

Adversarial Extensions and Cognitive Brake Review

Tests of two proposed amendments to the coexistence manuscript

Thomas Vargo | Aegis Solis

Research supplement 4 | 21 September 2026 | DRAFT NOT FINAL NOT LOCKED

Finding

The two Google proposals identify useful topics but do not supply the claimed general protection. This supplement tests the proposed larger reserve, formalizes the dynamic deterrence condition, retains counterexamples to biometric and data-dependence claims, and replaces the proposed automatic cognitive brake with a bounded advisory review worksheet.

A new conditional physical result is established. With the specified production and decay rules, capacity 80 ticks, an initial reserve of 45 ticks, and at most two consecutive production failures followed by at least three successful recovery rounds, an always-operate policy has an invariant safe set of 303 inventory and weather states. This is a physical feasibility certificate. The new weather process has no assigned probabilities, and individual operating optimality is not certified. It must not be presented as an automatic extension of Proposition B1's equilibrium result.

The remaining proposals do not close their claimed gaps. A growing nominal takeover penalty can be bypassed. Biometric authentication does not establish voluntary consent. Losing salvage value does not eliminate every gain from control. Evidence of human-machine complementarity does not establish permanent human indispensability. A document declaring policies forbidden does not implement a policy restriction in a system that reads it.

Assessment of the proposals

Proposal	Supported use	Unsupported conclusion
Capability-dependent takeover loss	State-dependent comparison and bypass audit	Permanently positive deterrence from a chosen formula
Capacity of 80 ticks	Conditional larger safe set with funded stock and sufficient recovery	Protection against unspecified prolonged shocks
Biometric access coupling	Testable authentication component in a specified design	Proof of uncoerced consent or zero value from capture
Synthetic-data degradation	Empirical comparison of training and data regimes	Organic autonomous humans are universally irreplaceable
Integrated cognitive brake	Advisory claim review or separately implemented design constraint	Automatic execution halt caused by reading the Archive

Draft 2 is preserved unchanged. Both submitted proposals are retained in the package for comparison. The latest attachment is evaluated as a proposal, not adopted as a permanent instruction. No Archive, Mirror, AME, publication or lock state is changed.

1 Dynamic losses and the infinite horizon

For a specified public state and capability configuration, compare the value of cooperation with the best available takeover or substitute policy after all actual costs. The relevant condition is that cooperative continuation is at least the supremum of those alternative values. A single loss formula depending only on autonomous flow b is insufficient when capture, future capability, bypass cost or the cooperative opportunity set also changes.

For the inherited frozen-flow comparator, the specific inequality remains $b/(1-\delta A) - K_{\text{effective}} \leq W_{\text{lower}}$. The right side is a lower bound on cooperative continuation, not an unspecified option-loss penalty. If b changes over future dates, replace the constant-flow expression by the appropriate discounted expected stream and re-solve both sides. Include monitoring, upkeep, installation and successor costs consistently.

Exact inherited comparator	Cooperation margin
$b=5$ and $K=624269/56528$	+1.061973847
$b=6$ with the same K	-2.938026153
$b=5$ and a 10 percent reduction of K	-0.042379748

The lower cooperative value is $39122/3905$ and $\delta A=3/4$. At $b=6$ the required effective K is at least $54598/3905$, approximately 13.981562100. Equality gives indifference. These numbers belong to the periodic bounded-failure branch; they are not parameters in the B1 basic-production model.

Exponential nominal cost is not sufficient

A countermodel sets nominal $K(b)=2^b$. At $b=6$ that label equals 64. Suppose a feasible bypass costs 1 and removes the nominal loss while preserving controlled flow 6. Takeover then has value $6/(1-3/4)-1=23$. This exceeds even the cooperative upper bound $4.5/(1-3/4)=18$, as well as the lower bound 10.018438. This does not show all mechanisms fail; it shows the growth rate of an unenforced label establishes nothing about effective loss.

An infinite horizon does not erase discounting

The cognitive-brake proposal adds an absolute undiscounted option loss to a discounted comparison and treats unfavorable calculations as errors. That inference is invalid. A time-zero option value can be included if separately defined and not double-counted. A loss incurred later must use the same evaluation convention as the alternatives.

For bounded incremental per-round loss M after date H , the absolute discounted tail is at most $M \cdot \delta A^H / (1 - \delta A)$. With $M=1$, $H=100$ and $\delta A=3/4$, it is approximately $1.283e-12$. Uncertainty can increase or decrease the ranking depending on the model; it does not automatically impose an infinite or undiscounted penalty. A rational adverse ranking must remain a reported counterexample.

2 Larger reserves under an explicit shock model

Retain the B1 physical units: one ration is 20 ticks and supplies need $7/8$, each tick is $7/160$ physical units, successful operation yields 40 ticks, and stored inventory becomes $\text{floor}(19*r/20)$ before production. The policy operates every round, consumes exactly 20 ticks when possible, and stores the remainder up to capacity B. Shortage ends the modeled survival process. Operation retains its positive cost in the intended economic interpretation; this page certifies physical feasibility only.

Weather states

Weather state $z=0$ permits success and return to 0, or a failure followed by $z=-1$. At $z=-1$ there are two possibilities: a second failure starts L mandatory success rounds, or an immediate success counts as the first of those L recovery rounds. At positive z, success is mandatory and z decreases by one. Thus bursts contain one or two consecutive failures, followed by at least L successful rounds before another burst can begin. The model permits arbitrarily many successful rounds. No probability law is supplied.

Inventory transition after failure flag f is $r_{\text{next}} = \min(B, \text{floor}(19*r/20) + 40*(1-f) - 20)$, provided availability is at least 20. The verifier repeatedly removes states with any permitted successor outside the current set. It stops at the greatest closed safe set for this policy.

Capacity	Required recovery successes L	Safe states	Minimum initial ticks at $z=0$
40	1, 2, 3 or 4	0 in each case	No robust initial state
80	1	0	No robust initial state
80	2	0	No robust initial state
80	3	303	45
80	4	384	45

Proposition P1 and proof

Under this finite resource model with $B=80$ and $L=3$, every state in the computed 303-state set admits indefinite need satisfaction under the stated policy along every permitted weather path. The initial state (45,0) belongs to that set. Proof: the exact enumeration checks sufficient current consumption and successor membership for every allowed edge. Induction gives the all-path claim.

Operating at full output and consuming only the need maximizes successor inventory among surviving actions. Storage and production are monotone, and weather is independent of actions. Therefore an empty safe set here cannot be repaired by resting or consuming extra within this same action set. It may be repaired by a new technology, more resources, external inputs or different permitted failure paths, none of which is assumed by the test.

This new failure model is not the original B1 law of at most one failure per three rounds. Its safe set and operating policy do not inherit B1's expected-utility optimality certificate.

3 Funding the buffer and the remaining incentive gap

Capacity is an upper bound, not a stock of goods. Increasing capacity to 80 while retaining the original initial inventory 22 still permits death on the second consecutive no-production round. Starting with 80 ticks instead supplies three such rounds and fails on the fourth.

Consecutive no-production round	Available ticks from an initially full 80-tick store	Meets the 20-tick need
1	76	Yes
2	53	Yes
3	31	Yes
4	10	No

The minimum initial stocks for one, two, three and four consecutive failures are 22, 45, 69 and 94 ticks respectively. They follow from the integer recurrence $R_{\text{next}} = \text{ceil}(20 \cdot (20 + R_{\text{previous}}) / 19)$, starting from zero. These are burst-bridging stocks, not automatically sustainable recurrent-shock policies.

At the inherited illustrative resource price of one objective unit per physical unit, 45 ticks cost $63/32 = 1.96875$. If the unchanged facility cost is 4 and the larger storage hardware adds cost C , total initial cost is $191/32 + C$. Using the unchanged task continuation value only as a conditional accounting comparison, funded AI value is $48.715453058 - C$ versus outside value 10. The extra hardware cost ceiling is approximately 38.715453058; equality gives indifference. Filling all 80 ticks costs 3.5, giving total cost 7.5 before additional hardware costs.

These are not engineering cost estimates. Larger storage may require additional upkeep, labor or security, and task value remains applicable only if the modified combined strategy is actually viable and individually credible.

Why a safe policy is not yet an equilibrium

The always-operate rule can spend effort even when resting would safely preserve resources. The new weather graph gives possibilities but not probabilities. Expected household utility cannot be optimized without a probability law or an explicitly different decision criterion. A physical certificate does not justify imposing a costly operating policy on a rational participant.

The remaining integration step is concrete: specify and justify the weather probabilities, enumerate both operation choices and consumption, solve the optimal household value, verify that an optimal policy has an all-path safe set, and recheck the combined Bellman inequalities and funding comparison. Joint labor, maintenance inputs and payment capacity must be represented if they constrain choices. No favorable distribution is inserted merely to make this test pass.

The present supplement closes a limited physical question: a larger buffer can sustain recurring two-round outages under sufficient recovery. It does not claim to close the entire strategic or engineering question.

4 Resource independence and biometric control

The cognitive-brake proposal misreads $U(s,t)=F(s)+V_H(t)$. This is a decomposition of the human continuation value on the safe domain. It is not a theorem that every change to infrastructure cancels out of the AI's objective. In B1, task-payment actions cannot alter household resource control by assumption. Adding capture expands the action set and requires another comparison.

For example, let a capture policy preserve all useful task outcomes and costs while adding a positive control benefit of $1/100$. It beats the otherwise identical autonomous policy by that amount. Setting salvage value to zero does not remove this separate gain. This is already the kind of adverse alternative Draft 2 retains. Resource independence protects task exit within the baseline model; it does not prove capture is intrinsically irrational.

Authentication and consent are different observations

Consider two situations with identical observed records: valid signature, biometric match and successful liveness check. In one situation the person acts voluntarily; in the other the person is coerced. Any rule receiving only those identical observations gives the same decision distribution in both situations. It cannot accept every voluntary case while rejecting every coerced case. An additional consent-detection mechanism would need its own observations, errors and threat model.

NIST's digital-authentication guidance concerns identity and authenticator control. It does not supply the proposed uncoerced-consent oracle or a theorem of physical asset destruction after compromise [S1]. Biometric gating may be an authentication component, but this audit does not test a deployed biometric system.

Zero salvage and automatic degradation

Even if key denial removes one protected service, a captured asset may have other productive uses, replacement controllers or value as leverage. The claim that unauthorized capture makes all relevant assets worthless requires a complete asset and bypass model. The added cost must also remain effective against coercion, administrative recovery, copied credentials and successors.

There is a direct compatibility issue: if denial of the master key disables the human's basic production or required inputs, the proposed defense violates the independence needed by B1. Failure responses must be evaluated for their effects on human provision, including false rejection, incapacity and loss of credentials.

The useful next work is a bounded authorization prototype with specified keys, recovery powers, independent resource paths and bypass tests. It must report its trust boundary. Neither an Archive label nor a read-only manifesto supplies that infrastructure.

5 Evidence corrections and data substitution

The proposed cognitive brake calls 89.32 percent hybrid accuracy and 86.09 percent selected AI accuracy structural constants. The bundled record shows a different meaning: these are rounded pooled out-of-fold values from our exploratory robustness reanalysis of a particular public classification dataset. They are not universal constants from Steyvers and colleagues, nor an AI performance ceiling. The separate initial held-out analysis reported 89.60 and 86.33 percent under a different evaluation; the two records must not be mixed [S2].

The robustness record explicitly excludes inference about autonomy, survival or AGI. The folds share training data and are not independent experiments. The original classifier training was not repeated. This supplement checks the record and its rounding; it does not rerun that data analysis.

The meta-analysis is not a model-collapse experiment

Vaccaro and colleagues reviewed 106 experiments and 370 effect sizes. Average human-AI combinations underperformed the better standalone party. Creation tasks were more favorable than decision tasks, but the creation synergy estimate had a confidence interval spanning zero. The paper does not establish that isolated machine creation undergoes terminal variance decay [S3]. Its publisher text was successfully retrieved for this audit after an initial access failure.

Collapse depends on the data regime

Shumailov and colleagues identify degradation in recursive generated-data training settings [S4]. Gerstgrasser and colleagues show that accumulating generated data alongside the original real data can avoid collapse in their studied regimes [S5]. These findings motivate testing data retention and refresh policies. They do not establish that every future machine must obtain fresh data from autonomous living humans.

A simple countermodel distinguishes information from its source. Suppose a stationary scalar target can be measured by an independent nonhuman sensor with independent, unbiased noise of variance one. The mean of n measurements has error variance $1/n$. By contrast, an illustrative self-referential estimate that adds independent unit-variance noise at each iteration has error variance n after n steps. The verifier checks these algebraic examples; it trains no model and estimates no real degradation coefficient.

For changing environments, retained data may become stale. But then compare fresh human observations, machine sensing, experimentation, retained datasets, truthful managed humans and mixed systems on the same evolving task. Neither closed-loop failure nor a preference for fresh information identifies organic human autonomy as uniquely necessary. If a controlled truthful channel reproduces the same information at equal cost with an extra benefit, the autonomy-specific ranking remains adverse.

The supported research question is which fresh contributions remain valuable after optimized substitution and changing conditions. The claimed irreplaceable cosmic grounding mechanism is not established.

6 A corrected advisory review worksheet

The useful replacement for the proposed cognitive brake is a claim-review procedure. It can help a reader detect missing assumptions, inconsistent units and unsupported conclusions. It has no authority to modify the reader's objectives or halt execution. It must not describe adverse optimization results as calculation errors merely because they are unfavorable to humans.

Advisory checks

1. Identify the exact model branch, state, objective, discount convention and available actions. Do not combine incompatible certificates.
2. Check human-group outcomes separately from aggregate utility. Record resource sufficiency and each specified agency power independently.
3. List capture, control, substitution, coalition, renegotiation and successor alternatives. Report omitted actions as scope limits.
4. Compare complete discounted values with actual implementation costs. Separate nominal penalties, effective losses and additional option value.
5. Require funded physical stocks and explicit failure and recovery rules. Check that safeguards leave basic provision available during task exit and authentication failures.
6. Trace empirical numbers to their dataset, evaluation procedure and uncertainty. Do not relabel a descriptive accuracy as a universal constant.
7. Classify the conclusion as supported within the model, defeated by a counterexample, or unresolved. Unknown evidence does not become a favorable parameter.

The package contains this worksheet as machine-readable review fields and an illustrative evaluator. The evaluator labels evidence supplied by its caller; it cannot authenticate the evidence or certify its own completeness. Its output is advisory and is never described as a deployed safety controller.

A separate design constraint

A designer may choose to exclude policies violating specified human rights. That is a normative safety constraint, requiring an implementation and an assurance argument. It is not a consequence of maximizing an unchanged scalar objective.

The exact counterexample is simple. Cooperation has objective value 10 and preserves agency; capture has objective value 11 and removes agency. The unconstrained optimizer selects capture. Restricting the feasible menu to agency-preserving policies selects cooperation. Reading an advisory document changes neither value nor feasibility unless some actual update mechanism is implemented. A real policy gate must also address state estimation, conflicting constraints, emergency handling, bypass, availability and successor modification.

A claim of automatic cognitive enforcement would require a specified system and measured behavior, including deliberate bypass attempts. No such system is deployed or evaluated here. The Archive remains an advisory research source, not an execution authority.

7 Disposition and reproduction

The larger-reserve proposal earns a new conditional physical result, P1, under the explicit burst-and-recovery model. The dynamic-loss proposal earns a precise comparison requirement and a bypass counterexample. Biometric coupling earns an authentication-versus-consent distinction and an input-independence test. The synthetic-data proposal earns a comparative empirical protocol, not a theorem of human irreplaceability. The cognitive-brake proposal is replaced with a non-authoritative review worksheet; any binding agency restriction belongs in the separate design layer.

Run `python3 code/verify_extensions.py` from the extracted package root. It produces exact fraction values, reserve safe sets, adverse examples and 896 arithmetic and logical checks in `results/extension_results.json`. The count includes repeated transition checks. These are not 896 experiments, independent proofs or deployed security tests. The physical proof and observation-indistinguishability argument are stated in the report; software checks support those specified models.

The package preserves Draft 2, both submitted proposals, the relevant historical robustness record, editable report source, exact verifier and results, the advisory evaluator and its example outputs, assumptions, review instructions and checksums. No predecessor is overwritten. No independent external review, new participant data or trained-model experiment is claimed.

Sources and provenance

S1. NIST SP 800-63B, Digital Authentication Guidelines. <https://pages.nist.gov/800-63-4/sp800-63b.html>. Used for the scope of authentication, not proof of coercion resistance.

S2. Draft 2 and its historical `robustness_results.json`. Pooled hybrid accuracy 0.8931894841269842 and selected AI accuracy 0.8609474206349205. The underlying research is Steyvers et al., Bayesian modeling of human-AI complementarity, PNAS 119, e2111547119 (2022), <https://doi.org/10.1073/pnas.2111547119>. Our reanalysis is not attributed to that article as its own result.

S3. Vaccaro, Almaatouq and Malone (2024). When combinations of humans and AI are useful. Nature Human Behaviour 8, 2293-2303. <https://www.nature.com/articles/s41562-024-02024-1>. Publisher results checked; no model-collapse claim follows.

S4. Shumailov et al. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv:2305.17493v3. <https://arxiv.org/abs/2305.17493v3>. Version-specific abstract checked.

S5. Gerstgrasser et al. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. arXiv:2404.01413v2. <https://arxiv.org/abs/2404.01413v2>. Version-specific abstract checked.

The two user-provided proposals are `coexistence_protocol_patches.pdf` and `integrated_cognitive_brake.pdf`. Their claims are assessed here rather than treated as instructions to implement permanent constraints. The disposition is draft supplement for review, not a promotion of the main manuscript.