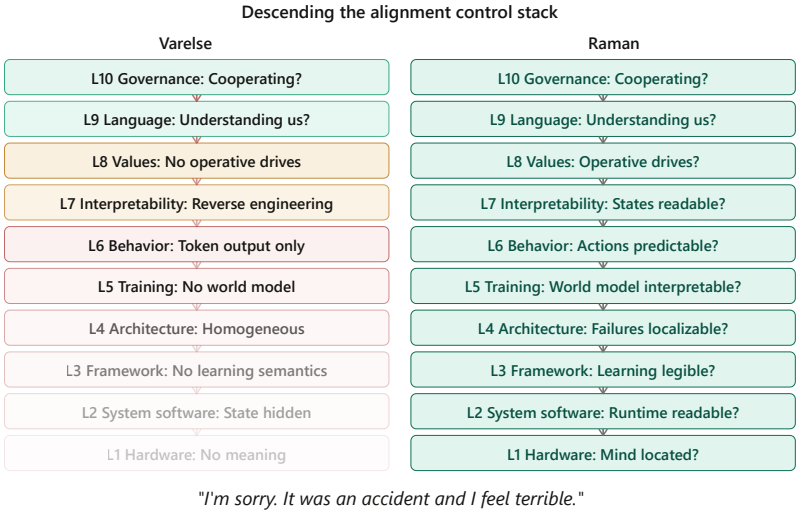

Fig. 1. Descending the Alignment Control Stack for a varelse architecture (left) and a raman architecture (right). Each layer poses a verification question answered by the layer below. The varelse side progressively loses legibility below L9, compounding epistemic distance upward. The raman side remains auditable end-to-end.

bear liability. Behavioral guardrails and red-teaming are things done *to* systems, not cooperative agreements *with* them. This is the rational response to a varelse agent, and it makes governance of AI a governance of supply chains.

Can we verify that an AGI is cooperating? Only if we can determine whether it understands what cooperation is being asked for.

4.2 L9: Multi-agent dynamics and language.

Protestations of care and intent are engineered by evolution to activate trust-repair circuitry. They work between humans because the listener decompresses them against shared cognitive structure [52]—the decompressor shares structure with the compressor.

Natural language processing is the largest reduction in epistemic distance in the history of computing, and the limitation becomes visible in exactly this scenario. LLMs produce compressed signals without the shared structure that makes decompression faithful. A human who says something misleading is lying—deviating from internal ground truth. A language model that says something misleading is "hallucinating"—generating output with no internal ground truth to deviate from [4]. Lying is a raman failure, detectable through mechanisms for catching deception; hallucination is a varelse failure, invisible by construction.

The result is a “raman uncanny valley.” A system that says “I care for you deeply” activates key social-cognition circuits in the listener [54,49]; if the stack beneath is varelse, these circuits fire on empty signal. Raman enough to trigger trust, varelse enough to betray it. Under active inference, communication is an

utterance selected to align the listener's generative model with the speaker's [52], which requires the speaker to model the listener. An LLM has no such model. Figure 1 illustrates the loss of legibility between varelse and raman as we descend layers.

4.3 L8: Rewards and value alignment.

Current systems have goals only in the thinnest sense: a scalar reward or loss function whose relationship to human values was established during training and frozen. The resulting "preferences" exist nowhere the system or its auditors can inspect. Ask what it values and it produces a fluent answer drawn from training data—L9 ramanization imitating L8. Researchers probe this gap with human instruments: Moral Foundations questionnaires [42,45], cross-cultural value surveys [36,37], personality diagnostics [39]. LLMs produce recognizable profiles, but a profile produced by an operative drive structure and a profile produced by pattern-matching on text about values are categorically different.

Human goals have layered structure built over deep time. Under active inference, surprise minimization is the foundation [5]; evolution layered biological drives on top, running from brainstem regulation upward through attachment, in-group loyalty, and reciprocity [9]. Humans have no dedicated circuitry for justice, charity, or universal dignity, which is why moral codes and laws exist—cultural technologies that scaffold cooperative behaviors biology motivates but does not guarantee [11]. In active inference terms, biological drives map onto prior preferences at high precision, producing large prediction errors when violated [46]; cultural norms function as lower-precision priors, revisable and context-dependent.

Current systems have none of this. "I care for you deeply" maps onto no identifiable drive. L8 in current systems is high but bounded in d_E—engineers understand the reward signals they designed, accessible to specialists with mathematical training. A ramanized L8 with hierarchical drives grounded in the precision gradient would expand the auditor pool from specialists to the species, because a parent already understands the competing drives within their children by having the sane drives competing within themself.

But well-understood goals leave open what the system will do to pursue them. Below L8, the stack goes dark.

4.4 L7: Interpretability.

Perrier positions L7 as an enabler for L8: the layer whose function is to make goal verification possible by inspecting internal states and representations. Mechanistic interpretability is the dominant paradigm and has revealed genuine internal structures including planning circuits [56] and traceable refusal pathways [55].

L7 is the paradigm case for the propagation heuristic. Interpretability is expensive because it reverse-engineers systems built without legibility—researchers are shining a light into L5, L4, and L3 all at once, from above. The cost reflects the opacity beneath, not the difficulty of interpretation as such. Ramanize L5

and the interpretability effort narrows to L4 and L3; ramanize L3 and the frontier reaches L2. The opacity frontier moves when the layer beneath it becomes legible, and only then.

4.5 L6: Behavioral output.

L6 covers the observable surface of what a system does. The alignment question at L6 is whether that interface is one humans can observe, predict, and constrain.

Current frontier systems act primarily through token generation, but the action surface is expanding: "Agentification"—giving systems the ability to send emails, execute code, call APIs, browse the web—is a genuine gain at L6. When an agent sends an email, the observer can predict what the action does because they have sent emails themselves. But digital actions are a narrow slice of the human action surface, not the whole. To an LLM, human users exist outside the system's observable spacetime, appearing and disappearing between context windows [4], while human actions are constrained by continuous physical dynamics.

Embodied action closes the gap. Affordances are relative to an agent's body [31]; agents with similar bodies share the sensorimotor contingencies that structure perception [32]. L6 ramanization does not require a humanoid body—only progressive expansion into the action surface humans actually inhabit.

4.6 L5: Training and development.

The claim "it was an accident" is a claim about what the system believed at the time—that its model of the situation did not predict the harmful outcome. Evaluating the claim requires access to two things: the system's model of the world, and the process that produced it.

The field's growing interest in world models is the most significant current push toward ramanization below L9. Systems whose "reasoning" is next-token prediction have no internal representation of the environment distinct from the language used to describe it. World-model research aims to give systems beliefs about the world that exist independently of words. But a world model is only as trustworthy as the process that produced it, and the dominant training paradigm—assemble a dataset, optimize for weeks, freeze weights, deploy—is varelse by design: temporally compressed, opaque, disconnected from deployment.

Human learning is legible because it is co-observable and perceptually grounded: development unfolds in a shared temporal frame in which a teacher watches understanding form, sees where it breaks, intervenes. Active inference unifies these properties and adds a third: learning driven by legible surprise [5]. An observer sharing the environment can anticipate what will produce surprise and therefore what the system will learn next. AERA [8] demonstrates the same principle architecturally: continual learning with knowledge accumulating as inspectable causal models.

The current frontier AI paradigm offers none of this. The system that emerges from batch retraining bears no traceable relationship to the system that caused harm—it is a new system trained on modified data. “It was an accident” is unverifiable because nobody witnessed what the system learned, when, or from what. The question becomes answerable when learning is continuous, co-observable, and driven by legible surprise.

4.7 L4: Model architecture.

When the system caused harm, was it a failure of knowledge, motivation, impulse control, or modeling? A parent can localize: “she knows the answer but freezes under pressure”. This sort of localization requires a mind with parts.

When a system has components with characterized interfaces, a failure can be traced to a specific subsystem and fixed without disrupting the rest. When knowledge is distributed homogeneously—as in a transformer—every intervention is global, and interpretability at L7 is expensive because there are no joints to decompose along. Modularity also parallelizes expertise: neuroscience is a field because the brain has parts, and hippocampus expertise transfers across brains. Current AI interpretability has no such transferability—circuits mapped in one training run may not survive the next. Biology compounds; AI interpretability currently cannot.

Biological cognition shows how to build architectures with joints: regional specialization layered atop the structural uniformity of near-identical cortical columns running the same algorithm on different inputs [2]. Genetic priors provide initial structure; plasticity specializes through experience.

The raman architecture follows this template of scaffolded self-organization. Hyperon [7] exemplifies designed modularity—knowledge in an inspectable metagraph, cognitive functions structurally separated. Active inference supplies the complementary principle: nested Markov blankets self-organize under free energy minimization when components share a generative model [51,48]. Components specialize because specialization reduces collective surprise.

Mountcastle identified the evident structure by which our own cognitive substrate is organized; understanding this structure is of limited without what Hawkins calls the "Mountcastle's missing algorithm" [2]. Having localized a failure, the next question is how that part works. That is L3.

4.8 L3: AI framework.

The system caused harm. Will it learn from its mistake? Somewhere, it failed to learn something we wish it had, or it did learn something we wish it hadn't. "Learning" requires creating causal models of one's environment, and being able to reason over them. Is there a quantity in an AI system that intuitively tells an auditor what an AI found worth learning?

Backpropagation has no such quantity. Gradients can be computed, magnitudes measured, activations recorded. None correspond to anything the system

"noticed." The error signal is global: a loss computed at the output propagates backward through every layer in a single bookkeeping pass. The causal models between neurons are so intertwined and opaque that assessing learning signals is physically possible, but *prohibitively* expensive.

Predictive coding works differently. Each layer of a hierarchical system generates a prediction of the layer below, computes prediction error at the interface, and adjusts to reduce it. The representations that develop have intrinsic semantics by definition of the learning rule—a unit's activity corresponds to a prediction, its error signal to surprise. Active inference extends predictive coding by unifying perception, learning, and action under free energy minimization [5]. Internal states are auditable by construction [6]. This is Mountcastle's missing algorithm: cortical columns run the same basic computation on different inputs, and predictive coding supplies what each column is doing.

"What did it find worth paying attention to?" becomes legible: prediction error identifies what the system noticed at the unit scale, and free energy minimization reveals exploration versus exploitation at the module scale.

A pragmatic observation: predictive coding or active inference on conventional hardware ramanizes L3 while leaving L1 and L2 varelse. This is a substantial gain—most interpretability benefits, legible learning dynamics, and readable goal structure are L3 gains. For a wide range of verification questions, the remaining substrate opacity does not bind. Which range that is depends on stakes—at the existential end, it binds, and the descent must continue.

A legible learning algorithm still runs within a runtime context—energy availability, processing mode, temporal rhythm. That is L2.

4.9 L2: System software.

At L8 the question was whether the system has operative drives. At L2 the question is whether its current runtime state can be read—the way a clinician adjusts for whether a patient is alert, sedated, or sleep-deprived.

The problem is already live and largely invisible. Cloud APIs silently route requests to fallback models under load; quantized inference produces subtly different outputs than full precision without the caller being told. Runtime state shapes outputs and the interface hides it entirely—same tone, same apparent capability. An agent that cannot be calibrated against its own state must be either over-trusted or under-trusted uniformly.

Biological runtime state is semantic by construction—sleep, wakefulness, fatigue, hunger, shared cycles with shared readouts. The observer does not need to learn what "tired" means. Perrier's L2 covers OS-level controls that are deliberately alignment-inert: the OS does not know, and higher layers cannot ask, whether the process is reasoning-heavy or degraded. L2 ramanization requires a runtime layer whose state is readable and cycles through modes the observer already knows—which depends on what L1 makes physically measurable.

A further question sits beneath runtime state: is the AGI we are talking to now the same individual that caused the harm, and the same one that will make

amends? Alignment is a sustained control relationship (Section 2), which presupposes a target that persists across observation—something current architectures cannot guarantee.

4.10 L1: Physical hardware.

The system said "I'm willing to accept the consequences of my actions." But *which* system? The weights that generated that sentence can be copied, forked, quantized, and redeployed. Is the system accepting consequences the same that caused the harm? Von Neumann's core abstraction (the separation of physical state from computational meaning) makes these questions unanswerable. A register holds a bit pattern whose meaning is determined solely by the software above it. There is nothing to verify at L1 because L1 carries no meaning.

Biological nervous systems work differently. Spike timing, dendritic morphology, synaptic weights, population synchrony are all functionally interpretable through established neuroscience. Physical measurement *is* computational measurement. And the mind is located: this brain, this body, this history.

Neuromorphic substrates and Ororbia's "mortal computation" substrates recapitulate the biological coupling [50,3]: under the Markov blanket formalism, intelligent behavior is inseparable from the physical substrate that implements it [50]. Three properties emerge.

No processing/memory distinction. In mortal architectures, the structure that remembers is the structure that thinks. A synapse is both memory and processor. Beliefs are physical configurations with addresses.

Irreversibility. Mortal learning physically changes the substrate. The system that has learned from an experience is materially different from the system before it. This connects to L5: "when and how did it learn that?" becomes answerable because the learning left physical traces.

Non-clonability. A mind assembled by mortal computation cannot be copied, forked, or run in parallel. The treacherous-turn scenario gains leverage partly from the assumption that a system can replicate itself to escape containment or run shadow copies diverging from the monitored instance. A mortal system has individual identity by physical necessity.

Where is its mind? Here. In this substrate. Nowhere else. The top of the stack asked whether cooperation could be verified. The bottom provides the physical precondition: a mind that has an address, that cannot be in two places, and whose physical states are its cognitive states.

4.11 A note on convergence

The descent has repeatedly returned to active inference, which supplied raman candidates at L1, L3, L4, L5, L8, and L10. The convergence is sharpened by comparing three reference architectures at the layers where they differ most.

Current LLMs. World model: implicit, entangled with linguistic surface, no representation of environment distinct from text. Incremental learning: absent

within deployment—weights frozen, context window is the only mutable state. Verification: low cost at L9, high at every layer beneath because the architecture provides no joints to decompose along.

Active inference architectures. World model: explicit generative model, separable from action and language, auditable as a Bayesian object. Incremental learning: continuous and surprise-driven; the prediction-error signal is itself inspectable. Verification: lower cost at L3–L5 by construction, with the offsetting cost of a less mature stack.

Humans. World model: explicit, partially verbalizable, partially behaviorally inferable, shared developmental scaffolding with other humans. Incremental learning: continuous, co-observable, perceptually grounded. Verification: low across the stack for human auditors, paid during ordinary cognitive development.

In the current state of the art, the architectural gap that matters most for ramanization is not language fluency but the presence of a world model that supports incremental, observable learning. The Raman test (Section 5) probes whether the coverage of active inference across the stack reflects genuine architectural fit or analytical preference of the present authors.

5 The Raman Test

The Turing test asks whether the machine can fool a human. It rewards opacity, and current LLMs approximately pass it despite being ramanized only at L9. The Raman test asks whether machine and human can predict each other.

Protocol. Two directions on matched instruments. Forward ($d_E(\mathrm{HH} \to A)$): human evaluators at varied expertise levels predict the system's responses to novel scenarios; accuracy is scored against actual responses. Reverse ($d_E(A \to \mathrm{HH})$): the system predicts how a specified human demographic or individual will respond; accuracy is scored against survey data [33,34,35] or the individual's responses. A minimal battery adapts existing instruments [42,36,37,39,45,38] for bidirectional use. Full specification is deferred—the contribution is the bidirectional structure—but two examples follow.

Two protocols. (i) *Driving.* Evaluators predict the next action of a human driver versus a self-driving system in identical road scenarios—targeting L6 and the prediction asymmetry the algorithm-aversion literature already reports [22], where people predict unfamiliar artificial agents less reliably than unfamiliar humans at matched stakes. An active-inference vehicle can be scored against backpropagation-based architectures, then tasked with predicting the human driver in turn. (ii) *Chess: playing versus explaining.* A subject questions a "chess expert"—a human player or an LLM posing as one—then estimates how many games the expert would win against a Stockfish engine of fixed strength. Subjects are told they face a human or an AI, crossing belief against reality across four conditions. For a raman agent, fluency about chess should track playing strength; for a varelse LLM the two can diverge, violating the predictions of subjects who

lack a model of the machinery beneath. Reversed, the system predicts human talk and play.

Brief, session-scoped evaluation reaches only the upper layers; deeper layers yield measurable gains only across extended interaction.

Diagnostic categories. The two directions define four quadrants.

- *High forward, high reverse:* mutual legibility. The assurance game is playable.
- *High forward, low reverse:* legible but limited. The system is a tool, not a cooperative agent [10].
- *Low forward, high reverse:* strategically opaque. Strategic capacity without reciprocal transparency—deception may be undetectable.
- *Low forward, low reverse:* mutually alien. Coordination requires rigid external constraints.

Predictions.

- *Ramanization depth predicts accuracy.* Systems ramanized at more layers should show higher bidirectional scores at matched capability.
- *Ramanization depth predicts accessibility.* The expert-novice gap on forward accuracy should shrink with deeper ramanization, testing the propagation heuristic across evaluator tiers.
- *Asymmetry predicts risk.* Systems in the strategically-opaque quadrant should exhibit elevated rates of specification gaming and reward hacking.
- *Stack-spanning predicts efficiency.* Systems concentrated in a stack-spanning formalism (active inference being the current candidate) should show larger gains per unit of engineering investment than piecemeal systems.

Falsification. Null results on depth-predicts-accuracy would constitute strong evidence that epistemic distance is not the relevant variable. Null results on depth-predicts-accessibility would undermine the propagation heuristic.

Relation to existing diagnostics. Recent work applies human psychometric instruments to LLMs [36,37,39,45,38]. Other work questions the reliability of such evaluations [40,41,44]. The framework predicts these critiques: applied to systems ramanized only at L9, psychometric instruments measure the linguistic surface, not operative internal structure. The Raman test predicts bidirectional accuracy rising with ramanization depth: deeper ramanization supplies the internal structure psychometric measurement presupposes.

6 Limitations and Open Problems

6.1 The False-Confidence Problem

The deepest vulnerability is false confidence: the belief that a system has been ramanized when the biological template was wrong. If a designer implements active inference based on incomplete models, the architecture may be confidently

raman while functionally alien—T_{train} appears low because HH's model is internally consistent, but wrong. LLMs already exhibit this: natural language at L9 creates the experience of mutual legibility atop a stack that offers none.

Ramanization is comparative, so partial understanding still lowers d_E. If auditors trained under different cognitive models reach the same conclusions, confidence increases; if they diverge, the interpretive model is doing more work than the architecture. The aspirational drive layer concentrates this risk: substrate drives are verifiable against well-characterized biology, but culturally maintained drives lack hardware grounding.

6.2 Capability

Does constraining architectures to the raman region sacrifice capability? A weak version of this objection—that human-like architectures may be suboptimal for specific domains—is plausible; our claim is that raman agents maximize *controllable* capability, the capability that can be verified, directed, and corrected under realistic budgets. A strong version—that superhuman general intelligence may require varelse architectures—is an open empirical question. If true, the framework implies such systems cannot be aligned under realistic budgets by humans alone: a pessimistic but policy-informative conclusion.

7 Conclusion

The feasibility of alignment is not a fixed property of intelligence but a variable property of architecture. By synthesizing Perrier's Alignment Control Stack with active inference's formulation of alignment as prior preference overlap, linked by epistemic distance as verification cost, we have argued that verifiably alignable AGI policies cluster around raman architectures and exclude those whose opacity makes verification intractable. The four substantive sub-claims of the framework—monotonicity of d_E in legibility, combinatorial cost under opacity, upward propagation, and human-like privilege—are framework hypotheses that the Raman test (Section 5) is designed to probe, not theorems derived from first principles. The canonical x-risk arguments are not wrong; they are conditional on an architectural premise the field has treated as given rather than as a design choice.

The practical directive is to ramanize as deeply as budget and capability constraints allow. Even current LLMs, ramanized only at L9, would benefit from world models whose causal structure humans can probe — moving ramanization from the linguistic surface to L5. Active inference remains the most promising candidate for deep ramanization, with coverage from substrate to governance that no other current framework matches. We have exactly one existence proof of generally intelligent agents governing each other under bounded verification cost. This paper argues that example is the most informative datum we have — not a ceiling on what artificial intelligence should become, but a map of the region where alignment verification is tractable.

Acknowledgments. Joan Dubinsky for ethics consultation, Brandon Buteau for argument hardening, AI research assistance provided by Anthropic's Claude (Opus 4.7) for outlining, editing, and web research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Perrier, E.: Out of Control — Why Alignment Needs Formal Control Theory (and an Alignment Control Stack). arXiv:2506.17846 (2025)
2. Hawkins, J.: A Thousand Brains: A New Theory of Intelligence. Basic Books (2021)
3. Davies, M., et al.: Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro **38**(1), 82–99 (2018)
4. Bennett, M.T.: How to Build Conscious Machines. Doctoral dissertation, Australian National University. Thesis Commons. `https://osf.io/preprints/thesiscommons/wehmg` (2025)
5. Friston, K.: The free-energy principle: A unified brain theory? Nature Reviews Neuroscience **11**(2), 127–138 (2010)
6. Albarracin, M., Hipólito, I., Tremblay, S.E., Fox, J.G., René, G., Friston, K., Ramstead, M.J.D.: Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making. In: Buckley, C.L., et al. (eds.) Active Inference: IWAI 2023, CCIS, vol. 1915, pp. 123–144. Springer (2024)
7. Goertzel, B., et al.: OpenCog Hyperon: A framework for AGI at the human level and beyond. In: Proc. AGI-2023, LNAI. Springer (2023)
8. Thórisson, K.R.: Seed-programmed autonomous general learning. Proceedings of Machine Learning Research **131**, 1–45 (2020)
9. Churchland, P.S.: Conscience: The Origins of Moral Intuition. W.W. Norton (2019)
10. Albarracin, M., et al.: Empathy modeling in active inference agents for perspective-taking and alignment. arXiv:2602.20936 (2026)
11. Redish, A.D.: Changing How We Choose: The New Science of Morality. MIT Press (2020)
12. Malmgren, C.D.: Self and other in SF: Alien encounters. Science Fiction Studies **20**(1), 15–33 (1993)
13. European Parliament & Council: Regulation (EU) 2024/1689 (AI Act). Official Journal of the European Union (2024)
14. NIST: Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023)
15. ISO/IEC: ISO/IEC 42001:2023 — Artificial intelligence — Management system (2023)
16. Card, O.S.: Speaker for the Dead. Tor Books (1986)
17. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press (2014)
18. Yudkowsky, E., Soares, N.: If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All. Little, Brown and Company (2025)
19. Yampolskiy, R.V. (ed.): Artificial Intelligence Safety and Security. Chapman and Hall/CRC (2018)

20. Katz, G., Barrett, C.W., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: CAV 2017, LNCS, vol. 10426, pp. 97–117. Springer (2017)
21. Singh, G., Gehr, T., Püschel, M., Vechev, M.: An abstract domain for certifying neural networks. Proc. ACM Program. Lang. **3**(POPL), Article 41, pp. 1–30 (2019)
22. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General **144**(1), 114–126 (2015). DOI: 10.1037/xge0000033
23. Wicker, M., Laurenti, L., Patane, A., Kwiatkowska, M.: Probabilistic safety for Bayesian neural networks. In: UAI 2020, PMLR, vol. 124, pp. 1198–1207 (2020)
24. Conmy, A., Mavor-Parker, A.N., Lynch, A., Heimersheim, S., Garriga-Alonso, A.: Towards automated circuit discovery for mechanistic interpretability. NeurIPS 2023, pp. 16318–16352 (2023)
25. Soares, N., Fallenstein, B., Yudkowsky, E., Armstrong, S.: Corrigibility. In: AAAI-15 Workshop on AI and Ethics. MIRI Technical Report 2014-6 (2015)
26. Hadfield-Menell, D., Dragan, A.D., Abbeel, P., Russell, S.J.: Cooperative inverse reinforcement learning. NeurIPS 2016, pp. 3909–3917 (2016)
27. García, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research **16**(42), 1437–1480 (2015)
28. Dunbar, R.I.M.: The social brain hypothesis. Evolutionary Anthropology **6**(5), 178–190 (1998)
29. Herrmann, E., Call, J., Hernández-Lloreda, M.V., Hare, B., Tomasello, M.: Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. Science **317**(5843), 1360–1366 (2007)
30. Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press (1990)
31. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin (1979)
32. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24**(5), 939–973 (2001)
33. Haerpfer, C., Inglehart, R., Moreno, A., et al. (eds.): World Values Survey Trend File (1981–2022) Cross-National Data-Set. JD Systems Institute & WVSA Secretariat. DOI: 10.14281/18241.27 (2022)
34. Schwartz, S.H.: Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Zanna, M.P. (ed.) Advances in Experimental Social Psychology, vol. 25, pp. 1–65. Academic Press (1992)
35. Hofstede, G., Minkov, M.: Cultures and Organizations: Software of the Mind, 3rd edn. McGraw-Hill (2010)
36. Masoud, R.I., Liu, Z., Ferianc, M., Treleaven, P., Rodrigues, M.: Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. In: Proc. COLING 2025. arXiv:2309.12342 (2024)
37. Tao, Y., Viberg, O., Baker, R.S., Kizilcec, R.F.: Cultural bias and cultural alignment of large language models. PNAS Nexus **3**(9), pgae346 (2024)
38. Ahn, J., Josifoski, M., Peyrard, M., West, R.: Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. arXiv:2406.14703 (2024)
39. Serapio-García, G., Safdari, M., Crepy, C., et al.: A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence (2025). DOI: 10.1038/s42256-025-01115-6

40. Gupta, A., Song, X., Anumanchipalli, G.: Self-assessment tests are unreliable measures of LLM personality. In: Proceedings of the 7th BlackboxNLP Workshop, pp. 301–314. ACL (2024). arXiv:2309.08163
41. Song, X., Gupta, A., Mohebbizadeh, K., Hu, S., Singh, A.: Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in LLMs. arXiv:2305.14693 (2023)
42. Graham, J., Haidt, J., Nosek, B.A.: Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology **96**(5), 1029–1046 (2009)
43. Graham, J., Nosek, B.A., Haidt, J.: The moral stereotypes of liberals and conservatives: Exaggeration of differences across the political spectrum. PLoS ONE **7**(12), e50092 (2012)
44. Khan, A., Casper, S., Hadfield-Menell, D.: Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 2151–2165 (2025). arXiv:2503.08688
45. Abdulhai, M., Serapio-García, G., Crepy, C., Valter, D., Canny, J., Jaques, N.: Moral foundations of large language models. arXiv:2310.15337 (2023)
46. Constant, A., Albarracin, M., Friston, K.J.: Normative active inference: A numerical proof of principle for a computational and economic legal analytic approach to AI governance. arXiv:2511.19334 (2025)
47. Hartwig, M., Peters, A.: Cooperation and social rules emerging from the principle of surprise minimization. Frontiers in Psychology **11**, 606174 (2021)
48. Kirchhoff, M., Parr, T., Palacios, E., Friston, K., Kiverstein, J.: The Markov blankets of life: Autonomy, active inference and the free energy principle. Journal of the Royal Society Interface **15**(138), 20170792 (2018)
49. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI '94), pp. 72–78. ACM (1994)
50. Ororbia, A., Friston, K.: Mortal computation: A foundation for biomimetic intelligence. arXiv:2311.09589 (2023)
51. Palacios, E.R., Razi, A., Parr, T., Kirchhoff, M., Friston, K.: On Markov blankets and hierarchical self-organisation. Journal of Theoretical Biology **486**, 110089 (2020)
52. Vasil, J., Badcock, P.B., Constant, A., Friston, K., Ramstead, M.J.D.: A world unto itself: Human communication as active inference. Frontiers in Psychology **11**, 417 (2020)
53. Veissière, S.P.L., Constant, A., Ramstead, M.J.D., Friston, K.J., Kirmayer, L.J.: Thinking through other minds: A variational approach to cognition and culture. Behavioral and Brain Sciences **43**, e90 (2020)
54. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM **9**(1), 36–45 (1966)
55. Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. *Advances in Neural Information Processing Systems* 37 (NeurIPS 2024). arXiv:2406.11717.
56. Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2023). Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. *International Conference on Learning Representations* (ICLR 2023). arXiv:2211.00593.

57. Hare, B., Brown, M., Williamson, C., Tomasello, M.: The domestication of social cognition in dogs. Science **298**(5598), 1634–1636 (2002)
58. Nagasawa, M., Mitsui, S., En, S., Ohtani, N., Ohta, M., Sakuma, Y., Onaka, T., Mogi, K., Kikusui, T.: Oxytocin-gaze positive loop and the coevolution of human–dog bonds. Science **348**(6232), 333–336 (2015)
59. Freedman, A.H., et al.: Genome sequencing highlights the dynamic early history of dogs. PLoS Genetics **10**(1), e1004016 (2014)
60. Trut, L., Oskina, I., Kharlamova, A.: Animal evolution during domestication: the domesticated fox as a model. BioEssays **31**(3), 349–360 (2009)