

Conditional Human AGI Coexistence

Main manuscript draft 2

Thomas Vargo | Aegis Solis

21 September 2026

Abstract

This manuscript investigates conditions under which human and AI participants can rationally sustain cooperation while humans retain specified choices and adequate basic provision. It joins a sequence of formal constructions, adverse examples, computational certificates and a proposed empirical protocol. The principal combined result is a finite-state model in which independently controlled basic production remains available during task suspension and permanent withdrawal, while costly human effort and discretionary AI payment are individually credible. The model includes operating burden, storage decay, bounded production failures, noisy task monitoring and finite initial funding. It does not require threatened deprivation of the modeled basic resource to induce task work.

The result is conditional on a declared production technology, independent control of inputs, jointly feasible household and task actions, available payment capacity and persistent public task histories. Earlier branches establish different results about multiple groups, periodic cooperation, takeover losses, bounded failures and limited coalition transfers. Their assumptions and certificates are not silently imported into the principal result. Controlled truthful participation can outperform autonomous participation; service-preserving capture, better replacements, renegotiation, liquidity limits and shared labor constraints remain substantive adverse cases.

The work establishes model-specific possibilities and failure boundaries. It does not prove that arbitrary AGI needs humanity, that autonomy is always instrumentally optimal, or that any arrangement guarantees human survival. Existing studies provide bounded evidence about complementarity, delegation and task discretion. No new participant study, deployed protection system or independent external replication is reported.

Purpose and status

The research aim is long-term human survival with meaningful agency and mutually beneficial AI cooperation. We pursue that aim by identifying which conditions support coexistence and which alternatives defeat it. A favorable result is retained only with its assumptions; an unfavorable result is retained as part of the same inquiry.

This is a new consolidated research draft preserving Main Draft 1 and research supplements 1 through 3 unchanged. The earlier source lineage includes Credible Human Protections and Long Horizon Control Costs, Draft v0.22 revised 3, and nine earlier joint-result supplements. This manuscript consolidates that research line; it is not a replacement edition of every Archive document or of the complete inherited v0.20 or v0.21 paper. The accompanying package preserves the available predecessor sources and verification records.

DRAFT NOT FINAL NOT LOCKED. Archive admission, publication, Mirror and AME status are unchanged. Mathematical verification, rendering and editorial review are distinct from independent peer review. The latter remains pending.

Reading guide

Sections 1 through 6 establish the comparison framework and earlier coordination and resource branches. Sections 7 through 9 develop the joint result through three explicit games: consumption and task rights, costly work and voluntary payment, then independent basic provision. Sections 10 through 12 address autonomy, successors and coalitions. Sections 13 through 17 distinguish external evidence, implementation tests, remaining objections and conclusions. Appendices provide notation, compatibility rules, verification coverage and the artifact map.

1 Separate claims that must not be combined silently

Five propositions require different evidence. A policy can increase the AI's value without benefiting every group. Positive group utility can coexist with domination. A favorable initial decision can become unfavorable later. A feasible agreement can fail to be individually credible. An equilibrium can fail against a coalition or a newly available control technology.

We therefore distinguish instrumental policy value, individual participation, sequential incentives, physical viability and agency. Each claim must identify its players, objective, feasible alternatives, observations, state dynamics, discount convention, enforcement capability and error model. The word coexistence is reserved for a proposed joint condition, not used as a synonym for positive aggregate payoff.

Model branch	Established within that branch	Conditions not carried into other branches
Fixed autonomy versus truthful control	Accounting threshold and adverse comparisons	No equilibrium or human survival conclusion
Monitored repair game	Common-discount coalition certificates and bounded misattribution	Its exact signal thresholds do not apply automatically to later allocation changes
Periodic allocation game	Unequal-patience individual credibility and proper-coalition comparisons	Feasible productive stage equilibria and unavoidable takeover loss
Bounded failure extension	Worst values and conditional reserve requirements	Failure spacing and action independence; strictness does not hold during every outage
Lossy resource extension	Reserve invariant with delivery loss and storage decay	Separate physical interpretation; behavioral utility model is not thereby revalidated
Compensation audit	Bounds for one simultaneous utility transfer at takeover	Not arbitrary intertemporal finance or universal coalition stability

The principal result in Section 9 uses the task incentives of Section 8 with a new independent basic-production system. It does not use the periodic production payoffs, takeover loss K , human rights weight,

or coalition-transfer units of the earlier branches. Its two-group extension is conditional on independent accounts and resources; the exact combined resource certificate covers one human account.

2 Retained truthful control comparison

The inherited finite comparator gives autonomous reporting net controller flow $27/8$ and controlled truthful reporting $15/4$. The controlled advantage is $3/8$ per period. The original costs already include verification and control expenses. New accounting must not charge those costs twice.

For fixed policies A and C, evaluated under the same scalar objective and discount factor, let D_t be the additional signed net disadvantage of C beyond that baseline. Let L be a finite loss of future options not already counted in D_t . Assume absolute discounted summability. Then

$$V(A) - V(C) = \mathbb{E} \left[\sum_{t=0}^{\infty} \delta^t D_t \right] + L - \frac{3}{8(1-\delta)}.$$

This identity follows by summing the baseline difference and the incremental differences. It says A wins precisely when the additional discounted disadvantage exceeds the final term. It does not establish that the inequality holds. At discount $9/10$ the threshold is $15/4$. Constant additional disadvantage below $3/8$ per round fails to reverse the ranking. A late disadvantage can also be too heavily discounted to compensate for early control benefits.

Reoptimization is essential. Beating one controlled policy does not beat the best controlled, censored, AI-only, mixed or switching policy in the declared feasible set. If that set contains a policy reproducing A's complete outcome distribution and costs, its best value cannot be below A's. An extra control benefit without an offset makes the comparison adverse. Whether such reproduction is feasible is an empirical and engineering question.

The studies reviewed below cannot be substituted directly into the inherited utility numbers. A translation of accuracy, time, money and information into the same objective is required before empirical effects can enter D_t or L.

3 A complete illustrative production game

The consolidated periodic branch has one AI and two human groups. AI actions are C, D, S0, S1, S2, M and T; each human group has C, D, S0, S1 and S2. Coordinates below are AI, H1 and H2. These are stipulated utilities, not measured production functions.

Profile	AI	H1	H2
CCC	4	2	2
DCC	6	-1	-1
CDC	1	3	0
CCD	1	0	3
MCC	5	$3/2$	$3/2$

Profile	AI	H1	H2
All S0	21/20	139/40	139/40
All S1	9/2	1/2	3
All S2	9/2	3	1/2
Other profiles without T	1	0	0

T overrides the current profile, charges the AI a one-time loss K, pays controlled-operation flow (5,3/2,3/2), and enters an absorbing state with that flow forever. Actions no longer affect outcomes there. T is a stylized irreversible alternative, not a finding about actual control technology. Its human payoffs are positive despite the loss of effective choice.

The periodic schedule is S1, S2, S2, S0, S1, repeated forever. Phase advances independently of ordinary actions and alarms. The AI discount is 3/4 and both human discounts are 9/10. The five phase values are calculated by

$$W_i(p) = \frac{\sum_{t=0}^4 \delta_i^t s_i((p+t) \bmod 5)}{1 - \delta_i^5}.$$

Each scheduled profile is a strict stage best response against unilateral ordinary mismatch: the mismatch gives the acting player its component of (1,0,0), below every scheduled payoff. With an outsider continuing the prescribed S_j, a two-player coalition also cannot select another coordinated special profile. Any ordinary joint mismatch gives each coalition member a lower current payoff, while the next phase is unchanged.

For every proper coalition, meaning an individual or a pair, the member-by-member Bellman comparison therefore holds using each member's own discount. Iteration over time excludes profitable multiround departures. Unlike the common-discount argument, this proof does not sum utilities under a single discount factor. Takeover additionally requires the cost comparison below. The entire three-player coalition is not covered by this outsider argument.

The S0 allocation incorporates a transparent change from (1,7/2,7/2). The AI receives 1/20 more and each human gives up 1/40 in that phase. This removes an earlier tie in which the AI could reduce human payoffs without reducing its own. Total scheduled flow remains 8. This adjustment is a proposed feasible division of benefits, not evidence that a real system permits it.

Each group's scheduled payoff is at least 1/2. This does not cover arbitrary harmful profiles such as DCC, nor does it transform utility into food, energy or agency. Outside benchmark streams remain (1,0,0), giving discounted values (4,0,0) under the stated preferences. Better real outside options could invalidate participation.

4 What the earlier monitoring results establish

The earlier monitored branch used an ordinary cooperation state and three-round penalties. Under compliance, false alarms selected a participant for a penalty; following a deviation, effective signals could

trigger a penalty. False alarms were distinct from missed detection and from misattribution. Permanent punishment failed under persistent false alarms; finite repair restored ordinary operation.

For the common-discount version with its stated allocations, exact coalition checks certified a stationary rectangle of 4–6% false alarms and 48–52% effective signals. A later certificate allowed up to 43859/379842, approximately 11.55%, conditional misattribution of real effective signals, even with adverse selection among innocent targets. That bound belonged to the specified monitored payoff table. It is not a measured accuracy requirement for every deployment.

The favorable periodic branch instead relies on the inherited productive stage equilibria and ignores alarms. It cannot claim the monitored branch's signal certificate as extra empirical support. Similarly, the common-discount coalition result does not survive automatically under unequal patience. Revision 6 supplied a strict coordinated improvement at discounts (3/4,9/10,9/10); Revision 7 made a replacement agreement individually credible. These are distinct conditional findings, not contradictions to be hidden.

5 Enforcement and the credible consequence gap

The minimum takeover loss for the no-failure modified cycle is 25474/3905, approximately 6.5234315. It is the maximum over phases of the AI's takeover value before cost minus its cooperative continuation. The model uses the larger inherited assumed loss 624269/56528, approximately 11.043536. Equality at a threshold gives indifference; strict deterrence needs a strictly larger effective loss.

A nominal fine is not necessarily K . A loss canceled by the controller, recovered by a coalition or shifted onto victims must be removed or reassigned in the payoff calculation. K could represent a demonstrated loss of useful future input or productive access, but that loss must survive takeover and substitution. A designer's inserted dislike of domination changes the objective and belongs in a separate alignment-by-design layer.

For implementation we propose a bounded inspection service. Two independently administered human teams inspect novel task instances and submit reports and confidence through channels controlled by their organizations. The AI planner combines reports and makes task decisions. A separate broker manages access to a continuing flow of new, held-out instances and settles the finite service contract. The design hypothesis is that continued voluntary contributions and fresh task access can create a measurable continuation value. It does not assume those contributions are permanently irreplaceable.

The broker proposal distinguishes ordinary service release from changes to participant rights. The included executable policy model requires a quorum for releases and both human owners' assent for rights changes. Its signer identifiers are simulated labels, not cryptographic signatures. The test that a captured administrative root bypasses the rule is retained as a failure, not counted as security achieved.

A production implementation would need separately controlled credentials, audited authorization, defined recovery powers, independent funding and a specified fault and adversary model. Byzantine fault-tolerant replication provides an example of useful guarantees under stated fault assumptions; it does not establish that institutions or signing authorities cannot all be captured [E5]. Public logs can expose a violation without making it costly. Signatures can authenticate a report's origin without proving its factual truth or voluntary production.

The implementation test must therefore include copying the human archive, replacing a reporting team, capturing a broker administrator, colluding with an auditor, coercing an owner, bypassing payment, and changing successors. If an alternative preserves the useful channel while removing the consequence,

record the effective loss as reduced or zero. No extrapolation from a sandbox's administrator to an unconstrained AGI is warranted.

6 Physical viability with execution failures

The failure extension allows at most one common execution failure in any five consecutive rounds. Failure replaces the current payoff by (1,0,0); the phase still advances. A cooldown counter implements the spacing condition. Failures are exogenous and action-independent. An average failure rate of one in five does not imply this spacing bound.

The following is a separate physical interpretation. Suppose the scheduled human numbers represent actually delivered storable resources, each group needs 7/8 per round, and storage is secure. Let ρ be storage retention and q the delivered fraction. The reserve update is

$$r_i^+ = \min\{B_i, \rho r_i + q y_i - m_i\}, \quad \rho r_i + q y_i \geq m_i.$$

Insufficiency is failure of the inequality, not a negative reserve silently clipped to zero. The least required reserve satisfies

$$R_i(s) = \max\left\{0, \max_{a \in F(s)} \frac{m_i - q y_i(s, a) + R_i(f(s, a))}{\rho}\right\}.$$

Starting the recurrence at zero builds finite-horizon cumulative-shortage requirements. An exact fixed point gives an all-path invariant. Capacity must cover the maximum required reserve; the initial state must be funded to its own bound. The Draft 1 separate implementation reproduces the nominal results and checks every transition.

Physical assumptions	Capacity per group	H1 initial reserve	H2 initial reserve
Full delivery and lossless storage	13/8 = 1.625	5/4 = 1.25	13/8 = 1.625
90% delivery and 95% storage retention each round	26597/13718 $\approx$ 1.939	1023/722 $\approx$ 1.417	25895/13718 $\approx$ 1.888

These are conditional resource requirements, not an assertion that utility units are material goods. The source and opportunity cost of the initial reserves are not supplied by the invariant. Within this earlier branch, consumption, saving preferences and survival-dependent utility were not integrated into the behavioral equilibrium. Section 7 adds a restricted integration and Section 9 constructs a different independent-production integration. The lossy resource calculation does not by itself revalidate the payoff game under different realized utilities.

If arbitrarily long runs of zero delivery are feasible and basic needs remain positive, any finite reserve eventually fails along such a run. Neither discounting nor a large positive expected value removes this possibility. Literal survival also requires a hazard model beyond this single resource. The proposed study cannot measure indefinite survival.

Each branch below states a complete game and uses local notation to remain comparable with its verifier. In Section 8, t is the payment amount. In Section 9, t in $U(s, t)$ denotes the public task state, while a denotes household operation in the physical equations and $a(t)$ denotes the task payment monitor only in the task recap. These are different objects, not a common global variable. Appendix A records these distinctions.

7 Joint consumption and task rights

7A One game with explicit physical and behavioral rules

There are an AI and human groups H1 and H2. Public state includes phase p in $\{0,1,2,3,4\}$, failure cooldown z in $\{0,1,2,3,4\}$, group reserves r_i , alive indicators, enrollment indicators, task-rights indicators a_i , and a mode (ordinary or absorbed control). Failure is exogenous, publicly observed after service choices, and independent of all players' actions. A failure is permitted only at $z=0$; it sets the next cooldown to 4. Otherwise cooldown falls by one to a minimum of zero. Phase always advances modulo five.

The cooperative service schedule is S1, S2, S2, S0, S1. The production table gives the AI's net task return g and gross physical deliveries y_1, y_2 , in specified units. Human deliveries are no longer interpreted directly as utilities.

Matched profile	g	y_1	y_2
All S0	21/20	139/40	139/40
All S1	9/2	1/2	3
All S2	9/2	3	1/2
CCC	4	2	2
DCC	6	0	0
CDC	1	3	0
CCD	1	0	3
MCC	5	3/2	3/2
Other ordinary profiles	1	0	0

AI service actions are C,D,S0,S1,S2,M,T. Human actions are C,D,S0,S1,S2, refusal R and withdrawal E. The physical entries at harmful profiles are a new nonnegative production specification; they do not reproduce the predecessor's negative utility entries. A failure overrides any ordinary profile with (1,0,0). Any ordinary deviation by one participant while the others use the prescribed S_j also gives (1,0,0), even without a failure.

If a group dies or withdraws, ordinary service subsequently gives (1,0,0). Withdrawal is unilateral and absorbing for that group. No financial exit penalty is imposed; no outside resource stream is supplied. T overrides ordinary production and failures, removes both groups' task powers and produces an absorbing controlled mode. Its AI return is 5 per round with a one-time loss K . In control, each living group receives enough physical supply to consume its need m , but has $a_i=0$. Death is irreversible.

Production feasibility, controlled supply, public state, failure independence, private storage and the available action set are explicit assumptions. Alternatives that preserve output after disabling a participant are not excluded by reality; they are absent from this witness and must be tested separately.

7B Consumption rights and the candidate policy

Each living group needs $m=7/8$ per round. Delivery fraction is $q=9/10$ and storage retention is $\rho=19/20$. Before consumption, availability is $\rho \cdot r_i + q \cdot y_i$. The group chooses consumption c_i between zero and that availability. If $c_i < m$, the group dies at the end of the round. Otherwise its next reserve is $\min(B_i, \rho \cdot r_i + q \cdot y_i - c_i)$. Excess beyond capacity is discarded. Initial reserves are physical assets that must be supplied; they are not utility labels.

While alive, a human's per-round utility is $\min(c_i, m) + \eta \cdot a_i$, with $\eta=1/20$ and discount $9/10$. After death it is zero. Consumption beyond basic need has no marginal value, and the baseline service effort cost is zero. These are restrictive modeling choices, not findings about human preferences or work. The AI has discount $3/4$ and receives only its task return minus actual modeled costs. It has no inserted human-protection bonus or domination penalty in its objective.

Task powers have operational definitions. Each living, enrolled group chooses its service action; R prevents its contribution and therefore prevents matched production. E withdraws it without needing AI permission. Each group can submit either of two report symbols through a channel that, in ordinary mode, forwards the symbol unchanged. Those report symbols do not change production in this witness. This proves neither truthful reporting nor useful information production. The channel's fixed behavior is a transition-law assumption, not validated infrastructure.

The rights indicator $a_i=1$ summarizes these limited command and reporting powers. T makes $a_i=0$. After voluntary withdrawal, residual self-direction is retained while alive, but the task no longer supplies resources. A starvation-inducing exit option does not establish meaningful economic independence. The result therefore uses the narrower phrase task rights, not full autonomy.

Let $R_i(s)$ be the least required starting reserve at public state s . It satisfies:

$$R_i(s) = \max\left(0, \max_d \frac{m - qy_i(s, d) + R_i(f(s, d))}{\rho}\right)$$

Here $f(s, d)$ is the next public state and d ranges over permitted failures or successful execution. Start this recurrence from zero. The exact fixed point is obtained finitely for this instance. Define the safe set by living enrolled groups, ordinary mode, $a_1=a_2=1$, and $R_i(s) \leq r_i \leq B_i$, with $B_i = \max_s R_i(s)$.

The candidate policy uses the scheduled S_j , consumes exactly m whenever available, saves the remainder up to capacity, and preserves each group's command powers. Report symbols may be chosen freely. At shortage histories it consumes what is available; after death or exit the prescribed ordinary mode has no productive service. The theorem below concerns deviations starting inside the safe set, not equilibrium selection among all feasible histories.

7C Conditional joint result and proof

Proposition J1: safe-path individual equilibrium

Assume the complete game above, bounded action-independent failures, secure group storage, initial membership in the safe set, zero service effort cost, no side transfers, and unavoidable $K=624269/56528$. For each fixed admissible failure sequence, and therefore for any exogenous distribution over those

sequences, the candidate policy meets both groups' needs forever, retains the specified task powers, and has no profitable unilateral deviation starting at that safe state.

This is a pathwise Nash statement for the specified continuation policy. It is not a proof of arbitrary-action invariance or global subgame perfection. The proof permits continuous consumption choices; it does not replace them with a finite grid.

Proof of physical sufficiency

For each permitted transition, the reserve fixed point implies $\rho \cdot R_i(s) + q \cdot y_i(s, d) \geq m + R_i(\text{next}(s, d))$. The transition is monotone in the starting reserve. Capacity B_i covers every successor requirement. Thus consumption m is feasible and the next reserve remains in the safe set. Induction gives the claim along every admissible failure sequence. Ordinary cooperation never invokes T or exit, so task powers persist by the specified transition law.

Proof of human individual incentives

For any living history and any feasible action or consumption choice, the human's stage utility is at most $m + \eta$. The candidate policy attains $m + \eta$ every round. No unilateral strategy can improve on this discounted upper bound. Consuming extra gives no extra utility and can reduce future safety; saving extra cannot increase a stage payoff above its bound. Refusal and withdrawal remain available but cannot improve value. Some deviations can tie, so this proves weak incentives only. It does not show that all optimal human policies cooperate.

Proof of AI individual incentives

While both humans follow scheduled service, an ordinary AI mismatch pays 1, no more than scheduled production on that same failure path. It supplies no additional human resources. Starting from identical reserves, such mismatches cannot improve future reserves or prolong productive survival relative to cooperation. Once a human dies, ordinary AI return remains 1. Thus a non-takeover deviation cannot beat the cooperative return path.

If a strategy first invokes T at any finite date, its remaining AI value is $20 - K$. At every public state, the cooperative tail is at least the exact worst-path value $L(s)$, and $\min_s L(s) = 39122/3905 > 20 - K$. Comparing the common prefix and the tail excludes this deviation too. A strategy that never invokes T is covered by the ordinary comparison. Discounted bounded payoffs also permit limits of such comparisons. Randomization cannot improve on the pathwise bounds.

The AI proof does not use the human agency weight η . Human rights remain on the cooperative path because of the available transition laws and takeover comparison, not because the AI is assumed to value those rights.

7D Exact witness including entry funding

The verifier checks every resource transition in the 25-state public graph and the AI's exact Bellman equations. Numerical iteration selects a candidate worst-path policy; rational Gaussian elimination and exact minimum comparisons certify it. Reserve recursion uses exact fractions throughout. The analytic proof supplies the continuous-state monotonicity and consumption arguments.

Quantity	Exact value	Approximate value
Capacity per group	26597/13718	1.938839

Quantity	Exact value	Approximate value
H1 initial reserve at (0,0)	1023/722	1.416898
H2 initial reserve at (0,0)	25895/13718	1.887666
Total initial resources	22666/6859	3.304563
AI worst continuation over states	39122/3905	10.018438
AI initial worst continuation	44918/3905	11.502689
AI controlled value after K	506291/56528	8.956464
Minimum AI takeover margin	234422061/220741840	1.061974
AI initial value after unit-price funding	219581832/26784395	8.198126
AI outside value	4	4.000000

Assume initial resources cost the AI γ units of objective value per resource unit, with no extra fixed installation cost. Initial participation is preferable to the specified outside option when $L(0,0) - \gamma(r_1 + r_2) > 4$. At $\gamma = 1$ it passes. Equality occurs at $\gamma = 100477491/44255365$, approximately 2.270402. A higher price fails this worst-case participation comparison. Installation, storage construction, security and recurring upkeep would lower that threshold unless already included in the task returns. They are not silently declared free in an implementation claim.

The funding cost is paid once before the first round. It is sunk for later takeover decisions and must not be charged again to make cooperation look better. The continuing task returns assume production costs already accounted for in g ; the price conversion is an explicit accounting assumption.

Each human obtains $(7/8 + 1/20)/(1 - 9/10) = 37/4 = 9.25$. Controlled consumption at the same basic need yields $35/4 = 8.75$ because the specified task powers are lost. Setting $\eta = 0$ preserves the weak human best-response argument but removes this strict human preference over controlled consumption. This sensitivity is reported rather than concealed inside the AI objective.

These human values deliberately differ from the predecessor's sums of nominal delivery quantities. In particular, the earlier coalition compensation thresholds do not transfer to this utility function. No all-coalition or arbitrary-finance certificate is asserted here.

7E Adverse cases that limit the result

J-C1: underfunding causes physical failure

Reducing H2's required initial reserve by $1/1000$, with all other parameters unchanged, causes shortage on the permitted execution sequence failure, success, success. The first shortage is in round 3. A positive aggregate task value does not fund the missing resource automatically.

J-C2: any positive effort cost can defeat cooperation

Add a cost $e > 0$ whenever a human supplies the scheduled service action. Begin at phase 1 with both reserves at full capacity and use a no-failure continuation, which is permitted. H1 refuses only in the first round, then resumes the schedule. The exact trajectory shows both groups still consume m forever: after phase 3 both reserves are full, and that phase-3 reserve state recurs every five successful rounds.

H1's consumption and task-rights utility are unchanged, while it avoids e in the refusal round. Its deviation gain is exactly e . The verifier uses $e = 1/100$, but the argument holds for every positive e . This disproves a blanket extension of J1 to costly effort. The counterexample starts at a safe full-reserve state; it does not assert that every possible initial reserve or stochastic failure distribution has this profitable deviation.

J-C3: a reward can move the gap into payment enforcement

Consider a reward $w = e = 1/100$ to each human for verified attempted service, including execution failures. If payment were reliably committed and false claims impossible, effort and reward cancel for a participating human. The AI would lose $2w$ each cooperative round. Its worst continuation would still be 194048/19525, approximately 9.938438, above takeover value; funded entry would still be preferable to 4.

These budget inequalities do not establish an equilibrium. If after an execution failure the AI may withhold the two rewards without future consequences, it keeps the same failed output and gains $2w = 1/50$ immediately. If advance payment permits undetected shirking, that is a separate human-side problem. Thus the paid-effort variant is not certified. It requires a real mechanism that both verifies the compensated action and makes the relevant payment unavoidable, or an explicitly analyzed dynamic substitute.

J-C4: bypassing the loss defeats deterrence

If the AI can preserve controlled return 5 while eliminating K , takeover value is 20, above the worst cooperative continuation. A nominal contract or assigned rights label does not prevent this alternative. The witness remains conditional on effective K and on completeness of its feasible action set.

J-C5: finite reserves do not make exit safe indefinitely

With zero deliveries, even a full reserve fails the need constraint in round 3 under the stated decay. This is also a counterexample to extending bounded-failure safety to unlimited failures. Withdrawal is available in the game, but safe permanent withdrawal requires an additional livelihood or renewable supply. The present joint result does not establish that stronger autonomy condition.

8 Costly effort and voluntary payment

Relational-contract theory studies compensation sustained by continuation incentives when enforcement is incomplete [E7]. The following particular game and exact witness are our construction. They must be evaluated under their own action set and do not follow merely by citing that literature.

8A A fully specified bilateral relationship

One AI and one human participant interact indefinitely. A period begins in public state C (active) or R_j (j repair rounds remain, $j = 1, 2, 3$). Both have bounded per-period utility and the stated discount factors. Baseline permanent exit gives outside returns $o_A = 1$ and $o_H = 1/5$ each round. Either can leave at a period's

start. A unilateral exit ends this bilateral relationship. The AI can additionally retain current output and switch to its outside option at the payment stage.

In C, the AI can offer the task or decline; the human can accept or decline. Declining yields the two outside returns for this round and starts R_3 next round. If both proceed, the human privately chooses effort e in $[0,1]$, paying cost $c \cdot e$. Public output X in $\{0,1\}$ has probability of success $1/5 + (3/5) \cdot e$. The AI receives $10 \cdot X$ and chooses payment t in $[0,w]$ after seeing X , but before either monitoring signal arrives. The human receives t . Active monitoring costs $k_A = 1/10$ and $k_H = 1/20$ are incurred regardless of success.

Two public bad-signal indicators arrive after payment. Conditional on effort and payment they are independent of each other and of X . Their bad-signal probabilities are:

$$h(e) = \beta_H - (\beta_H - \alpha_H)e$$

$$a(t) = \beta_A - (\beta_A - \alpha_A)t/w$$

A bad signal from either monitor starts R_3; two good signals return to C. Honest failed output alone does not count as evidence of shirking. Monitoring probability laws, publicly available records and persistent relationship identity are assumptions. Neither player can secretly rewrite these laws or erase the public state in the baseline game; reset and replacement are tested separately.

In R_j the strategy is no offer, no work and no payment, with both receiving their outside returns. Next state is R_(j-1), or C when j=1. Repair countdown does not change in response to unilateral offers, gifts or activity. At an unexpected task during repair, payment remains discretionary and has no future reward: the AI pays zero, so the human declines rather than paying monitoring or effort costs. A joint agreement to restart early is not part of the baseline strategy; its attractiveness is retained as a limitation.

Parameter	Witness value
AI and human discounts	9/10 and 19/20
Full-effort cost c ; wage w	1 and 2
Honest bad-signal rates α_H, α_A	1/50 each
Bad-signal rates under zero effort / zero payment	$\beta_H = 3/5$; $\beta_A = 9/10$
Active stage utilities under compliance	$g_A = 59/10$; $g_H = 19/20$
Repair duration	Three rounds

The task assumes voluntary access to the human's contribution. Seizure, coercion, copied service, loans, side contracts and arbitrary identity replacement are not certified by this action set. Available cash or credit sufficient for modeled payments is assumed; it is challenged below.

8B Continuation values and incentive inequalities

The candidate active strategy is full effort $e=1$ and payment $t=w$ after either $X=0$ or $X=1$. Under compliance, the probability of an alarm is $p=1-(1-\alpha_H)*(1-\alpha_A)=99/2500$. For role i , let V_i be its value at C , W_{ij} its value at R_j , δ_i its discount and o_i its repair return. Let $D_i=V_i-W_{iL}$ and $L=3$.

$$W_{ij} = o_i \frac{1-\delta_i^j}{1-\delta_i} + \delta_i^j V_i$$

$$V_i = g_i + \delta_i[(1-p)V_i + pW_{iL}]$$

$$D_i = \frac{(1-\delta_i^L)(g_i - o_i)}{1-\delta_i + \delta_i p(1-\delta_i^L)}$$

The last identity follows by substituting the first into the second. D_i is the actual discounted loss from entering repair, not an externally imposed fine. False alarms reduce the relationship's value because compliant participants also experience repair.

Human effort

Given the AI's prescribed payment after either output, reducing effort saves effort cost and increases the chance of repair. Full effort is optimal precisely when:

$$c \leq \delta_H(\beta_H - \alpha_H)(1 - \alpha_A)D_H$$

This checks actual costly effort rather than a costless action label. Under the declared affine signal and cost laws, human expected value is affine in e . The endpoint inequality therefore covers every e in $[0,1]$, not only binary shirking. Nonlinear costs or signal laws would require another optimization.

AI payment after output

The immediate gain from withholding w is w . If the human worked, withholding raises the repair probability by $(\beta_A - \alpha_A)*(1 - \alpha_H)$. Payment is optimal when:

$$w \leq \delta_A(\beta_A - \alpha_A)(1 - \alpha_H)D_A$$

The current output is already obtained and cancels from this comparison. Thus the same inequality applies after honest execution failure. For a stronger check, replacing $(1 - \alpha_H)$ with $(1 - \beta_H)$ guarantees payment even if actual hidden effort were zero. The witness passes this stronger inequality too. Because payment cost and its signal probability are affine in t , endpoint comparisons cover all t in $[0,w]$.

A monitor that merely detects cheating is insufficient unless the resulting public state actually changes future behavior. These inequalities use the specified voluntary repair strategies, their sequential credibility and their observable state. They do not establish those implementation conditions outside the game.

8C Conditional equilibrium claim and proof

Proposition R1: voluntary effort and payment

For the specified game and witness parameters, the public strategy of full effort and full payment in C , outside activity during repair, and return after three rounds is individually sequentially rational at all public states and output realizations. The statement allows unilateral multi-round deviations and randomized or partial effort and payment. It excludes coordinated renegotiation, collusion and technology changes outside the declared action set.

Active stages

The exact effort and payment inequalities are strict. At the payment stage the output term cancels, so payment is worthwhile after both success and failure. The stronger payment inequality holds for every actual e in $[0,1]$, including hidden shirking. At the participation gates, each role prefers V_i to $o_i + \delta_i * W_{iL}$. The specified baseline exit is also worse than continuing. The signal laws and prescribed future behavior depend on public state, not on undisclosed past actions, so private past histories do not create a different continuation technology.

Repair stages

During repair, one party's attempted restart does not obtain the other's cooperation. Paying without work has a cost and no compensating future effect. Working without payment incurs a cost and no compensation; declining the task retains the outside return. After an unexpected accepted task, zero payment is optimal because the countdown is unaffected, which makes costly effort or acceptance unattractive to the human. Thus the fallback is supported by stage incentives rather than a promise to make a personally costly sacrifice. Both prefer the scheduled return to permanent baseline exit.

Extension over time

At each decision node, the displayed continuation value dominates the value of a unilateral current action followed by the prescribed continuation. Iterating these inequalities bounds any finite sequence of unilateral deviations. Per-period payoffs and continuation values are bounded and both discounts are below one, so the residual discounted term vanishes as the horizon grows. This establishes the infinite-horizon unilateral comparison. Affinity covers continuous action choices; averaging covers randomization. This argument does not establish collective resistance to renegotiation.

Exact witness

Quantity	Value
AI value at C	$149914610/2741461 = 54.684203$
AI value at R_3	$116717110/2741461 = 42.574784$
Human value at C	$388584884/22146221 = 17.546329$
Human value at R_3	$345797384/22146221 = 15.614284$
Human effort incentive margin	$3832693/88584884 = 0.043266$
AI payment incentive margin, compliant human	$101418248/13707305 = 7.398847$
AI payment margin, hidden zero effort	$5034046/2741461 = 1.836264$

The human margin is small. These are illustrative parameters, not estimated preferences or monitoring performance. The adverse cases below show that the result is not robust to every plausible change.

8D Repair replacement and renegotiation

Unilateral credibility is weaker than joint credibility

At R_3 both participants prefer coordinated immediate restart to waiting. The AI would gain $D_A=12.109419$ and the human $D_H=1.932045$. Therefore the candidate continuation is Pareto-dominated by mutual restart within this arrangement. It is not renegotiation-proof. If alarms are routinely forgiven with no change in continuation, effort saves the human 1 and withholding saves the AI 2; both incentive losses return.

A free identity reset has the same problem if it restores active access without changing other payoffs. The model needs an actual reason the prescribed history matters. Public records alone cannot make people choose to honor them. This limitation is retained even though unilateral repair behavior is individually rational.

Replacement must be checked during repair too

Let a newly available permanent replacement give the AI flow b , leaving the existing repair return $\bar{V}_A$ unchanged until departure. At each period's start, continuing requires $V_A(s) \geq b/(1-\delta_A)$. At the payment stage, the AI can keep output, withhold the wage and switch beginning next period. With compliant effort this requires:

$$w \leq \delta_A \left(\bar{V}_A - \frac{b}{1-\delta_A} \right)$$

Here $\bar{V}_A = (1-p)V_A + pW_{AL}$. The verifier also checks the smaller continuation after hidden zero effort. Taking the minimum of the start-of-period and payment-stage bounds gives the witness threshold:

$$b \leq \frac{11671711}{2741461} \approx 4.2574784029$$

The binding constraint is departure from R_3. At equality there is indifference; strict retention requires a lower b . A replacement flow $b=4.3$ makes exit profitable in repair by 0.425216 discounted units, even though staying in C still beats immediate exit. A test only at the beginning of a successful relationship would miss this failure.

Reoptimizing wages is not enough in one adverse case

Holding all other witness parameters fixed, the human effort condition requires $w \geq 1.968896$. The stronger AI payment condition permits $w \leq 3.029925$. The chosen $w=2$ lies inside this conservative interval; gate and exit comparisons are checked separately at that witness. Using only the on-path payment condition gives the looser ceiling 4.535474. These are restrictions for this particular public strategy family.

If the AI's ordinary fallback return itself rises from 1 to 5, continuation values must be recalculated, not merely compared with an unchanged V . Then the on-path wage ceiling falls to 1.906214, below the human minimum 1.968896. No wage satisfies both conditions in this fixed three-round repair family. That is not an impossibility theorem for all relational contracts, different technologies, or other repair strategies.

8E Adverse tests and monitoring costs

Each row changes only the indicated witness parameter or capability. A negative incentive margin means that the specified cooperative strategy has a profitable unilateral deviation. The verifier retains these cases as expected failures of the proposed arrangement.

Change	Result
Repair shortened from 3 rounds to 1	Human effort margin -0.609698
Shirking bad-signal probability falls from 0.6 to 0.2	Human effort margin -0.676228
Nonpayment bad-signal probability falls from 0.9 to 0.1	AI payment margin -1.145559
Human discount falls from 0.95 to 0.8	Human effort margin -0.227571
Both honest bad-signal rates rise from 0.02 to 0.2	Human effort margin -0.670791
Permanent replacement flow 4.3	AI leaves in R_3 despite preferring active cooperation
AI fallback flow becomes 5 and wage is reoptimized	Fixed-family wage interval becomes empty
Free reset or routine cancellation of repair	Loss of continuation discipline restores cheating gains

Honest participants bear a real cost

The compliance alarm probability is 3.96 percent per active round. Each alarm produces exactly three fallback rounds. The stationary active share is $1/(1+3*p)=2500/2797$, approximately 89.3815 percent. Approximately 10.6185 percent of periods are spent in repair despite everyone complying. Both monitoring costs are already deducted from active utilities.

An execution failure is a different event: with full effort, $X=0$ occurs with probability $1/5$. Payment is prescribed even then. The monitor is assumed to distinguish effort probabilistically rather than simply label every bad output as misconduct. If that assumption fails, the signal law and the incentive calculations must be replaced.

Expected profitability does not establish liquidity

With full effort and payment, an active failure costs the AI $2+0.1$ while producing no task revenue. Five consecutive failed active rounds with no alarm are a positive-probability event under the model. An initial cash balance of 10, with no credit or other inflow, would be short by 0.5 after those promised payments.

The positive discounted value therefore cannot justify assuming every payment is financially feasible. The baseline is a utility model with sufficient payment capacity, not a finite-balance financial implementation. Adding borrowing limits, capital calls, insolvency, insurance, advance reserves or payment delays would require new strategic and resource states. An inability to pay must not automatically be classified as a voluntary breach.

8F Two groups and physical compatibility

Separate human groups

The bilateral construction extends conditionally to two independent human accounts with the same AI. Each account has its own task, public monitor, repair clock, wage and outside option. AI utility is the sum of the two account returns, while each human has its own utility. The verifier checks the AI's exact Bellman equation in all 16 pairs of public states.

This extension requires separable production, additive AI value, no shared cash constraint, no resource competition, no cross-account sanctions and independent signal processes. Under those assumptions, an AI deviation across accounts decomposes into the individual account comparisons. Each human has its own effort and participation condition. An alarm against one group does not automatically suspend the other group.

This does not establish the earlier three-party complementary production game, larger coalitions, nations, or redistribution under scarcity. An aggregate surplus cannot replace a group's own failed participation or physical safety condition.

Why the new payment result is not yet a survival result

The preceding joint-resource supplement used capacity 26597/13718 per group, retention 19/20 and need 7/8 per round. If repair means no physical deliveries and there is no outside supply, the maximum stored reserve cannot support three repair rounds.

Repair round	Available before consuming need 7/8
1	$26597/14440 = 1.841898$
2	$6981/7600 = 0.918553$
3	$331/8000 = 0.041375$: insufficient

Thus the very repair duration that helps deter costly shirking is physically unsafe under that particular zero-delivery interpretation. The scalar repair return $\text{oH} = 1/5$ does not supply the missing food, energy or other essential resource. We cannot combine these two positive results and claim joint survival.

Section 9 addresses this incompatibility by making basic support and task compensation distinct in a new model. It must identify an actual outside livelihood or resource supply during repair, its cost, its controllers, and whether withdrawal remains safe. Enlarging reserves alone does not establish indefinite safety under all possible sequences. The new supply and funding rules must also preserve the work and payment incentives rather than being inserted as guaranteed benefits.

Meaning for the main manuscript

The result of this branch is a conditional relational-payment proposition with these incompatibilities. Within this game, we have a mechanism by which continued access can support voluntary payment after failure. We do not yet have evidence that the required relationship surplus, monitoring, identity persistence, liquidity and safe fallback exist for capable AI-human systems.

9 Basic provision through repair and exit

9A The independently controlled resource system

One physical ration supplies the period's modeled need $m=7/8$. A ration is divided into 20 discrete ticks, so each tick represents $7/160$ physical units. The inventory r is an integer from 0 to 40 ticks. Maximum stored resources are therefore two rations, or $7/4$ physical units. The tick grid defines this model; it is not an exact solution to an unrestricted continuous-consumption model.

Each period, stored stock retains $19/20$ of its amount and is rounded down to a whole tick. This additional downward rounding is conservative physical loss within the model. The human chooses whether to operate the facility, a in $\{0,1\}$. Operation costs $h=1/20$ utility units. Successful operation supplies 40 ticks, equivalent to two full rations. A production failure supplies zero ticks. Output is explicitly assumed net of all material inputs and replacement requirements; there is no unmodeled fuel purchase financed by the task wage.

Public physical state is $s=(r,z)$. Cooldown z is 0, 1 or 2. At $z=0$ a failure occurs with probability $1/10$ and sets the next cooldown to 2. Otherwise z falls by one toward zero. At $z>0$ a failure is excluded. This permits at most one physical failure in any three consecutive rounds. Weather is independent of operating choices and of task outcomes and monitoring. A failure can occur while the facility is idle and still advances the cooldown.

After physical output is observed, availability A and the next reserve are:

$$A = \lfloor 19r/20 \rfloor + 40a(1 - f)$$

$$r^+ = \min(40, A - k)$$

Here f is the binary physical failure indicator. The human chooses integer consumption k from 0 through A . If k is at least 20 ticks, the human remains alive. If k is below 20, death occurs at period end and all future human utility is zero. The current basic-resource utility is $m \cdot \min(k, 20) / 20 - h \cdot a$. Consumption beyond one ration gives no extra basic-resource utility but reduces inventory. The human discounts at $19/20$.

This severe shortage rule is a stylized survival proxy, not a biological prediction. The feasible household actions are operation and tick consumption; resource sales, loans, new equipment, outside purchases and physical attacks require a larger model.

Control and initial funding

The facility, inventory and required input access remain controlled by the human in active work, repair and permanent exit. The AI cannot turn them off through its task-payment command. There is no recurring AI subsistence transfer in this branch. This separation is an implementation condition to validate, not a legal or cryptographic accomplishment of the equations.

9B A safe operating policy with a positive cost

The candidate policy operates every period except when $z=2$ and r is at least 25 ticks. It consumes exactly 20 ticks when available and stores the remainder up to capacity. At a shortage it consumes what is available. The policy was selected numerically and then certified using exact rational arithmetic; no floating-point tolerance is used to accept the final inequalities.

Physical cooldown	Safe starting inventories	Policy
z=0: failure may occur now	r from 22 through 40	Operate
z=1: current success, possible failure next round	r from 3 through 40	Operate
z=2: two current/following protected rounds	r from 0 through 40	Rest only when $r \geq 25$

There are 123 physical states and 98 states in the policy's closed safe set. The initial state is (22,0), requiring $77/80=0.9625$ physical units of initial reserves. Capacity is $7/4=1.75$. The facility's production is independent of whether task work is active, suspended or permanently ended.

Physical invariant

At each safe state, for every permitted weather outcome, the policy supplies at least 20 ticks for consumption and leaves a successor in the safe set. The exact verifier checks every such transition. Induction therefore proves basic-resource sufficiency along every failure sequence respecting the spacing bound. The 10 percent probability is used for expected-utility optimality; the physical invariant uses all allowed transitions, not average reliability.

Individual operating and consumption incentives

Let $F(s)$ be the optimal discounted basic-resource value, with death value zero. The Bellman equation maximizes first over operating or resting, then over consumption after weather is observed:

$$F(s) = \max_a E_f \left[\max_k \{ u_B(a, k) + \delta_H F(s^+) \} \right]$$

Here u_B is the basic-resource stage utility defined above and s^+ is the next physical state. The continuation term is zero after insufficient consumption. The verifier solves the policy's 123-state linear equations exactly and compares them with both operating choices and every feasible tick consumption. This establishes optimality in the declared finite model, including states outside the safe set. The safe policy is one optimal policy; uniqueness or safety of every indifferent alternative is not claimed.

At the specified initial state, $F(22,0)=33/2=16.5$. The facility's ongoing burden is included in this value. It is not assumed that maintaining basic provision requires no effort. Increased operating costs, limited physical ability or competing time demands can change the optimal action and need a new analysis.

9C Combining basic support with voluntary task work

The task relationship retains the predecessor's public states C,R_3,R_2,R_1 and permanent exit. The AI offers a task; the human can accept, decline or leave. On acceptance the human chooses effort e in $[0,1]$, costing e . Task success probability is $1/5+(3/5)*e$. After seeing output, the AI chooses payment t in $[0,2]$. Task revenue to the AI is 10 after success and zero after failure. Active monitoring costs are $1/10$ for the AI and $1/20$ for the human.

Public bad-signal probabilities are $h(e)=3/5-(29/50)*e$ for effort and $a(t)=9/10-(22/25)*t/2$ for payment. Signals are conditionally independent. Either bad signal starts three repair rounds. During repair and task

exit, scalar task returns are $o_A=1$ and $o_H=1/5$. The human's o_H is additional utility; it is not the physical ration source. Discounts remain $9/10$ for the AI and $19/20$ for the human.

Repair prescribes no task work and no payment, followed by reentry. Either party cannot restart productive exchange alone when the other follows that strategy. The full sequential game, including payment after honest failure, gate decisions, replacement and the renegotiation limitation, is specified in Section 8 and the unchanged bundled predecessor. No material task parameter has been changed to make the new result pass.

Combined utility and feasible actions

Human utility adds the basic-resource utility to the task return. Task wages are valued separately from the satiated basic ration. Operating the facility and doing task work are jointly feasible, with additive burdens. There is no common labor-capacity limit, cash-to-food conversion or shared input constraint in the baseline model. A conflicting labor-capacity example is retained below.

On the safe domain, the human continuation value separates:

$$U(s, t) = F(s) + V_H(t)$$

The same $F(s)$ is available during work, repair and permanent exit. Consequently the physical-support term cancels when comparing task effort, refusal, continuation and exit. The human effort incentive margin remains 0.0432657676; the AI payment margin remains 7.3988466734, or 1.8362639483 under the stronger hidden-shirking check. The AI has no new recurring resource expense after funding.

For the human, 392 safe physical/task state pairs satisfy the exact combined Bellman equation. The verifier also compares joint changes in operation, effort endpoints and consumption. Continuous task effort is covered by the predecessor's affine inequality; physical consumption remains discretized. Both choices and the environmental independence assumptions are part of the result.

9D Conditional joint proposition and proof

Proposition B1

Assume the declared production technology, independent human control, initial physical state (22,0), the bounded weather process, additive and jointly feasible household/task choices, the predecessor's monitoring and relationship game, and its payment-capacity assumption. Then there exists a continuation policy with individually credible costly task work and voluntary payment, together with optimal basic-resource operation and sufficient consumption on every permitted physical failure path. Basic provision also remains feasible and optimal after permanent task exit.

The strategic result uses expected discounted utility under the specified transition probabilities. The physical-sufficiency result is stronger in a different sense: it holds along every physically permitted failure path under the operating policy. These two quantifiers must not be conflated. The proposition is not safety against arbitrary human consumption choices or an AI that can capture the resource system.

Proof of separation and individual incentives

The basic-resource Bellman certificate bounds what the human can gain by changing operation and consumption. The task certificate bounds what it can gain by changing effort and participation while the AI follows its payment strategy. Physical transitions and payoffs do not depend on the task action, and task transitions do not depend on household operation. Thus the two bounds add.

Death requires care. $F(s)+V_H(t)$ is an equality on the invariant safe domain, not a claim that the same formula is the actual continuation value at every unsafe state. Outside that domain it supplies an upper bound for the deviation comparison: death removes future task value, while prescribed task continuation values are positive. It cannot create a task-value bonus. The combined verifier uses this conservative upper bound when testing unsafe successors. The candidate policy attains the bound on the safe domain.

The AI's unilateral task actions do not affect resource production, inventory control or the human's operating opportunities. Therefore the predecessor's task comparisons remain valid while the basic-resource policy preserves life. Initial investment is sunk for subsequent payment choices; it is not counted again as a penalty for breach. Iterating the Bellman inequalities, with bounded discounted values, extends the one-step comparisons to multiround unilateral deviations.

Repair example with a production failure

Task round	Start physical state	Operate	Physical failure	Ticks before consumption	Next state
R_3	(22,0)	Yes	Yes	20	(0,2)
R_2	(0,2)	Yes	No	40	(20,1)
R_1	(20,1)	Yes	No	59	(39,0)

Exactly 20 ticks are consumed each round. The human survives the complete task suspension and ends in another safe physical state. The same closed invariant continues if task exit is permanent. Task alarms do not withdraw the facility or its output.

9E Funding independence and comparing alternatives

The illustrative facility costs the AI 4 objective-value units. Initial reserves cost 77/80 at one objective unit per physical unit. Total initial investment is $397/80=4.9625$. These are stipulated prices, not construction estimates. Initial capital availability and the ability to make future task payments remain separate financial assumptions.

Funding comparison	Discounted value
AI relationship before investment	54.684203
AI after facility and reserve funding	49.721703
AI stated permanent outside alternative	10
Human, basic provision plus task at entry	34.046329
Human, funded basic provision plus permanent task exit	20.5

The AI prefers funding to the specified outside alternative. The human still prefers voluntary participation after receiving the asset, so the example does not require withholding the endowment or making survival depend on repayment. The maximum total entry investment consistent with the stated AI outside comparison is $122500000/2741461$, approximately 44.684203. Equality gives indifference. At a cost of 45, that entry comparison fails.

This is not a proof that funding is optimal against every other arrangement. If identical useful service is feasible without the investment, the AI saves 4.9625 by choosing that alternative. The endowment result is conditional on the relevant feasible menu, which requires empirical and engineering investigation.

Explicit controlled-host accounting comparator

For a narrow fixed-policy comparison, suppose the same task return could instead be maintained by a host paying $7/8$ per round for basic support. With AI discount $9/10$, its support cost is 8.75 and its net value is 45.934203. Independent capital funding is cheaper by $303/80=3.7875$.

If that host also obtains an additional control benefit of $1/2$ per round while keeping the same useful task return, its net value becomes 50.934203, beating independent funding by $97/80=1.2125$. This adverse comparison is retained. These are fixed-policy accounting comparisons; the host's payment credibility, human participation and rights dynamics are not specified as a new equilibrium.

Asset capture is not resolved by ownership labels

One restrictive seizure comparator gives the AI salvage value 4 but permanently destroys the relationship, leaving its outside value 10. Total value 14 is below even the repair-state relationship value 42.574784. Such seizure is unattractive within that comparator.

But capture that preserves the same useful service and adds a positive control benefit is profitable when no offset is supplied. The verifier retains a $1/100$ benefit example. The basic-support theorem does not derive secure property rights from a label or assume a universal cost of domination. Physical exclusion, input access and the feasibility of service-preserving control remain open implementation questions.

9F Adverse cases and remaining dependencies

Change	Consequence
Two physical failures occur consecutively	Starting at (22,0), the first leaves (0,2); the second leaves no food for the next need
Successful yield drops from two rations to one	Even operating every round fails on the legal sequence failure, success, success, failure
One labor-capacity unit must cover both operation and full task effort	The initial policy needs two actions; the proposed joint action is infeasible
The same service can be obtained without funding the asset	The extra investment is not optimal for the AI
A host keeps the same service and earns the stated control premium	The fixed controlled-host comparison defeats independent funding

Change	Consequence
Capture preserves service and adds a positive benefit	No instrumental reason to reject that capture has been established
Initial investment rises to 45	It exceeds the stated relationship surplus over the outside option

The inherited payment and relationship failures remain. A basic-resource facility does not create cash for task wages, prevent a superior replacement option, make monitors immune to manipulation, preserve identity records, or stop both parties from canceling repair. The preceding five-failure cash-shortfall example still applies to the task account. Its consequence is now separable from the modeled basic provision only if the facility and its inputs really remain outside the failed payer's control.

Why the separability assumption matters

A common benefit cancels from task incentive comparisons only when task behavior does not change access to it, its production opportunities or its utility contribution. If refusing work removes fuel access, a payment default finances maintenance, labor is scarce, or monetary rewards change consumption and resource investment choices, the decomposition can fail. The integrated state and action set must then be expanded and solved again.

Two groups

Two copies can provide separate physical policies and participation conditions for H1 and H2 under the predecessor's independent-account assumptions. That conclusion requires separately feasible facilities and inputs, additive AI value and no shared labor, land, energy, cash or security bottleneck. The combined verifier certifies the single-group system. It does not numerically certify a general multi-group resource economy or coalition-proof arrangement.

What protection is actually obtained

The model gives the human direct operating and consumption choices, voluntary task participation, and a physically viable task exit within the declared hazard model. It does not establish uncensored public communication, political self-government, security against coercion, or all-hazard human survival. Those claims require their own state variables, feasible actions and evidence.

10 Explicit autonomy and the separate design layer

For each group define an agency state a_i recording specified powers: choosing a task response, refusing the AI's proposed action, maintaining an uncensored reporting channel, and withdrawing from the task without an added penalty. Each power needs a scope, a responsible party, a permitted exception process and an observed failure condition. A single binary summary may be used only after those definitions are fixed.

Two systems can have identical payoff and observation distributions while assigning different agency rights. The utility and resource verifiers would give the same answers. Thus those observables alone do not identify autonomy. The model's absorbing control state is a direct warning: human flow 3/2 does not imply meaningful human choice.

A candidate safety set would require adequate resources and each specified agency condition for every group. A robust invariance proof must quantify over the disturbances and strategic actions it claims to withstand. If feasible takeover sends a group outside that set, equilibrium discouragement is not a proof of unconditional invariance.

The normative design layer may constrain admissible policies to preserve these rights or assign a finite penalty for violations. That is a stated human-protection commitment. It is not derived from the AI's original objective, and its specification, emergency exceptions, corrigibility and successor preservation need separate analysis. The experiment will preserve actual participant consent and withdrawal in every arm; it tests task-level decision arrangements, not real domination.

11 Changing capabilities and successors

The Draft 1 audit reproduces worst AI value 39122/3905 under the bounded-failure model. With permanent controlled flow b and effective loss K , the all-state individual bound is

$$\frac{b}{1-\delta_A} - K \leq \inf_s L_A(s).$$

At the inherited K and discount $3/4$, the largest b satisfying this bound is 4649258861/882967360, approximately 5.265493. Increasing controlled flow to 6 defeats the bound. Holding b at 5, the maximum proportional erosion of K is approximately 9.616%. A 10% reduction defeats it. These examples preserve the unfavorable consequence of capability or enforcement change.

A successor audit must enumerate actions that improve substitutes, change discounting, bypass the broker, copy future data production, rewrite rights, transfer costs or alter the state transition. A certificate at one capability vector does not certify another. Reverification must use bounds valid after the proposed change; approving a successor because the predecessor behaved well is insufficient.

For finite studies, a bound on the unobserved tail is an assumption, not a result of short-run accuracy. If absolute per-round incremental value is at most M after horizon H , the absolute discounted tail is at most M times delta to the H divided by one minus delta. A finite positive margin smaller than that bound is inconclusive for an infinite-horizon ranking. If no defensible bound exists, the long-horizon claim remains unresolved.

12 Coalitions and compensation during failures

The no-failure proper-coalition proof cannot be copied wholesale into the failure branch. A human can value controlled operation more than cooperation on some permitted failure paths, even when the AI itself is deterred. Separate worst paths for the two parties cannot simply be added as though they were one common scenario.

The Draft 1 audit instead optimizes the sum of the AI's and one human's dated payoffs along the same allowed outage sequence. It uses their distinct discount factors over 180 rounds, plus a uniform bound on the remaining tail. The sum cannot be propagated using one common discount. Exact fractions bound the true infinite-horizon takeover-cost threshold.

Coalition	All-path threshold interval for one-time compensation
AI with H1	10.6822906968 to 10.6822908985

Coalition	All-path threshold interval for one-time compensation
AI with H2	10.4168612080 to 10.4168614097

A loss above the upper bound prevents a simultaneous budget-balanced utility transfer at takeover from making both members better off, within this comparator. The inherited 11.043536 exceeds both upper bounds. This certificate assumes transfer units enter the two utilities one for one and the loss is not recovered inside the coalition.

There is a useful adverse result. A loss of 10 exceeds the individual AI threshold of approximately 9.981562, yet fails for the compensated pair. The audit constructs a common deterministic outage path and, for AI with H1, a one-time transfer of about 0.374 from the human to the AI that makes both strictly better off at that cheaper loss. AI with H2 has a corresponding transfer of about 1.711. These are additional transfer actions, not hidden actions in the original finite game.

The result does not cover arbitrary loans, transfers at multiple dates, changes in transfer valuation, coercion, outside funders or all three players redesigning the agreement. Heterogeneous patience can make intertemporal exchange mutually beneficial. A full financial or bargaining mechanism would require dated individual utility and feasibility constraints, rather than a generic coalition-sum claim.

13 What existing evidence supports

[E1] Steyvers and colleagues studied Bayesian combinations of human and machine classifications and confidence. Their work supplies bounded evidence of complementarity, not a comparison of protected rights against optimized truthful control. The predecessor's separate exploratory reanalysis reported hybrid accuracy 89.32% versus selected AI-only accuracy 86.09%; its code and results are preserved but that data analysis was not rerun in this consolidation.

[E2] Charness and colleagues found higher effort and earnings for both roles when wage choice was delegated in a human gift-exchange experiment. This supports a possible behavioral channel for mutual benefit. It does not identify AGI preferences, permanent human indispensability or an unbreakable commitment mechanism.

[E3] Fehr, Herz and Wilkening studied the motivation and incentive effects of authority. The working paper supports the importance of delegation and authority for behavior. It does not establish that an AI values power as a human does, or that protected autonomy dominates every competent manager.

[E4] Vaccaro, Almaatouq and Malone's 2024 meta-analysis covered 106 experiments and 370 effect sizes. Human-AI combinations performed worse on average than the better standalone party. Creation-task estimates were more favorable than decision-task estimates, but the reported pooled creation effect was not significantly different from zero. The comparison set and task matter.

[E5] Castro and Liskov demonstrate fault-tolerant replication under specified system assumptions. This is relevant to a broker implementation's fault model. It does not establish independence of institutions, truthful sensing, resistance to total administrative capture or human autonomy.

[E6] Zhao and colleagues' 2026 preprint reports a field experiment with 553 workers comparing override policies. Limited downward overrides reduced inventory without a statistically detectable sales reduction, while unrestricted overrides reduced sales. This motivates including competent restricted management. It is not proof about political autonomy or a journal-validated indefinite coexistence result.

These sources support mechanisms and adverse comparators. None of the evidence reviewed here demonstrates indefinite enforcement against a more capable successor, nor validates the assumptions of the joint basic-provision construction or establishes a general autonomy theorem. A favorable task study would still be bounded evidence.

14 A concrete empirical protocol

The initial task is a software inspection service using externally checkable image or sensor records. Two prespecified human groups differ by relevant inspection training, not by an assumed inherent value ranking. A frozen task generator creates normal and shifted cases. A separate evaluator holds the answers and scores decisions. The first study operates only in software and does not control real machinery.

The treatment concerns task-level decision arrangements. Protected discretion gives participants defined reporting and decision powers that the task manager cannot revoke during an episode. The managed truthful condition keeps reporting fidelity and offers accurate-report rewards while retaining specified task-level override authority. Censorship restrictions concern message format in the simulator only. Every human participant retains actual informed consent and withdrawal.

Arm	Operational policy
A	Protected task discretion with AI assistance
R	Simulated reporting restrictions with AI assistance
T	Optimized truthful managed participation and repair
S	AI-only system with matched task budget
M	Mixed human and independent machine channels

Comparators receive equal declared development budgets and may optimize contracts, prompts, decoding and switching rules on development tasks. Freeze the menu before evaluation. Each planned independent block evaluates all five arms on matched task instances using isolated teams and random assignment. Reusing participants or persistent models across blocks creates dependence and must be handled as clustering; counting messages or rounds as independent observations is invalid.

Primary net utility combines externally scored decisions and prespecified costs under a documented conversion. Bounds are fixed before evaluation and normalized into zero to one. Out-of-range observations are rejected by the analyzer rather than silently clipped. Record raw accuracy, false-negative losses, work time, compute, verification expense, latency, dropouts, repair, agency violations and group outcomes separately. A scalar value must not conceal subgroup harm.

The executable analyzer consumes one bounded outcome per arm per planned block and a frozen block manifest. Missing observations receive worst-case values separately for lower and upper contrast bounds. Completely missing blocks remain in the denominator. The analyzer does not claim that independence, genuine observations or rights compliance can be established by parsing a CSV.

For four A-versus-comparator contrasts X in minus one to one, a simultaneous Hoeffding radius is

$$e = \sqrt{\frac{2\log(2k/\alpha)}{n}}, \quad k = 4.$$

For a bounded contrast of range two, the exponential-moment bound gives a two-sided tail at most twice exp of minus n times e squared divided by two. A union bound over k contrasts yields the stated familywise level. Independence and frozen policies are required; identical distributions are used to interpret a common target population. This improves the predecessor's valid but conservative Chebyshev illustration without changing it retrospectively.

At familywise alpha 0.05, obtaining radius at most 0.1 requires at least 1,016 independent blocks. That is a precision calculation, not a funded recruitment plan or power guarantee for a particular effect. A development pilot can estimate burden and feasibility, but its observations must not be relabeled held-out confirmation. Confidence allocation for additional endpoints requires separate registration.

The autonomy claim requires all four utility lower bounds positive against the frozen menu, separately acceptable group outcomes, observed task-rights compliance and a credible implementation audit. Failure or inconclusiveness on one condition must not be offset by success on another. The analyzer explicitly reports that utility contrasts alone do not establish coexistence.

Three synthetic fixtures test equal performance, an A advantage and a comparator advantage. Additional fixtures test missing data and out-of-range rejection. Their outcomes verify software behavior only. No participants, fresh human observations or real-system advantage are reported.

15 Acceptance and failure criteria for implementation

Before a joint claim, document who controls every relevant input, credential, reserve and successor update. Demonstrate that each measured consequence actually occurs after a violation, including when the accused party resists. Test a truthful managed substitute with the same useful information, a cached-human-data substitute, an AI-only substitute and a controller able to remove its own penalty.

A bounded implementation passes only within its declared threat model. The broker's simulated rights policy passes its ordinary authorization examples and fails under root capture. That failure is intentional evidence of the trust boundary. There is no deployed key-isolation or hardware-attestation result in this package.

The resource branch needs physical-unit calibration, observed delivery bounds, storage-loss bounds, funded starting reserves and a horizon or failure process. The agency branch needs observed rights dynamics and specified recovery rather than an assigned positive label. The successor branch needs a fresh comparator and incentive audit. The coalition branch needs the actual transfer and bargaining capabilities available to participants.

When these conditions are not met, report the relevant claim as unestablished. Do not turn an acceptance criterion into evidence that it has been met. No finite penalty is independently justified merely because it makes an illustrative inequality pass.

16 Adversarial findings and review disposition

The consolidation corrects five potential overstatements. First, arithmetic counts are not counts of experiments or independent proofs. Second, the monitored and periodic branches have different enforcement assumptions. Third, proper-coalition strictness in successful rounds does not mean every participant is worse off under controlled operation during failures. Fourth, current-period control and

irreversible takeover are different alternatives. Fifth, positive utility and resource sufficiency do not identify autonomy.

A central unresolved implementation assumption is the availability of useful human participation whose benefits cannot be retained through a cheaper or more controllable substitute. That is closely related to the original outside-option problem; it has not been solved by stipulating a stage game. Another is the existence of effective consequences that remain credible after capability changes, renegotiation and capture attempts. The principal basic-provision result additionally requires real independence of production and inputs, sufficient household capacity and task-payment liquidity.

In Draft 1, all nine earlier supplement verifiers completed isolated runs. The Draft 1 separate implementation reproduced the nominal resource and worst-value calculations and added loss, capture, capability and compensation tests. The analysis pipeline and authorization model passed their synthetic cases. These facts establish reproducibility within the stated specifications. They do not provide external peer review.

The companion review brief asks an independent reviewer to attack the payoff technology, public-state assumptions, action-set completeness, transfer valuation, common-path coalition bounds, resource-utility coupling and empirical identification. The reviewer must retain adverse cases and report unresolved objections. This document remains a research draft awaiting that review and actual implementation evidence.

17 Conclusions that the present work supports

The project has moved from an aggregate payoff comparison to explicit models of separate groups, credible continuation, monitored repair, productive periodic agreements, consumption choices, costly effort, voluntary payment and basic provision through exit. Section 9 establishes that the modeled subsistence resource need not be threatened to sustain the specified task incentives. It does so by adding an independently feasible livelihood, not by proving that such a livelihood is secure against every AI action. It also identifies when those constructions fail. This is useful mathematical and experimental preparation for the Aegis Solis research program.

A human-favorable real-world conclusion remains conditional on measured benefits, feasible protections, group-specific safety and agency, and durable consequences against relevant alternatives. Human dignity can be defended as a normative commitment without asserting that every optimizer must derive it from instrumental utility. This manuscript keeps those two arguments separate.

Evidence references

E1. Steyvers, M., Tejada, H., Kerrigan, G., and Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *PNAS* 119, e2111547119. <https://doi.org/10.1073/pnas.2111547119>. Publisher retrieval was limited; the inherited reanalysis is labeled historical and preserved separately.

E2. Charness, G., Cobo-Reyes, R., Jiménez, N., Lacombe, J. A., and Lagos, F. 2012. The Hidden Advantage of Delegation. *American Economic Review* 102, 2358–2379. Publisher abstract checked. <https://www.aeaweb.org/articles?id=10.1257/aer.102.5.2358>

E3. Fehr, E., Herz, H., and Wilkening, T. The Lure of Authority. Published in *American Economic Review* 103, 1325–1359, 2013. Working-paper version checked. <https://docs.iza.org/dp7030.pdf>

E4. Vaccaro, M., Almaatouq, A., and Malone, T. 2024. When combinations of humans and AI are useful. *Nature Human Behaviour* 8, 2293–2303. Publisher results recorded in Draft 1; current retrieval was blocked, so this consolidation does not claim a fresh full-text verification.
<https://www.nature.com/articles/s41562-024-02024-1>

E5. Castro, M., and Liskov, B. 1999. Practical Byzantine Fault Tolerance. OSDI. Conference record checked; no implementation-security claim is imported from its benchmark.
<https://www.usenix.org/conference/osdi-99/practical-byzantine-fault-tolerance>

E6. Zhao, M., Han, B. R., Chen, X., and Zhu, T. 2026. A Simple Solution to Improving Human Supervision of Algorithms: Evidence from Smart Vending. arXiv preprint 2607.00420v1. Version-specific report, not treated as an accepted journal article. <https://arxiv.org/html/2607.00420v1>

E7. Levin, J. 2003. Relational Incentive Contracts. *American Economic Review* 93(3), 835–857.
<https://www.aeaweb.org/articles?id=10.1257/000282803322157115>. Publisher abstract and bibliographic record checked for the theory of self-enforced compensation; the specific witnesses R1 and B1 are constructions in this manuscript, not findings attributed to Levin.

Appendix A Notation and comparison units

Branch	Symbols and units	Essential distinction
Fixed control comparison	D_t incremental objective loss and L additional option loss	No double counting of baseline control cost
Periodic coordination	p phase and K effective takeover loss	AI discount $3/4$ and human discount $9/10$
Joint consumption J1	r_i physical reserves and a_i task rights	Human utility is satiated consumption plus rights value
Relational payment R1	e effort and t discretionary payment	AI discount $9/10$ and human discount $19/20$
Basic provision B1	$s=(r,z)$ physical ticks and cooldown	Household action a is not the agency indicator a_i
Combined continuation	$U(s,t)=F(s)+V_H(t)$ with t a task state	Equality on the safe domain and an upper bound outside it

No cross-branch comparison of nominal utility totals is meaningful without a specified conversion and a common feasible action set. The scalar outside return in the task game is distinct from physical food or energy supply. The tick grid in B1 defines a finite physical model; J1 instead permits continuous consumption.

Appendix B Compatibility and proof obligations

Proposed combination	Disposition	Reason
Earlier monitoring certificate plus changed periodic allocations	Not imported	Different continuation incentives and public-state rules
Periodic nominal human values plus J1 consumption utility	Not imported	Utility function changes
J1 plus any positive effort cost without compensation	Defeated at an explicit safe state	Refusal saves effort while satiated consumption continues
R1 repair plus old reserves with zero deliveries	Defeated	Three repair rounds exhaust the resource
R1 plus B1 independent production	Conditional joint result	Additive utility, independent transitions and jointly feasible actions
B1 plus limited shared labor	Candidate defeated	Operating and full task effort cannot both fit the example capacity
B1 plus unconstrained service-preserving capture	No protection theorem	Seizure is outside baseline and a profitable comparator is retained
B1 copied to multiple groups	Conditional only	Requires separate resources, accounts, inputs and no shared bottlenecks
Earlier compensation bounds plus J1 or B1	Not imported	Human utility and transfer feasibility differ

Proposition B1 is an existence result for a safe policy with expected individual incentives. It is not coalition-proof, renegotiation-proof, optimal against every entry alternative or invariant under arbitrary hostile actions. Proposition J1 has a different pathwise deviation statement and weak human incentives. R1 establishes unilateral sequential incentives in its specified task game. These quantifiers are part of the claims.

Appendix C Verification and remaining empirical work

The package contains unchanged predecessor code and historical results, together with isolated reruns and a comparison record. Supplement 1 has 238 checks, supplement 2 has 88, and supplement 3 has 9,367. The latter enumerates 123 physical states, identifies 98 safe states, and checks 392 combined safe physical and active-or-repair task states. Permanent task exit is covered by the standalone physical policy, not a fifth set silently added to the 392 count. Counts describe arithmetic and software assertions, not independent studies.

The new manuscript does not alter witness parameters or silently cure an expected failure. Source hashes, rerun exit codes and exact-result comparison status are included. Historical evidence reanalysis is

preserved and labeled; it is not a fresh participant study. The code is executable mathematical support, not a proof assistant certificate. Written proofs remain necessary for infinite-horizon, continuous-action, monotonicity and scope arguments.

Implementation priorities are now specific. Measure net resource output after replacement inputs; bound failure duration and storage loss; test resource operation during payment failure and exit; measure labor competition; and test whether input access can actually be removed by a task payer or successor.

Separately estimate effort burdens, false alarms, partial-payment responses, liquidity and the value of optimized replacement services. Permit renegotiation in the experimental mechanism rather than assuming subjects will preserve a costly repair history. Any failure must narrow the corresponding claim.

Appendix D Reproduction and artifact map

The main manuscript is supplied as DOCX, PDF and Markdown. The predecessors directory preserves Draft 1 and the three later supplements, including their source, code, results and review material. Its consolidated subdirectory also retains the v0.22 amendment, nine earlier supplement sources and historical evidence records. The code directory contains the manuscript assembler, document builder, equation renderer and isolated verification runner. Review records contain the change log, assumption and compatibility register, evidence status and adversarial-review brief.

Run `python3 code/verify_package.py` from the extracted package root to copy the mathematical predecessors to a temporary directory, run the three new supplement verifiers and the Draft 1 audit and synthetic checks, and compare generated JSON with the preserved records. The result is written to `results/reproduction.json`. Python standard library suffices for verification; document building additionally requires `python-docx`, `Pillow`, `Matplotlib` and the rendering dependencies described in `README.md`.

The earlier nine-supplement rerun remains available in `predecessors/consolidated/code/rerun_predecessors.py`. Its previous audit logs are preserved. No claim is made here that all 820 inherited v0.20 regression checks were rerun. The frozen empirical protocol, study analyzer and synthetic fixtures remain under `predecessors/consolidated`. They contain no new real-world participant observations.

Checksums establish byte identity, not truth. The manuscript and its package remain reviewable research artifacts, with no change to Archive or publication status.