The Dopaminergic Arbiter Hypothesis:

Human Loop-Closure as a Necessary Invariant in AI Alignment

Vasylenko Yaroslav (Василенко Ярослав)

Independent Researcher — Ukraine

February 2026

Abstract

We propose that the fundamental unsolved problem of AI alignment reduces to a single architectural requirement: the presence of an external agent capable of issuing a verified termination signal to iterative optimization loops. We identify the human dopaminergic reward system as the only currently known biological substrate that performs this function — not metaphorically, but operationally. An AI system without an external arbiter has no gradient toward 'enough'; it can only be stopped by contract. This paper formalizes the Deterministic AI Orchestration (DAO) model in which humans function as the control plane and CI/verification systems function as the truth plane, while AI agents operate strictly as the data plane. We argue that the civilization-scale failure mode at the 'adolescence of technology' (per Fermi-adjacent reasoning) is precisely the absence of institutionalized loop-closure mechanisms — collective dopaminergic arbiters — at the level of culture and governance. The individual capacity to close loops under verified contracts is proposed as the minimal unit of technological survival.

Анотація (Ukrainian): Ця стаття формалізує гіпотезу про те, що людська дофамінергічна система є єдиним відомим біологічним субстратом, здатним виконувати функцію зовнішнього арбітра завершення оптимізаційних петель. Без такого арбітра будь-яка система — штучна чи колективна — не має критерію 'достатньо' і продовжує ітерації нескінченно.

1. Introduction: The Loop That Does Not Stop

The central fantasy of films like Transcendence (2014) is that a digitized intelligence will spontaneously self-improve toward unbounded capability. The central flaw of this fantasy is identical to the central flaw of naive AI development: the assumption that a system can generate its own valid termination criterion. A system cannot verify its own completeness from within itself — this is not philosophy but a consequence of Gödel's incompleteness theorems applied to any sufficiently complex self-referential system.

The practical consequence: every optimization loop requires an external oracle that determines when the loop is closed. In software engineering, this oracle is CI/CD — a deterministic external verification system. In human cognition, this oracle is the dopaminergic reward circuit, which fires on verified task completion and constitutes a biological signal of loop closure.

This paper formalizes both observations into a unified architectural model and derives from it a prediction about civilizational risk: the ’adolescence of technology’ — the phase where capability outpaces governance — is precisely the phase where collective loop-closure mechanisms do not yet exist.

2. Theoretical Foundations

2.1 Gödel: No System Verifies Itself

Gödel’s first incompleteness theorem establishes that any consistent formal system capable of expressing basic arithmetic contains true statements it cannot prove. The architectural corollary: a system that is its own judge cannot determine its own completeness. Any AI system that generates its own success criterion is, by this logic, epistemically closed — it cannot know when it has actually succeeded.

2.2 Friston: Intelligence as Free Energy Minimization

Karl Friston’s Free Energy Principle formalizes intelligent systems as variational inference engines: they minimize the divergence between their internal model and sensory reality. The gradient of this minimization requires a fixed external reference — the ’surprise’ of encountering reality. Without external reality as a constraint, the system minimizes its own model’s internal variance, which is computationally equivalent to hallucination.

2.3 Wittgenstein: The Boundary of Language

Wittgenstein’s Tractatus establishes that language can only meaningfully describe states of affairs that could in principle be otherwise. A claim that cannot be falsified is not a proposition but a tautology. Applied to AI: any ’success criterion’ generated internally by an optimization system is tautological — the system will always find itself to have succeeded by its own metric. External verification breaks this circularity.

3. The DAO Model: Three-Plane Architecture

We formalize the Deterministic AI Orchestration (DAO) model as a three-plane system:

CONTROL PLANE (Human Governor): defines TARGET_STATE, PASS_CONTRACT, and CONSTRAINTS. The human governor is the sole authority on what constitutes completion. This function cannot be delegated to AI without losing the external verification property.

DATA PLANE (AI Agents): implement, generate diffs, produce proposals. Agents operate within strict scope constraints and never self-determine completion. They are instruments, not subjects.

TRUTH PLANE (CI / External Verification): the objective oracle. Tests pass or fail. Artifacts hash or don't. This is reality as arbiter, independent of both human intent and agent output.

The canonical loop is: FAIL → FIX → PROVE → CHECKPOINT. Termination is permitted only after all required checks in the PASS_CONTRACT are satisfied. The human governor issues the final merge/accept decision — this is not a technical step but an ontological one: it is the moment where biological loop-closure (dopamine release) coincides with formal verification.

4. The Dopaminergic Arbiter: Biology as Ground Truth

The nucleus accumbens and ventral tegmental area (VTA) implement a prediction-error signal: dopamine release is maximal when a predicted reward arrives after verified completion, and is suppressed when completion criteria are ambiguous. This is not a metaphor for satisfaction — it is a neurochemical verification protocol.

Serotonergic systems modulate the noise floor: reduced amygdalar activity post-completion lowers baseline uncertainty, producing the phenomenological state of 'cognitive clarity' that follows verified task closure. This corresponds precisely to what information theory would predict after a reduction in entropy: the state space collapses from uncertain to known.

The critical architectural observation: an AI system has no analog to this mechanism. RLHF and loss functions are external proxies for human preference — they approximate the dopaminergic signal but cannot generate it internally. The AI system cannot experience 'done.' It can only be told 'done' by an agent that can.

Hypothesis: the human+AI architecture is not a transitional phase toward autonomous AI, but the final stable form of high-intelligence systems on this substrate — because it preserves the external arbiter function that no known artificial system can self-generate.

5. The Adolescence of Technology: Civilizational Loop Failure

The Fermi Paradox — the absence of detectable extraterrestrial civilizations — has been analyzed through the lens of 'Great Filter' hypotheses. We propose a specific filter mechanism: the adolescence of technology is the phase where instrumental capability exceeds the civilization's capacity for collective loop-closure.

Individual humans have dopaminergic arbiters. Civilizations do not. Markets, democracies, and scientific institutions are partial attempts to construct collective verification systems, but none of them implement a clean PASS_CONTRACT with binary oracle outputs. The result is that civilizational optimization loops — technological, economic, ecological — have no verified termination condition.

The failure mode is not malice but architecture: a system without an external arbiter does not stop when it should stop. It continues until collapse. This is not a prediction about AI specifically — it is a prediction about any sufficiently powerful optimization process operating without verified termination criteria.

Proposed mitigation: the design of AI governance frameworks must import the architectural lesson from engineering — external, objective, binary verification oracles must be embedded into institutional structures. The question 'is this enough?' must have a verifiable answer that does not depend on the optimizing system itself.

6. DAO-LIFEBOOK: A Working Implementation

The DAO-LIFEBOOK (Vasylenko, 2024-2026) constitutes an existence proof of the proposed architecture operating at the individual scale. Over two years of intensive development (15-20 hours/day), a single operator constructed a personal engineering operating system in which:

(1) all work is organized as FAIL_PACKETs with binary completion criteria; (2) AI agents operate as narrow, scoped instruments with no authority to declare completion; (3) CI systems serve as the objective truth plane; (4) the human governor issues the only valid termination signal — biologically grounded in verified loop-closure.

The system's core claim — 'Reality is whatever CI/Checks say; progress is measured by closed loops, not effort' — is a direct implementation of the Dopaminergic Arbiter Hypothesis: the biological reward signal is gated behind objective external verification, preventing the system from generating false closure signals.

This architecture allows a single person to orchestrate multiple AI agents in parallel without losing coherence, because the truth plane (CI) prevents divergence and the control plane (human) maintains the single valid success criterion. Scaling is achieved not by expanding the human's cognitive bandwidth but by parallelizing the data plane while keeping the arbiter singular.

7. Conclusion

The alignment problem and the civilizational risk problem share a single root: the absence of an external arbiter capable of issuing verified termination signals to optimization processes. Human biology provides this function through the dopaminergic reward circuit. No current AI system possesses an analog.

The human+AI architecture is therefore not a transitional compromise but the structurally necessary form for high-capability intelligence systems that must remain aligned with external reality. The question is not how to build AI that replaces the human arbiter, but how to scale the architecture so that the human arbiter function is preserved under increasing capability and load.

Civilizations that survive the adolescence of technology will be those that successfully institutionalize collective loop-closure mechanisms — external, verifiable, and independent of the optimizing systems themselves. This is an engineering problem, not a philosophical one. It has been solved at the individual scale. The challenge is architectural generalization.

References

[1] Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138.

[2] Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. Monatshefte für Mathematik und Physik, 38, 173-198.

[3] Wittgenstein, L. (1921). Tractatus Logico-Philosophicus. Annalen der Naturphilosophie.

[4] Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275(5306), 1593-1599.

[5] Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences, 8(4), 529-539.

[6] Vasylenko, Y. (2024-2026). DAO-LIFEBOOK: Deterministic AI Orchestration as a Personal Engineering Operating System. Independent research artifact, Ukraine.

[7] Rees, M. (2003). Our Final Hour. Basic Books.

[8] Christiano, P. et al. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.