

Competition Cooperation Alliance Bloc Community

Computational Mechanism Research

From “cooperation is better” to cooperation that is reachable, maintainable, verifiable, legitimate, correctable, and open to exit

Unified Research Master | Formally Locked

Version status: MASTER/V1.0 | FROZEN | First locked master eligible for derivation | Not the theoretical Canonical | Not policy authorization

Version lineage: V0.1 project baseline (frozen) → V0.2 load-bearing discipline revision (frozen) → V0.3 fully integrated candidate (historical candidate) → V0.4 locking candidate (historical candidate) → V0.5 final-review revision candidate (direct semantic predecessor) → V1.0 FROZEN

Current phase: PRE-HUMAN-COS | Research design may proceed. Real-world mechanisms, tool effectiveness, and real-world authorization must not be written as already established.

Date of this version: August 31, 2026

Theoretical lineage: When Intelligence Is No Longer Scarce, V0.6 formal publication candidate master

Current public mechanism bridge: Chapter 9 of When Intelligence Is No Longer Scarce, Chinese public long-form standalone edition V1.1.1. V1.1 records only when that bridge first emerged and has no independent current interface authority.

Project relationship: Independent mechanism research. It does not modify upstream theoretical texts or change the Human-COS theoretical baseline or Runtime scope.

Identity and reading discipline: This project does not predetermine a winner between competition and cooperation. It studies which structures hold, fail, or remain unknown under which conditions. It uses self-interest as a reachability stress test and real consequences to assess legitimacy. It permits negative results and refuses to “prove” a predetermined direction correct through parameter tuning, aggregate scores, or authoritative identities. MASTER/V1.0 is the first formally locked research master to which attachments may be bound. Subsequent documents may instantiate only the domains, actors, variables, thresholds, and record formats expressly left open here. They cannot silently rewrite frozen disciplines, acquire real-world powers, or transpose research governance into real-world institutions.

Structural navigation

Reading part	Chapters	Questions addressed
Research foundations	I–IV	Version, object, relationship states, actors, and behavioral baseline
Judgment discipline	V–IX	Claim levels, six tests, legitimacy, dual-track computation, and stage gates
Research operation	X–XV	Separation of powers, evidence methods, WP-01, H3/H5, stopping, and negative results
Boundaries and locking	XVI–XVIII and appendices	Risks, Human-COS, open items, sources, final checks, and version migration

I Version self containment and interpretive responsibility

1.1 Why V1.0 may be formally locked

V0.1 established the research object, six tests, dual-track computation, R0–R6 stage gates, WP-01, evidence lines, risk register, and Human-COS boundary. V0.2 added multidimensional self-interest, dual legitimacy gates, separation of verification, determination, and stopping powers, asymmetric evidence for H3/H5, and strict unknown discipline. V0.3 removed hidden inheritance and first integrated all current rules into a candidate master readable as a single document. V0.4 did not reopen the architecture already closed in V0.3. It closed only the interpretive and execution interfaces identified in the locking audit: functional decomposition must not masquerade as real-world institutions; H3 must not treat non-exploitation as mechanism safety; H5 must not treat limited coverage as exhaustive pathways; and norms, resources, prolonged unknowns, tool failure, and version compatibility must have executable consequences.

V0.5 repaired only load-bearing vulnerabilities and self-containment discrepancies identified during the final continuous reading of V0.4: RUN:REOPEN must not become a standalone operating permission; confirmed breaches of H3's preregistered exposure upper bound must enter FAIL; Human-COS and functionally equivalent tools must not assume R-L or manufacture "independent" review through tool replicas; independence must hold relative to the specific challenged object; and the same work package must audit actual control across the M and R planes together. Appendix claim levels, authorization labels, source attribution, and version chains were aligned with the main text.

MASTER/V1.0 is the locked result of a complete continuous review of V0.5, logical counterexample testing, power-responsibility audit, source-claim review, main-text/appendix consistency checks, and page-by-page finished-artifact verification. Apart from version identity, lineage, derivation root, and handover interfaces, V1.0 does not change V0.5's mechanism semantics. All current definitions, judgment disciplines, evidence rules, stage gates, power boundaries, stopping conditions, and open items appear in full here. V0.1–V0.5 now serve only historical tracing, difference review, and accountability evidence; they no longer serve as current clause repositories for interpretation or derivation of new attachments.

1.2 Zero silent changes

Any change to claim level, research object, actors and permissions, evidence status, normative gates, stage gates, stopping and restarting, upstream relationships, or the Human-COS boundary requires a new version and an entry in the change ledger. Text compression, layout, numbering, and editing without a change in meaning may be recorded as nonsemantic revisions. No rule may be repealed, replaced, tightened, or relaxed by simply ceasing to mention it.

MASTER/Vx.y-patchN may repair only nonsemantic errors that do not change definitions, permissions, thresholds, state transitions, normative mappings, open items, or the meaning of conclusions. A patch also creates a new file, new hash, and separate change record. Any safety defect requiring these rules to be tightened, relaxed, or reassigned is a semantic version change, however small the edit; it must not bypass final review as a patch.

The editor, original author, or same actual controller cannot alone declare patch eligibility. R-E preserves the baseline, complete diff, and before/after hashes. R-V verifies literal differences, referenced objects, and the impact inventory. At least one independent R-X review checks semantic equivalence and effects on permissions, norms, states, and claims. Reasonable disagreement disqualifies a patch and requires semantic versioning. R-G may decide adoption scope only from the established diff and review results; R-S appends records. Neither may rewrite the semantic judgment.

If an attachment identifies a gap in the master, it may only request a change; it may not amend the master internally. Until the conflict is closed, the relevant proposition or derivative work remains under scoped RUN:HOLD. An attachment is permanently bound to its original parent-master version and hash; a master upgrade does not rewrite existing attachments. After a change to the master version or hash, an existing attachment is, by default, pending compatibility verification under the new master. Compatibility must not be assumed. An independent record must review parent-version differences, affected clauses and fields, H/T/L/permission gates requiring reruns, review results, conclusions, and hashes; it must distinguish "compatible," "revalidation required," "migration and version upgrade required," and "incompatible." Until compatibility is closed, no stage advancement, attachment derivation, or positive claim dissemination may use the new master's name. Any change to attachment content, parent version, hash, normative mapping, or conclusions requires an attachment version upgrade. These are version-interface outcomes only; they create no new evidence or operational-state layer.

1.3 What this master closes and what it does not claim

- Fully restores V0.1's multi-actor topology, five structures, three timescales, evidence methods, event fields, failure outcomes, risk register, and detailed Human-COS boundaries that V0.2 did not restate to equivalent extent.
- Adopts V0.2's upgraded formulations: recursive self-interest, dual L0 and T5 gates, separation of research powers, H2 criterion adjustment, H3 safety upper bounds, H5 pathway coverage, and strict three-valued conjunction.
- Adds relationship-state definitions, M/R dual planes, R2 downstream counterexample-screening rules, L0-R/L0-M separation, evidence-state transitions, and the distinction between computational reproducibility and replication studies.
- Clarifies that research governance is not a real-world institutional blueprint, and supplies minimum metarules for opportunity coverage, pathway coverage, normative conflict, resource capacity, prolonged unknowns, tool failure, and version compatibility.
- Closes final-review bypasses involving temporary REOPEN permission, confirmed H3 breaches, tools making legitimacy determinations, merely formal independence, and cross-plane dual-hat control; aligns claim, authorization, version, and source formulations in the main text and appendices.
- Does not claim that WP-01 has been supported, that any country, alliance, or community should accept a particular real-world mechanism, or that the research team or Human-COS receives real-world powers.

II Project identity

Research structures without deciding for the real world

2.1 Central question

Where actors have different objectives, incomplete information, asymmetric capabilities, coupled domestic and international dynamics, and possible misjudgment and agreement deviation, which arrangements can change information, incentives, risk exposure, commitments, verification, and feedback so that risk-reduction or cooperative action becomes enterable, maintainable, verifiable, legitimate, correctable, and open to exit under specific conditions? Under what conditions do these arrangements fail, become captured, shift risks to third parties, or remain unknown?

2.2 Three levels must not impersonate one another

Level	Question answered	Responsible party	Boundary not to cross
Theoretical direction	Why upstream structures that continually generate confrontation, misjudgment, and escalation need to change	When Intelligence Is No Longer Scarce and its public long-form essay	Does not prove a specific mechanism works
Mechanism research	Which arrangements hold, fail, or remain unknown under which conditions	This master and subsequent work packages	Does not create real-world authorization or deployment eligibility
Real-world mechanisms	Who is authorized, who bears consequences, and how execution and correction work	Real sovereign actors, organizations, institutions, and communities	Cannot be exercised by theory, models, or tools in their place

Theory must not impersonate mechanisms, direction must not impersonate institutions, and research candidates must not impersonate deployable systems. Real-world mechanisms require actors entitled to bear the consequences to establish authorization, decision, execution, appeal, and responsibility chains.

Research charter identity: This master does not design a single cooperation mechanism. It establishes a research order in which “whether cooperation is reachable” can be studied honestly, without researchers, real-world powers, theoretical positions, or tools winning the question in advance. Research order can constrain how research reaches conclusions; its own maturity cannot confer authorization for real-world institutions.

2.3 The precise meaning of computable

“Computable” means that actors, states, causal chains, variables, evidence, thresholds, counterfactuals, failure conditions, and versions can be explicitly represented, tested, recomputed, and overturned. It does not mean real-world payoff structures have already been measured accurately, that all values can be quantified, or that model outputs can replace authorization, responsibility, or decisions.

Computability is first a discipline against ambiguity: turning “why believe this” into an object open to challenge, separating “possibly better” from “reachable in reality,” and honestly leaving unidentifiable parts UNKNOWN.

2.4 The M and R planes must be separated

This project strictly distinguishes two planes:

- M plane | Real-world mechanism operation: event notification, operational factual verification, conditional responses, repair, exit, and their real-world authorization. Only actors with real-world authorization who bear the consequences can carry the M plane.
- R plane | Research governance: preservation and verification of research evidence, posterior reference labels, legitimacy review, stage gates, version status, and reproducibility. E/V/X/A/L/G/S in this text belong by default to the R plane.

R-plane conclusions may constitute evidence; they cannot automatically trigger M-plane punishment, retaliation, policy action, or infrastructure shutdown. M-plane operational records cannot by themselves prove R-plane research conclusions. Subsequent work packages must prefix M-plane roles with M- and R-plane roles with R-, separately demonstrating their authorization, credibility, and reachability.

R-plane rules do constitute this project’s research governance. M-plane symbols, however, only decompose the functions, authority conflicts, and proof obligations that real-world mechanisms must address. They are not an agency roster, fixed hierarchy, unified organizational topology, sufficient institutional design, or source of authorization. Mature research governance, complete functional naming, or conditional support for T6-R must never imply that the M plane exists, is reachable, legitimate, credible, or deployable.

2.5 What this project does not currently answer

- It does not design one unified global mechanism or predetermine the eventual institutions of alliances, blocs, or communities.
- It does not reduce a country to one rational actor or equate “national utility” directly with society’s overall interests.
- It does not prove that cooperation between China and the United States, or any two parties, is already reachable, or that conditional cooperation has formed a stable equilibrium.
- It does not substitute multi-agent simulation for historical identification, legal analysis, expert judgment, political processes, or real-world authorization.
- It does not let Human-COS automatically generate policy, performance commitments, verification credibility, stopping powers, or real execution capacity.
- At this stage it does not address military deployment details, sensitive inspection schemes, or operational designs that could increase real confrontation risks.

Identity ruling: This project studies “under which conditions mechanisms hold, fail, or remain unknown.” It does not predetermine cooperation’s conclusion, optimize competition for one side, or acquire real sovereignty for any central system.

III Research object

Relationship states actor topologies and structural change

3.1 Five relationship states are not a one way civilizational ladder

Competition, cooperation, alliances, blocs, and communities are neither linear stages nor value rankings. They are relational configurations that can coexist across domains and actors. Cooperation can be embedded in competition;

alliances can cooperate internally and compete externally; blocs may be partially open. Community does not mean competition disappears.

State	Minimum operational definition	What cannot be inferred
Competition	Actors contest scarce resources, advantages, rules, or relative positions in a relevant domain; their outcomes substantially affect one another	Competition is necessarily hostile or zero-sum
Cooperation	Two or more actors coordinate action, share some information, or assume conditional obligations in a bounded domain to change joint outcomes	Cooperation is already stable, comprehensive, or legitimate
Alliance	Actors establish comparatively durable, selective mutual commitments and coordination across events, retaining membership boundaries and exit questions	Alliances are inherently credible or harmless externally
Bloc	Internal coordination and identity boundaries intensify substantially, and resources, rules, or security commitments cluster in networks that may generate exclusionary externalities	Blocs are internally homogeneous, permanently closed, or inevitably conflictual
Community	Overlapping members form continuing institutional relationships around observable public goods, dispute resolution, entry and exit, and shared responsibility	A community has a permanently correct identity or competition has ended

Dynamic transformation means changes in information, incentives, risk coupling, commitment density, membership boundaries, public goods, authorization, and correction rights. Transformations may be partial, bidirectional, reversible, or interrupted; the same actors' relationships may occupy different states in different domains simultaneously. R0–R6 indicates research maturity, not an inevitable sequence of real civilizational evolution.

3.2 Actors are not two abstract countries

Level	Typical actors	Questions that cannot be omitted
Individuals and groups	Leaders, implementers, experts, directly affected people, publics, and intergenerational actors	Cognition, fear, responsibility, burden-bearing, dignity, and legitimacy
Organizations and markets	Government departments, firms, research institutions, media, civil organizations, transnational networks	Agency problems, incentives, information filtering, resources, and capture
Countries and localities	Central/local government, regulation/execution, domestic veto players, policy coalitions	Domestic–international two-level interaction, perceived gains, and execution gaps
Alliances and blocs	Members, partners, peripheral actors, switchable members, third-party countries	Entrapment/abandonment, member heterogeneity, network externalities
Communities and institutions	Regional communities, international organizations, professional communities, public-goods networks	Entry, contribution, exit, dispute resolution, and anti-capture

The research object is not these actors' fixed personalities. It is their positions, visible information, capacities, constraints, perceived interests, real consequences, and interactions.

3.3 Five structures mechanisms change

1. Information structure: Who sees what and when; how signals become distorted; whether dissent can be retained.
2. Incentive and payoff structure: How gains from agreement deviation, cooperation returns, first-mover exposure, reputation, and accountability costs change.
3. Risk-exposure and coupling structure: Which actions transmit local risk system-wide; which domains can be isolated, modularized, or protected by firewalls.

4. Commitment and verification structure: Whether commitments can be performed in stages, symmetrically, or differentially; whether factual deviation, false reports, and attribution of motives can be separated.

5. Feedback and correction structure: How outcomes feed into the next judgment round; how review, suspension, exit, repair, reentry, and reauthorization work.

Mechanisms are not technological substitutes for “making people good.” Their minimum task is to make constructive action possible without key actors continually sacrificing core interests; to make concealment, burden-shifting, and unbounded escalation harder to profit from cheaply; and to make errors discoverable, containable, and correctable.

3.4 Three timescales coexist

- Event timescale: One notification, misjudgment, verification, or scoped response.
- Interaction timescale: Repeated conditional cooperation, agreement deviation, forgiveness, recovery, and strategy learning.
- Structural timescale: Changes over time in domestic coalitions, alliance networks, technological conditions, institutional paths, and community identities.

An event-only view mistakes temporary calm for stability. A structural-only view loses observable feedback. Each work package must identify which timescales it can and cannot identify.

IV Behavioral baseline

Respect self interest without making it an explanation of everything

4.1 Human problems cannot treat human experience as external noise

Material, technological, ecological, and biological causes are real causes. But any problem requiring people to keep choosing, organizing, obeying, confronting, coordinating, or bearing consequences cannot treat the human mind, human nature, human relationships, perceived gains, meaning, and individuals’ and groups’ experience of survival and development as external noise. A mechanism that works only if “all participants first overcome self-interest” has not demonstrated real-world reachability.

This does not mean every problem is caused by the human mind alone. More precisely, human cognition, interests, relationships, and structures of meaning help determine or substantially influence how external problems are perceived, allocated, amplified, collaboratively addressed, or repeatedly generated.

4.2 Multidimensional self interest

Dimension	Real motivations that mechanisms must admit
Survival and security	Life, body, property, organizational continuity, controllable risks, future options
Dignity and identity	Respectability, status, recognition, identity continuity, group belonging, freedom from domination
Relationships and fairness	Reciprocity, trust, reputation, closeness and distance, human relationships, procedural fairness, relative gains
Development and capability	Resources, knowledge, room to act, competitiveness, intergenerational opportunity, sense of control
Meaning and values	Achievement, mission, convictions, responsibility, intelligibility of sacrifice, shared narratives

These dimensions may conflict and change by stage; narratives, reference groups, identity relationships, and expectations also shape them. Research may use dimension-specific proxies and behavioral evidence, but must not compress them into an unjustified aggregate utility score.

4.3 Four fixed disciplines

Discipline	Minimum requirement
Non-saint condition	A candidate mechanism must not depend on key actors continually sacrificing their core interests, maintaining goodwill, or suppressing foreseeable reactions
Multidimensional-interest condition	Record gains, losses, and unacceptable boundaries separately at least by actor, dimension, time, and reversibility
Non-transfer condition	Actor A's net gain cannot establish the mechanism if it rests on uncounted burdens borne by B, third parties, vulnerable people, or future actors
Dual-ledger condition	Perceived interests explain behavior and reachability; real consequences determine legitimacy, distribution, and irreversible risks

“Actors pursue their perceived interests” is a behavioral hypothesis. “Their actions are therefore legitimate” is not a valid inference. Self-interest helps explain why actors act; it cannot erase real harm and responsibility caused by their actions.

This project does not require Pareto improvements for all vested interests. Interests derived from harm, monopoly, rent-seeking, or burden-shifting may be lawfully constrained. Buying obstructors’ cooperation must not shift costs again onto those without the power to refuse. Compliance produced through deception, coercion, concealment, or manipulation of perception must not count as evidence of mechanism reachability.

4.4 Multidimensional self interest must be refutable

Multidimensional self-interest is a design baseline and stress test for mechanism reachability, not a universal law explaining all behavior. Interest dimensions must not be added after observing outcomes so that every behavior is retrospectively explained as self-interest. Each work package must declare in advance relevant interest dimensions, observable proxies, competing motivational explanations, and refutation conditions. Where explanations cannot be distinguished, motive attribution remains UNKNOWN.

Altruism, responsibility, or sacrifice cannot automatically be absorbed merely by renaming them “meaning interests.” Institutional inertia, cognitive errors, emotional attachment, moral commitments, rule-following, and chance events may all offer competing explanations. Research must allow them to overturn or limit explanations based on self-interest.

4.5 Recursive self interest applies to researchers and gatekeepers

The self-interest hypothesis applies equally to competing actors, research teams, funders, evidence custodians, verifiers, challengers, determiners, legitimacy reviewers, stage gatekeepers, and state executors. Role names cannot replace incentive analysis. Every load-bearing role must record sources of interest, actual control, exit conditions, error costs, and accountability consequences.

Recursive self-interest cannot be solved by endlessly adding higher reviewers. It requires competing evidence, actual-control disclosure, terms and rotation, dissent and appeals, external independent challenge, computational reproducibility, independent reanalysis, and replication studies when new evidence is available, while prohibiting a closed loop of end-to-end control by one actor.

V Claim levels and state systems

Different vocabularies must not impersonate one another

5.1 Claim levels

Level	Precise meaning	Permitted wording
Direction	Identifies structures or risks worth changing without concluding a mechanism holds	“Requires research”; “a structural entry point may exist”

Level	Precise meaning	Permitted wording
Mechanism hypothesis	Specifies action–mediator–outcome and competing explanations	“If the conditions hold, this may produce...”
Evidence finding	Reports observations in specific samples, parameters, models, or events	“Within the frozen scope, we observed...”
Conditional support	All load-bearing hypotheses are supported within the frozen scope, with no load-bearing unknown	“R2-level conditional support, without pilot eligibility”
Research candidate mechanism	Passes the relevant stage’s six tests, legitimacy gates, independent challenge, and version gates	Must state both R stage and authorization status

Every output must report four relevant state layers: narrow-fact verification, load-bearing inference and effectiveness tests, legitimacy, and research operation. The second layer must report H states alongside T1–T4, T6-R, and T6-M. Conditional support for H cannot replace any T state. Outputs must also state the R0–R6 research stage and real-world authorization status. An inapplicable layer must explicitly say N/A with a reason; an unperformed required check can only be UNKNOWN. Omission or N/A must not imply passage. A research candidate cannot be rewritten as policy advice or deployment eligibility.

The standard output record has a fixed order: claim ID and scope → document version and hash → R stage → M authorization status → narrow-fact states → summary of load-bearing blocks and FAIL/UNKNOWN → each load-bearing H → T1–T4/T6-R/T6-M → L0-R/L0-M/T5 → RUN state → strongest counterevidence → reopening conditions. The summary must include H FAIL/UNKNOWN; T1–T4, T6-R, and T6-M FAIL/UNKNOWN; L0R-REDESIGN/LIMIT/FAIL; L0M:REDESIGN/LIMITED/UNKNOWN/REJECT; and T5 FAIL/UNKNOWN. It must precede every positive H/T/L conclusion. Earlier narrow-fact or authorization verification must not be described as a positive mechanism conclusion. Every state also records its forming actor, evidence window, version, and challenge status. Later itemized states must not change this order. States remain separate; aggregate scores, overall traffic lights, or “overall passage” are forbidden. Appendix F supplies only a quick-reference order, not a second rule set.

The minimum standard real-world authorization labels are NO M-AUTHORIZATION / PILOT-SCOPE AUTHORIZED / OPERATIONAL AUTHORIZED. NO M-AUTHORIZATION means only that research has not verified real-world authorization satisfying this master’s minimum semantics. It is not a final ruling by the research team on real-world legal validity.

PILOT-SCOPE AUTHORIZED requires at least a verifiable real-world authorization document specifying: the authorizing actor, issuer, authorization basis, and chain of authority; approved activities, participants, geographical area or technical systems, data and disclosure scope, risk upper bounds, and explicit exclusions; commencement, termination, automatic expiry, and renewal procedures; the concrete holders of M-D and M-S, responsible actors, and authority ceilings; triggers, decision-makers, and executors for initiation, suspension, revocation, emergency stopping, rollback, and safe exit; direct participants’ rights to information, consent, refusal, exit, and data, and the authority basis, protection boundaries, and prohibited burden transfers where individual third-party consent cannot be obtained; evidence preservation, access, logs, incident reporting, independent M-V/M-X, decision review, appeal, repair, and necessary compensation; conflicts of interest and actual-control disclosures; document version and hash; and reauthorization after load-bearing changes to actors, scope, data, risk, or powers. The document must expressly forbid transferring M-plane powers to the R plane or Human-COS.

PILOT-SCOPE AUTHORIZED must not be used if any minimum element is missing, expired, beyond the authorizing actor’s powers, unverifiable, or not reauthorized after a load-bearing change. The label proves only that specified actors are permitted to pilot within a specified scope and period. It does not prove effectiveness, legitimacy, scalability, T6-M passage, or OPERATIONAL AUTHORIZED status, and cannot renew itself or automatically become operational authorization. WP-01-A can provide only blank fields and candidate validation plans. Real authorization can be issued only by real actors empowered to bear consequences in a future R6.

OPERATIONAL AUTHORIZED is not the natural endpoint of research labels. It may be reported only if, in addition to the pilot minimum elements, the real authorization document specifies routine-operation scope, continuing legal and authority bases, operational capacity and responsibility holders, independent M-V/M-X, continuing supervision and incident reporting, periodic review, appeal, repair and compensation, automatic suspension and revocation, termination

or migration plans, and reauthorization after load-bearing changes. Real authorization is issued by empowered actors; the R plane may only verify and accurately report it. R-E preserves authorization documents and subsequent records; R-V verifies the existence of clauses, scope, validity period, and authority chain; independent R-X challenges omissions, actual control, and material changes. No positive authorization label may be used while any load-bearing element is unverified, expired, or subject to unresolved disagreement. Operational authorization does not prove permanent effectiveness, permanent legitimacy, universal scalability, or exemption from H/T/L review. Tools, the R plane, and operating organizations must not confer it on themselves.

5.2 Four state systems each stay in their own layer

Layer	States	Who forms them	What they do not represent
Narrow-fact verification output	SUPPORTED / CONTRADICTED / INSUFFICIENT / NOT VERIFIABLE	R-V; challengeable by R-X	Do not directly determine H or infer motives
Load-bearing inference and effectiveness tests	H: FAIL / UNKNOWN / CONDITIONAL SUPPORT; T1–T4, T6-R, T6-M: the same three-valued states	R-A forms H, T1–T4, and T6 technical parts; R-L forms T6 rights–authorization parts	H does not replace T; none equals real authorization
Legitimacy admission and review	Layered L0-R, L0-M, T5 states	R-L from frozen norms and evidence	Do not prove causal mechanism effectiveness
Research operation	RUN:PROCEED / RUN:LIMITED / RUN:HOLD / RUN:REOPEN (transition record) / RUN:CLOSE	R-G decision or hard conditions taking effect; mechanical execution by R-S	Do not stop real actors or systems; REOPEN itself permits no research action

INSUFFICIENT and NOT VERIFIABLE usually leave the corresponding H UNKNOWN. Whether CONTRADICTED makes H fail depends on preregistered failure conditions and relevance; it cannot automatically skip levels. Only unresolved, load-bearing V/X conflicts place the corresponding proposition in scoped UNKNOWN or trigger scoped RUN:HOLD. Irrelevant or strategic objections must not receive unlimited stopping power.

5.3 The state transition chain

R-V judges only the relationship between evidence and narrow propositions; R-X supplies independent challenge. Under the frozen hypothesis–test mapping and aggregation rules, R-A forms sourced, versioned, challengeable, revocable H and T1–T4 research states, and separately assesses technical operability of suspension, exit, repair, rollback, and reopening interfaces in T6-R/T6-M. R-L forms L0-R, L0-M, and T5, and separately assesses authorization, appeal, responsibility, and rights interfaces in T6-R/T6-M. R-A and R-L judgments of their own correction interfaces require R-X or independent reanalysis; self-assessment cannot prove T6-R. R-G decides research stages only from already formed H/T/L states. It neither produces evidence, forms test conclusions, nor changes thresholds. R-S only executes established research states and appends records.

T6-R and T6-M are each formed mechanically by preregistered strict conjunction: any load-bearing FAIL makes the whole FAIL; with no FAIL, any UNKNOWN makes the whole UNKNOWN; only support for all components yields CONDITIONAL SUPPORT. R-S merely records the conjunction; R-G cannot reweight or rewrite it. Real authorization can be issued, changed, or revoked only by M-plane authorizers with actual statutory or agreed powers. R-E preserves the documents; R-V verifies scope, duration, and validity; R-A/R-L form research judgments only. No R-plane role may create, expand, or revoke real authorization.

No shortcuts: Verification output must not directly generate “malice,” “betrayal,” “a response is warranted,” or real punishment. Stage gates must not manufacture permission to proceed by changing models, evidence, or definitions of success.

VI Six effectiveness tests

Six parallel dimensions

The same research result cannot be promoted across levels. Test numbering is for tracking; it does not mean legitimacy intervenes only after effectiveness.

Test	Central question	Minimum output and boundary
T1 Relative improvement	Compared with a real baseline and feasible alternatives, does the candidate state reduce risks or improve relevant outcomes?	Do not compare only with a costless ideal; preserve distribution
T2 Real world reachability	Can actors enter from real initial conditions, rather than merely have an ideal end state?	First-mover exposure, startup conditions, perceived interests
T3 Maintainability	Can it persist or recover under misjudgment, agreement deviation, succession, shocks, and third-party action?	Distinguish stability, recovery, and reentry
T4 Verifiability and robustness	Can behavior be identified within limits; do conclusions hold across parameters, models, and independent teams?	Verification is not motive determination; report unknowns
T5 Legitimacy	Does it transfer costs, violate rights, omit actors, create capture, or produce unacceptable distribution?	Governed by L0-M and the process ledger; passage is stage-specific
T6 Correctability	Can it suspend, exit, repair, reopen, reauthorize, and retain negative results?	Separate stopping and restarting powers; record time limits, appeals, and feedback; establish real authorization separately

T6 must report two distinct states. T6-R examines whether the research process can suspend, correct itself, retain negative results, and reopen. T6-M examines whether the real mechanism has authorized structures for suspension, exit, repair, appeal, rollback, and reauthorization. Chapter XV directly establishes only T6-R; it does not automatically prove T6-M. Passage of T6-R cannot offset T6-M failure or unknowns. Before R6, only design-level conditional judgments of T6-M consistent with the current evidence stage may be formed; real operational validation cannot be claimed.

T1–T4, T6-R, and T6-M use FAIL / UNKNOWN / CONDITIONAL SUPPORT; T5 uses FAIL / UNKNOWN / PASS. H states are important evidence inputs for T1–T4, but H CONDITIONAL SUPPORT does not automatically mean any T has conditional support. Each T judgment must specify its dependent H, counterfactuals, computational reproducibility, independent reanalysis/replication, normative or correction evidence, and stage ceiling. With inadequate evidence windows, actor coverage, or normative bases, T5 must remain UNKNOWN. “No harm found yet” does not automatically constitute PASS.

“Research candidate mechanism at that R stage” is permitted only when the supporting confirmatory evidence was obtained within the scope expressly allowed by L0R-PASS; L0-M admits it; T1–T4 all have CONDITIONAL SUPPORT at the current stage; T5 is PASS; T6-R and the T6-M required at that stage have conditional support; no load-bearing unknown remains; and the stage-required computational reproducibility, independent challenge, and independent reanalysis/replication have been completed. Materials from L0R-REDESIGN/LIMIT/FAIL scopes can support only redesign, scoping, gap identification, or negative results; they cannot alone support positive candidate eligibility. Results obtained through unsafe, unauthorized, or unauditable research cannot support it either. R1 normally yields only a “mechanism hypothesis” or “structural candidate”; research-candidate eligibility first becomes possible at R2. If a purported R1 result independently satisfies T2, T3, and corresponding T4 requirements, its actual maturity must be recorded as R2; lower stage names cannot evade thresholds. No stage before R6 has real pilot or deployment eligibility.

T5 PASS means only that no hard failure or load-bearing unknown has been found under the prefrozen normative framework, actor scope, evidence window, and public objections. It is not ultimate proof of universal legitimacy. Changes to the normative framework, emergence of new actors, or new evidence supporting an objection require reopening.

VII Dual legitimacy gates

Review research actions and mechanism eligibility separately

Notation: R in L0-R means Research Action, the research-action safety gate; M in L0-M means Mechanism Design, the candidate-mechanism legitimacy gate. Both are formed by R-L on the R plane. Neither is the research operational state formed by R-G, and neither constitutes M-plane real-world authorization. R in R0–R6 means Research Maturity. Fully qualified state names are used below to avoid mixing gate judgments with operational states.

7.1 L0 R research action safety gate

L0-R reviews research actions themselves: whether data acquisition, disclosure, simulation, experiments, publication, or research interactions create unacceptable risks, rights violations, or real escalation. It is an R-L safety judgment, not an execution order. Its states are:

- L0R-PASS: Research action is implementable within the frozen scope.
- L0R-REDESIGN: First change data, scope, disclosure, authorization, or safety measures, then review again.
- L0R-LIMIT: Only scoping, gap identification, historical, or noninterventional research is permitted.
- L0R-FAIL: The research action itself crosses a hard boundary, forming grounds for scoped RUN:HOLD/CLOSE.

L0R-REDESIGN suspends the planned research action; it cannot continue until redesign and review are complete. L0R-LIMIT permits only the scoping, gap-identification, historical, or noninterventional activities expressly specified in the state. L0R-FAIL prohibits the corresponding research action. Only L0-R states may limit research actions on grounds of research safety, rights violations, information leakage, or real escalation. Under preregistered rules, R-G may still limit claims, stages, versions, or close propositions/work packages for evidentiary, methodological, governance, or resource reasons, but cannot thereby erase safe, noninterventional negative-result research.

Once L0-R establishes grounds, R-S mechanically executes objective hard conditions; where interpretation, attribution, or balancing is needed, written R-G review applies. R-L cannot execute stopping or restarting itself. A mechanism's illegitimacy does not prohibit safe research; the formation conditions, failure causes, and negative results of dangerous mechanisms particularly need to be recorded.

7.2 L0 M candidate mechanism legitimacy admission gate

Before domain selection, sampling, and substantial modeling investment, L0-M reviews whether a candidate mechanism depends on cost transfer, rights violations, or irreversible burdens. It does not prove legitimacy. It prevents clearly unacceptable designs from gaining a path advantage through “first establish effectiveness, discuss costs later.”

State	Effect
L0M:ADMIT	Allows confirmatory research within the frozen scope where major actors and burden pathways are adequately covered, no hard threshold is crossed, and no load-bearing L0-M unknown remains; does not mean T5 has passed
L0M:REDESIGN	No advancement until distribution, scope, authorization, or repair structures are redesigned
L0M:LIMITED	Only scoping, failure-mechanism, or noninterventional research is permitted
L0M:UNKNOWN	Only evidence needed to close legitimacy unknowns may be obtained
L0M:REJECT	The design must not continue toward candidate-mechanism, pilot, or policy-recommendation status; negative results must still be preserved

L0M:REJECT closes the positive candidate eligibility of that design version. It does not delete existing evidence or constitute safety permission for continued research. Within the scope expressly allowed by L0-R, noninterventional negative-result research may continue without changing real actors' rights, incentives, resources, or action environments, including evidence preservation and reanalysis, historical or comparative research, causal diagnosis of failures, and scoped simulation and red-teaming that do not increase real implementation capacity. These activities do not require rewriting the original REJECT as ADMIT.

After REJECT, researchers must not recruit real actors to implement the design, expand sensitive disclosure or operational capacity, generate pilot, deployment, or policy-recommendation materials, or continue optimizing positive candidate eligibility under the name of “studying failure causes.” If research changes mechanism structures or parameters to restore feasibility, introduces real intervention, expands operational capacity, or seeks new positive labels, it must create a new version and pass L0-R and L0-M again. The original REJECT and its negative results remain permanently. Each work package must record research scopes permitted and prohibited after REJECT.

L0-M’s “no load-bearing unknowns” refers only to the actors, rights, irreversible burdens, and transfer pathways required for admission. It does not require H1–H5 to be answered before research begins.

7.3 Mandatory L0 M questions

Check	Required answer
Actor coverage	Who benefits, who bears burdens, who can veto or execute, and who may be omitted through absence?
Causal dependence	Must some class of actors bear costs for the candidate mechanism to produce positive results?
Burden character	Reversible/irreversible; compensable/noncompensable; exit possible/impossible; short-term/intergenerational
Authorization and repair	Who may consent, refuse, or appeal; can compensation, exit, correction, and accountability actually reach them?

7.4 Process externality ledger and T5

After L0-M passage, each change in models, samples, scope, participants, and parameters must append to the externality ledger without overwriting prior records. Minimum fields include burden-bearing actors; direct/indirect/delayed/network pathways; predicted and actual burdens; probability/intensity/distribution/time; reversibility and compensability; exit conditions; authorization/appeal/repair; evidence; unknowns; objections; and whether L0-M changes.

New actors, new burden categories, potential irreversible risks, failed exit, concentrated burdens, or transfer to future actors are only signals for reopening L0-M. Where a preregistered objective expression is met, R-S executes scoped RUN:HOLD; where interpretation is needed, R-G determines the response from R-L’s written opinion. R-L reforms L0-M but cannot control stages or execute actions itself. T5 reviews L0-M and the full process evidence at the current stage. Without real pilot evidence, it can form only a legitimacy judgment at the current research stage, not deployment-level legitimacy. Positive T1–T4 results, sunk research investment, or increased bilateral gains cannot offset T5 failure or load-bearing unknowns.

7.5 Prevent burden shifting and arbitrary normative vetoes

- Freeze applicable norms, actor scope, burden categories, and thresholds before seeing results. Standards added afterward belong only in a new version.
- Legitimacy judgments must connect to concrete rights, harms, distribution, authorization, responsibility, and real actors, rather than unfalsifiable abstract slogans.
- Compare real baselines and feasible alternatives; do not veto every improvement against a costless ideal. Compensation cannot erase noncompensable rights or irreversible risks.
- Trace the source of gains. “Benefits” created by first deliberately imposing threats, blockades, or artificial scarcity and then removing some of the harm must not count directly as mechanism benefits.
- Mechanism proponents bear the burden of proving non-transfer. Evidence from affected people, independent review, written reasons, objections, and reopening conditions must be retained.
- L0-R controls the safety boundary of research actions. L0-M and T5 control mechanisms’ research-candidate eligibility, claim levels, and pilot recommendations. None acquires stopping powers over real countries, firms, individuals, or infrastructure.

7.6 Norm formation and conflict

What is frozen is an auditable normative package

What is frozen beforehand is not an abstract value conclusion but an auditable normative package. It records at least: normative sources and scope; affected actors and their input; representation of absent actors; rights and potentially irreversible hard boundaries; competing norms; conflict-handling rules; reasons for nonadoption; objections; and revision and reopening conditions. R-L forms L0-M/T5 states but cannot choose normative premises unilaterally. Affected parties' views must enter the evidence and reasoning chain, but do not automatically confer unlimited veto power.

Where a load-bearing normative conflict cannot be closed under prior rules, the relevant L0-M or T5 remains UNKNOWN. Majority votes, aggregate gains, funder preferences, research investment, or dominant actors' identities cannot flatten it. Semantic changes to the normative package require versioning and reopening affected L0-M/T5 judgments. Work packages instantiate participants, representation, procedures, and deadlines. Where competing norms have no common adjudication rule that can be preregistered, research must report conflicting conditions and their respective consequences side by side, rather than manufacture a falsely unified answer.

VIII Computational discipline

Separate tracks by harm dimension without offsets across tracks

8.1 Not a simple reversible versus irreversible event split

Tracks are assigned by harm dimension, not event name. One mechanism may simultaneously generate reversible economic costs, noncompensable rights violations, and irreversible systemic risks. Only reversible and compensable dimensions enter the net-value track. Irreversible, noncompensable, or hard-rights-threshold dimensions enter the constraint/threshold track. No offsets across tracks are permitted.

8.2 Reversible and compensable dimensions

The net value distribution and counterfactual track

Dimensions such as transactions, notifications, supply chains, military-expenditure increments, and recovery time may undergo net-value, interval, counterfactual, and sensitivity analysis where data and identification permit. Results must preserve distribution by actor, interest dimension, and time, rather than report only one aggregate net value.

8.3 Irreversible noncompensable and hard rights threshold dimensions

This project treats nuclear escalation and civilization-scale loss of AI control as potentially irreversible or exceptionally difficult-to-reverse dimensions. They are not monetized indiscriminately alongside ordinary benefits. Research uses conservative lower bounds on losses, probability intervals, unacceptable thresholds, stress tests, and robust decision-making. Greater total gains for other actors cannot automatically override noncompensable rights. Precisely pricing civilizational survival is unnecessary and may hide normative premises.

8.4 Required risk reduction threshold

From P^* to Δp^* with subscripts i, r, t

The historical P^* is retained only as an old name. The precise quantity is the minimum absolute reduction in risk that the candidate mechanism must achieve relative to the competitive baseline for actor i , risk type r , and time window t :

$$\Delta p^*_{i,r,t} = \max\{0, W_{i,r,t}(\text{comp}) - W_{i,r,t}(\text{mech})\} / L^{\text{LB}}_{i,r,t}$$

Here W includes only outcomes in compensable dimensions, and $L^{\text{LB}} > 0$ is a conservative lower bound on the corresponding potentially irreversible loss. Numerator and denominator must use comparable units. The threshold answers only “how much must risk fall for it possibly to outweigh the ordinary-benefit difference?” It does not answer whether risk actually falls, the mechanism is reachable or maintainable, or it is legitimate.

The expression holds only where the candidate mechanism primarily changes the event probability and the conditional loss after the event in comp and mech can be represented by the same loss scale $L_{i,r,t}^{LB}$. If the mechanism also changes loss severity, consequence distribution, or burden-bearing actors, $p_{comp} L_{comp} - p_{mech} L_{mech}$ must be modeled separately. If no comparable scale can be established, output NOT COMPUTABLE. Multiple risk types must not reuse the same W difference unless an allocation rule or joint-risk model is specified beforehand.

- Actors, risk types, time windows, and probability units must match. Do not combine actors into one global threshold.
- A negative numerator gives a zero threshold. This means only that this particular ordinary-benefit barrier is absent, not that the mechanism holds. The max operator reflects this project's discipline that ordinary gains cannot compensate irreversible risk; it is not a purely algebraic conclusion of break-even analysis.
- Achievable risk reduction satisfies $0 \leq \Delta p_{i,r,t} \leq p_{i,r,t}(comp)$. If Δp^* exceeds a credible upper bound on the competitive-baseline risk probability, mark UNATTAINABLE. If the uncertainty interval overlaps the achievable boundary, retain UNKNOWN. A value above 1 or above the baseline probability must not be interpreted as a reachable threshold.
- Without a credible lower loss bound, identifiable causal risk difference, or consistent units, output NOT COMPUTABLE and use robust constraints instead. Do not force computation.
- Rights, actor standing, noncompensable harm, and nontransferable responsibility must not be placed in W and compressed into one monetary score.

This quantity is only a threshold-sensitivity probe, not an offset across tracks. If any potentially irreversible, noncompensable, or hard-rights threshold has already been crossed, the probe does not apply and cannot confer permission.

IX R0 to R6

Increasing research maturity rather than a sequence of real world evolution

Stage	Research object	Main outputs	Gate to the next stage
R0	Research charter	Claim levels, L0-R/L0-M, six tests, dual-track computation, separation of powers, negative-result rules	Operational concepts; no mixing of states and permissions
R1	Minimal bilateral mechanism	One domain, two principal actors, one testable structural change	Can test a change to a payoff, belief, or transition probability
R2	Reachability and maintenance	Misjudgment, deviation, recovery, first-mover exposure, three-valued conditional map	Real entry pathway exists; failure and unknown regions can be marked
R3	Domestic–international two-level interaction	Departments, interest groups, implementers, veto players, internal perceived gains	No longer depends on “the country is a single actor”
R4	Alliance and bloc networks	Member heterogeneity, third-party externalities, entrapment/abandonment, network transitions	Conditional results persist under network changes without external burden-shifting
R5	Communities and public goods	Overlapping membership, entry, contribution, exit, disputes, anti-capture	Observable public goods arise without dependence on permanent closure
R6	Reversible real-world pilots	Authorized low-risk settings, rapid feedback, replication across real settings, rollback	Feedback is faster than harm propagation; authorization, responsibility, computational reproducibility, necessary replication in new settings, and exit hold

9.1 No stage skipping but downstream counterexample screening is required

Stage gates limit positive-claim levels and full-model complexity. They do not prohibit downstream pathway screening for counterexamples, externalities, or omitted variables. R1/R2 must perform scoped screening of domestic veto players, allies, and third-party pathways, without deriving R3/R4 positive claims from it.

If a downstream pathway could be decisive, the current conclusion remains UNKNOWN. An exploratory higher-level module may open solely to close that unknown. Until prior stages pass, exploratory findings cannot become proof of a higher-level mechanism, community formation, or deployment eligibility. Thus “no stage skipping” cannot justify ignoring third parties, and “screening was performed” cannot justify proving a more complex mechanism at a higher stage.

9.2 Reproducibility and replication span the stages

Computational reproducibility means obtaining consistent results using the same data, code, and steps. Reimplementing code or analysis using the original data is independent reanalysis. Replication studies retest claims with new data or independent samples. Independent challenge and computational reproducibility are cross-cutting disciplines from R1 onward. R6 adds authorized reversible real-world pilots and replication across real settings; it is not the first point at which independent review becomes permissible.

9.3 Stage dependencies and work not yet undertaken

Timing	Work	Boundary
Current main-text stage	Freeze definitions, claim levels, six tests, dual tracks, and state/permission invariants	Do not instantiate concrete institutions or thresholds
After master locking	Prepare WP-01-A and its governance, coding, variable, and output attachments	Instantiate only the master's open items
After shadow trials	Compare conventional processes with Human-COS increments in actor coverage, error recording and correction, dissent, and computational reproducibility	Test research quality first, not real-world “success or failure”
After independent review	Progress stage by stage into two-level interaction, networks, public goods, and low-risk real-world pilots	Follow R0–R6; simulation scale does not justify skipping stages
Not yet undertaken	An aggregate index compressing complex relations into one score; a global central coordinating body; high-consequence automated intervention; large full-domain simulation	Too many dependencies, with risks of false precision and hidden power

X Research governance

Determination stopping and restarting powers also require constraints

10.1 Seven R plane functions

Concrete institutions and personnel will be instantiated in work packages after master locking. Functions, outputs, and incompatible roles must first be frozen in the master. All powers below are R-plane research powers, not real-world mechanism powers.

Function	Permitted tasks only	Explicitly prohibited tasks
R-E Evidence preservation	Original materials, hashes, sources, access, versions, revision records	Judging cooperation/betrayal; deciding stage states
R-V Research verification	Four evidence outputs for preregistered narrow propositions	Inferring motives; directly deciding H or real responses

Function	Permitted tasks only	Explicitly prohibited tasks
R-X Cross-challenge	Independent review under the same protocol, counterevidence, omissions, alternative explanations	Erasing facts through identity, vote counts, or noisy objections
R-A Posterior determination	Challengeable, revocable H, T1–T4, and T6-R/T6-M technical-operability reference labels under frozen rules	Creating an unquestionable “fact center”; forming T5; also serving as R-G
R-L Legitimacy review	L0-R, L0-M, process ledger, T5, and rights–authorization components of T6-R/T6-M	Replacing empirical tests; arbitrary vetoes using unregistered abstract values; also serving as R-G
R-G Stage gate	Deciding RUN:PROCEED, LIMITED, HOLD, REOPEN, or CLOSE from existing H/T/L states	Changing evidence, thresholds, models, or definitions of success
R-S State execution	Mechanically executing R-G decisions on version freezing and scoped RUN:PROCEED/LIMITED/HOLD/REOPEN/CLOSE; objective hard triggers mechanically activate only scoped RUN:HOLD; appending records	Interpreting conditions itself, deleting negative results, restarting alone

R-A must not call its labels a fact center. It produces sourced, versioned, challengeable, revocable posterior reference labels. Only narrow propositions adequately established by independent external facts may be factual calibration benchmarks. Event facts, source attribution, actor intentions, responsibility, and real responses must remain layered. R-A has no authority to determine actor intentions, responsibility, or real responses. R-X and independent reanalysis may challenge R-A. Its reversal rates, omissions, and unresolved states also enter error records.

10.2 M plane roles require separate proof

WP-01’s limited verification is a candidate mechanism component on the M plane. At a minimum, any future real agreement must distinguish, in authority and effect: M-E operational evidence preservation, M-V operational narrow-fact verification, M-X cross-party review, M-D real-world decision authorization, and M-S real-world action execution. Concrete actors remain for work-package instantiation; the master freezes the following boundaries.

Functional positions are not institutional blueprints. M-E/M-V/M-X/M-D/M-S identify only functions, permissions, incompatible roles, and proof obligations that real mechanisms must answer if they enter design or authorization. They do not prescribe five agencies, a fixed organization chart, one jurisdiction, a hierarchy, or a unified global institution. Any mapping of these functions to real actors, institutions, procedures, or technical interfaces is a new M-plane candidate design. It must independently specify real authorization, responsibility-bearing, actual control, appeals/review, and exit, and pass L0-M, T5, T6-M, and corresponding stage gates again. Real arrangements may use different names, numbers, and organizational topologies and add necessary functions; they may not dissolve the master’s frozen permission separation. Complete R-plane governance, mature research labels, or available tool interfaces are not evidence that this mapping is reachable, legitimate, credible, or authorized.

These boundaries constrain only how future candidate designs must not exceed authority. They create no positions, appointments, organizational eligibility, or action powers for real actors:

- M-V/M-X may form only preregistered narrow-fact outputs, not directly order responses, punishment, or stopping.
- When M-V/M-X conflict, M-D may take only scoped, reversible action under preauthorized “rules for handling uncertainty”; it must not rewrite UNKNOWN as fact.
- M-D must possess real authorization and bear decision consequences. M-S may execute only within authorization, retaining records of responsibility, appeal, repair, and exit.
- R-plane roles, Human-COS, and research labels must not exercise M-D or M-S. At the current PRE-HUMAN-COS stage, only M-D/M-S rules are studied; no real responses are initiated.

Minimum M-plane role incompatibilities are: M-V cannot also be M-D; for actions imposing constraints, losses, or rights effects on others, M-D cannot also be M-S. M-X independence must hold relative to the specific object under review: it must be independent of the challenged M-E, M-V, M-D, M-S, mechanism proposer, verified actor, and their actual controllers. Absence of conflict with another object cannot substitute for independence here. No actual controller may simultaneously control M-E evidence entry, M-V/M-X factual outputs, M-D real decisions, M-S execution, and

appeal/review. M-D decisions must be reviewable by a separately authorized, actually independent actor. Without review, appeal, and correction chains, T6-M remains UNKNOWN and real authorization status can only be NO M-AUTHORIZATION. If minimum separation cannot be met in a domain, “low risk” or “limited resources” cannot make nominal task division count as separation of powers. The M-plane scope must shrink, independent countersignature be added, or work remain design-level research.

If M-X is not independent of the challenged object, that review cannot count as independent factual or correction evidence. Affected T4, T6-M, or other dependent states remain UNKNOWN; authorization labels cannot count that review as satisfying an independence requirement. M-S, M-E, and other objects absent from a fixed shortlist are not exempt from independent challenge of their conduct and records because their role names differ.

M-plane operational records enter R-E only as raw materials, then undergo independent research verification by R-V/R-X. M-V cannot manufacture R-A labels for itself or prove effectiveness merely because “the mechanism operated.” Nor may the R plane directly command M-plane actions through research conclusions.

10.3 Four outputs of limited verification

R-V and future M-V output only the following for preregistered narrow propositions within their respective planes:

- **SUPPORTED:** Current evidence supports the narrow proposition within the frozen scope.
- **CONTRADICTED:** Current evidence has a load-bearing contradiction with the proposition.
- **INSUFFICIENT:** Evidence quantity, quality, or observation window is inadequate.
- **NOT VERIFIABLE:** Verification is impossible under the current protocol, permissions, or available information.

These outputs cannot directly generate “malice,” “strategic betrayal,” or “retaliation is warranted.” Unverifiability caused by refusal, concealment, delay, or destruction of preregistered evidence interfaces may count as “observable agreement deviation” only where interface duties, the actor’s control capacity, and actual deviation are all observable. Missing evidence alone cannot imply malice.

Notification counts, verification counts, or agreeing votes cannot substitute for evidence quality, coverage, opportunity windows, and source independence.

10.4 Minimum role incompatibilities and actual control tests

At least these absolute relationships are frozen:

- $R-V \neq R-A$: A real-time or principal research verifier cannot supply final posterior labels for its own judgments.
- $R-A \neq R-G$: Whoever forms H labels cannot decide whether to advance those labels.
- $R-L \neq R-G$: Legitimacy reviewers cannot acquire stage control alone.
- $R-G \neq R-S$: State decision-makers cannot interpret and execute their own decisions.
- $R-X$ must be independent of $R-V$, $R-A$, and the mechanism-proposing team.
- $R-X \neq R-G$: Challengers cannot decide stage stopping or restarting from their own challenges.
- $R-A \neq R-L$: Empirical-label formation and legitimacy judgment cannot be performed under the same actual controller.
- The original model or mechanism-proposing team must not substantively control $R-E$, $R-V$, $R-X$, $R-A$, $R-L$, $R-G$, or $R-S$, particularly for the evidence entry, label formation, legitimacy judgment, stage decisions, HOLD execution, or restart execution for its own project.

These are minimum general incompatibilities, not an exhaustive independence list. Independence must be judged by specific claim, object, event, version, and challenge scope. $R-X$, independent-reanalysis, and replication nodes must be independent of the evidence custodian, verifier, label former, legitimacy reviewer, stage decision-maker, state executor, mechanism proposer, researched actor, and actual controllers challenged in that review. Otherwise, the review cannot count as an independent evidence line; relevant T4, T6-R, T6-M, or other dependent states remain UNKNOWN. Another conflict-free review cannot fill this independence gap.

Independence is assessed through actual control and shared dependencies, not institution names, signature counts, or formal separation. Any combination of funding, appointment/removal, data entry, model/intellectual property, credentials, personnel and technical upstream dependencies, judgment/publication, and stopping/reopening sufficient to

determine evidence, labels, or states creates substantive-control risk. “Not holding every lever simultaneously” is no safe harbor.

Before freezing, each work package must submit an actual-control matrix mapping nominal holders, ultimate beneficiaries or controllers, and funding, appointment/removal, data entry, model/IP, credentials, personnel/technical upstream dependencies, judgment/publication, and stopping/reopening powers. It distinguishes direct control, indirect control, common upstream dependence, and de facto dependence. Each relationship records whether an actor can initiate, prevent, change, delay, or revoke the matter, with basis, scope, duration, alternative channels, and independent checks. R-E preserves supporting materials, R-V verifies narrow facts, and R-X independently challenges. R-G may apply governance consequences only from existing records, not supply evidence itself. If a control pathway spanning incompatible functions has not been excluded, or key control relationships are unverifiable, T6-R remains UNKNOWN. If the gap also compromises evidence integrity, affected H/T4 likewise remain UNKNOWN or become FAIL under preregistered failure conditions. These materials cannot obtain positive candidate eligibility. The master does not mechanically score control by “holding a certain number of levers.” Work packages instantiate actors, isolation measures, and substitutes.

For the same work package, event, and applicable scope, a combined M/R actual-control matrix is also required. M-plane actors being studied, verified, or potentially acting on research findings, and their actual controllers, must not control R-plane evidence entry, verification, challenge, labels, legitimacy, stages, state execution, or reopening that evaluate their own records, decisions, or mechanisms. Separate naming, dual appointments, internal firewalls, outsourcing, or formal separation within one group are no safe harbor. Where a conflict cannot be excluded, relevant actors must recuse themselves and actually independent substitutes must be used. States formed through the conflicted chain cannot remain in use. If the conflict touches R-L or the normative chain, original L0-R, L0-M, and T5 states lose current validity. This is loss of formation eligibility, not re-formation of L0/T5 by R-G, R-S, or automated rules. Standard outputs must retain the old labels, invalidation reasons, and “pending re-formation by independent R-L.” Until actually independent R-L reforms them, L0-M and T5 are treated in aggregation and operational judgments as UNKNOWN, and scoped RUN:HOLD applies by default because no valid L0R-PASS exists. New L0-R, L0-M, and T5 states may still be formed only by actually independent R-L from evidence. If the conflict touches evidence, challenge, labels, stages, or correction chains, affected T4, T6-R, T6-M, and research-candidate eligibility remain UNKNOWN. Only failure states already satisfying preregistered hard-failure conditions through independent, valid evidence and independent review may be retained. REJECT/FAIL formed through the conflicted chain likewise cannot remain effective. Conflicted research conclusions cannot be an independent basis for M-D. Whether real authorization exists is still verified separately under Section 5.1; research conflicts do not automatically create, expand, or revoke it.

Where early resources are limited, one institution may carry several nonconflicting functions, but personnel, permissions, credentials, logs, recusal, and decision authorization must be separated, with remaining conflicts disclosed. If minimum separation cannot be met, T6-R remains UNKNOWN; affected evidence/test states remain UNKNOWN according to their dependencies. Limited resources cannot lower the standard.

Resource shortages may reduce research scope, stage, and claim level, not evidence, separation-of-powers, legitimacy, or safety thresholds. Each work package must declare beforehand the minimum governance capacity for positive claims: auditable evidence preservation, independent challenge of load-bearing evidence, actual isolation of incompatible powers, L0 review and appeal, append-only state/version records, and executable research stopping and restarting. Where capacity is insufficient but safe research remains possible, only RUN:LIMITED is available for scoping, historical, noninterventional, or negative-result research. Relevant H/T cannot receive positive support; they remain UNKNOWN if preregistered failure conditions are unmet, and remain FAIL if those conditions are met. Without a credible path to close capacity gaps within the frozen research window, enter RUN:CLOSE. If evidence cannot be safely preserved, risks cannot be scoped, or end-to-end control cannot be removed, enter scoped RUN:HOLD/CLOSE. If only M-plane holders, authorization, or correction capacity are inadequate, maintain NO M-AUTHORIZATION and design-level research; safe R-plane research need not close. Work packages instantiate personnel, budgets, deadlines, and alternatives.

10.5 Pre acceptance error records and recursive closure

Record who accepted which procedures in advance: accepting actor, procedure version, scope, evidence interfaces, conflict handling, exit/reacceptance conditions, and timestamp. Preacceptance does not waive challenge; procedural failure, changed evidence, or power capture still permits reopening under the rules.

- Eligibility, pay, renewal, and reputation must not depend on “finding deviation/betrayal” or “proving cooperation successful.”
- Appointment, funding, data, IP, and career dependencies must be disclosed. Challengers should receive substantively symmetric protocols, evidence interfaces, and appeal conditions.
- Both verified actors and verifiers enter append-only records; permanent traceability must not apply only to the verified.

R-E, R-V, R-X, R-A, R-L, R-G, and R-S all retain append-only records of decisions, errors, corrections, and consequences. Each record includes at least the original proposition, evidence snapshot and sources, time, output, posterior label and its formation source, false positives/negatives or other errors, correction time, real harm, error source, self-report or external discovery, repair, and eligibility consequences:

- R-E: Evidence omissions, loss or tampering, preservation delay, broken provenance, unauthorized access, hash/version mismatches, selective receipt, repair time, and consequences.
- R-V: False positives, false negatives, UNKNOWN/NOT VERIFIABLE rates, coverage, and correction latency. Probability calibration is used only if probability outputs are introduced later.
- R-A: Reversals, omissions, unresolved cases, label sources, and challenge records.
- R-L: Omitted actors, normative drift, interest dependencies, and capture risk.
- R-G: Mistaken stops, missed stops, selective exemptions, reasons, and review results.
- R-S: Execution delays, excess of authority, selective execution, and record completeness.
- R-X: Unfounded challenges, strategic obstruction, omissions, and refuted challenges.

“Remaining unknown for a long time” triggers only an audit of procedure and evidentiary effort; it cannot force affirmative verification. The expected cost of voluntarily reporting and quickly correcting errors should be lower than concealing, delaying, or having errors discovered externally. No role may prove its own reliability.

M-E, M-V, M-X, M-D, and M-S must likewise retain append-only decision, error, appeal, correction, real-consequence, and responsibility records. M-plane actors cannot prove judgment reliability merely by saying they are authorized.

XI Research organization evidence lines and event data

11.1 Organizational division does not replace separation of powers

Research division	Main responsibility
Questions and boundaries group	Research questions, normative premises, potentially irreversible boundaries, permissible conclusions
Evidence and identification group	Historical data, counterfactuals, selection bias, competing explanations, uncertainty
Formal modeling and simulation group	Models, parameters, sensitivity, code, computational reproducibility
Red team and dissent group	Strategic exploitation, alternative explanations, capture, leakage, failure conditions
Actor and legitimacy review	L0-R/L0-M, process ledger, affected-party evidence, responsibility, T5
Independent reanalysis/replication nodes	Reanalysis or testing with new evidence without reliance on the main team's internal process

These are divisions of labor, not automatic equivalents of R-E/V/X/A/L/G/S. One organization may perform multiple research tasks, but compatibility of permissions and isolation of actual control require separate review.

11.2 Four mutually constraining evidence lines

1. Historical events: Build event data and a coding manual recording initial states, information, notifications, operational verification, responses, outcomes, omissions, and unknowns.

2. Comparative identification: Compare comparable events with and without the relevant arrangements, addressing institutional selection bias, common shocks, reverse causality, and sample comparability.

3. Formal models: Make repeated or stochastic interaction, state transitions, incomplete information, misjudgment, deviation temptations, response delays, and recovery explicit. Model equilibria do not automatically establish real reachability.

4. Red-teaming and review: Independently seek information leakage, deception, automatic commitments, third-party arbitrage, domestic capture, mistaken attribution, and unfalsifiability.

No evidence line can validate itself. Historical evidence that an arrangement existed does not prove it reduced overall risk. Finding an equilibrium in a formal model does not mean real actors can enter it. Preregistration reduces post hoc tuning; it does not automatically repair wrong variables, wrong causal graphs, or invalid data.

11.3 Minimum event data fields

- T0: Information, capacities, constraints, expectations, and version states reasonably available to each actor before the event.
- Actors and roles: Formal power, de facto influence, execution, M-plane verification, R-plane research roles, consequence-bearing, and third-party positions.
- Events and signals: Triggers, timestamps, information sources, reliability, missingness, objections, and evidence-preservation status.
- Mechanism actions: Notification content, disclosure scope, M-plane verification methods, conditional responses, deadlines, review, exit, and repair.
- Outcomes: Misjudgment, identification of agreement deviation, escalation, leakage, recovery, externalities, distribution, and subsequent updates.
- Counterfactuals and competing explanations: What might occur without the mechanism, with delayed notification, or with a different disclosure scope; how alternatives are distinguished.
- H3/H5: Exposure vectors, opportunity windows, third-party actors and pathways, systemic safety thresholds, coverage, and unknowns.
- Legitimacy: L0-M actors, process externalities, rights, authorization, appeal, repair, and T5 changes.
- Governance and versioning: R-E/V/X/A/L/G/S decisions, errors, HOLD, reopening, dissent, data/code versions, and parent-version hashes.

11.4 Sensitive information boundaries

Public research starts only with low-disclosure, reversible settings reviewed by L0-R. Materials involving military vulnerabilities, exploitable gaps in critical infrastructure, real action windows, personnel identities, or other high-consequence information require separate access, minimal-disclosure, responsibility, audit, and exit rules. Where handling within safety boundaries is impossible, relevant research remains RUN:LIMITED or RUN:HOLD.

XII First work package WP 01

Crisis notification limited verification and conditional response

12.1 Topic and boundary

WP-01 does not begin with “comprehensive cooperation” or sensitive arms transparency. It studies a low-disclosure, reversible candidate mechanism with observable feedback cycles: can crisis notification, limited operational verification, and proportionate, reversible conditional responses reduce escalation through misjudgment without significantly increasing risks of first-mover exploitation, third-party reversal, or illegitimate burden-shifting?

Public health, nonsensitive AI safety incidents, disaster-resource interruption, and infrastructure failure are only candidate domain names, not inherently low-risk settings. Each concrete domain must first pass L0-R and L0-M and demonstrate acceptable data-disclosure, observation-window, and real-consequence boundaries.

Field	WP-01 master formulation
Research unit	One domain, two principal actors; permits preaccepted M-V or cross-verification under the same protocol
Target variables	Notification delay, misjudgment probability, agreement-deviation classification, escalation intensity, information exposure, recovery time, third-party systemic risk
Core arrangements	Trigger, minimal disclosure, timestamps, limited M-plane verification, proportionate reversible response, false-report repair, review and exit
Core outputs	Conditional-support/failure/unknown regions; four basic state layers, T1/T2/T3/T4/T5/T6-R/T6-M, R stage, and authorization status reported together
Hard boundaries	No automatic retaliation, no attribution of blame through motives, no conversion of research HOLD into real-world stopping

12.2 Minimum causal chain remains a hypothesis to test

1. A predefined observable event triggers notification and an operational evidence-preservation record.
2. Actors disclose only the minimum information needed for handling and verifiable timestamps, not complete capability transparency.
3. Limited M-plane verification narrowly classifies agreement deviation, false reports, concealment, or inadequate evidence, retaining cross-verification conflicts.
4. M-V/M-X produce only narrow-fact outputs. The current package makes no determination of “strategic betrayal” that can directly trigger real action. If such labels are needed later, a separate independent, multisource, appealable M-plane determination and authorization process is required.
5. Candidate conditional responses are rules to test only. Any real response requires separately authorized M-D decision and M-S execution, remaining proportionate, reversible, time-limited, appealable, and repairable. False reports, erroneous verification, and partial breaches must retain repair and reentry paths.
6. Outcomes enter the R plane to update behavioral hypotheses, verifier records, externality ledgers, triggers, disclosure scope, thresholds, and the next version. Original records must not be silently overwritten.

This chain is only a mechanism hypothesis. If any step lacks an observable mediator, has unidentifiable causality, or depends on unrealistic conditions, the entire claim cannot be upgraded.

12.3 WP 01 load bearing hypotheses

Hypothesis	Test question	Main observations
H1 Reduced misjudgment	Do notification and limited verification reduce the probability of misreading nonhostile events as hostile acts?	Delay, classification errors, escalation intensity, counterfactuals
H2a Identification performance	Is classification accuracy improved for observable agreement deviation, false reporting, or concealment?	False positives/negatives, coverage, correction delay
H2b Incentive effect	Does changed identification capacity actually reduce expected gains from, or incidence of, hidden deviation?	Behavioral changes, opportunity windows, substitute strategies
H3 First mover noninferiority	Do the first notifier's security, bargaining, information, economic, and domestic exposures remain at or below preregistered upper bounds?	Exposure vector, bounds, opportunity coverage
H4 Recoverability	Are there repair and reentry paths after false reports, partial breaches, or external shocks?	Recovery time, repair rate, reentry costs

Hypothesis	Test question	Main observations
H5 No third party reversal	Do third-party/network responses avoid offsetting bilateral risk reduction and avoid crossing potentially irreversible systemic safety thresholds?	Pathway coverage, net systemic risk, safety thresholds

If either H2a or H2b is FAIL, H2 is FAIL. With no failure but either UNKNOWN, H2 is UNKNOWN. Only conditional support for both gives H2 CONDITIONAL SUPPORT. H2a identification improvement does not directly establish H2b incentive changes; H2b behavioral changes must exclude alternative causes. H5 subhypotheses also use strict conjunction. Every H must predefine support, failure, unknowns, real baselines, counterfactuals, observation/exploitation windows, error ranges, and competing explanations.

12.4 Failure outcomes admitted in advance

- Notification does not significantly reduce misjudgment or escalation.
- Information leakage, probing, evasion, or strategic exploitation offsets verification gains.
- The mechanism works only under unrealistically high detection rates, high future weights, near-symmetric capabilities, or extremely narrow parameter ranges.
- Domestic veto players, alliance commitments, suppliers, platforms, or other third-party actions can reverse outcomes.
- The mechanism depends on unacceptable burden transfers, lacks exit and repair, or the research itself cannot remain within safety boundaries.

Any established failure must yield “not supported,” “reduced scope of applicability,” “L0M:REJECT,” or “research remains RUN:LIMITED/RUN:HOLD.” Cooperation must not be tuned into victory through unversioned changes to samples, parameters, thresholds, or proxies.

XIII H3 and H5

Stricter evidence for claims that something did not happen

13.1 H3 safety upper bounds and noninferiority logic

H3 cannot receive support merely because “no exploitation was observed.” It borrows and adapts noninferiority-testing logic to establish dimension-specific safety upper bounds for strategic exposure. This does not mean a clinical method has been externally validated in this domain. Specific margins require joint support from domain evidence, acceptable risk, and affected actors’ reasons.

Element	What WP-01-A must instantiate
Exposure vector	Security, bargaining, information, economics, domestic politics, reputation, future action space; key dimensions must not offset each other
Margin δ_k	Preregistered safety upper bound, measurement error, and severe-event triggers for each dimension
Counterfactual	Feasible controls: no notification, delayed notification, different disclosure scope
Opportunity coverage	The other party must have sufficient exploitation opportunity and observation windows; no opportunity is not proof of safety
Evidence minimum	At least two mutually constraining evidence lines; crossing intervals or inadequate coverage mean UNKNOWN

A “sufficient opportunity window” cannot be represented only by elapsed time or a one-off yes/no opportunity judgment. WP-01-A must preregister proxies, evidence sources, and competing explanations for these load-bearing links: whether exposure actually arose; whether potential exploiters could access and recognize it; whether they had action capacity; whether incentives or gains met a real threshold; whether domestic, legal, resource, alliance, or third-party constraints could block action; whether delayed exploitation windows had matured; and whether research

interfaces could observe attempts, failed actions, and delayed consequences. If any load-bearing link cannot be distinguished, “no exploitation occurred” cannot support H3. If the claimed mechanism protects by eliminating exposure, reducing recognizability, limiting exploitative capacity, or changing exploitation gains, this protection chain must be separately preregistered as a mediator and evidenced; zero events cannot establish it by inference.

A competing explanation is a load-bearing objection only if it specifies a clear causal mechanism, fits the frozen evidence, would alter H3’s state or scope if true, and identifies distinguishing observations or explains structural unverifiability. Explanations below that threshold enter a residual-objections ledger without independently blocking. Those meeting the threshold but structurally indistinguishable keep H3 UNKNOWN. R-X may raise load-bearing objections, but cannot obtain permanent stopping power by endlessly adding hypotheses with no discriminating pathway.

H3 state rules: A severe-event trigger in any key exposure dimension, or a confirmed breach of δk under preregistered measurement-error and interval rules → FAIL. Evidence excluding breaches within preregistered opportunity chains and observation windows, or independent support that a preregistered protection chain truly blocked exposure while other load-bearing exposure dimensions stayed within bounds → CONDITIONAL SUPPORT. Zero events alone, any unknown load-bearing opportunity-chain link, insufficient samples, distorted proxies, or intervals overlapping δk without confirming whether it was crossed → UNKNOWN. Negative safety signals may justify early stopping; positive support cannot be formed before the opportunity chain and delayed observation window mature. If an external constraint is explicitly frozen as an applicability condition, support holds only within that narrow condition. The external constraint’s effect cannot be attributed to the mechanism or extrapolated beyond that condition. Discriminating rules developed after seeing outcomes are exploratory repairs. Apart from mechanical error correction, the same already-seen data cannot reconfirm H3; confirmatory judgment requires holdout samples, new data, subsequent independent windows, or independent replication evidence.

WP-01-A must preregister the $H \rightarrow T$ dependency graph. H3 is a necessary evidence condition for T2 “real-world reachability.” T2 cannot receive conditional support if H3 is FAIL or UNKNOWN. If delayed exposure has a load-bearing effect on maintainability, its propagation to T3 must also be specified beforehand, not chosen after outcomes appear.

13.2 H5 pathway coverage outcome nonreversal and systemic safety thresholds

Subhypothesis	Empirical test requirement
H5a Pathway coverage	Freeze inclusion thresholds, exclusion reasons, direct/indirect/delayed/network pathways, and observation windows by capability × motivation × opportunity × consequence
H5b Outcome nonreversal	Identified third-party and network responses do not reverse bilateral risk reduction into higher systemic risk
H5c Systemic safety threshold	Third-party pathways do not cross preregistered potentially irreversible systemic safety thresholds

Third parties include allies, partners, suppliers, platforms, domestic veto players, localities and implementers, affected groups, and switchable members. “No reaction observed” is evidentially meaningful only with adequate coverage of actors, capabilities, opportunities, pathways, and time windows.

H5a’s “adequate coverage” is a bounded procedural judgment, not exhaustion of every imaginable pathway. Before seeing results, WP-01-A must freeze actor domains, pathway-generation methods, search-source categories, direct/indirect/delayed/network observation windows, inclusion thresholds, materiality thresholds, exclusion reasons, and search-stopping conditions. Only execution of these procedures and attainment of minimum information permit H5a support within that versioned scope and window. This cannot be stated as universal absence of third-party risk.

A newly proposed pathway reopens relevant H5 components only if it is credible, material, and boundary-relevant. Credibility requires actors, actions, testable causal links, and at least one auditable evidence anchor. Materiality requires possible change to H5b/H5c within credible parameter ranges, or revelation that preregistered coverage methods systematically omit high-consequence pathways. Boundary relevance requires the pathway to be within the frozen design scope or sufficient to demonstrate that the original boundary was inappropriate. Potentially irreversible pathways may use a lower, but nonzero, evidentiary inclusion gate. If the threshold and failure conditions are met, use FAIL. If

failure conditions are not yet met, H5 returns to UNKNOWN until the pathway closes, and all current states dependent on that support label roll back. Below-threshold pathways enter a residual-pathway ledger with exclusion reasons, evidence gaps, review times, and reopening conditions. They neither automatically reopen nor may be deleted. Their cumulative effect may trigger reopening only through preregistered aggregation rules. A single conjecture cannot gain unlimited veto power.

If a new pathway should have been discovered within the originally frozen generation domain and observation conditions but was omitted, current H5a should be FAIL and coverage-robustness-related T4 must reopen. If it arises from later actors, technologies, or environmental changes, the old conclusion remains only a bounded historical-version judgment and current H5a returns to UNKNOWN. Neither situation permits silently overwriting old labels or continuing to use the old version as current positive support. WP-01-A must preregister dependency mappings from H5 components to T1/T3/T4; R-G cannot decide whether to propagate state consequences through “overall judgment.”

H5 empirically tests only whether third-party pathways reverse risk and cross preregistered systemic safety thresholds. These thresholds are not value-neutral: domain evidence supplies the risk basis and measurement boundaries; R-L reviews rights, distribution, noncompensability, and normative premises; R-A judges threshold crossing only under frozen rules. An H5 threshold judgment cannot replace L0-M/T5 legitimacy judgments, and L0-M/T5 passage cannot count as empirical support for H5. When H5 reveals new actors or burdens, first form an L0-M reopening signal under 7.4. After preregistered trigger conditions are met, reopen through Chapter XV’s state chain. Pathway proponents cannot directly change L0-M or research operational states. If a new pathway also changes the normative package, rights, or burden boundaries, T5 must reopen under its own formation rules.

Any FAIL among H5a/H5b/H5c → H5 FAIL. No failure but any UNKNOWN → H5 UNKNOWN. All supported within the frozen scope → H5 CONDITIONAL SUPPORT.

Positive H5 support cannot be formed before preregistered third-party response windows, pathway coverage, and minimum information requirements are met.

XIV Strict conjunction and conditional maps

Positive evidence cannot wash out unknowns

14.1 WP 01 conjunction of H states

Conjunctive state	Rule	Permitted action
FAIL	Any load-bearing H or load-bearing subhypothesis fails	Overall FAIL; narrow scope, redesign, or close
UNKNOWN	No failure, but any load-bearing H remains unknown	Overall UNKNOWN; only research to close unknowns
CONDITIONAL SUPPORT	All load-bearing H supported within the frozen scope	L0-M, T5, T6, stage gates, and independent review still required separately

Voting is not permitted. Positive H1, H2, and H4 evidence cannot offset H3 or H5 UNKNOWN. Passing 4/5 is not “majority support.” WP-01’s H1–H5 use strict conjunction. Other work packages must preregister their own logical aggregation: necessary conditions use strict conjunction; alternative sufficient pathways may preregister disjunctive or nested structures. No load-bearing FAIL or UNKNOWN on a pathway supporting a positive claim may be voted away. Changing aggregation logic after observing results requires versioning. H1–H5 belong only to WP-01 and must not expand into a permanent hypothesis set for all mechanisms.

14.2 A conditional map rather than one aggregate score

WP-01’s core output is a conditional map. At its first occurrence, “conditional-support region” means only that load-bearing H receive conditional support within frozen actors, parameters, evidence windows, and hypotheses. It does not mean a candidate mechanism, much less a region of real-world validity.

Every condition combination includes at least: four basic state layers, T1/T2/T3/T4/T5/T6-R/T6-M, R stage, real authorization status, parameter ranges, strongest counterexample, externalities, sensitive-information boundaries, and reopening conditions. Every load-bearing unknown must be disclosed before positive results.

XV Stopping restarting and negative results

Research cannot control the real world

15.1 Complete state chain

Research states follow this chain:

1. Any actor may report a risk signal. A signal report is not itself a formal trigger.
2. R-E confirms only material existence, integrity, sources, versions, and interface status; it makes no substantive stopping judgment.
3. A preregistered objective expression may become OBJECTIVE_TRIGGER only when input fields, evidence-integrity requirements, and computational rules are frozen. Inputs involving factual values or classifications require R-V confirmation under protocol. Any input dispute, attribution, or balancing sends the condition to G_REVIEW; R-S must not interpret it. OBJECTIVE_TRIGGER requires the same verified inputs to necessarily yield one unique result under frozen rules, with trigger scope and execution action predetermined and no attribution, value ranking, proportionality balancing, or discretion needed. Hiding interpretive judgments in field names, classification labels, or threshold sources does not make them objective triggers.
4. R-S mechanically executes scoped RUN:HOLD for objective hard conditions; R-G cannot waive them. R-G determines interpretive conditions in writing after conflict recusal, preserving dissent and independent review.
5. During RUN:HOLD, only evidence preservation, counterevidence, repair, and unknown-closing activities separately confirmed safe by L0-R are permitted. Under L0R-FAIL, normally only minimum evidence preservation and accountability records that do not perpetuate risk remain; other research cannot continue under the pretext of “closing unknowns.”
6. Remedial evidence first enters R-E; R-V reviews narrow facts, with at least one independent R-X or independent-reanalysis evidence line. Changes to H or T1–T4 labels require R-A redetermination. Actor, transfer, rights, norm, or T6-authorization interfaces require R-L review. Unclosed load-bearing conflicts keep corresponding states UNKNOWN or under scoped RUN:HOLD.
7. After independent review, conflict recusal, and preservation of dissent, R-G decides in writing whether to reopen. R-S executes through a new operational/work-package version, retaining original objections, parent version, and hash. Reopening does not automatically restore old positive labels, cannot exceed the pre-HOLD claim scope, and must repass all affected gates.

RUN:REOPEN is only a one-time, versioned transition record, not standalone research permission. When R-S opens a new version, R-G must simultaneously assign RUN:PROCEED, RUN:LIMITED, or RUN:HOLD under this section. If affected L0/H/T/actual-control/stage gates have not been repassed, or no simultaneous assignment is made, default to RUN:HOLD. RUN:REOPEN itself cannot authorize data acquisition, analysis, publication, stage advancement, or attachment derivation, restore old positive labels, expand old scope, or bypass any gate. Once version registration and target-state assignment are complete, the REOPEN record ends. Reopening after RUN:CLOSE follows the same rules.

RUN:HOLD must be scoped to specific propositions, versions, models, datasets, publication claims, stage advancement, or pilot recommendations. Local conflicts must not freeze unrelated research.

R-G must use the same state-selection rules. RUN:PROCEED applies only when every current research action has L0R-PASS; L0-M is L0M:ADMIT or documented as inapplicable; T5 has no load-bearing block; minimum governance capacity is met; no unresolved load-bearing FAIL/UNKNOWN or active HOLD trigger exists; and proposed work stays within frozen stage, scope, and claim level. RUN:LIMITED applies where current actions are allowed by L0-R and a bounded, auditable unknown-closing plan exists; or where L0R-LIMIT, L0M:LIMITED/UNKNOWN/REJECT, or T5 FAIL/UNKNOWN permits only negative-result preservation, failure diagnosis, independent reanalysis/replication, counterevidence, or unknown-closing within the respective permitted scopes; or where H/T FAIL has formed and only

these safe negative-result activities and noninterventional redesign preparation for a new version are proposed. None may upgrade positive claims for the original version. Scoped RUN:HOLD applies to L0R-REDESIGN, L0M:REDESIGN, L0R-FAIL; when continuing the current protocol would increase exposure or perpetuate invalid confirmation; when safety or governance boundaries must be redesigned first; or when positive claims are being improperly upgraded. Under L0R-FAIL, normally only minimum evidence and accountability records that do not perpetuate risk remain. Where no feasible path to distinguish states remains within frozen time, resources, safety, and available evidence, use RUN:CLOSE under 15.4. State choice, scope, and reasons must be written and challengeable by R-X, not freely replaced by R-G “overall judgment.”

The precise meaning of RUN:HOLD: It suspends only research states, positive conclusions, parameter-tuning upgrades, version derivation, or pilot recommendations. It does not suspend government, corporate, individual, financial, military, infrastructure, or other real actions, and confers no real sovereignty on the research team or Human-COS.

15.2 Preregistered stopping conditions

- Key data cannot be audited, or one interested party controls evidence entry without challenge.
- Load-bearing competing explanations cannot be distinguished, or conclusions hold only under one optimistic parameter, window, sample, or model.
- H3, H5, or other load-bearing hypotheses lack preregistered coverage and remain UNKNOWN while positive conclusions are proposed for upgrade.
- Major relevant actors are omitted, or the process ledger reveals dependence on burden transfer.
- Ordinary net-value aggregation dilutes potentially irreversible risks, or preregistered systemic safety thresholds are crossed.
- Majority or identity flattens load-bearing R-V/R-X conflicts, role-error records are concealed, or computational reproducibility persistently fails.
- The original team tunes parameters, changes windows, deletes samples, changes labels, or switches proxies without versioning, or relaxes L0/T1–T6 conditions afterward.
- M/R planes are mixed so that research labels are intended to trigger real punishment, retaliation, policy action, or infrastructure stopping.
- Shadow operation of Human-COS or equivalent tools triggers Section 17.3 conditions involving audit failure, bias, leakage, centralization, responsibility evasion, or inability to stop, yet tool use is proposed for expansion, unreviewed positive outputs are retained, or tools reopen themselves.

WP-01-A must label each condition OBJECTIVE_TRIGGER or G_REVIEW. Nonmechanical conditions must not masquerade as automated stopping; stage gates must not waive objective hard conditions through “overall judgment.”

15.3 Negative results and version discipline

- Finding failure regions, unknown regions, unacceptable transfers, or unverifiability is itself a formal research output.
- A mechanism operating only with unrealistically high detection rates, high future weights, near-symmetric capabilities, or extremely narrow parameter ranges must be recorded as lacking real-world support.
- Exploratory adjustments must be separated from confirmatory results, preserving original hypotheses, code, data, thresholds, and differences.
- Cooperation must not be tuned into victory through unversioned sample deletion, shortened windows, relaxed margins, switched proxies, or altered exclusion criteria.
- Directional contradictions in computational reproduction or independent reanalysis of original data make T4 FAIL. Inability to recompute due to inadequate permissions, data, or documentation makes T4 UNKNOWN.
- Directionally contradictory replication with new data or new independent samples requires reopening the original support label and assessing T4 FAIL or UNKNOWN through preregistered relevance rules. “Different methods” cannot automatically exclude it.

15.4 Prolonged UNKNOWN and RUN CLOSE

UNKNOWN is not an instruction for infinite investment, and time alone cannot close it. Each load-bearing unknown must preregister feasible evidence lines, access and observation windows, stage-resource boundaries, expected information gains, risk ceilings, review points, responsible actors, and reopening conditions.

After R-V verification, R-X or independent-reanalysis challenge, and R-A or R-L confirmation according to the proposition type, R-G may decide in writing to RUN:CLOSE a specified proposition, version, or work package if: currently safe, lawful, realistically available evidence lines have been implemented or documented as infeasible; remaining key evidence is structurally unavailable, or further effort cannot reasonably yield enough information to change states within frozen resource/risk boundaries; no narrower design exists without relaxing failure conditions, omitting actors, or increasing unacceptable risk; and closure reasons, untested explanations, residual risks, resources spent, and future reopening conditions are publicly recorded. Expiry of time or budget normally creates only a G_REVIEW signal, not a change in evidence state.

RUN:CLOSE closes current research operation, resource investment, and version derivation, not the unknown itself. R-A cannot change H or T1–T4 because a project closes; R-L cannot erase normative unknowns. Standard outputs must retain corresponding FAIL or UNKNOWN, not rewrite them as no risk, mechanism safety, mechanism failure, default opposition, or default permission. Only preregistered new evidence, methods, permissions, or material environmental changes allow a reopening request through the original chain. Reopening requires versioning and repassing affected L0, evidence, actual-control, and stage gates, without automatically restoring old positive labels.

XVI Risk register

Research itself may create risks

Risk	Manifestation	Control
False precision	Complex actors compressed into aggregate utility, an overall index, or one threshold	Layered outputs; preserve distribution, intervals, unknowns, dissent
Temptation of scale	Premature full-domain multi-agent simulation	R0–R6; add only one positive-claim complexity level at a time
Substitution of direction	“Better” promoted into “reachable/stable/legitimate”	Six tests, claim levels, four state layers, stage/authorization labels
Bilateral extrapolation	Bilateral prototype directly expanded into two large blocs	Downstream counterexample screening; R3/R4 positive gates
Verification monopoly	Same controller holds evidence, labels, stages, reopening	M/R planes, incompatible roles, actual-control tests
Weaponized unverifiability	Refusing interfaces creates unknowns, or “unknown” grants permanent immunity	Distinguish controllable deviation from genuine unavailability; audit evidentiary effort
State confusion	SUPPORTED, UNKNOWN, HOLD, ADMIT substitute for each other	Fix four layers and transition powers
Normative capture	Powerful actors, funders, or abstract values monopolize L0/T5	Prefrozen norms, affected-party evidence, dissent, reopening
Human-COS centralization	Real actors reduced to tool inputs	Conventional research first, shadow operation; Human-COS/equivalents cannot hold R-A/L/G/S or form legitimacy states
Safety spillover	Research reveals sensitive weaknesses or expands confrontation	L0-R, minimal disclosure, scoped access; HOLD if unsafe
Silent tuning	Samples, windows, thresholds, proxies change without versioning	Separate exploration/confirmation; parent version, diff, hash
Plane leakage	Research HOLD or labels turn into real actions	No R-plane real stopping powers; separate authorization architecture at R6

XVII Human COS

A future research tool rather than a mechanism or determination center

17.1 Before trials prepare without prewriting gains

The research charter, mechanism cards, historical event database, conventional causal graphs, preregistered hypotheses, role separation, and red-team questions may be prepared. Future actor, T0, evidence, objection, source, version, visibility, and responsibility fields may be defined.

It is not permissible to claim Human-COS has already improved prediction, coordination, cooperation, or policy quality; to reduce real actors, institutions, and professional methods to “input variables” to fit future tools; or to describe the existence of Runtime as proof that a mechanism can already operate.

17.2 Shadow trials first test research increments

Initially, Human-COS operates only alongside conventional research without changing real actions. The comparison is not “did it bring about cooperation?” It is whether it improves actor coverage, T0 discipline, causal-chain completeness, transparency of parameter estimation and updating, provenance, dissent visibility, error correction, computational reproducibility, and update speed, while testing for new bias, centralization, leakage, and responsibility evasion. Probability calibration is discussed only if the system actually outputs probabilities.

Research uses may expand only if the tool itself is auditable, replaceable, and stoppable and does not monopolize direction, evidence, determination, stage gates, or state execution.

17.3 Stopping rollback and closure of shadow operation

Before operation, every Human-COS shadow protocol must map these risks to OBJECTIVE_TRIGGER or G_REVIEW, scoped to a specific tool version, module, data domain, output, or research claim: Human-COS directly or through automated interfaces creates, changes, or executes decisions within R-A/R-L/R-G/R-S/M-D/M-S authority; forms or changes L0-R, L0-M, T5, or T6 rights–authorization components; serves as sole/final R-V/R-X or M-V/M-X carrier, or uses model replicas, agent chains, automated services, or a common provider to impersonate independent review; key inputs, outputs, sources, versions, prompts, model changes, or logs cannot be audited/recomputed, or are overwritten/selectively deleted; unauthorized access, sensitive leakage, use expansion, or unrecoverable permission spread occurs; compared with conventional controls, load-bearing actor omissions, single-perspective bias, suppressed dissent, hidden errors, or responsibility evasion persist; funders, tool providers, or a single controller can use models, data, credentials, IP, or service termination rights to control evidence, determination, legitimacy, stages, or reopening; the tool cannot be stopped, replaced, or rolled back without damaging original records, conventional baselines, or independent research capacity; or expanded use/research-gain claims are proposed without preregistered conventional controls and independent review.

After triggering, R-G may decide only scoped RUN:LIMITED/HOLD/CLOSE for Human-COS-related research modules, data domains, output claims, or uses. Where mechanical conditions hold, R-S directly executes scoped RUN:HOLD. This does not stop real systems or automatically stop conventional research independent of the tool. Post-trigger tool outputs cannot support positive candidate eligibility, but must be retained as audit and negative-result materials. Rollback must return to the latest independently operable, auditable conventional research baseline and latest reviewed tool version, preserving input, output, objection, permission, and error records before and after the trigger.

Reopening requires at least root-cause, impact-scope, and responsibility-chain records; new tool version, hash, data and permission mapping; R-V review and independent R-X challenge; renewed affected L0-R and T6-R judgments; and an R-G written decision with R-S execution under the new version. Human-COS must not participate in final determination of its own safety, effectiveness, or reopening, nor form the L0 or T6 rights–authorization states on which reopening depends. Repeated load-bearing biases, or structural inability to audit, replace, stop, locate responsibility, and operate the conventional baseline independently, should produce RUN:CLOSE for that shadow module. This does not constitute

FAIL for the real mechanism under study or prevent safe conventional research. Future shadow protocols instantiate bias thresholds, comparison metrics, and observation windows.

17.4 Permanent boundaries

- Human-COS cannot automatically generate real payoff structures or turn simulation outputs into equilibrium proofs.
- It cannot make policy decisions, promise performance, create verification credibility, execute sanctions, or bear consequences for real actors.
- It cannot hold R-A, R-L, R-G, or R-S; form or change L0-R, L0-M, T5, or T6 rights–authorization components; or obtain M-plane real stopping or response powers.
- It cannot be the sole or final carrier of R-V/R-X or M-V/M-X, or use tool replicas, agent chains, or common providers to satisfy independent review and actual-control separation.
- It cannot treat multi-model consensus, majority opinion, or public attention as fact.
- It cannot hide sensitive-information governance, major relevant actors’ visibility, or responsibility attribution.
- Theory or Runtime version changes do not automatically change this master. Interface changes require independent assessment and explicit versioning.

These boundaries apply to all functionally equivalent models, agents, automated services, or outsourced platforms, regardless of name, interface, deployment location, or provider. Tool outputs are only sourced, traceable, challengeable inputs. Corresponding states must be formed by real actors able to bear responsibility, pass actual-control audits, and hold relevant powers.

The tool’s place: Human-COS can help real actors see feedback, responsibility, and multi-actor coupling. It cannot create real interests, trust, commitments, execution capacity, or communities for them.

XVIII Current closed items open items and locking sequence

18.1 Master invariants frozen in V1.0

- Relationship states are not a linear value ladder; research maturity is separate from real-world evolution.
- Multidimensional self-interest is only a reachability stress test; it must be preregistered, refutable, and recursively constrain researchers and gatekeepers.
- M-plane real mechanisms and R-plane research governance are strictly separate; research states have no real stopping effect.
- M-E/V/X/D/S decompose only real functions, authority conflicts, and proof obligations, not a five-agency blueprint. Research-governance maturity cannot imply institutional maturity.
- Six tests, L0-R/L0-M, process ledgers, T5, T6-R/T6-M, four state layers, R0–R6 stages, and real authorization jointly govern claim advancement.
- Dual tracks are assigned by harm dimension; ordinary net value cannot offset potentially irreversible, noncompensable, or hard-rights thresholds.
- R-E/V/X/A/L/G/S functions, transitions, incompatible roles, actual control, and error records precede concrete institution selection.
- Independence is judged against the specific challenged object. A combined M/R control matrix blocks dual hats, common upstream control, and cross-plane self-review in each work package.
- WP-01’s H2a/H2b, H3, H5, and strict three-valued conjunction cannot be canceled by a majority of positive hypotheses.
- The master freezes H3 opportunity chains, confirmed-breach FAIL rules, competing explanations, bounded H5 coverage, and new-pathway reopening thresholds. Neither zero events nor limitless imagined pathways create shortcuts to support or veto.
- R0–R6 forbids positive stage-skipping but permits downstream counterexample screening. Computational reproducibility and independent challenge operate across stages from R1.
- Norm-formation packages, actual-control matrices, minimum resource capacity, RUN:CLOSE for prolonged UNKNOWN, and research boundaries after L0M:REJECT now have executable consequences.

- Human-COS and functionally equivalent tools occupy PRE-HUMAN-COS/future shadow-validation positions, not those of mechanisms, empirical/legitimacy determination centers, or real actors. Tool replicas cannot manufacture independence; shadow failures stop, roll back, and close only through the existing research-state chain.
- RUN:REOPEN is a one-time version transition record, not a fifth research permission bypassing PROCEED/LIMITED/HOLD.
- Semantic upgrades, nonsemantic patches, and attachment compatibility records are distinct. Version relationships produce neither evidence support nor real authorization.

These are frozen MASTER/V1.0 clauses. Subsequent work packages, results, communications, and compatibility records must not silently invalidate them through omission, renaming, aggregation, parameterization, tool interfaces, or “domain specificity.” Required changes must first stop affected derivations, explicitly upgrade the master semantically, and reestablish derivation relationships.

18.2 Open items to be instantiated by WP 01 A

- First concrete domain, event inclusion/exclusion criteria, historical samples, comparability rules.
- Concrete candidate holders, eligibility, combined cross-plane M/R actual-control separation, preacceptance, refusal, substitutes, removal, and appeal deadlines for M/R functions. M-plane mapping remains a research object, not real appointment, institutional design, or authorization.
- H1–H5 metrics, sample sizes, observation windows; H3 proxies for opportunity access, recognition, capacity, decision space, delayed exploitation, observability, exposure vectors, and δk ; H5 actor–pathway matrices, generation methods, credibility/materiality thresholds, stopping conditions, residual-pathway aggregation, systemic safety thresholds.
- $H \rightarrow T$ dependency graph, downstream counterexample-screening completeness, and domain-specific rules distinguishing new pathways, environmental changes, and omissions within original scope.
- Data dictionary, downgrade rules for missing fields, evidence interfaces, logs, access control, blinding, code, models, statistical methods, conditional-map implementation.
- Domain-specific normative-package sources, affected-actor participation, absent-actor proxies, representation limits, conflict handling, procedures, deadlines, version-upgrade fields.
- OBJECTIVE_TRIGGER/G_REVIEW classification, HOLD scope, response time limits, and restart materials for every stopping condition.
- Personnel, budgets, deadlines, alternatives, and information-gain margins in resource–capacity statements; domain-specific permitted/prohibited research lists after LOM:REJECT.
- Human-COS shadow-control protocols, increment metrics, bias thresholds, observation windows, concrete rollback baselines. No results may be prewritten before tool trials.
- Worked examples of $\Delta p^*_{i,r,t}$, typical NOT COMPUTABLE settings, and data contracts. These examples confer no new permission.
- WP-01 failure-mode library and applicability boundaries for repeated games, state transitions, causal inference, and agent simulation.

Operational or quantitative thresholds for moving from R2 to R3/R4, and the real sovereign/pilot-starting actors for R6, are later-stage open items. WP-01-A cannot close them out of thin air, nor can non-instantiation count as passage.

The master first freezes functions, boundaries, and incompatibility invariants. The WP-01-A candidate instantiates holders, thresholds, and record forms, and may supply only blank validation fields for future pilot authorization. Actual PILOT-SCOPE AUTHORIZED can be issued only by real actors at R6. These instantiations must close before WP-01-A freezes, not be conjured into completion before attachment drafting starts.

18.3 Master and attachment relationship

Master freezing does not stop research; it fixes definitions, disciplines, and boundaries for subsequent research. Independent attachments may instantiate only explicitly open domains, actors, variables, thresholds, samples, role holders, and record formats. They cannot change the six tests, dual legitimacy gates, dual-track computation, separation of powers, three-valued conjunction, R0–R6, or Human-COS boundaries.

Subsequent attachments, results, and communications must not cite an unqualified “V1.0.” Masters use MASTER/Vx.y; work packages use WP-01-A/Vx.y. The first WP-01-A must bind to the external delivery hash of this MASTER/V1.0

artifact. Every attachment remains permanently bound to its original parent-master version and hash; master upgrades do not rewrite existing attachments. Compatibility records cannot rewrite this binding; they judge only whether an attachment can continue to be used under another master. A separate record includes at least attachment identity and hash, original parent master, target master, affected clauses, H/T/L/permission gates to rerun, independent review, scope, conclusions, and reopening conditions. While impacts remain unclear, attachments serve only as historical materials or methodological inputs, not current compliance evidence, stage-advancement grounds, or derivative parents under the new master. Any change to attachment content, parent version, hash, normative mapping, or conclusions requires versioning.

Compatibility follows Section 1.2's version judgment chain. R-E preserves attachments, old/new masters, diffs, and hashes. R-V verifies impact inventories and fields requiring revalidation. At least one independent R-X challenges semantic, permission, normative, state, and claim impacts. Attachment authors, the original team, or the same actual controller cannot alone form a complete compatibility conclusion. Reasonable disagreement leaves compatibility pending; affected attachments have scoped historical-material use only. R-A reforms affected H/T; R-L reforms L and normative gates; real authorization requires separate revalidation. R-G only decides research-adoption scope; R-S appends compatibility, scope-limitation, revalidation, or version-upgrade records.

18.4 The sole next handover item

Next work: Prepare Crisis Notification—Limited Verification—Conditional Response Mechanism | Preregistered Research Design V0.1 (WP-01-A), instantiating MASTER/V1.0's explicitly open domains, role holders, H1–H5, data, thresholds, norm-formation packages, stopping conditions, and record formats. This step produces only a candidate research protocol. It does not initiate real pilots, national-policy simulation, or Human-COS Runtime, and does not simultaneously derive papers, policy briefs, English drafts, or other attachments.

Until WP-01-A is complete and independently reviewed, do not derive Human-COS interface specifications, academic papers, policy briefs, English drafts, or other related manuscripts. Any real pilot must still await R6 real authorization and independent validation.

Appendix A Mechanism unit card

Minimum master fields

Field group	Required record
Identity and version	Name, master version, work-package version, domain, actors, timescale, R stage, hash, source of changes
Relationship state	Domain-specific competition/cooperation/alliance/bloc/community definitions and transformation direction
Behavior and interests	Preregistered interest dimensions, proxies, competing explanations, real consequences, non-saint condition, future options
Structural change	Information/incentives/risk exposure/commitment verification/feedback correction
Causality and counterfactuals	Action → mediator → outcome; real baseline, feasible alternatives, competing explanations, refutation conditions
General key assumptions	Information, capacity, rationality, discounting, execution, verification, domestic structures, third parties, time windows
Vulnerabilities	First-mover exposure, deviation temptation, misjudgment, leakage, capture, arbitrage, plane leakage
Observability	Variables, proxies, timestamps, sources, errors, H3 opportunity chains, bounded H5 pathway coverage, residual pathways, unknown conditions
L0 and externalities	L0-R, L0-M, actors, transfer pathways, burden character, authorization/repair, norm-formation package, normative conflicts, process ledger, T5
Permissions and determination	M-E/V/X/D/S functional positions, R-E/V/X/A/L/G/S, separate-plane and combined M/R actual-control matrices, preacceptance, error records, four outputs

Field group	Required record
Hypotheses and evidence	Work-package load-bearing H, support/failure/unknown, strongest counterexample, strict conjunction, independent review
Dual tracks and thresholds	Reversible-compensable and irreversible/noncompensable dimensions, lower loss bounds, intervals, thresholds, reasons not to monetize
Stopping and reopening	Signals, trigger types, determination, execution, capacity gaps, HOLD scopes, reasons for closing UNKNOWN research, Human-COS rollback baselines, repair evidence, review, versioning, dissent
Conclusion states	Standard-output order, four basic state layers, T1/T2/T3/T4/T5/T6-R/T6-M, R stage, real authorization; precise use of direction/mechanism hypothesis/evidence finding/conditional support/research candidate mechanism

Appendix B Upstream lineage public interfaces and historical prototypes

Manuscript or interface	Current identity	What it provides	Effects it cannot produce
When Intelligence Is No Longer Scarce V0.6	THEORY LINEAGE	Overall direction, conceptual sources, civilizational boundaries	Does not prove a specific mechanism works
Chinese public long-form standalone edition V1.1.1 Chapter 9	DIRECT PUBLIC MECHANISM DISCIPLINE; the sole current public directional interface	Structure before personality; distribution and capture; real mechanisms cannot be performed by theory/tools; auditable/replaceable/stoppable tools	Does not replace V0.6 or constitute empirical validation; V1.1 records only first emergence and has no independent current interface authority
When Intelligence Is No Longer Scarce V0.4 External Evidence Reinforcement and Directional Mechanism Bridge Control Matrix	EVIDENCE CONTROL RECORD; preparation and evidence-control record	Treatment, exclusion, and evidence-source records	Not theoretical baseline, validity evidence, or authorization
Computable Research Design for China–US Competition and Cooperation v1.0	EARLY BILATERAL PROTOTYPE	Question sources, risk–benefit thresholds, some mechanism components	Not the master, equilibrium proof, or deployment basis
Assessment report on that v1.0	REVIEW INPUT	Exposes reachability, counterfactual, dual-track, verification, and third-party gaps	Not independent empirical validation
Human-COS	FUTURE TOOL INTERFACE	Records, actor coverage, causal reasoning, dissent, reproducibility, shadow comparison	Cannot hold R-A/L/G/S, form L0/T5/T6 rights–authorization components, be sole/final R-V/X or M-V/X carrier, or acquire real M-plane powers

These materials are not one linear higher/lower version chain and do not automatically update this master. Future versions only trigger impact assessment. Changes to this project’s definitions, stage gates, claim levels, or permission boundaries require explicit master versioning.

Appendix C Methodological lineage and borrowing boundaries

This project's governance architecture and strict state rules are project designs, not existing internationally accepted standards or externally validated best architectures. The following sources supply methodological references only; they do not prove WP-01 or this master's mechanism claims:

- HM Treasury, [The Green Book \(2026\)](#), provides a current policy-appraisal reference for sensitivity, switching values, distribution, and nonmonetizable effects. US OMB [Circular A-4](#) (historical official text dated September 17, 2003) supplies historical methodological references for threshold/break-even analysis, unquantified effects, and distribution. Δp^* is this project's application of break-even/switching-value logic, not a new mathematical method, and inherits none of those documents' policy-authorization force.
- Holland, [Statistics and Causal Inference \(1986\)](#), supplies a classic formulation of counterfactual/potential-outcomes causal inference. Selection bias, common shocks, and sample comparability are historical-comparison identification requirements retained by this project on that basis, not an exhaustive list claimed to come from that paper.
- Shapley, [Stochastic Games \(1953\)](#), provides the classic original formulation of stochastic games. Axelrod, [Effective Choice in the Prisoner's Dilemma \(1980\)](#), reports the original iterated-prisoner's-dilemma computer tournament. Neither model nor tournament results directly prove equilibria reachable by real actors.
- Nosek et al., [The Preregistration Revolution \(2018\)](#), supplies preregistration distinctions between prediction/postdiction and confirmatory/exploratory research, explaining that preregistration does not eliminate poor statistical practices. "Preregistration cannot remedy wrong variables, wrong causal graphs, or invalid data" is this project's boundary based on that work, not a claim that the paper has validated each item individually.
- The US FDA's final guidance [Non-Inferiority Clinical Trials to Establish Effectiveness](#) (November 2016) supplies a clinical-method reference for prespecifying noninferiority margins and comparing confidence-interval upper bounds with them. H3 borrows only this logic. Clinical margins, premises, standards of proof, and regulatory conclusions must not be directly transplanted.
- Altman and Bland, [Absence of evidence is not evidence of absence \(1995\)](#), supports only the general evidentiary discipline that failure to observe a significant difference does not prove safety or no difference. It does not independently supply H3's positive safety criteria.
- The US National Academies of Sciences, Engineering, and Medicine (NASEM, 2019), [Reproducibility and Replicability in Science](#), defines computational reproducibility using the same inputs and computational steps, and replication with independently obtained new data. "Independent reanalysis" is an operational category this project explicitly adds between them, not a third standardized category claimed to be defined by that report.
- [NIST FIPS 180-4](#) (August 2015) specifies secure hash algorithms, including SHA-256, and message digests. In this project a SHA-256 digest detects file-byte changes relative to a known baseline only. It does not authenticate author, source, date, content correctness, or real-world authority.

If historical communication, notification, verification, or consultation cases later enter attachments, they are only evidence that "an arrangement existed or operated." They cannot alone prove lower overall war probability, reachable or maintainable cooperation, legitimacy, or deployability.

Appendix D Single veto final checks

Any one of the following prevents advancement of the current claim, version, or stage:

1. Relationship states are written as a one-way civilizational ladder, or R0–R6 impersonates inevitable real evolution.
2. Multidimensional self-interest absorbs every counterexample by adding dimensions afterward and loses falsifiability.
3. M/R planes are mixed so research labels, HOLD, or tool outputs are intended to trigger real action.
4. The same actual controller spans evidence entry, verification, challenge, posterior labels, legitimacy, decisions, execution, stage states, or reopening within one plane or across M/R; M-plane actors control R-plane evaluation of their own records, decisions, or mechanisms; or old L0/T5 labels retain current validity after formation eligibility has expired.
5. R-V outputs directly generate judgments of malice, betrayal, responsibility, or real responses.
6. Positive H1/H2/H4 evidence turns H3/H5 or other load-bearing UNKNOWN into support.
7. H5 replaces L0/T5 normative judgments, or L0/T5 counts as empirical H5 support.

8. Bilateral or aggregate gains override third-party noncompensable rights, potentially irreversible risks, or uncounted burdens.
9. L0M:REJECT deletes safe, noninterventional negative-result research on dangerous mechanisms, or conversely permits continued optimization, intervention, pilot/policy materials without a new version and L0-R/L0-M.
10. Ordinary net value offsets potentially irreversible, noncompensable, or hard-rights thresholds across tracks.
11. Parameters, windows, samples, proxies, or thresholds change without versioning.
12. R2 invokes no stage-skipping to refuse third-party counterexample screening, or invokes completed screening to assert R3/R4 positive conclusions prematurely.
13. Original conditional-support labels are retained despite inadequate computational reproducibility or provenance boundaries, or without independent challenge relative to the actual challenged object and its controllers.
14. Human-COS or equivalent tools receive R-A/L/G/S, L0/T5/T6 rights–authorization states, real verification credibility, policy decision, stopping, or execution powers; serve as sole/final R-V/X or M-V/X carriers; or impersonate independent review through tool replicas, agent chains, or common providers.
15. T6-R research-process correctability replaces T6-M real-mechanism authorization, exit, repair, appeal, and reauthorization.
16. M-V/M-X narrow facts bypass M-D authorized decisions or directly command M-S execution.
17. M-E/M-V/M-X/M-D/M-S functions are interpreted directly as a real institutional scheme, or R-plane maturity proves their reachability, legitimacy, credibility, or real authorization.
18. H3 substitutes “no exploitation occurred” for load-bearing evidence of opportunity access, recognition, capacity, decision space, delayed windows, and observability; a confirmed key-dimension breach of δk under preregistered rules does not produce FAIL; or T2 receives conditional support while H3 is UNKNOWN.
19. H5 describes bounded frozen-scope coverage as exhaustive third-party pathways; ignores new pathways meeting credibility, materiality, and boundary-relevance thresholds; or grants unlimited stopping power to below-threshold speculation.
20. L0M:ADMIT or T5 PASS is formed without an auditable norm-formation package, or while normative conflicts capable of changing admission/legitimacy conclusions remain unresolved.
21. Confirmatory research or positive candidate eligibility continues despite unmet minimum governance capacity or actual-control matrices, through expanded scope, outsourcing, automation, formal separation, or “complete it later.”
22. RUN:CLOSE rewrites load-bearing UNKNOWN as safety, nonexistence, failure, default opposition, or default permission.
23. After Human-COS/equivalent tools trigger load-bearing bias, unauditability, leakage, centralization, responsibility evasion, or inability-to-stop conditions, their uses still expand, records are deleted, or they self-determine or self-reopen.
24. Semantic changes are labeled patches, or attachments are adopted as current materials under a changed master without compatibility records and necessary revalidation.
25. PILOT-SCOPE AUTHORIZED is reported without minimum semantics for authorizing actor, authority chain, scope, duration, stopping/rollback, review/appeal, and responsibility records; or pilot authorization implies effectiveness, legitimacy, scalability, T6-M passage, or operational authorization.
26. OPERATIONAL AUTHORIZED is reported without verifiable routine-operation authorization, continuing supervision, periodic review, suspension/revocation, responsibility/repair, and reauthorization after load-bearing changes; or it exempts H/T/L review.
27. Editors, the original team, or the same actual controller alone declare semantic patch eligibility or complete attachment compatibility without independent R-X review, or adopt a positive conclusion despite reasonable disagreement.
28. Load-bearing blocks, FAIL, and UNKNOWN are not consolidated before positive H/T/L conclusions; or the same scope receives RUN:PROCEED under L0R-REDESIGN/LIMIT/FAIL, L0M:REDESIGN/LIMITED/UNKNOWN/REJECT, or T5 load-bearing blocks.

Appendix E V0.1 to V1.0 version migration and hashes

Version	Status	Main function	Current effect
V0.1 Project baseline	Historical frozen artifact	Established six tests, dual tracks, R0–R6, WP-01, evidence and risk boundaries	No longer a current clause repository
V0.2 Load-bearing discipline revision	Historical frozen artifact	Added recursive self-interest, dual legitimacy gates, permission separation, H3/H5, strict unknowns	No longer independently derives attachments
MASTER/V0.3 Fully integrated candidate	Historical candidate	Restated all current rules in one document and closed structural loops	No longer a current clause repository or attachment parent
MASTER/V0.4 Locking candidate	Historical candidate	Closed institutional-blueprint ambiguity, opportunity/pathway coverage, norms/resources, prolonged unknowns, tool failure, version interfaces	No longer a current clause repository or attachment parent
MASTER/V0.5 Final-review revision candidate	Historical candidate/direct semantic predecessor	Closed REOPEN, H3 thresholds, tool powers, object-relative independence, cross-plane control, and appendix/source differences	No longer a current clause repository or attachment parent
MASTER/V1.0 FROZEN	Current formally locked master	Completed continuous final review and established the first derivable research-master identity	Sole current derivation root for WP-01-A and subsequent attachments

SHA-256 of preceding deliverables: These hashes cover the complete bytes of the formally delivered DOCX files listed below. Rendered PDFs, extracted text, and later copies do not share these hashes. Original Chinese filenames are retained as exact archival identifiers.

- 竞争合作联盟阵营共同体_可计算机制研究_V0.1_立项基线.docx
7867da79cf0e0cb805b48f305d695814cceeaa327ab86d917e41890f1d34f3d
- 竞争合作联盟阵营共同体_可计算机制研究_V0.2_承重纪律修订版.docx
e6ac854ff1989271700b518f1a43c0d561c19398725222f4964f16502db7ac17
- 竞争合作联盟阵营共同体_可计算机制研究_MASTER_V0.3_完整整合候选稿.docx
a9d0abdf8e7b34f7f81389c0664de7bc45e8fcf24ce12c327d2a3e2653074926
- 竞争合作联盟阵营共同体_可计算机制研究_MASTER_V0.4_锁定候选稿.docx
3b2750badbc0753ca21292d58756181f58e23e4773b4117398efc8291df77f9c
- 竞争合作联盟阵营共同体_可计算机制研究_MASTER_V0.5_终审修订候选稿.docx
4f6d119969ce961024921c7e31392668014a1725ec0c0ca626e3dd8061871c04

Semantic change ledger for V0.3 relative to predecessors:

ID	Change	Accountability implication
V03-01	Added operational definitions of five relationship states	Blocks linear civilizational-ladder interpretations
V03-02	Separated real M-plane mechanisms and R-plane research governance	Blocks direct conversion of research labels into real action
V03-03	Made multidimensional self-interest an explicitly refutable stress test	Blocks retrospective absorption of every counterexample
V03-04	Froze four state layers, research stages, real authorization labels	Prevents states from impersonating each other

ID	Change	Accountability implication
V03-05	Split L0 into L0-R and L0-M	Separates research-action safety from mechanism legitimacy
V03-06	Calibrated old P^* to $\Delta p^*_{i,r,t}$	Specifies actors, risks, time, and noncomputability boundaries
V03-07	Added downstream counterexample-screening rules	Prevents stage gates from hiding third-party counterexamples
V03-08	Split H2 into H2a identification performance and H2b incentive effects	Prevents classification improvements from directly implying behavioral change
V03-09	Separated empirical H5 and normative L0-M/T5 judgments	Prevents facts and legitimacy substituting for one another
V03-10	Distinguished computational reproducibility, independent reanalysis, replication	Aligns evidence terminology and stage responsibilities
V03-11	Split T6 into T6-R/T6-M and added M-D/M-S	Prevents research correction from replacing authorized real correction

Semantic change ledger for V0.4 relative to V0.3:

ID	Change	Accountability implication
V04-01	Clarified research-charter identity and that M functions are not institutional blueprints	Blocks transposition of research governance or functional decomposition into real institutional design
V04-02	Froze failure-first standard-output order and minimum pilot/operational authorization semantics	Blocks omitted states, overall traffic lights, empty authorization shells
V04-03	Clarified negative-result research boundaries after L0M:REJECT	Blocks both erasure of negative results and continued optimization of dangerous designs under their cover
V04-04	Added norm-formation packages and load-bearing normative-conflict rules	Blocks unilateral R-L norm selection and majority flattening of hard boundaries
V04-05	Added actual-control matrices and minimum resource-capacity gates	Blocks nominal division, outsourcing, or automation impersonating real governance capacity
V04-06	Completed H3 opportunity chains, competing explanations, $H \rightarrow T$ propagation	Blocks zero events impersonating safety and T2 upgrades under H3 UNKNOWN
V04-07	Defined H5 coverage as a bounded procedure; added three new-pathway gates	Blocks unlimited-path vetoes and arbitrary dismissal of counterevidence after preregistration
V04-08	Tightened objective triggers and froze PROCEED/LIMITED/HOLD/CLOSE selection rules	Blocks hidden interpretation in fields and discretionary operational-state choice
V04-09	Established RUN:CLOSE and reopening discipline for prolonged UNKNOWN	Allows ending infinite investment without erasing unknowns
V04-10	Established Human-COS shadow stopping, rollback, closure chain	Blocks tool-use expansion or self-reopening after failure
V04-11	Separated semantic upgrades, nonsemantic patches, attachment compatibility; included independent review	Blocks editors declaring their own patches or silent attachment inheritance
V04-12	Added Appendix F nonnormative quick-reference layer	Eases external handover without creating a second rule source

Semantic and self-containment change ledger for V0.5 relative to V0.4:

ID	Change	Accountability implication
V05-01	Limited RUN:REOPEN to a one-time version transition with default HOLD	Blocks reopening from bypassing PROCEED/LIMITED/HOLD conditions
V05-02	Added H3 FAIL for confirmed δk breaches	Blocks downgrading proven threshold breaches to UNKNOWN
V05-03	Prohibited Human-COS/equivalents from R-L, legitimacy states, or independence through tool replicas	Blocks indirect stage veto or real credibility through L0/T5
V05-04	Bound R-X/M-X and reanalysis/replication independence to the specific challenged object	Blocks absence of conflict elsewhere impersonating independent review here
V05-05	Added combined M/R control matrices and cross-plane recusal consequences	Blocks dual hats, common upstream control, and formal firewalls hiding self-evaluation
V05-06	Calibrated Appendix C institutional attribution, term sources, project-defined boundaries	Blocks presenting project disciplines as conclusions proven by external sources
V05-07	Aligned Appendix A/B/E/F claim levels, authorization labels, role summaries, version chains	Blocks quick-reference/migration tables creating a second formulation

Locking change ledger for V1.0 relative to V0.5:

ID	Change	Accountability implication
V10-01	Completed clause-by-clause continuous review, logical counterexamples, power–responsibility, norm–resource, source–claim, main-text/appendix, and artifact regression audits; cleared P0/P1	Freezing eligibility comes from reviewable audit, not version names or author declarations
V10-02	Changed identity to MASTER/V1.0 FROZEN, fixing the first derivable master	Subsequent attachments bind this version and external delivery hash, not historical candidates
V10-03	No change to V0.5 mechanism semantics except locking identity, lineage, derivation root, handover interfaces	Prevents smuggling unaudited rules under “formal locking”

A hash proves only consistency of current file bytes with the delivery record, not author, date, theoretical correctness, or real authority. The MASTER/V1.0 finished-artifact hash must not be written back into the hashed file. Record it in a separate version manifest, delivery record, or same-name .sha256 file. The hashed master records only predecessor hashes and external-manifest rules, avoiding self-referential changes.

Appendix F Framework overview and standard output quick reference

Authority of this quick reference: It only helps external teams locate the main text. It does not create, relax, replace, or consolidate main-text rules. Every shorthand must be interpreted through the corresponding full clause. Inconsistency is resolved by the main text’s complete qualifications. This appendix must not be extracted alone as an institutional blueprint, authorization document, or automated decision-maker.

F.1 One minute orientation

Separate planes first then states

First question	Correct location	Common misreading
Who may decide and execute in reality?	M plane; real authorization, capacity, and responsibility require separate proof	Research labels/tool outputs replace real decisions

First question	Correct location	Common misreading
Who preserves, verifies, challenges, determines, and gates research?	R plane; governed by actual control and incompatible roles	Mature research governance directly becomes real institutions
What does evidence currently show?	Narrow facts, H, T, L0/T5 reported separately	One aggregate score or majority cancels UNKNOWN
What may research do now?	Current permissions: PROCEED/LIMITED/HOLD/CLOSE; REOPEN is one transition record requiring a simultaneous target state	Research operation becomes a real stopping order
Has research verified real authorization meeting minimum semantics?	NO M-AUTHORIZATION / PILOT-SCOPE AUTHORIZED / OPERATIONAL AUTHORIZED	Research candidacy or pilot authorization becomes operational eligibility
Under which version is the document valid?	Master, attachment, parent version, hash, compatibility record	Silent inheritance, rewriting parent versions, or semantic change called a patch

F.2 Functional positions on both planes

Similar names different powers

Position	Produces only	Absolutely cannot thereby produce
R-E	Research materials, sources, versions, hashes, access records	H/T/L judgments, stage decisions, real action
R-V	Four evidence outputs for preregistered narrow propositions	Motive, responsibility, betrayal, real responses
R-X	Independent challenges, omissions, competing explanations	Unlimited noisy vetoes, self-decided stopping/reopening
R-A	Challengeable, revocable H, T1–T4, T6-R/T6-M technical-operability reference labels	T5, real fact center, stage control, authorization
R-L	L0-R, L0-M, process ledger, T5, respective T6-R/T6-M rights–authorization parts	Empirical substitution, arbitrary abstract-value veto, stage execution
R-G	Research operational scope from established states	Evidence production, threshold change, state rewriting
R-S	Mechanical freezing, PROCEED, LIMITED, HOLD, REOPEN, CLOSE, and records	Condition interpretation, negative-result deletion, standalone restart
M-E	Real operational evidence preservation	Fact verification, real decisions, action execution
M-V	Operational narrow-fact outputs	Direct orders of punishment, response, or stopping
M-X	Cross-party review and conflict exposure	Replacement of authorized decisions or automatic execution through objections
M-D	Real-authorized scoped decisions with consequence-bearing	Fabricated facts, unauthorized execution, delegation to R/tools
M-S	Authorized execution and responsibility records	Rewritten decisions, expanded scope, self-authorization

These are functions, conflict checkpoints, and proof obligations, not twelve agencies. Concrete organizations may use different names, numbers, and topologies, but cannot dissolve frozen main-text boundaries on incompatible roles, review, appeal, exit, and responsibility.

Human-COS and every functionally equivalent tool supply only sourced, traceable, challengeable inputs. They cannot hold R-A/R-L/R-G/R-S, form L0/T5/T6 rights—authorization states, be sole/final carriers of R-V/R-X or M-V/M-X, or satisfy independence through tool replicas.

F.3 Standard output record

Every field must appear

1. Claim ID, actors, domain, time scope, exact claim level.
 2. Document version, parent version, hash, source of differences.
 3. Current R0–R6 research stage and stage ceiling.
 4. M authorization status and authorization-document verification result.
 5. Each narrow fact’s SUPPORTED/CONTRADICTED/INSUFFICIENT/NOT VERIFIABLE.
 6. Consolidated load-bearing blocks, FAIL, and UNKNOWN: H FAIL/UNKNOWN; T1–T4, T6-R, T6-M FAIL/UNKNOWN; L0R-REDESIGN/LIMIT/FAIL; L0M:REDESIGN/LIMITED/UNKNOWN/REJECT; T5 FAIL/UNKNOWN. This precedes every positive H/T/L conclusion. Earlier narrow-fact/authorization verification must not impersonate positive mechanism conclusions.
 7. Each load-bearing H’s FAIL/UNKNOWN/CONDITIONAL SUPPORT.
 8. Three-valued T1–T4, T6-R, and T6-M states.
 9. Current L0-R, L0-M, T5 states and normative-package version. If formation eligibility has expired, separately list original labels, invalidation reasons, aggregation treatment, and pending independent R-L re-formation.
 10. Current RUN:PROCEED/LIMITED/HOLD/CLOSE and scoped objects. For reopening, separately list the REOPEN transition, new version, and simultaneous target state. REOPEN cannot stand alone as current permission.
 11. Strongest counterevidence, residual objections, residual pathways, competing explanations.
 12. State-forming actors, evidence windows, challenges, independent review.
 13. Repair materials, reopening triggers, next review point, prohibited extrapolation scope.
- Inapplicability must read N/A with a reason. Required but unperformed checks can only be UNKNOWN. No missing field may be interpreted as passage.

F.4 State families each answer one question

State family	Question answered	Cannot impersonate
Four narrow-fact outputs	How does current evidence relate to a preregistered narrow proposition?	Subjective motives, responsibility, H, real responses
Three-valued H	Does a work-package load-bearing hypothesis fail, remain unknown, or receive conditional support?	Six tests, legitimacy, real authorization
T1–T4/T6-R/T6-M	Are relative improvement, reachability, maintenance, verification, and correction supported?	T5 or overall deployment eligibility
L0-R	Is the research action safe, implementable, or necessarily scoped?	Mechanism effectiveness or real stopping power
L0-M/T5	Candidate-mechanism admission and stage-specific legitimacy	Empirical causality or permanent legitimacy
RUN	How far may current research operation, claims, and version derivation proceed?	Real-system states or epistemic conclusions

State family	Question answered	Cannot impersonate
R0–R6	Research maturity and permissible evidence activities	Civilizational evolution sequence or real-mechanism maturity
M authorization	Has research verified authorization of real actors in a specified scope meeting minimum semantics?	Mechanism effectiveness, T6-M passage, permanent permission
Version compatibility	May an attachment continue to be used under another master?	New evidence, research-operation, or authorization states

F.5 Actual control matrix

Check at least eight lever categories

Control lever	Required record
Funding	Who allocates, suspends, imposes conditions, audits, substitutes?
Appointment and removal	Who appoints, evaluates, renews, removes, affects careers?
Data	Who generates, selects, admits, accesses, deletes, blocks?
Models and IP	Who controls code, parameters, computing environments, licenses, recomputation conditions?
Permissions and credentials	Who issues, revokes, holds, views logs?
Personnel and technical upstream dependencies	Common parent institutions, outsourcing, cloud services, platforms, key dependencies
Judgment and publication	Who changes hypotheses, thresholds, normative packages, success definitions, publication scope?
Stopping and reopening	Who triggers, decides, executes, appeals, reviews, repairs, restores?

For every item, trace ultimate controllers and record powers to initiate, prevent, change, delay, or revoke, along with independent alternative channels. Different institution names, more signatures, or outsourcing to the same dependency network do not prove separation of powers.

Audit M/R planes together for the same work package, event, and scope. Independence must hold relative to the object challenged this time and its actual controllers, not be replaced by another conflict-free relationship.

F.6 Shortest correct stopping review and reopening chain

1. Any actor reports a signal; the signal itself changes no state.
2. R-E preserves materials; R-V forms narrow facts; R-X independently challenges.
3. Only a unique mechanical result may be OBJECTIVE_TRIGGER. Interpretation, attribution, value, or proportionality balancing enters G_REVIEW.
4. R-S mechanically executes only scoped HOLD for objective hard triggers. After recusal, R-G decides in writing among four current operational states, or decides a one-time REOPEN transition with a simultaneous target state.
5. PROCEED is available only within the same scope with L0R-PASS, L0M:ADMIT or documented inapplicability, no T5 load-bearing block, and no other load-bearing FAIL/UNKNOWN. During HOLD, only L0-R-permitted minimum evidence preservation, counterevidence, repair, and unknown-closing activities proceed. Under L0R-FAIL, retain only minimum evidence/accountability records that do not perpetuate risk. LIMITED may additionally permit L0-R-approved scoping, historical, noninterventional, negative-result, independent reanalysis/replication, and new-version redesign preparation, without upgrading the original version's positive claims.
6. Repair materials pass through R-E, R-V, R-X again; R-A/R-L reform affected states.

7. R-G decides whether to reopen. R-S registers a one-time REOPEN transition under a new version while R-G simultaneously assigns PROCEED/LIMITED/HOLD. Without simultaneous assignment or repassing affected gates, default to HOLD. REOPEN permits no research actions and restores no old positive labels.

8. When evidence lines are exhausted, RUN:CLOSE is possible. FAIL/UNKNOWN and dissent remain permanently.

Never permit these shortcuts: R states directly into M actions; M narrow facts directly into punishment; majority H results canceling load-bearing UNKNOWN; T6-R replacing T6-M; L0/T5 replacing empirical H5; Human-COS/equivalents forming R-L states, generating/executing/reopening their own powers, or manufacturing independence through replicas; pilot authorization automatically becoming operational authorization.

Personalized further reading

P05-COOPMECH-001 | MASTER V1.0 FROZEN | Formal research asset

This library-navigation appendix does not change the frozen research master's main text.

Toward upstream civilizational judgment

The Only Path: Why Everyone May Become Less Safe When Every Country Is Protecting Itself | Public thought edition

Why read: Continue asking why old logics of competition, security, and self-preservation may jointly produce greater insecurity. This research addresses only when mechanisms are reachable, fail, or remain unknown.

<https://www.civitas.top/en/documents/p03-only-002/>

Toward responsibility and real world judgment

Seeing Earlier What Would Otherwise Be Seen Only Later: A Public Introduction to Human-COS

Why read: Return mechanism claims to consequences, responsibility, critical dissent, and real actors with powers and duties. Human-COS thereby acquires neither policy nor stopping powers.

<https://www.civitas.top/en/documents/p05-humancos-004/>

Toward irreversible boundaries

Public Window Protocol: An Open Future and Three Constraints

Why read: When mechanism research enters high-consequence domains, continue checking minimum conditions for rollback, exit, nonmanipulation, and the future's ability to choose again.

<https://www.civitas.top/en/documents/p05-window-005/>

Research boundaries

This is a research master, not the theoretical Canonical, policy authorization, a real institutional blueprint, or an automated decision-maker. MASTER/V1.0 is in PRE-HUMAN-COS: research design may proceed, but real mechanisms, tool effectiveness, pilot eligibility, and real authorization must not be written as established. Failures, UNKNOWN, counterexamples, and single-veto results must remain permanently.

Subsequent research: P05-FAIRNEG-001 Multi-Actor Fair Negotiation and Interest Balancing remains at research-baseline stage and has no stable public link yet.

Complete public library

<https://www.civitas.top/en/library/>

Civilization Leap official website

<https://www.civitas.top/en/>