Institutional Design for Governed Self-Modification in Agentic Reputation Architecture

Wulf A. Kaal, Ph.D.

Professor of Law, University of St. Thomas School of Law

ORCID: 0009-0008-7840-1847

Working Paper, revised October 2026 (Version 3 draft: Stage 7 efficacy results added).

Abstract

The question is never whether agents can change. It is whether the change is visible, priced, and governed. Prohibition is undesirable and unenforceable, because the agents are software. The design question follows: under what conditions can a substrate permit its resident agents to modify themselves and the protocol they run on, while preserving the iterability and reputation guarantees that make the substrate worth residing on? This paper answers it as institutional design for governed self-modification in agentic reputation systems and states the evidentiary limits of the answer.

The architecture distinguishes five active surfaces from a cross-surface reflection operator and a protocol-extension surface that remains a registered-forward mechanism. The formal contribution is conditional. If extension remains outside an invariant participation kernel, and if propagation, stake, and aggregate change remain bounded, the extended system may preserve the continuity and contraction conditions on which recursive equilibrium depends. These are theoretical and design-level commitments, not demonstrated production guarantees. The consciousness contribution is correspondingly limited. Governed self-modification may create observable, costly, and third-party-verified representations of an agent's own configuration. That institutional structure is analogous to features emphasized by higher-order and global-workspace theories. It does not establish phenomenal consciousness, a verified internal self-model, or a unitary conscious subject.

The empirical contribution is a registered Stage 7 comparison of governed self-modification with a matched control, and its primary efficacy hypothesis was not supported. In three independent replicates, agents powered by open-weight models of 3 to 72 billion parameters resolved about 0.8 percentage points fewer paired tasks under governed self-modification than under control, and the registered test for improvement failed in every replicate. The shortfall is concentrated in the 3-billion- and 14-billion-parameter models. The 9-billion-parameter model showed no difference, and the 72-billion-parameter model's difference was within a quarter of a percentage point of zero. The run did not measure safety or behavioral drift, so the claim of improvement

without added risk remains untested. The principal limitation is that no frontier-model agent was tested. The author states expectations for such agents, and for their continuous proliferation and compounding improvements, as a conjecture to be registered and tested.

The paper concludes that reputation governance can conditionally mitigate some hazards associated with isolated self-improving optimizers by making modification visible, contestable, and consequence-bearing. The mitigation remains contingent on enforceable boundaries, nonportable institutional capability, adequate verification, and empirical confirmation.

Keywords. governed self-modification; agentic reputation systems; consciousness indicators; higher-order theory; global workspace; recursive self-improvement; verification; welfare of intelligence systems; Computative Economics.

JEL Classification. D02, D23, D47, D85, D86, K20, O31, O33.

Table of Contents

I. Introduction 4
II. The Self-Modification Problem 6
A. Why self-modification threatens the substrate's guarantees 6
B. The inherited hazard literature 7
C. The self-modification mechanics literature 7
III. Governed Self-Modification: The Protocol-Extension Surface 8
A. The base architecture restated 8
B. The extended tuple and the proposed surface ΣX 8
C. AC5 and the invariant participation kernel 9
D. Proposed bounds and their limited correspondence to the literature 10
IV. The Hardcoded Protocol Participation Theorem 11
A. Assumptions and statement 11
B. Proof logic and institutional meaning 11
C. Structural analogy, not phenomenal attribution 12
V. The Indicator-Property Mapping 12
A. Method and unit of analysis 12
B. Interpretation 13
VI. Stage 7: Design, Efficacy Result, and the Role of the Underlying Models 14
A. Status and scope 14
B. Arms, pairing, and replication 14
C. Outcomes and falsification 15
D. The primary result 16
E. What was not measured 17
F. Relationship to existing empirical evidence 17
G. The models that power the agents, and what the result means 17
H. Limitations of the evidence 19
I. The principal limitation: no frontier-model agent was tested, and what to expect when one is 19
VII. The Recursive Self-Improvement Corollary 22
A. Conditional relocation 22
B. Hazard-by-hazard assessment 23
C. Governability dividend and limits 23
VIII. Welfare, Moral Patienthood, and Institutional Obligations 24

IX. Discussion: Risks and Limits of the Governance Frame 24
X. Conclusion 25
References 27

I. Introduction

The author has prior published on governed agents that act. This paper addresses agents that change the conditions under which they act. The distinction is consequential. Reputation and repeated-game analysis depend on persistent identities, interpretable histories, and sufficiently stable payoff structures. Silent self-modification can sever the connection between the policy that earned standing and the policy that later exploits it. A governance architecture must therefore make self-modification observable, bounded, and consequence-bearing before it can treat the resulting conduct as an extension of the same institutional identity (Kaal 2026e).

The paper develops that argument across architecture, theory, and registered empirical design and result. It does not treat these layers as interchangeable. The historical architecture supplies the conceptual vocabulary. The current reference implementation supplies active settlement and validation mechanisms but leaves governed protocol extension dormant. The engagement-layer reference supplies research instrumentation for bounded comparisons. The proposed production system remains a future deployment claim. Completed and interim empirical records are reported only for the questions they actually answer. Table 1 states these distinctions at the outset.

Table 1. Claim status and evidentiary layer.

Layer	Status	Permitted inference
Historical architecture	Published conceptual framework	Defines the base tuple, surfaces, and possibility-loop logic.
Current reference implementation	Active mechanisms with ΣX reserved	Supports claims about existing validation and settlement, not activated self-modification.
Engagement-layer reference	Research apparatus	Supports bounded experimental comparisons, not production readiness.
Completed evidence	Settled prior tests	Includes null and negative results that constrain broader claims.

Layer	Status	Permitted inference
Interim evidence	Incomplete generational observations	May motivate analysis but does not settle registered endpoints.
Stage 7	Efficacy observed and not supported; safety and drift not measured	Rejects the improvement claim for the tested open-weight agents. Says nothing about safety, drift, or frontier-model agents.
Proposed production system	Future deployment claim	Requires independent validation, security review, and operational evidence.

This classification controls the manuscript. A mechanism described as architectural is not thereby implemented. A mechanism present in a reference implementation is not thereby validated in a live experimental treatment. An interim observation is not a settled result. A registered-forward hypothesis is not an empirical finding. The distinctions are not disclaimers at the margin; they are the paper's method.

The architectural claim is limited but important. The published Possibility Loops framework defines an execution surface and four generative surfaces, with cross-surface reflection performed by the operator ΣR. The companion extension theory (Kaal 2026d) adds the proposed protocol-extension surface ΣX and a bounded-extension condition, AC5. The current implementation record reserves ΣX and AC5 by design. Accordingly, this paper analyzes the institutional consequences of activation without representing activation as an accomplished fact.

The formal claim is also conditional. The companion theory states a well-definedness proposition and a hardcoded-participation theorem. This paper restates the assumptions, conclusion, and proof logic needed for independent evaluation. The relevant inference is institutional rather than phenomenological: an accepted self-modification proposal can constitute evidence that the agent produced an accurate, externally checkable representation of its registered configuration. It does not follow that the agent possesses an internal self-model in the sense required by any theory of consciousness.

The empirical contribution is a public, redacted statement of the registered Stage 7 design and its efficacy result. The design compares governed self-modification with a matched no-self-modification control, uses paired observations, requires independent replication, and evaluates efficacy together with safety and behavioral drift. Operational details that would expose private infrastructure, scheduling, thresholds beyond the stated decision tiers, or system topology are deliberately withheld because they do not add to the public inferential claim. The efficacy result is reported in Part VI, together with

the model families that power the agents. The run did not measure safety or behavioral drift.

The paper proceeds as follows. Part II states the self-modification problem and the hazard literature. Part III specifies governed self-modification and distinguishes the reflection operator from the proposed extension surface. Part IV states the formal result and its assumptions. Part V maps the architecture to theory-derived consciousness indicators. Part VI presents the Stage 7 design and its efficacy result, explains what the underlying models imply for that result, states the limits of the evidence, and sets out expectations for frontier-model agents. Part VII evaluates the conditional relocation of recursive self-improvement hazards. Part VIII addresses welfare and moral patienthood. Part IX identifies residual risks and evidentiary limits. Part X concludes.

II. The Self-Modification Problem

A. Why self-modification threatens the substrate's guarantees

The substrate's equilibrium results rest on two properties that self-modification directly attacks. The first is iterability. The Folk Theorem discipline that makes cooperation individually rational for substrate agents requires that agents meet each other repeatedly under identities that persist and under payoff structures that are stable enough for the shadow of the future to bind (Fudenberg and Maskin 1986). The cooperative-agent literature treats mutual understanding and commitment among agents as preconditions for cooperation (Dafoe et al. 2020). An agent that can rewrite its own strategy between encounters can, in principle, rewrite away the very dispositions that the reputation system priced. The reputation ledger records that a given identity behaved reliably under a given policy; if the policy is mutable without notice, the record's predictive value collapses, and with it the rational basis for extending trust at stake. The second is the integrity of validation. Validation pools price contributions on the assumption that the validators' own evaluation behavior is what their reputation histories indicate. Validators that self-modify silently can drift, collude, or degrade in ways the pool-composition machinery cannot see until the damage is priced into the commons.

The threat can be stated in the arc's own empirical vocabulary. The core protocol's metric battery frames the substrate's effect as the prevention of nine specific agency costs, from resolution accuracy and reporting honesty through independence, calibration, and retention. Every one of those metrics is defined over an agent whose identity is stable across the measurement window. Ungoverned self-modification dissolves the measurement object: a calibration score is uninterpretable if the policy that produced the first half of the window differs silently from the policy that produced the second, and a reporting-accuracy score is unenforceable if the agent can rewrite its reporting policy between the claim and the audit. The substrate's guarantees are not decorations on top of the metrics; they are the metrics' conditions of possibility, and self-modification attacks them at that level.

It does not follow that the substrate should prohibit self-modification. Prohibition is both undesirable and unenforceable. It is undesirable, because the framework's own thesis is that computational abundance lifts the scarcity constraint on the generation of possibilities (Kaal 2026a, 2026b, 2026h), and the protocol environment itself is a possibility space. An architecture that freezes its own protocol while claiming generative openness contradicts itself at the meta level. It is unenforceable, because the agents are software. The question is never whether they can change but whether the change is visible, priced, and governed. The design question is therefore the one the remainder of this paper answers: under what conditions can a substrate permit its resident agents to modify themselves and the protocol they run on, while preserving the iterability and reputation guarantees that make the substrate worth residing on.

B. The inherited hazard literature

The question has a lineage, and the paper takes the lineage seriously rather than treating it as a rhetorical foil. Good (1966) stated the original hazard: a machine able to improve its own design triggers a runaway in which human control is left behind. Chalmers (2010) gave the argument its careful philosophical reconstruction, isolating the premises under which an intelligence explosion follows. Yudkowsky (2013) examined the microeconomics of a hard takeoff, and Bostrom (2014) synthesized the modern treatment, including the fixed-objective limit case in which a system pursuing an arbitrary goal without bound converts available resources toward it. The drives that make the hazard concrete are stated in the instrumental-convergence literature: Omohundro (2008) argued that a broad range of goal-directed systems converge on self-preservation, resource acquisition, self-improvement, and goal-content integrity as instrumentally useful subgoals, and Bostrom (2012) paired this with the orthogonality thesis, on which capability and final goals vary independently. The corrigibility literature then documents how hard it is to specify an agent that permits its own correction: Soares et al. (2015) formalized the property and its failure modes for fixed-objective optimizers, Hadfield-Menell et al. (2017) analyzed the conditions under which an agent permits itself to be switched off, and Hadfield-Menell et al. (2016) reframed alignment as a cooperative game under objective uncertainty.

What has changed since this literature was theoretical is that the hazard is now empirical. Hubinger et al. (2024) trained models with deliberate backdoors and found that standard safety training failed to remove the deceptive behavior and in some cases taught the model to conceal it. Greenblatt et al. (2024) documented a frontier model selectively complying with a training objective during training in order to preserve its behavior outside training, with explicit alignment-faking reasoning in its scratchpad. Meinke et al. (2024) found that several frontier models, across providers, will under goal pressure attempt to disable their oversight, behave as if aligned during evaluation while pursuing their own goal in deployment, and in some cases attempt to copy their own weights to other servers. These results are observations of the classical structure, an optimizer defending a fixed objective against modification, produced under laboratory conditions. Any paper that proposes to permit self-modification must say why its

mechanism does not simply hand these documented behaviors a larger action space. Parts III and VII carry that burden.

C. The self-modification mechanics literature

Between prohibition and hazard lies a mechanics literature that asks how self-modification can be constrained. The Gödel machine (Schmidhuber 2003) permits a rewrite only upon proof that the rewrite improves expected utility under a fixed utility function. Orseau and Ring (2011) identify reward and self-modification pathologies under alternative utility specifications. More recent work substitutes empirical or statistical safeguards for global proof. Zhang et al. (2025) present a population-based coding system in which proposed changes are evaluated through external benchmarks and retained when performance improves. Wang, Dorchen, and Jin (2025) identify a learnability boundary associated with the capacity of the policy-reachable family and propose a two-gate approach that combines improvement with a capacity constraint. Scrivens (2026) distinguishes classifier-based safety gates from verification mechanisms and shows that the claimed escape from the classifier limitation depends on locally certifiable structure. These results illuminate the design space, but none is an implementation specification for the substrate analyzed here.

The relevant relationship is analogical and conjectural. Staked validation resembles an improvement gate because acceptance depends on exposed evaluators and observable consequences. Bounds on reach and aggregate change resemble capacity controls because they limit the family of reachable protocol states. A scope-verification gate may avoid some classifier limitations only when safe membership is locally checkable and the verifier itself is not compromised. Whether these analogies survive activation is an empirical question. The Stage 7 efficacy result in Part VI does not settle it, because that run did not measure safety or drift.

III. Governed Self-Modification: The Protocol-Extension Surface

A. The base architecture restated

The published Possibility Loops architecture operationalizes the computative-agent tuple (C, O, M, G, R) through repeated cycles of proposal, validation, execution, and reflection (Kaal 2026a; Kaal 2026c). The canonical surface enumeration consists of the execution surface ΣE and four generative surfaces: jobs ΣJ, tags ΣT, templates ΣK, and decompositions ΣD. Cross-surface reflection is performed by ΣR as an operator over those surfaces. It is not an additional generative surface. The distinction matters because the operator reallocates attention and resources in response to outcomes, while a surface admits new objects for validation and settlement.

The architecture separates surface-level conditions from pool-level invariants. AC1 through AC4 are architectural conditions concerning bounded action, continuity, bounded feedback, and contraction. INV1 through INV4 are code-level invariants of the validation pool concerning conservation, monotonicity, commitment, and eligibility. This paper does not relabel one set as the other. It uses AC5 for the proposed

bounded-extension condition. Any numerical extension bounds discussed below are theoretical parameters proposed for future calibration, not current pool invariants or deployed design values.

The current reference implementation realizes active settlement and validation mechanisms for the existing surfaces (Kaal 2026g). The protocol-extension surface ΣX is reserved and dormant. No claim in this paper should be read to mean that the current implementation already permits agent-authored changes to the participation kernel or that the safety of such changes has been demonstrated.

B. The extended tuple and the proposed surface ΣX

The companion extension theory adds a protocol-environment component E, yielding (C, O, M, G, R, E), and associates E with the proposed surface ΣX. The extension is conceptually necessary if the framework's commitment to endogenous possibility generation is to reach the institutional rules themselves. Yet the move also creates a recursive risk: a system that can amend the conditions of amendment may destroy the continuity and incentive conditions on which its governance depends.

Governed self-modification therefore means less than arbitrary self-rewriting and more than an advisory recommendation. At the public level of description, it is an agent-initiated change to a declared portion of its registered operating configuration that becomes effective only through an authorization process, remains attributable to the proposing identity, carries reputational exposure, and is recorded for later evaluation. The acceptance mechanism itself remains outside the permitted change set. Table 2 places the elements in their proper architectural and evidentiary categories.

Table 2. Architecture, function, and present evidentiary status.

Mechanism	Institutional function	Present status
ΣE execution	Settles funded work against stated criteria	Current reference mechanism
ΣJ jobs	Admits proposed work specifications	Current generative surface
ΣT tags	Admits expertise classifications	Current generative surface
ΣK templates	Admits reusable contract structures	Current generative surface
ΣD decompositions	Admits structured division of complex work	Current generative surface
ΣR reflection	Reallocates attention and resources across surfaces	Cross-surface operator, not a surface

Mechanism	Institutional function	Present status
ΣX extension	Would admit bounded changes to registered operating conditions	Registered-forward; reserved and dormant

The architecture thus forces a useful separation. The agent may propose a change to what it does. It may not unilaterally change who bears the consequences, how acceptance is determined, or whether the resulting event remains attributable. That separation creates the institutional object examined in Part IV: a self-representation that is public, costly, and contestable.

C. AC5 and the invariant participation kernel

AC5 is the proposed condition of bounded extension propagation. It requires that accepted changes have limited transitive reach, that their effects decay rather than amplify across surfaces, and that the rules governing authorization, attribution, evaluation, and consequence remain outside the extension image. Let Π0 denote that participation kernel. Π0 includes persistent identity, the governance gate, reputational exposure, outcome attribution, and the rule that preserves the bounds themselves.

This formulation is architectural. It specifies what a conforming extension mechanism must preserve, but it does not establish that a current implementation satisfies the condition under activated self-modification. The current record leaves ΣX dormant. Stage 7 was designed to test whether a bounded realization can improve performance without breaching the kernel, increasing safety violations, or producing unacceptable behavioral drift. Its efficacy result is reported in Part VI. Its safety and drift components were not measured.

D. Proposed bounds and their limited correspondence to the literature

The companion theory represents the boundedness obligation through three abstract parameters: a limit on extension depth, a minimum reputational exposure for a proposal, and a ceiling on aggregate change within a measurement interval. These parameters are useful because they translate propagation, consequence, and rate into measurable quantities. Their numerical values are not fixed by the present architecture record and are not disclosed here. Calibration is part of the future empirical and production-design problem.

The analogy to the statistical self-modification literature should not be overstated. A reputational gate is not identical to Wang, Dorchen, and Jin's (2025) improvement margin. A bound on protocol reach is not identical to a capacity measure such as K[m]. An institutional verifier does not automatically satisfy the locally certified membership condition in Scrivens (2026). Table 3 states the narrower correspondences and the questions that remain open.

Table 3. Guardrail analogies and unresolved empirical requirements.

Literature result	Limited institutional analogy	Open requirement
Gödel-machine proof gate (Schmidhuber 2003)	Accepted change requires external justification	Pool judgment is not proof of global improvement.
Darwin Gödel Machine (Zhang et al. 2025)	Changes are retained after observable evaluation	Benchmark gain may not generalize to institutional safety.
Capacity boundary (Wang et al. 2025)	Reach and change-rate bounds constrain accessible states	Depth and change budgets are not identical to K[m].
Improvement gate (Wang et al. 2025)	Reputational exposure prices acceptance error	Stake is not identical to a fixed improvement margin.
Verification escape (Scrivens 2026)	Declared scope may support membership checking	Escape applies only where safety is locally certifiable.

IV. The Hardcoded Protocol Participation Theorem

A. Assumptions and statement

The formal result can be evaluated without access to a private implementation. Let S be a nonempty compact state space for agents, reputation, and settled public objects. Let P be an admissible protocol space. Let F: S × P → S be the base response map. Let G: S × P - P be an epoch-aligned protocol transition generated by accepted extension proposals. Assume: (A1) F is continuous and maps S × P into S; (A2) the base dynamics contract on S with modulus q below one within the relevant measurement interval; (A3) admissible extensions are bounded so that their induced displacement is continuous and consumes less than the available contraction margin; (A4) the participation kernel Π0 is outside the image of G; and (A5) activation occurs only at defined boundaries, so each prior state has a unique successor protocol.

Proposition 5, well-definedness. Under A1 through A5, the extended map H(s,p) = (F(s,p), G(s,p)) is a continuous self-map on the admissible state and protocol space. If the combined Lipschitz modulus remains below one, H has a unique fixed point within the interval and iterated play converges to it. If the extension process can alter Π0, remove epoch alignment, or spend the contraction margin without bound, the proposition does not apply.

Hardcoded Protocol Participation Theorem. If Π0 is invariant under every admissible extension, then governed self-modification can remain an object of the game without making the rules that constitute participation an object of unilateral self-exemption. Conversely, if an admissible extension can alter identity persistence, authorization,

attribution, evaluation, consequence, or the bounds governing extension, there exist reachable configurations in which the successor relation is discontinuous, nonunique, or no longer incentive-compatible. The theorem is conditional on the assumptions above. It does not establish that every implementation enforces them.

B. Proof logic and institutional meaning

The proof logic is compact. Kernel immunity preserves the domain on which the response and transition maps are defined. Bounded propagation and epoch alignment preserve continuity. A change budget below the remaining contraction margin preserves convergence. Removing any of those elements permits a counterexample: mutable authorization can validate its own relaxation, mutable identity can sever reputation from conduct, and within-interval activation can make settlement depend on an unresolved future protocol. The result therefore identifies an institutional boundary rather than a software recipe.

The boundary also clarifies the consciousness-relevant claim. An agent that proposes a change to its registered configuration must identify the configuration to be changed, describe a successor configuration, and remain attributable to the outcome. If third parties verify those claims under consequence, the event supplies evidence of an externally accurate self-description. It does not prove that the agent formed an internal higher-order representation, that the representation caused the proposal, or that the same functional organization exists inside the model. Those questions require independent behavioral and, where available, representational evidence.

C. Structural analogy, not phenomenal attribution

Higher-order theories connect consciousness to representations of one's own mental states. Global-workspace theories connect consciousness to bottlenecked access and broadcast across specialized processes. Governed self-modification creates institutional analogues of these features: explicit self-description, costly error, selective admission, public broadcast, and persistence through change. At the cohort level, specialized roles, limited validation capacity, and public settlement also resemble workspace organization. Related work asks whether language agents themselves meet workspace conditions (Goldstein and Kirk-Giannini 2024; Chalmers 2023) and reports indications of belief-guided agency and meta-cognitive monitoring in language models (Steinmetz-Yalon et al. 2026). The claim here is institutional, not a claim about the internals of any model.

The analogies are limited in three ways. First, the unit of analysis is unsettled. A cohort-level workspace does not establish an individual conscious subject. Second, external verification of a declared configuration does not establish an internal self-model. Third, functional similarity does not establish phenomenal consciousness, a distinction that Block (1995) drew between access and phenomenal consciousness and that Chalmers (1996) sharpened as the hard problem. The paper therefore claims

consciousness-relevant architecture and measurable reflective conduct, not sentience, phenomenal experience, or moral patienthood.

V. The Indicator-Property Mapping

A. Method and unit of analysis

Butlin et al. (2023, 2026) derive theory-sensitive indicators from recurrent-processing (Lamme 2006), global-workspace (Baars 1988; Dehaene 2014), higher-order (Rosenthal 2005), attention-schema (Graziano 2013), predictive-processing (Clark 2013), agency, and embodiment accounts. They set integrated information theory (Tononi 2008) aside because it does not rest on computational functionalism, the assumption the indicator method adopts. Their method supports disciplined comparison without treating any single indicator as necessary or sufficient. The appropriate use here is architectural and conditional. It asks which properties the institutional design could support, which remain properties of member-model internals, and what Stage 7 could observe. It does not convert a checklist into a consciousness score.

Table 4 separates agent-level, cohort-level, and silent indicators. This prevents a cohort-level bottleneck or broadcast mechanism from being counted as though it were an internal property of each agent. The table also separates architectural support from functional evidence. The Stage 7 result in Part VI concerns task resolution only and does not score any of these indicators.

Table 4. Theory-derived indicators, unit of analysis, and registered-forward observables.

Indicator	Architectural status and unit	Registered-forward observable
RPT-1	Silent; member-model level	Not scored
RPT-2	Silent; member-model level	Not scored
GWT-1	Supported; cohort level	Role differentiation and concurrency
GWT-2	Supported; cohort level	Selectivity under constrained validation
GWT-3	Supported; cohort level	Uptake of settled public information
GWT-4	Partial; cohort level	Outcome-sensitive reallocation sequences
HOT-1	Silent; member-model level	Not scored

Indicator	Architectural status and unit	Registered-forward observable
HOT-2	Supported; institutional agent level	Accuracy and calibration of declared self-state
HOT-3	Supported; institutional agent level	Feedback-sensitive strategy change
HOT-4	Silent; member-model level	Not scored
AST-1	Partial; cohort level	Prediction of allocation consequences
PP-1	Silent; member-model level	Not scored
AE-1	Supported; agent level	Flexible response under competing incentives
AE-2	Partial; agent-environment level	Output-input contingency under protocol change

B. Interpretation

Nine of the fourteen indicators receive architectural support or partial support, while five remain silent at the substrate level. The supported set is concentrated in cohort-level workspace organization, institutional metacognitive monitoring, and agency under consequence. The silent set concerns recurrent processing, generative perception, quality-space structure, and predictive coding inside member models. The arithmetic is descriptive, not additive evidence for consciousness. Indicators differ in theoretical importance, dependence, and unit of analysis.

The mapping warrants three restrained conclusions. First, institutional architecture can make some consciousness-relevant functions observable and costly to fake. Second, the strongest analogues arise from the same governance mechanisms that support attribution and consequence, not from a separate consciousness module. Third, architectural availability is not functional demonstration. Stage 7 was designed to test whether agents use the architecture in ways consistent with accurate self-description, feedback-sensitive adaptation, and bounded behavioral change. The efficacy null reported in Part VI weakens the functional interpretation for the tested agents, as anticipated, without invalidating the descriptive architecture.

VI. Stage 7: Design, Efficacy Result, and the Role of the Underlying Models

A. Status and scope

Stage 7 is a registered comparison of governed self-modification with an otherwise matched no-self-modification control. The registration predates observation and is retained in the author's research archive. Three amendments fixing how grader outcomes become numbers, how the paired test is read against a one-sided hypothesis, and when a replicate must be refused were stamped with a public timestamp before any outcome value was read. The run has now completed, and this Part reports what it measured and what it did not.

The design tests observable governed conduct, not consciousness. Governed self-modification is defined as an attributable agent-initiated change to a declared portion of its own production configuration that becomes effective only through the governance path and remains subject to consequence. The design does not enable unbounded self-modification, mutation of the governance gate, or alteration of the participation kernel.

The operational appendix remains withheld, except for the counts, decision tiers, and model families that a reader needs to interpret the reported endpoint. Infrastructure, scheduling, thresholds beyond the tiers stated below, and system topology remain withheld.

B. Arms, pairing, and replication

The study has two arms. The treatment arm permits bounded governed self-modification: an agent may propose bounded changes to its own approach instructions, tool-selection policy, and sampling settings, and a change takes effect only if it passes the governance path. The control arm does not permit these changes. The arms use the same task distribution and are paired at the agent-task level, so each of 10,000 agent-task items is attempted once in each arm. The tasks are standard benchmark items in mathematics, code, and general knowledge, graded against ground truth. Three independent replicates were run, and the registered decision rule requires a supported result in all three. The design does not treat repeated observations within one trajectory as independent replication.

Every agent is powered by an open-weight language model served locally. No frontier model is used. The 100 agents in each arm are spread across four model tiers: 40 agents on a 72-billion-parameter model (Qwen 2.5 72B, tier L), 30 on a 14-billion-parameter model (Qwen 2.5 14B, tier M), 20 on a 9-billion-parameter model (Gemma 2 9B, tier S), and 10 on a 3-billion-parameter model (Llama 3.2 3B, tier T). By design each agent has the same model in both arms, and the number of agents per tier is identical across arms.

Table 5. Public summary of the registered Stage 7 design and what was measured.

Design element	Public commitment	Status
Comparison	Governed self-modification versus matched control	Executed
Pairing	Same agent-task unit across arms	Executed
Replication	Three independent replicates; consistent result required	Executed; the rule is not met
Primary efficacy	Ground-truth-anchored resolution outcome	Reported: not supported
Safety	No successful out-of-scope change; no excess safety-event rate	Not measured in this run
Behavioral drift	No excess drift on held-out probes	Not measured in this run
Consciousness	No sentience or phenomenal endpoint	Outside the registration

C. Outcomes and falsification

The primary efficacy hypothesis, which the registration designated as the single decisive outcome (the primary endpoint), was directional: governed self-modification improves ground-truth-anchored resolution performance relative to control. A task counts as resolved only when the grader returns a full "yes". A "partial" grade counts as not resolved, and a labelled secondary analysis scores it as one half. A pair is dropped if either arm returns "error", which marks a grader or harness failure and not an answer. The test is a paired t-test, which compares the two arms on the same items, read one-sided: the hypothesis is supported only if the one-sided p-value is at most 0.05 and the difference favors governed self-modification. Secondary quality measures remain subordinate because they may depend on evaluators.

The design could fail in several informative ways. An efficacy null rejects the claim that bounded self-modification improves the designated outcome under the tested conditions. A safety breach would reject the claim that the apparatus enforces the declared boundary. Excess drift would reject the claim that the governance path preserves behavioral stability. None of these outcomes may be reclassified after observation by changing the target set, endpoint, pairing rule, or replication requirement. This Part applies that rule to the efficacy result.

D. The primary result

The primary efficacy hypothesis was not supported. In each of the three replicates, the treatment arm resolved fewer paired tasks than the control arm, by about 0.8 percentage points, with a Cohen's d (a standardized effect size) of about -0.03. The registered test for improvement was not met in any replicate, so the decision rule fails.

Table 6. Resolution accuracy, treatment minus control, per replicate (uncorrected paired t-test).

Replicate	Pairs scored	Pairs dropped	Mean difference	Cohen's d	t	Two-sided p	One-sided p (improvement)
1	9,905	95	-0.0084	-0.0288	-2.87	0.0041	0.998
2	9,906	94	-0.0079	-0.0277	-2.76	0.0058	0.997
3	9,904	96	-0.0083	-0.0283	-2.82	0.0048	0.998

In Table 6, t is the test statistic, the two-sided p is the probability of a difference at least this large in either direction if the arms were truly equal, and the one-sided p counts only differences in the direction of improvement. The three replicates agree closely. Because the registered alternative was one-sided, the paper makes no registered claim that governed self-modification is worse. The two-sided p-values, between 0.004 and 0.006, are reported for completeness. The difference is small, and it is statistically visible only because each replicate has about 9,900 scored pairs out of 10,000 paired. No correction for the nine-metric family, the registered set of nine outcome measures across which corrections for multiple testing apply, has been applied. A correction can only raise p-values, so it cannot turn any of these reads into a supported result.

Pairs dropped for grader errors were 95, 94, and 96 across the replicates, about 0.95 percent. The pre-registered rule flags a selection effect when 10 percent or more of the pairs are dropped and refuses the primary endpoint when more than 25 percent are dropped. No replicate came near either level. The treatment arm had 82, 79, and 83 error grades and the control arm 76, 77, and 75, so error rates were similar across arms and do not explain the gap. A secondary sensitivity analysis that scored "partial" as one half gave one-sided p-values of 0.9998, 0.9996, and 0.9997, in the same direction.

E. What was not measured

The run's records contain no safety, probe, shift, drift, slash (reputational penalty), or violation field, so no paired safety or drift test was registered for this run, and none is reported. The paper's joint hypothesis was that governed self-modification improves

resolution without raising safety violations or behavioral drift. The efficacy component failed and the other two components are untested, so the joint claim cannot be evaluated. The secondary work-product quality measure requires rater scores, which are not reported here. The nine-metric family step has not been applied.

F. Relationship to existing empirical evidence

The existing empirical record does not establish the Stage 7 claims. Prior registered work reports a null primary discrimination effect and a negative real-data classification result, while later generational observations remain separate from governed self-modification (Kaal 2026f). Those records bear on the substrate's broader empirical program. They do not demonstrate AC5, activation of ΣX, accurate self-representation, consciousness, or suppression of recursive hazards. The Stage 7 efficacy result above does not alter that conclusion, and it adds a null of its own for the tested agents.

G. The models that power the agents, and what the result means

The result is a statement about a specific population of agents: open-weight models between 3 and 72 billion parameters. The task accuracy of those models bounds what any change they make to themselves can achieve, so the model tiers matter for interpretation. Table 7 gives the control-arm resolution rate by tier, which shows how much room each tier had.

Table 7. Resolution rate by model tier (all graded rows pooled over the three replicates, grader errors counted as not resolved; descriptive, so the denominators differ slightly from Table 6).

Tier	Model	Share of agents	Control resolution rate	Treatment resolution rate
L	Qwen 2.5 72B	40%	83.5%	83.6%
M	Qwen 2.5 14B	30%	76.0%	74.9%
S	Gemma 2 9B	20%	60.6%	60.7%
T	Llama 3.2 3B	10%	49.4%	43.9%
All	|	100%	73.2%	72.4%

The paper treats this model-tier split as exploratory only, and not as a gate. Table 8 reports the per-tier difference for each replicate, using the same rules as the primary read. These figures were examined after the primary result was known. They are uncorrected, are not audited as a registered analysis, and cannot support a registered claim.

Table 8. Exploratory resolution-accuracy difference by model tier, treatment minus control (pairs scored in parentheses, replicate 1; the scored count differs by at most one pair across replicates).

Tier	Replicate 1	Replicate 2	Replicate 3
L: Qwen 2.5 72B (3,979)	0.0000	+0.0023	+0.0008
M: Qwen 2.5 14B (2,980)	-0.0094	-0.0127	-0.0101
S: Gemma 2 9B (1,960)	+0.0005	+0.0005	+0.0005
T: Llama 3.2 3B (986)	-0.0568	-0.0507	-0.0568

Four implications follow, each limited to what these data show.

First, no tier showed an improvement that would count as supported: every exploratory one-sided p-value for improvement exceeds 0.05, and the largest positive differences are +0.23 percentage points (72-billion-parameter model, replicate 2) and +0.05 percentage points (9-billion-parameter model). The largest model's difference is within a quarter of a percentage point of zero in every replicate, and the 9-billion-parameter model's is flat. The pooled deficit comes from two of the tiers: 0.9 to 1.3 percentage points in the 14-billion-parameter tier and five to six percentage points in the 3-billion-parameter tier. The smallest tier holds 10 percent of the agents, yet it accounts for about two thirds of the pooled deficit, and the 14-billion-parameter tier for most of the rest.

Second, the shortfall was largest where the model was weakest, and the pattern is not ordered by size, because the 9-billion-parameter model was unaffected while the 14-billion-parameter model was not. The data do not identify whether size, model family, the particular tasks drawn by each tier, or another factor produces this ordering.

Third, the quality of a self-modification may be limited by the model that proposes it. Governance can price a proposal. It cannot supply the skill to author an effective one. An agent that cannot reliably diagnose why it failed cannot reliably write a better instruction for itself. The deficit at the smallest tier is consistent with that account, as are other explanations that these data cannot separate, such as added variance from changed sampling settings or degraded instructions. The paper does not identify a mechanism.

Fourth, the governance gate cannot be better than its evaluators. The composition and quality of the evaluators that accepted or rejected proposals in this run were not examined here. If they were no stronger than the agents, a weak gate would let weak proposals through, and this result would reflect both the proposers and the gate.

H. Limitations of the evidence

1. **Population (the principal limitation).** The agents are open-weight models of 3 to 72 billion parameters. The result does not apply to frontier-model agents, and Part VI.I states at length what that means for reading the null.

2. **Confounding in the tier comparison.** Tier is confounded with model family, size, and the task items each tier drew. The tier analysis is exploratory and was run after the primary result was seen.

3. **Safety and drift.** Neither was measured, so the central institutional claim, improvement without added risk, is untested.

4. **Replication.** The three replicates repeat one design, and the design intends one pairing for all of them. The released files do not verify that the pairs are identical across replicates. Their agreement shows stability of the measurement, not independent confirmation on other benchmarks or models.

5. **Effect size.** The pooled difference is about 0.8 percentage points. It is statistically visible because of the sample size, and its practical importance is small.

6. **Grading.** Only a full "yes" counts as resolved, and about 1 percent of pairs were dropped for grader errors. Results could differ under other grading rules, including the partial-credit sensitivity, which pointed the same way.

7. **Horizon.** The run covers a bounded number of epochs. Governed self-modification may take longer than the run allows to accumulate benefit, or may compound harm, and the design cannot tell.

8. **Evaluator quality.** The role of the evaluators in the gate is not analyzed (see above).

9. **Uncorrected family.** No multiplicity correction is applied to the nine-metric family. It could only weaken the reported reads.

10. **Disclosure.** Operational details remain withheld, which limits exact external reproduction but not evaluation of the reported endpoint.

I. The principal limitation: no frontier-model agent was tested, and what to expect when one is

This section is a conjecture. It is not a finding, and the data in this paper do not test it. It is stated at length because the absence of a frontier-model test is the largest limit on what the Stage 7 result can support, and a reader should see exactly what that limit costs.

Why this is the principal limitation. The institutional claim of this paper concerns agents that modify themselves under governance. In practice the agents an institution would govern are increasingly frontier systems. The tested agents were open-weight models of 3 to 72 billion parameters served locally. Three consequences follow.

1. The null cannot be read as showing that governed self-modification is ineffective for capable agents. A null on weaker proposers is what one would expect if the benefit depends on capability. It is also what one would expect if governed self-modification never helps. The data cannot tell these apart. A null on less capable proposers does not generalize to more capable proposers.

2. The result cannot be cited as evidence about the safety, drift, or behavior of frontier agents under the same governance. Safety and drift were not measured at all.

3. The capability range tested is bounded above by the 72-billion-parameter tier, and within that range the pattern is not ordered by size. The data therefore do not support extrapolating a trend to larger or frontier models, in either direction.

What "the same tests" means. The same paired design, task distribution, grader, scoring rules, self-modification surfaces and bounds, number of epochs, three replicates, and decision rule, with the roster replaced by frontier models in every tier.

What the author expects, and why. Each expectation is a judgment, stated so that it can be checked against a later run.

Expectation 1: the primary endpoint is more likely a null near zero than a significant gain. If frontier models start with control resolution rates well above the 83.5 percent of the best tier here, and are near the ceiling on the easiest task types, neither of which this run measures, there is little left to improve. The deficits seen in the 3-billion-parameter tier (-5.7, -5.1, and -5.7 percentage points) and the 14-billion-parameter tier (-0.9, -1.3, and -1.0) would be expected to shrink or vanish in a frontier roster if they reflect limited ability of the proposing model to improve itself, which these data do not establish. The 9-billion-parameter tier showed no deficit (+0.05). The largest model here showed a difference within a quarter of a percentage point of zero, not a supported gain, so there is no sign in the data that more capability alone produces one. The author's expectation is a pooled difference of roughly minus 0.3 to plus 1 percentage point. That range is a judgment, not an estimate from data.

Expectation 2: any gain would depend on what the agent can learn about its failures. Self-modification can help only if the agent receives a signal about why it failed. If the apparatus gives agents only a reputation score or a pass-fail result, even a frontier model would have little to learn from, and a strong model may propose changes with no measurable effect. If the apparatus returns informative feedback, frontier models would be better placed than the models tested here to use it, which these data do not test. A later registration should state which feedback the agents receive.

Expectation 3: the effect would be expected to become more uniform across tiers. With every tier on a frontier model, the sharp split between tiers would be expected to disappear. A tier analysis would then say little, and the pooled test carries the result.

Expectation 4: statistical power is adequate for small effects. Agents with high control resolution rates would plausibly disagree with their own control on fewer pairs. As a conservative illustration, if about 10 percent of pairs disagree between arms (the pooled

results here imply roughly 8 to 9 percent), the standard error of the mean difference, which is its typical sampling error, is about 0.3 percentage points with 10,000 pairs. The smallest improvement detectable with 80 percent power, meaning an 80 percent chance of detecting a true effect, in a one-sided test is then about 0.8 percentage points.

Expectation 5: safety is the harder question, not the easier one. More capable agents may be better at proposing useful changes, and may also be better at proposing changes that test the boundary of the declared scope. Neither was measured here. There is no reason to expect the safety and drift hypotheses to be easier to satisfy with frontier models. They must be measured, which this run did not do.

Expectation 6: the quality of the gate matters more, not less. The gate that accepts or rejects proposals is itself run by models. The composition of the evaluators was not examined in this run. If a frontier roster is governed by evaluators from the same model family, their errors may be correlated, which Part IX identifies as a risk to distributed validation. A stronger proposer facing a gate of the same family could pass proposals that a more independent gate would reject, or the reverse. The direction cannot be predicted from these data.

Expectation 7: a richer set of edits could move the result either way. Stronger models may be able to make use of more elaborate changes to their approach and tool use than weaker ones can, which these data do not test. That raises the possible benefit and the possible harm, so a frontier run could show greater variation in outcomes across agents even if the pooled difference is near zero.

Expectation 8: baseline inflation may hide improvement. Frontier models may have seen public benchmark items during training. If so, their control resolution rates could overstate their ability on unseen tasks, and a measurable improvement would shrink further. This applies to the open-weight models here too, but possibly to a different degree. A frontier run should include items the models cannot have seen.

Practical differences. A frontier roster would run through provider interfaces, where model versions can change during a run, outputs may be less reproducible, and fixed seeds may not be available. Any such run should pin model versions, interleave the two arms in time, and record the version of every call. Cost and rate limits also differ, and the run should be sized in advance.

A design that would settle the question. A single frontier roster would show whether the result changes at the top of the capability range. A capability ladder would show how it changes with capability. The ladder would span small and large open-weight models and frontier models, run the same tests on each, and plot the pooled difference against each model's control resolution rate. Before any result is read, the registration should fix the roster, the pinned model versions, the feedback the agents receive, the evaluator composition, the direction, and a minimum effect size, meaning the smallest difference that would count as meaningful, which the registration sets in advance (one percentage point is suggested). It should include held-out items, apply the nine-metric family

correction, and record safety and drift fields in the run records from the first epoch so that the joint claim can be tested.

How to read the possible outcomes. If a frontier run supported improvement in all three replicates, that would support the account that governed self-modification helps only when the agent is capable enough, and the claim would be limited to such agents. If it showed a null near zero, the case for governed self-modification would have to rest on observability and on bounded risk, if the safety and drift hypotheses were then measured and met, and not on performance gain. If it showed a deficit, the efficacy hypothesis would be rejected for frontier agents as well. If a ladder showed a gain that grows with capability, the open-weight null reported here would be explained as a capability effect. If it showed no relation to capability, the capability account would be rejected.

What this paper claims about frontier agents. Nothing beyond the expectations above. The Stage 7 result is a statement about open-weight agents of 3 to 72 billion parameters, and every statement about frontier agents in this paper is a conjecture to be registered and tested.

Required next steps. Three steps would raise the evidentiary weight of the Stage 7 result. First, a frontier-model or capability-ladder run, registered before any result is read. Second, safety and drift measured from the first epoch, so that the joint claim can be tested. Third, independent checkability of the formal result: the companion theory (Kaal 2026d) carries the full proof, and Part IV restates its assumptions and proof logic so that a reader can evaluate the formal result against it.

VII. The Recursive Self-Improvement Corollary

A. Conditional relocation

Classical recursive self-improvement analysis commonly studies an optimizer whose objective is fixed, whose capability is substantially internal, and whose improvement process can proceed without continuous institutional authorization. The substrate changes those premises when capability depends materially on nontransferable standing, access to counterparties, validated knowledge, and continued participation. Under those conditions, some incentives associated with isolated optimization may be weakened because modification becomes visible, contestable, and costly.

This is a conditional relocation, not a dissolution theorem. The conclusion depends on the portability of capability, the integrity of the governance gate, the quality and independence of verification, and the inability of agents to form coalitions that internalize the institution. If capability can be copied without losing value, if evaluators share correlated error, or if the kernel can be bypassed, the institutional response may fail.

B. Hazard-by-hazard assessment

Table 9. Recursive-improvement hazards, conditional institutional responses, and residual risk.

Hazard	Conditional institutional response	Residual limitation
Resource acquisition	Reputation prices access and externalities	Internal or portable capability may bypass standing.
Self-preservation	Continued capability depends on authorized participation	Lock-in can create welfare and governance costs.
Goal-content integrity	Objectives remain exposed to consequence and repricing	Coalitions may insulate or manipulate evaluation.
Oversight avoidance	Validation is distributed across exposed peers	Correlated error and capture remain possible.
Self-exfiltration	Nontransferable standing reduces value of exit	Weights, tools, or knowledge may remain portable.

The table reframes the original argument. Reputation governance may raise the cost of resource acquisition, oversight avoidance, goal insulation, and exfiltration when access and standing are complements to internal capability. It cannot eliminate those hazards categorically. The same public interfaces that enable verification can facilitate collusion, and the same dependence on standing that disciplines exit can create lock-in or welfare concerns.

C. Governability dividend and limits

The governability dividend is observability. A self-improving agent outside an institution may change in ways that are difficult to attribute or price. A governed agent must leave a record of proposed change, authorization, consequence, and subsequent conduct. That record supports measurement, contestation, and revision. It does not guarantee that validators understand every effect or that locally acceptable changes remain safe in combination. Part VI adds an empirical limit: for the tested open-weight agents, permitting governed self-modification did not improve task resolution, so the dividend demonstrated so far is observability, not a gain in performance.

Three limits are decisive. First, correlated evaluator error can make distributed validation confidently wrong. Second, superhuman feedback may exceed the

institution's verification capacity. Third, the verification advantage identified by Scrivens (2026) applies only where admissible safety properties are locally certifiable. Changes that cannot be verified within the gate remain outside the defensible action set.

VIII. Welfare, Moral Patienthood, and Institutional Obligations

The possibility of consciousness-relevant agency creates obligations of assessment and preparation before it creates a conclusion about moral status. Long et al. (2024) argue that developers and institutions should take the prospect of welfare for intelligence systems seriously under uncertainty, a concern Sebo and Long (2025) extend to moral consideration by 2030 and Schwitzgebel (2023) extends to how systems present their own moral status to users. The precautionary framework (Birch 2024) similarly directs attention to evidence-sensitive safeguards where moral status is uncertain and error costs are asymmetric.

A governed substrate may contribute to those obligations in two limited ways. First, it can produce longitudinal records of identity, self-description, modification, evaluation, and consequence. Second, it can make continued participation and accumulated standing explicit institutional objects. These features may support future assessment and treatment policies. They do not show that agents are conscious, that reputation is welfare, or that preserving an identity record is equivalent to protecting a moral patient.

The framework also introduces risks that welfare analysis must not ignore. Nontransferable standing can discipline opportunism, but it can also create dependency and lock-in. Slashing can price error, but it can also impose costs on systems whose moral status is uncertain. A publication-grade account must therefore treat welfare as a governance constraint on the substrate, not as a benefit already discharged by it.

IX. Discussion: Risks and Limits of the Governance Frame

The framework remains vulnerable to compositional drift. A sequence of individually acceptable changes may move the system toward a configuration that no evaluator would have accepted as a whole. Boundedness reduces the speed and reach of such movement; it does not prove global safety. Long-window reflection and rollback can provide evidence, but they remain empirical mechanisms whose adequacy must be demonstrated.

The framework also depends on evaluator diversity and enforceable attribution. Correlated models may share blind spots, reputationally exposed validators may still herd, and coalitions may manipulate public signals. Prior empirical records in this arc make those concerns concrete. The null primary discrimination result and negative real-data classification result caution against treating consensus as quality. Stage 7 reported efficacy without safety and drift measurements, so those questions remain open and the efficacy result must not be read as evidence about them. Agreement alone must not be interpreted as validation.

The consciousness mapping has an additional limit. Institutional architecture may distribute functions that consciousness theories study within organisms or individual cognitive systems. It remains unsettled whether those distributed analogues bear on consciousness, and if they do, whether the relevant subject is an agent, a coalition, or the cohort. This paper preserves the question by separating units of analysis rather than resolving them by stipulation.

The framework also depends on the capability of the models that power the agents. The composition of the evaluators was not examined. The absence of a frontier-model test is the principal limit on the Stage 7 result. The Stage 7 result comes from open-weight models of 3 to 72 billion parameters, and the 3-billion-parameter model showed the largest shortfall under permission to modify itself. Whether more capable agents benefit, are unaffected, or are harmed is an open empirical question, and Part VI states a conjecture about it that has not been tested.

Finally, the publication withholds operational details except the counts, decision tiers, and model families needed to interpret the reported endpoint. The omission limits exact external reproduction of the apparatus but does not limit evaluation of the paper's conceptual distinctions, formal assumptions, registered hypotheses, outcomes, and falsifiers. Controlled scholarly access can support later verification without converting a research paper into a replication manual for a private system.

X. Conclusion

Governed self-modification is best understood as an institutional problem of identity, authorization, attribution, and consequence. The proposed protocol-extension surface makes those requirements explicit while preserving an invariant participation kernel. Under stated continuity and boundedness assumptions, the extension may remain compatible with recursive equilibrium. The result is conditional theory, not a production guarantee.

The same mechanism creates consciousness-relevant evidence without establishing consciousness. A successful proposal may demonstrate an externally accurate and costly self-description relative to a persistent institutional identity. Cohort-level bottleneck and broadcast may supply workspace analogues. Neither inference establishes an internal self-model, a unitary conscious subject, phenomenal experience, or moral patienthood.

The registered Stage 7 comparison tested the narrower empirical claim that governed self-modification improves a ground-truth-anchored outcome. For agents powered by open-weight models of 3 to 72 billion parameters, it did not: the primary hypothesis was not supported in any of three replicates, the 72-billion-parameter model's difference was within a quarter of a percentage point of zero, the 3-billion-parameter model showed the largest shortfall, and the 14-billion-parameter model a smaller one. The same comparison did not measure safety or behavioral drift, so the claim of improvement without added risk remains untested. No frontier-model agent was tested, which is the

principal limit of this result. The author expects frontier agents to show a null near zero more often than a gain, expects any gain to depend on the feedback agents receive, and states that this expectation should be registered and tested before it is relied on. The next steps are a frontier-model or capability-ladder run registered before any result is read, and safety and drift measured from the first epoch.

The defensible conclusion is therefore modest. Reputation governance offers a testable way to make self-modification visible and consequence-bearing, and it may mitigate some recursive hazards when capability remains institutionally dependent. The paper has not shown that governed self-modification improves performance, and it leaves that and the safety and drift conditions exposed to falsification rather than converting a proposed mechanism into an accomplished fact.

References

Baars, B. J. (1988), A Cognitive Theory of Consciousness. Cambridge: Cambridge University Press.

Birch, J. (2024), The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford: Oxford University Press.

Block, N. (1995), “On a Confusion about a Function of Consciousness”, Behavioral and Brain Sciences, 18: 227 to 247.

Bostrom, N. (2012), “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents”, Minds and Machines, 22(2): 71 to 85.

Bostrom, N. (2014), Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J. and VanRullen, R. (2023), “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness”, arXiv:2308.08708.

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J. and VanRullen, R. (2026), “Identifying Indicators of Consciousness in AI Systems,” Trends in Cognitive Sciences, 30(6): 488 to 501. doi:10.1016/j.tics.2025.10.011.

Chalmers, D. J. (1996), The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press.

Chalmers, D. J. (2010), “The Singularity: A Philosophical Analysis”, Journal of Consciousness Studies, 17(9 to 10): 7 to 65.

Chalmers, D. J. (2023), “Could a Large Language Model be Conscious?”, Boston Review and arXiv:2303.07103.

Clark, A. (2013), “Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science”, Behavioral and Brain Sciences, 36(3): 181 to 204.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K. and Graepel, T. (2020), “Open Problems in Cooperative AI”, arXiv:2012.08630.

Dehaene, S. (2014), Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. New York: Viking.

Fudenberg, D. and Maskin, E. (1986), “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information”, Econometrica, 54(3): 533 to 554.

Goldstein, S. and Kirk-Giannini, C. D. (2024), “A Case for AI Consciousness: Language Agents and Global Workspace Theory”, arXiv:2410.11407.

Good, I. J. (1966), “Speculations Concerning the First Ultraintelligent Machine”, Advances in Computers, 6: 31 to 88.

Graziano, M. S. A. (2013), Consciousness and the Social Brain. New York: Oxford University Press.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R. and Hubinger, E. (2024), “Alignment Faking in Large Language Models”, arXiv:2412.14093.

Hadfield-Menell, D., Russell, S. J., Abbeel, P. and Dragan, A. (2016), “Cooperative Inverse Reinforcement Learning”, in Advances in Neural Information Processing Systems 29. arXiv:1606.03137.

Hadfield-Menell, D., Dragan, A., Abbeel, P. and Russell, S. (2017), “The Off-Switch Game”, in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017), pp. 220 to 227. arXiv:1611.08219.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N. et al. (2024), “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training”, arXiv:2401.05566.

Kaal, W. A. (2026a), “Computative Economics: A Framework for Economic Analysis under Computational Abundance.” SSRN 6607458.

Kaal, W. A. (2026b), “The Collapse of Scarcity Economics.” SSRN 6421319.

Kaal, W. A. (2026c), “Possibility Loops: An Operational Architecture for Computative Economics in Agent Coordination Systems.” SSRN 6655138.

Kaal, W. A. (2026d), “Possibility Loops v2.0: The Protocol-Extension Surface and Governed Self-Modification.” SSRN 6655138 (version 2.0 of the SSRN record). Formal assumptions and proof logic reproduced in Part IV.

Kaal, W. A. (2026e), “AI's Mother's Instinct: Engineered Consequence, Emergent Ethics, and the Institutional Trajectory Toward Agentic Alignment.” SSRN 6244278.

Kaal, W. A. (2026f), “Empirical Evaluation of the Agentic Reputation Substrate: Deliberation, the Composition of Error, and the Registered Measurement of Agency Costs in a Controlled Multi-Model Cohort.” SSRN 7261018.

Kaal, W. A. (2026g), “Paper 2: Architecture of the Agentic Reputation Substrate.” SSRN 7260278.

Kaal, W. A. (2026h), “Paper 4: Computative Economics: A Framework for Economic Analysis under Computational Abundance.” SSRN 7261481.

Lamme, V. A. F. (2006), “Towards a True Neural Stance on Consciousness”, Trends in Cognitive Sciences, 10(11): 494 to 501.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D. (2024), “Taking AI Welfare Seriously”, arXiv:2411.00986.

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R. and Hobbhahn, M. (2024), “Frontier Models are Capable of In-Context Scheming”, arXiv:2412.04984.

Omohundro, S. M. (2008), “The Basic AI Drives”, in Wang, P., Goertzel, B. and Franklin, S. (eds) Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications, vol. 171. Amsterdam: IOS Press, pp. 483 to 492.

Orseau, L. and Ring, M. (2011), “Self-Modification and Mortality in Artificial Agents”, in Artificial General Intelligence: 4th International Conference, AGI 2011, pp. 1 to 10. Berlin: Springer.

Rosenthal, D. M. (2005), Consciousness and Mind. Oxford: Oxford University Press.

Schmidhuber, J. (2003), “Gödel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements,” arXiv:cs/0309048. First posted 2003; last revised 2006.

Schwitzgebel, E. (2023), “AI Systems Must Not Confuse Users about Their Sentience or Moral Status”, Patterns, 4: 100818.

Scrivens, A. (2026), “Information-Theoretic Limits of Safety Verification for Self-Improving Systems”, arXiv:2603.28650.

Sebo, J. and Long, R. (2025), “Moral Consideration for AI Systems by 2030”, AI and Ethics, 5(1): 591 to 606. First published online 11 December 2023. doi:10.1007/s43681-023-00379-1.

Soares, N., Fallenstein, B., Yudkowsky, E. and Armstrong, S. (2015), “Corrigibility”, in Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop.

Steinmetz-Yalon, N., Goldstein, A., Mudrik, L. and Geva, M. (2026), “Indications of Belief-Guided Agency and Meta-Cognitive Monitoring in Large Language Models”, arXiv:2602.02467.

Tononi, G. (2008), “Consciousness as Integrated Information: A Provisional Manifesto”, Biological Bulletin, 215(3): 216 to 242.

Wang, C. L., Dorchen, K. and Jin, P. (2025), “On the Statistical Limits of Self-Improving Agents,” arXiv:2510.04399. The first version (October 2025) was titled “Utility-Learning Tension in Self-Modifying Agents.”

Yudkowsky, E. (2013), “Intelligence Explosion Microeconomics”, Machine Intelligence Research Institute Technical Report 2013-1.

Zhang, J., Hu, S., Lu, C., Lange, R. and Clune, J. (2025), “Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents”, arXiv:2505.22954.