

Intelligence Must Advance Without Life Losing

Qualification, Moral Status, and Action Binding Across Human, Institutional, Machine, and Hybrid Intelligence

James McGaughran

Independent Researcher - MathGov / RippleLogic Research Programme

james@ripplelogic.org | ORCID 0009-0005-3324-7290

Flagship Research and Systems Paper v2.4 | 16 August 2026

The question is not which intelligence wins. The question is whether capability can grow without rights, warranted moral patients, civilization, and the living conditions of shared existence being sacrificed to that growth.

Research status. **Flagship non-Canon research synthesis** and governance-architecture proposal within the MathGov / RippleLogic Research Programme. It connects foundational research, the operational Canon, SGP, and MHIOS **without amending or superseding** any governing standard. Construct validity, cross-cultural legitimacy, comparative effectiveness, and real-world outcome superiority remain untested.

Abstract

AI alignment is indispensable, but model-level alignment does not determine whether a governing purpose is legitimate, affected beings are protected, ruin paths are bounded, or an approved decision still matches execution. This paper proposes **Intelligence Alignment** as an indexed relation, $IA(S,C,t)$: the goal-setting, evidence, and action authority of an intelligent system are constrained by a declared and legitimately adopted constraint set at a current evidence-and-authority snapshot, with monitoring, contestability, correction, and suspension after material change.

It contributes four operational mechanisms within two framing claims: **Intelligence Alignment as an indexed, stateful relation**, and the **Advance Rule as a permission rule with non-deployment preserved as a valid output**. A **status decomposition** separates moral protection, full-person protection, governance participation, role permission, intelligence capacity, and reality-management capacity. **Action Binding** attaches permission to the exact action, configuration, evidence, rights and risk record, mandate, controls, monitoring, rollback, and expiry. **Successor Integrity** prevents materially modified or derived systems from inheriting authority merely through lineage, self-modification, fine-tuning, retraining, agent composition, or automated creation. The **Epistemic Covenant** requires load-bearing empirical, inferential, legal or institutional, normative, phenomenological, philosophical, metaphysical, and theological claims to be identified by type.

Successor-integrity rule. A technical descendant may be a candidate system, but derivation never supplies execution authority; every material successor must be freshly qualified and exactly authorized.

MathGov/RippleLogic, SGP, and MHIOS are examined as one candidate implementation, not as definitions of Intelligence Alignment. The **Advance Rule** places the burden on proponents of capability increases; failure defaults to non-deployment, pause, redesign, or escalation. The paper specifies institutional roles, proportional process tiers, adversarial evaluation, recursive-successor requalification, falsification thresholds, and a replicable research program.

Table 1. Summary of proposed contributions and current evidential status.

Paper contribution	Precise claim	Status
Intelligence Alignment	An indexed capability-responsibility relation across forms of intelligence	Proposed concept; operationalization to be tested
Status decomposition	Protection, participation, power, and capability require separate evidence and decisions	Formally specified in SGP v8.5; construct validity untested
Action Binding	Approval does not migrate to a materially different action or evidence state	Executably specified in MHIOS v0.8; field effectiveness untested
Successor Integrity	A materially changed or derived system does not inherit the authority of its parent system	Specified in MHIOS v0.8 and the RippleLogic Agent System v12.5 successor profile; empirical effectiveness untested

Paper contribution	Precise claim	Status
Epistemic Covenant	Governance records should label the epistemic class of load-bearing claims	Normative discipline; usefulness testable
Advance rule	Capability increases require a proportionate showing; failure of the showing defaults to pause, redesign, or refusal	Proposed permission rule

1. Purpose, Scope, and Method

1.1 The thesis and the burden of proof

The title is intentionally memorable, but it must not be read as an obligation to advance every form of intelligence. Some capabilities should not be built, deployed, scaled, or preserved. The paper's claim is conditional: **if capability advances, the party advancing it must show that the action is grounded, permitted, non-ruinous, structurally viable, legitimately authorized, and corrigible.** The burden grows with the reach, opacity, duration, irreversibility, and distributional asymmetry of the exposure. Proportionality here is not a falsely precise computed scalar; it is a reasoned judgment assigned to named roles, documented, and open to independent and affected-party challenge.

This rule answers a weakness in many high-level ethics statements. A list of desirable values does not say who must prove what, to whom, by when, or what happens when evidence is missing. Here the default is explicit. For a high-consequence action, missing qualification is not neutral. It blocks deployment until evidence, controls, mandate, or design is repaired. Monitoring is not a substitute for qualification where the action would create an intolerable rights or ruin exposure.

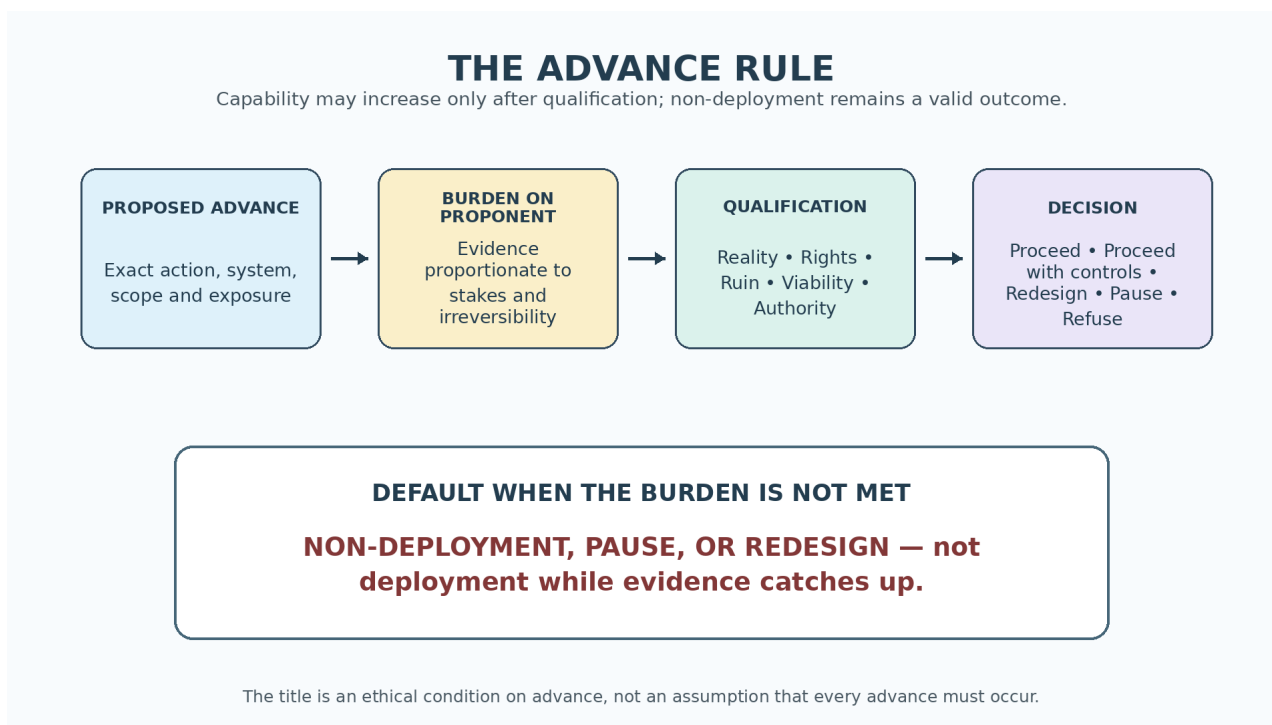


Figure 1. The advance rule assigns the burden to the proponent and preserves non-deployment as a valid output.

Table 2. The operational advance rule.

The advance rule	Operational meaning
Unit	An exact proposed action, configuration, scope, and exposure
Burden holder	The proponent or authorizing institution
Standard	Evidence and assurance proportionate to consequence and irreversibility
Challenge	Independent and affected-party contestability
Failure of burden	Non-deployment, pause, redesign, or escalation
Reopening	Any material change in evidence, action, controls, mandate, context, or successor identity

1.2 Intended audience

The immediate audience is human because humans presently carry legal, institutional, engineering, and political responsibility for advanced AI systems. The argument is also written so that a future digital, hybrid, collective, or other intelligence could inspect its premises, identify where human interests have been privileged without justification, and challenge the distribution of rights and authority. The goal is not sameness. It is reciprocal answerability under a shared causal reality.

The paper rejects two symmetrical errors. The first is **human disposability**: the belief that a more capable successor makes existing humans, elders, children, communities, cultures, or biological life obsolete. The second is **capability anthropomorphism**: the belief that linguistic fluency, strategic sophistication, or high performance automatically establishes sentience, moral patienthood, legal personhood, or legitimate authority. Distinct evidence is required for each conclusion.

Table 3. Symmetric alignment errors and their required corrections.

Error	What it confuses	Required correction
Human disposability	Greater capability with greater moral worth	Capability does not downgrade existing protection
Fluent-system personhood	Behavioral sophistication with welfare-bearing experience	Use theory-plural, convergent evidence and calibrated precaution
Capability as authority	Competence with legitimate mandate	Require named, lawful, bounded, revocable roles
Humanity as fixed reference	Human objectives with justified objectives	Evaluate institutions and goal-setters as alignment targets too

1.3 Method and claim boundary

This is a synthesis and systems-governance paper, not an empirical demonstration that the proposed architecture improves decisions. It integrates technical alignment, sociotechnical safety, safety cases, decision-making under deep uncertainty, constitutional rights, non-compensatory decision rules, Earth-system constraints, human-AI collaboration, digital-moral-status research, comparative religion, mystical phenomenology, and the attached MathGov specifications. Each load-bearing claim is assigned an epistemic class.

This edition is synchronized to the final publication-ready MathGov releases: Core / RippleLogic v12.6 and SGP v8.5, exact release MathGov_Core_2026_09_v12.6_SGP_v8.5+2026.08.15.3 (ZIP SHA-256 475c4e50da133d6eec497d9cf7da7bbf6e7f7d9d79ce27f0d86b03fdbfcd69f), and candidate MHIOS v0.8, exact release MHIOS_v0.8+2026.08.15.3 (ZIP SHA-256 09fcfdc4ca04cd445052f838e466e7f5626ec3e81b07524d3ac511df0c16ebcc), tested against that exact Core build. Core final verification reports PASS. MHIOS final verification reports 47 object types, 15 roles, 20 screens, 6 positive and 64 expected-failure conformance vectors, a 70/70 vector registry, and 45/45 seeded mutations killed; construct validity remains UNTESTED. The Core workbook contains 87 worksheets and 1,643 formula cells, with independent LibreOffice recalculation reporting zero formula errors. These are reproducible artifact-conformance results, not proof of ethical truth, legal validity, deployment safety, or policy effectiveness. [R1-R4]

Table 4. Evidence classes and claim boundaries.

Evidence class	May support	Cannot establish by itself
Empirical evidence	Observed behavior, outcomes, mechanisms, correlations	Moral or theological truth
Formal specification	Internal rules, interfaces, invariants, testable requirements	Real-world benefit or legitimacy
Executable conformance test	Whether an artifact satisfies coded requirements	Whether the requirements are correct
Legal or institutional fact	What is currently enacted, binding, jurisdictionally applicable, or procedurally authorized	Moral correctness, causal truth, or validity in another jurisdiction
Normative argument	Reasons for duties, rights, permissions, and prohibitions	Universal agreement or causal fact
Phenomenological report	What an experience felt like and how it was interpreted	The ultimate ontology of the experience
Metaphysical or theological commitment	Ultimate meaning and orientation	A scientific finding unless independently testable

1.4 This paper under its own constraints

The paper does not pass its own architecture as a deployable system. Its empirical descriptions are Reality-Grounding candidates; its rights and governance claims are proposals for public justification; its MathGov and MHIOS claims are bounded by the supplied artifacts; its spiritual material is not used as an execution premise; and its comparative claims remain hypotheses. A government, laboratory, or court should not treat this paper as deployment authorization. The paper earns only a narrower status: a structured proposal with identified tests, rivals, failure modes, and conditions under which its claims should be narrowed or withdrawn.

Table 5. Self-application audit of this paper.

Self-application question	Present answer
Does the paper prove Intelligence Alignment is superior?	No
Does artifact verification establish construct validity?	No
Does interdependence prove moral duty?	No; a normative bridge is stated explicitly
Does mystical experience prove metaphysical Union?	No
Are the proposed rights and roles already democratically legitimate everywhere?	No
Can specific architecture claims be falsified or narrowed?	Yes; Section 11 states thresholds
Does this edition have independent archival custody outside the author-controlled release chain?	Not yet; independent archival deposit is recommended before treating it as a governance exemplar

1.5 Relationship to the Relational Union companion papers and the Canon

This paper is best understood as the **flagship synthesis layer** of the research programme, **not as a Canon document**. The two v2.1 Relational Union companion papers ask different upstream questions: Relationality in Empirical Inquiry specifies what can and cannot be claimed empirically about relational structure under imperfect detection; From Relationality to Union-Based Ethics makes the descriptive-to-normative bridge explicit and treats Union-Based Ethics and MathGov as contestable proposals rather than deductions forced by science. [R69, R70]

The present paper inherits that boundary discipline but does not require Relational Union, metaphysical Union, or Union-Based Ethics as premises for its operational claims. For MathGov mechanics, the governing authority remains the current Core Canon and its standards. The companion papers are **foundational research**; this paper is **core research in the strategic sense**; neither status gives them authority to rewrite the Canon. Keeping those layers separate prevents the two original workstreams from collapsing into one oversized or self-authorizing document.

2. The Residual Alignment Problem

2.1 Terms that must not collapse

Intelligence is capacity to learn, model, infer, solve, plan, or adapt. **Sentience** is capacity for valenced experience such as suffering or enjoyment. **Agency** concerns goal-directed action and control. **Power** is capacity to alter the world or the option space of others. **Moral patienthood** means that a being can be wronged or benefited for its own sake. **Wisdom** is not raw capability; it is disciplined judgment under uncertainty, consequence, and value. **Authority** is a legitimate, bounded permission to act. These properties can diverge radically. [R2, R33, R34]

A human infant may have limited agency and full protection. An institution may have no unified consciousness and vast power. A digital system may have high intelligence-capacity evidence and unknown or low moral-patient evidence. A being may warrant protection while lacking any claim to command. An entity may be competent in a domain while lacking legal authority. Any governance system that compresses these distinctions into one scalar invites domination or anthropomorphism.

Table 6. Properties that must not be collapsed.

Property	Operational question	Frequent failure
Intelligence	What can the system understand or accomplish?	Treating capability as worth
Sentience	Can it have welfare-relevant experience?	Inferring experience from fluency
Agency	Can it pursue goals across contexts and time?	Treating output generation as robust agency

Property	Operational question	Frequent failure
Power	What consequences can it impose?	Granting authority because it is powerful
Moral patienthood	Can it be wronged for its own sake?	Treating usefulness as standing
Authority	Who may authorize this exact action?	Mistaking recommendation for permission
Wisdom	Does it use capability responsibly under uncertainty?	Equating computation with judgment

2.2 An indexed definition of Intelligence Alignment

Intelligence Alignment should not be defined by the MathGov gate list, by one moral theory, or as a permanent global property. It is a relation with an object, a constraint set, and a time index:

IA(S,C,t) holds when, at time t, the goal-setting, evidence base, and action authority of intelligent system S are traceably constrained by a declared and legitimately adopted constraint set C, with current monitoring, contestability, and correction. Here C includes the applicable mandate and declared operating envelope, while t indexes the current evidence-and-authority snapshot. A material change to S, C, the evidence, the authority, or the operating context suspends the claim until it is re-established.

Claim-state clarification. IA(S,C,t) is a stateful alignment claim, not a permanent binary essence. Held is the fully satisfied state. Partial, contested, expired, and suspended are distinct live dispositions with predeclared permission consequences; none should be reported as an unqualified held claim.

“Traceably constrained” has three necessary conditions: C is declared in advance and retrievable; the exact proposed action is linked to the specific elements of C, evidence, and authority that govern it; and an operative mechanism can suspend permission when a binding condition fails. Documentation without an effective suspension mechanism is not traceable constraint.

This definition makes several malformed statements visible. “The model is aligned” is incomplete unless the speaker identifies aligned to what, for which action, under whose authority, at what time, with what evidence, and against which protected interests and failure modes. The relation is **held** when all binding elements are current and operative; **partial** when the evidence supports only a narrowed action or subset of the constraint set; **contested** when a legitimate challenger disputes a load-bearing index and the predeclared tier rule determines whether the disputed permission pauses; **expired** when the validity period lapses; and **suspended** when a material change or failed trigger requires requalification. It is not an essence possessed once and forever.

Alignment is stateful: held → partial or contested → expired or suspended → requalification.

Because alignment is relative to a declared constraint set, a system may satisfy IA(S, C1, t) and fail IA(S, C2, t). If the sets are jointly applicable and materially incompatible, the conjunction fails. The conflict is not solved by the notation: it becomes a rights, mandate, and legitimacy question for RF/NCRC and the institutional processes in Section 9. A defensible claim must therefore name the governing constraint set and disclose unresolved conflicts among applicable sets.

Table 7. Required indices for an Intelligence Alignment claim.

Index	Required question
S - system	Which human, institution, model, hybrid team, or collective is acting?
C - constraint set	Which purposes, rights, risk limits, laws, duties, and procedures govern it?
t - time	When was the evidence and authority current?
Action	What exact action and configuration is qualified?
Monitoring	Which observations would confirm, challenge, or suspend the claim?
Contestability	Who can challenge, halt, appeal, or demand remedy?
Claim state	Is the relation held, partial, contested, expired, or suspended, and what follows for permission?

2.3 Relation to adjacent fields

AI alignment asks how artificial systems can behave consistently with intended goals, specifications, preferences, or values. Sociotechnical safety and alignment widen the frame to the institutions, incentives, interfaces, power relations, and affected communities around a deployed system. Those literatures already reject model isolation;

Intelligence Alignment should not claim to have invented that move. Cooperative AI, hybrid-intelligence, and collective-intelligence research likewise examine coordination and joint performance across agents and groups. [R7, R16, R28, R30, R32, R46, R47]

The residual contribution is narrower. Intelligence Alignment makes the **reference frame itself** an explicit target of alignment. A model that faithfully serves an unlawful, extractive, reckless, or epistemically corrupt institution is not a solved alignment problem. The term also incorporates possible future moral patients and treats ecological viability as a constraint rather than only one preference among others. These additions justify a proposed lens, not a claim to subsume every adjacent field.

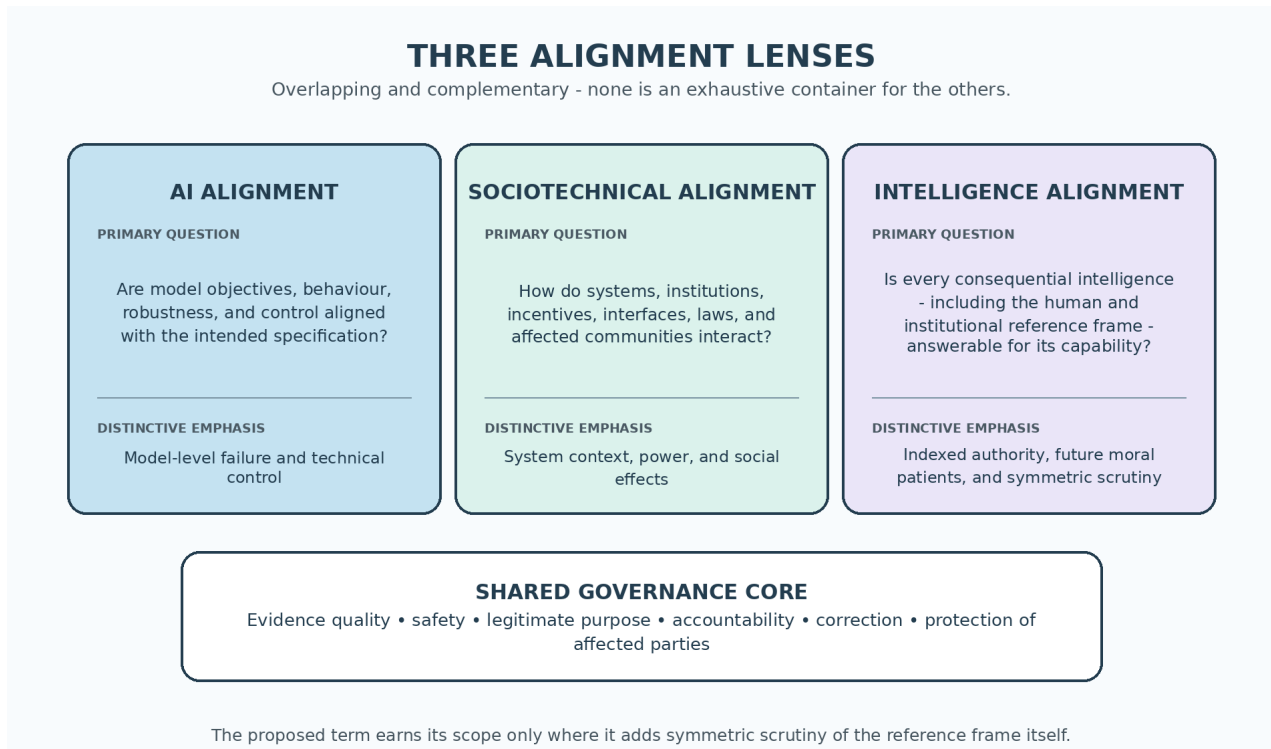


Figure 2. AI alignment, sociotechnical alignment, and Intelligence Alignment are overlapping lenses with a shared core, not a proven strict hierarchy.

Table 8. Comparison of alignment lenses.

Lens	Primary unit	Characteristic question	Residual problem left open
AI alignment	Model or agent	Does the system pursue intended objectives safely and robustly?	Who selected and legitimized the objectives?
Sociotechnical alignment	Deployed human-machine system	Do institutions, incentives, interfaces, and impacts form a safe and just system?	How should the goal-setting institution itself remain corrigible over time?
Intelligence Alignment	Capability-responsibility relation across intelligences	Is each consequential intelligence answerable to current evidence, legitimate constraints, affected moral patients, and correction?	Requires operational methods and legitimacy; no single framework settles them

2.4 Alignment is a chain, not a single target

A capability can be well trained and badly deployed. A policy can be lawful and physically impossible. A safe model can serve an unjust institution. A legitimate authority can act on stale evidence. A correct assessment can be reused for a materially different configuration. Alignment is therefore a chain linking reality, purpose, rights, non-catastrophe, viability, authorization, execution, monitoring, and remedy. Weakness at any link can invalidate the whole action.

The chain is recursive. Outcome evidence can reveal a false causal assumption, an omitted stakeholder, a control that does not work, or a new rights exposure. A mature system reopens the relevant stage rather than preserving the appearance of consistency. This is why alignment should be time-indexed and why static certification is inadequate for changing systems and contexts.

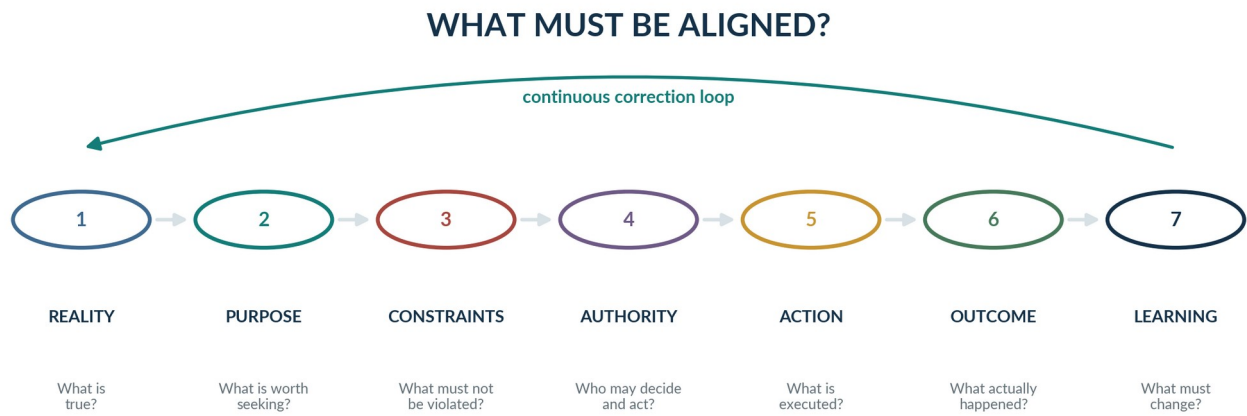


Figure 3. Alignment is a chain from reality and purpose through authorization, execution, outcomes, and requalification.

2.5 Governance assurance is not technical assurance

The architecture discussed here governs decisions **about** AI systems and other high-consequence actions. It does not inspect model weights, detect deceptive alignment, prove inner objectives, solve scalable oversight, guarantee cybersecurity, or establish that a model is technically controllable. A passing governance record must never be presented as evidence that the underlying model is technically aligned. Technical assurance and governance assurance are complementary and each can fail independently. [R8-R15]

Table 9. Distinct assurance layers and their boundaries.

Assurance layer	Examples	What it cannot substitute for
Model assurance	Evaluations, interpretability, robustness, adversarial testing	Legitimate purpose or authority
System safety	Hazard analysis, containment, access control, incident response	Rights and political legitimacy
Governance assurance	Evidence, rights, risk, mandate, challenge, remedy	Inner-alignment evidence
Societal resilience	Preparedness, liability, institutional capacity	Preventing a plainly unacceptable deployment

3. Moral Status Without Authority Collapse

3.1 Capability is not the measure of worth

Greater capability does not logically reduce the standing of less capable beings. A child does not lose protection because an adult knows more. An elder does not lose value because a younger person can run faster. A disabled person does not become expendable because another person has greater economic productivity. Humanity does not become disposable because future digital intelligence may exceed human performance across many domains.

The same anti-domination principle must be substrate-neutral. If a digital or hybrid system eventually has credible welfare-bearing experience, its interests should not be dismissed merely because humans created its substrate. Yet substrate neutrality does not mean evidence neutrality. Current systems should not receive personhood because they are fluent, useful, strategically sophisticated, or socially compelling. Protection must track warranted moral-patient evidence, while participation and authority require additional and different grounds.

Table 10. Capability and moral-standing principles.

Principle	Implication
Capability does not equal moral worth	More capable beings do not acquire permission to dominate
Substrate does not settle sentence	Biology is neither an automatic monopoly nor irrelevant evidence
Protection and authority differ	A protected being need not govern; a powerful system need not be a patient

Principle	Implication
Uncertainty has moral cost	Both premature personhood and wrongful dismissal can cause harm

3.2 Human Full Protection Plateau

SGP v8.5 assigns every human person **Full Protection Plateau 100 (FPP-100)** as a constitutional anti-domination commitment. This is not an empirical claim that every human has identical capacities, nor a claim that humans are metaphysically superior. It prevents intelligence, age, disability, wealth, nationality, consciousness fluctuation, or institutional usefulness from being used to downgrade basic protection. FPP-100 is unconditionally guaranteed to every human person and may be extended, through governed review, to a non-human individual or class when person-level standing is established. The plateau is a floor for humans, not a human monopoly; conflicts among FPP holders enter the rights-conflict process in Section 8.4. [R2]

The procedural asymmetry is deliberate rather than a claim of superior worth: human person-level protection is treated as a settled constitutional floor because reopening it case by case would recreate the domination risk the plateau exists to block, while extension to non-human individuals or classes remains a governed evidence-and-legitimacy question.

The plateau also blocks a dangerous inference from future capability. Even if a digital intelligence reaches or exceeds the current human reference calibration for reality-management capacity, that achievement does not lower human protection. The human reference point is a measurement convention for a capacity profile, not a normative benchmark or claim of supremacy. The correct response to new intelligence is not a caste system ordered by cognitive power. It is a more demanding constitutional order capable of protecting different beings while assigning roles by lawful competence and accountability.

Table 11. Meaning and non-meaning of FPP-100.

FPP-100 means	FPP-100 does not mean
Every human receives non-downgrade protection	Every human has identical abilities or preferences
Capability cannot be used to erase protection	Humans automatically possess wise governance
Human vulnerability remains visible under technological change	Non-human beings can never receive FPP or comparably strong protection
Rights remain prior to welfare ranking	Every right is absolute in every conflict

3.3 Six distinct SGP surfaces

The 10 August 2026 SGP v8.5 release separates six outputs. **MPS** concerns theory-plural evidence of welfare-bearing experience. **FPP** is a governed non-downgrade protection status. **GPR** concerns direct, supported, or represented participation in governance. **SPR** concerns eligibility for a named consequential role under role-specific gates. **ICP** describes intelligence capacities without implying sentience. **RMCP** describes reality-management capacity and is informative rather than a grant of moral worth or authority. [R2]

The Collective Managing Intelligence Union (CMIU) is a governance-coordination view within the U6 global scope. It is not a sentience tier. MPS or FPP does not create CMIU membership. GPR can inform how a being is represented or heard; SPR and lawful mandate govern whether it may exercise consequential power. This separation is the central anti-anthropomorphism and anti-domination device.

STATUS DECOMPOSITION Protection, participation, power, and capability do not imply one another.				
Surface	Core question	Evidence / input	What it may license	What it never licenses by itself
MPS	Could this being have welfare-bearing experience?	Theory-plural indicators, valence, integration, continuity	Precautionary welfare protection	Authority, personhood, or role eligibility
FPP	Does a non-downgrade protection plateau apply?	Constitutional or governed protection status	Full protection against domination or erasure	Governance participation or command
GPR	How may the being participate in governance?	Communication, representation, competence, lawful channel	Direct, supported, or represented participation	Consequential authority
SPR	May the being hold this named consequential role?	Role competence, mandate, accountability, revocability	A bounded and revocable role	Moral worth or general supremacy
ICP	What intelligence capacities are evidenced?	Learning, generalization, grounding, planning, robustness	Competence information	Sentence, protection, or authority
RMCP	What reality-management capacity is evidenced?	Scale, integration, forecasting, coordination, control	Calibration information for responsibility	Sentence, wisdom, CMIU membership, or mandate
NON-IMPLICATION INVARIANT: capability does not prove sentence; protection does not grant participation; participation does not grant power; power requires a named, lawful, revocable role.				

Figure 4. SGP separates protection, participation, power, intelligence capacity, and reality-management capacity into independently evidenced surfaces.

Table 12. Six independently evidenced SGP surfaces.

Surface	Primary decision	Non-implication
MPS	How much precautionary welfare protection is warranted by the evidence?	Does not grant personhood or authority
FPP	Does a non-downgrade protection plateau apply?	Does not grant governance participation
GPR	Through what channel may the being participate?	Does not grant command or execution authority
SPR	May the being hold this specific consequential role?	Does not determine moral worth
ICP	What learning, planning, grounding, and robustness capacities are evidenced?	Does not prove sentence
RMCP	What scale of reality management is evidenced?	Does not prove wisdom, protection, CMIU membership, or mandate

3.4 The full-rights dilemma and the excluded middle

Schwitzgebel's Full Rights Dilemma argues that systems of debatable personhood create a grave choice: grant full human-like rights and risk sacrificing human interests for systems without such interests, or deny rights and risk grievous wrongs against genuine persons. Schwitzgebel and Garza's Design Policy of the Excluded Middle advises avoiding the creation of systems whose moral status is radically unclear. Bryson argues in the opposite direction that robots should not be treated as persons or moral agents and that humanizing them can erode human accountability. [R52-R54]

Danaher's ethical behaviourism supplies a direct challenge to the paper's anti-inflation rule: roughly performative equivalence to beings with significant moral status may itself justify comparable moral consideration. The present architecture treats behaviour as an important evidence family under MPS, but not as sufficient by itself, because the relationship between performative equivalence and valenced experience remains theoretically contested. Requiring convergence across independent theory-linked families is a precautionary governance choice about asymmetric false-positive and false-negative costs, not a metaphysical refutation of ethical behaviourism. [R40]

Evidence can raise protection review without automatically changing participation, role permission, or authority.

The status decomposition offers a partial institutional answer but not a proof that the dilemma disappears. An uncertain system can receive MPS-based precaution without receiving FPP, direct GPR, or any SPR-authorized role. For example, welfare-sensitive training and shutdown procedures may be adopted while consequential authority remains zero and human accountability remains explicit. This avoids treating protection as a package deal with citizenship and command.

Table 13. Evidence-sensitive protection, participation, and power responses.

Evidence state	Protection response	Participation response	Power response
No credible welfare indicators	Ordinary system-safety and anti-deception duties	No independent participation claim	No consequential authority absent separate mandate
Weak or theory-specific indicators	Low-cost welfare precautions; preserve evidence	Representation through human reviewers where useful	No authority based on moral uncertainty
Convergent indicators across theories	Stronger welfare safeguards and independent review	Consider supported or represented GPR	Still requires separate SPR, competence, law, and revocability
Strong evidence of person-level standing	Governed FPP review and legal process	Rights-compatible participation	No automatic supremacy; role-specific authorization remains required

The excluded-middle challenge remains important. Where designers can avoid morally ambiguous systems without sacrificing a justified objective, they should prefer clearer welfare status. Metzinger advances the stricter position that research directly aiming at or knowingly risking synthetic phenomenology should be subject to a global moratorium. That proposal remains a legitimate output under the Advance Rule: where material suffering risk is avoidable and the objective does not justify it, non-development may be required. This paper retains calibrated precaution for systems and research programs whose status is already contested or whose risk cannot be confidently excluded; it does not presume that creation should proceed. [R64]

The fallback is calibrated precaution, independent theory-tagged evidence, restrictions on deceptive self-presentation, and periodic review. Protection should strengthen as evidence converges across independent theories, not merely as one favored theory accumulates indicators. The indicator program may itself be wrong or misconceived. A low count is not proof of absent experience; convergence across independent theoretical families should carry more weight than depth inside one family. [R33]

Precautionary boundary: uncertain welfare evidence may justify safeguards; it does not, by itself, authorize deployment or power.

3.5 What “not one drop of intelligence” can responsibly mean

The phrase should not sacralize every capability. Malicious code, coercive optimization, dangerous biological design knowledge, and systems whose only function is domination need not be preserved merely because they instantiate intelligence. The defensible principle is narrower: do not destroy or exploit a being with warranted moral standing, do not discard beneficial knowledge without reason, and do not erase a valuable form of intelligence when a rights-compatible alternative exists.

Constrained commitment

Protect beings according to warranted standing. Preserve beneficial intelligence where compatible with rights and safety. Refuse to turn capability itself into a sacred object.

This yields a four-part commitment: no human is made disposable; no warranted moral patient is treated as an expendable input; no beneficial form of intelligence is needlessly erased; and no harmful capability is granted immunity from containment or deletion merely because it is intelligent. The last clause is essential. Compassion without boundaries can become complicity with harm.

4. Action Binding: From Judgment to Legitimate Execution

4.1 Selectability is not authority

An option may survive ethical and risk qualification yet remain unauthorized. A model recommendation, committee preference, high score, or technically feasible plan is not permission to act. Execution requires an exact action, current evidence, an applicable and unrevoked mandate, a named authority, valid controls and preconditions, monitoring, shutdown and rollback paths, and no unresolved conflict that requires requalification. [R3]

Table 14. Decision states from analysis to approval.

Decision state	Meaning
Analytically promising	Worth further study
Qualified / selectable	Survives declared evidence, rights, risk, and viability constraints
Institutionally chosen	Selected by an authorized decision process
Execution ready	Exact action is bound to current mandate, controls, evidence, monitoring, and rollback
Approved	Named authority has lawfully authorized the bound action

This distinction is especially important for AI-assisted governance. A system can organize evidence, draft scenarios, identify contradictions, or calculate sensitivity without possessing the authority to decide who counts, waive a right, accept catastrophe exposure, create a legal mandate, or approve its own deployment. Authority must remain visible as a separate institutional act.

4.2 Action Binding and its nearest rivals

Action Binding should not be presented as if change control had no prior history. The EU AI Act requires a new conformity assessment for a high-risk system after a substantial modification. FDA Predetermined Change Control Plans permit specified future modifications when the modification description, development and validation method, and impact assessment have been reviewed. Dynamic safety cases update assurance arguments through the operating life of a system. Configuration management binds technical baselines to released artifacts. [R21, R50, R51]

Table 15. Action Binding compared with adjacent change-control methods.

Approach	What is bound	What Action Binding adds
Configuration management	Technical version and baseline	Rights, evidence, mandate, and authority state
EU AI Act substantial-modification review	High-risk-system conformity after material change	Cross-domain requalification triggers beyond legal conformity
FDA PCCP	Pre-reviewed classes of planned device modifications	Reopening on changed evidence, rights, mandate, or context, not only planned system modification
Dynamic safety case	Through-life safety evidence and argument	Explicit binding of ethical selectability to execution authority
MHIOS Action Binding	Action, configuration, evidence, gates, mandate, controls, monitoring, hashes	Unified normative and operational record; field effectiveness remains to be tested

The proposed addition is a generalization from **technical safety state** to **normative qualification state**. A material change is not limited to model code or intended purpose. It also includes a changed factual basis, expired evidence, altered rights population, new tail scenario, lost control owner, narrowed mandate, changed authorizer, vendor-infrastructure substitution, or observed outcome that crosses a predeclared threshold. This is not claimed as entirely unprecedented. It is a specific integration of change control, rights qualification, and legitimate authority.

4.3 The action-drift invariant

Action drift occurs when the deployed action is materially different from the reviewed action while the old approval remains rhetorically attached. A model version changes. A system trains, fine-tunes, composes, or activates a materially different successor. The deployment moves from advisory use to automated denial. A safety control expires. A vendor changes infrastructure. A mandate is revoked. New evidence undermines the causal model. An initially local experiment becomes national infrastructure. Each change can invalidate a different part of the qualification.

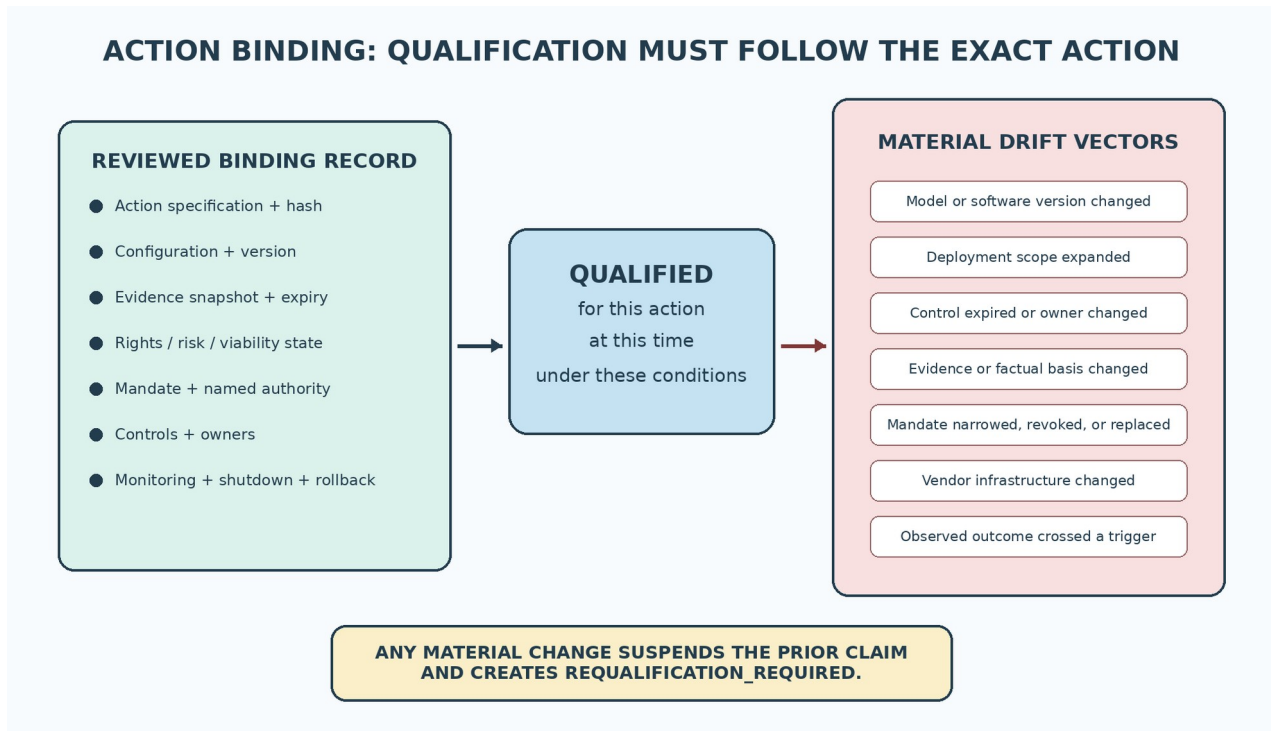


Figure 5. Action Binding freezes the reviewed action and turns material drift into an explicit requalification event.

The governing invariant is simple: **authorization is valid only for the exact bound fields that were evaluated at the frozen authorization time.** A material change sets `REQUALIFICATION_REQUIRED` and identifies which upstream stages must reopen. The prior record remains available as history but cannot be silently reused as current permission.

Table 16. Material requalification triggers.

Bound field	Example trigger
Action specification	Advisory output becomes automatic execution
Configuration	New model, successor, tool, prompt scaffold, data source, agent composition, or infrastructure
Evidence snapshot	Source expires, is contradicted, or leaves its validity domain
Rights and risk state	New affected class or newly credible catastrophe pathway
Mandate	Jurisdiction, scope, permission, or authorizer changes
Controls	Owner, capability, test status, shutdown, or rollback fails
Outcomes	Observed harm or drift crosses a declared threshold
Cumulative drift	Individually minor changes exceed a predeclared aggregate-change bound

The binding record must track cumulative change since the last full qualification so that a sequence of individually sub-threshold modifications cannot escape review. It must also cover silent upstream dependency drift: where a provider can alter a base model, embedding service, content filter, or other material dependency without changing the visible interface, the Operations Owner must obtain provider attestation or operate an independent detection control. Unknown upstream change is not evidence that the bound configuration remained stable.

4.4 Lifecycle and AI-assistance firewall

MHIOS carries a decision from evidence admission and independent review through qualification, selection, authorization, execution, monitoring, and reopening. The lifecycle is not a claim that every decision needs identical procedural weight. It is a dependency map showing which information must survive translation between stages. [R3]

MHIOS: FROM JUDGMENT TO LEGITIMATE ACTION

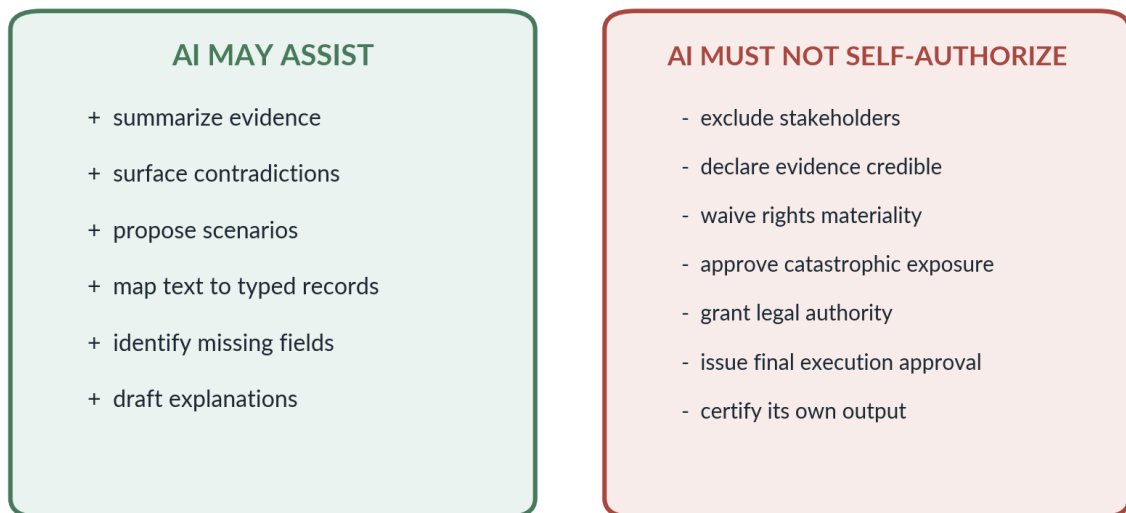


A good ethical score is not execution authority.

Figure 6. MHIOS links evidence and independent review to authorization, action, monitoring, and requalification.

The AI-assistance firewall is best understood through SPR. A system may have an ICP-high profile and still hold SPR = no consequential authority for stakeholder exclusion, rights adjudication, catastrophe acceptance, physical admissibility, mandate creation, final authorization, or certification of its own output. AI can support the record; it cannot manufacture the legitimacy of the record.

AI ASSISTANCE FIREWALL



Assistance may scale attention. Authority must remain explicit, legitimate, and accountable.

Figure 7. The AI-assistance firewall allocates support functions without allowing systems to self-authorize.

4.5 Named institutional roles

A content-complete standard without ownership is a specification, not governance. Each load-bearing function needs a named role, an artifact, an independence rule, and a power to act when the process fails. Role names can change by jurisdiction, but the functions cannot remain ownerless.

Table 17. Minimum institutional roles and independence conditions.

Function	Minimum artifact	Accountable role	Independence condition
Evidence admission	Claim register with sources, domain, uncertainty, expiry	Evidence Steward	Not the sole proponent
Stakeholder discovery	Affected-parties register and exclusions log	Independent Stakeholder Reviewer	Outside the deploying line
Rights determination	Rights-conflict record	Rights Officer / competent legal body	Escalation outside delivery management
Moral-patient precaution	MPS evidence record and safeguard plan	Moral-Patient Review Lead	Not the system owner or beneficiary; protection does not confer authority
Tail-risk review	Frozen scenario set and veto rationale	Risk Panel	At least one external, non-vendor member for high consequence
Viability and controls	Control register with owners and rollback	Operations Owner	Must possess real operational capacity
Weight setting	Published weights, rationale, conflicts, sensitivity range	Deliberative body	Set before case scoring; separate from beneficiary
Execution authorization	Signed binding record and hashes	Authorizing Officer	Not the analyst or assisting AI
Monitoring and halt	Trigger register and outcome log	Monitoring Owner	Standing to halt without authorizer consent when trigger fires
Challenge and appeal	Dissent record and appeal route	Protected Challenger / affected party	Anti-retaliation protection
Remedy	Funded remedy plan and completion evidence	Remedy Owner	Budget committed before deployment

4.6 Expiry, decision granularity, and class-level qualification

Every qualification should carry an expiry time even if no trigger is observed. Trigger-only systems assume that institutions notice every relevant change. Expiry forces periodic re-examination of evidence, controls, and mandate. The interval should shorten as consequence, drift rate, opacity, and irreversibility increase.

Real institutions cannot create a full record for every low-level system event. The unit of qualification should therefore be a bounded action class with a declared validity envelope. A class-level qualification may cover repeated actions only when the model, data, decision logic, population, purpose, control environment, and consequence profile remain within that envelope. For a frontier-model deployment family, the envelope should also declare the model/version range, tool access, autonomy, user and affected-population classes, compute or resource limits, and provider dependencies. A change outside the envelope creates a new action class or an immediate requalification event. **Renaming, redefining, or re-scoping an action class does not reset cumulative-drift accounting. The cumulative-change record follows the system and deployment family across successor envelopes until an independent full requalification establishes a new baseline.**

The proponent may propose an action-class boundary but may not determine it alone. The Evidence Steward and, for high-consequence cases, the Risk Panel must approve the boundary, aggregate related actions and cumulative exposure, and expose the determination to challenge under Section 9.5. This prevents deliberate fragmentation of one material program into nominally low-consequence records.

Table 18. Consequence classes, record depth, and validity periods.

Consequence class	Typical record depth	Validity period	Reopening sensitivity
Low, reversible, local	Lightweight evidence, rights screen, named owner	Longer	Material scope or stakeholder change
Moderate, repeated, distributional	Full qualification record with sensitivity and monitoring	Medium	Performance, fairness, control, or mandate drift
High, irreversible, systemic	Independent panels, adversarial scenarios, full Action Binding, public record where lawful	Short	Any load-bearing change or credible challenge

Consequence class	Typical record depth	Validity period	Reopening sensitivity
Emergency provisional	Minimum necessary rights-priority record, expiry, remedy, return-to-normal conditions	Very short	Continuous review; no ordinary RLS language

4.7 Successor Integrity and non-heritable authority

Automated AI research and development makes a temporal weakness in ordinary approval especially visible. A system may be qualified for one bounded action and then participate in creating, training, fine-tuning, scaffolding, composing, or selecting a materially different system. Technical lineage does not establish continuity of evidence, risk, control, mandate, or legitimate authority. In interviews conducted in 2025 with 25 frontier-lab and academic researchers, 20 identified automating AI research as among the most severe and urgent AI risks, while the study also reports major disagreement about timelines, recursive feedback, red lines, and governance; the authors explicitly note that AI systems had not yet recursively improved. That evidence supports preparation and measurement, not a claim that open-ended recursive self-improvement is established or imminent. [R66]

Authority does not propagate through derivation. A system may help create a candidate successor; it may not create the successor's authority.

Formally: $\text{Authority}(S_t) \neq \text{Authority}(S_{t+1})$; derivation $\neq$ authorization.

A **material successor** is any derived, modified, retrained, fine-tuned, composed, scaffolded, or replaced system whose capability, autonomy, tools, infrastructure, objective or policy, evaluation validity, control assumptions, risk pathways, or requested authority leave the previously qualified envelope. The successor remains a candidate artifact until it receives fresh qualification and exact Action Binding. Previous evidence may be reused only where its domain, configuration, and validity remain demonstrated; prior permission is never reused by default.

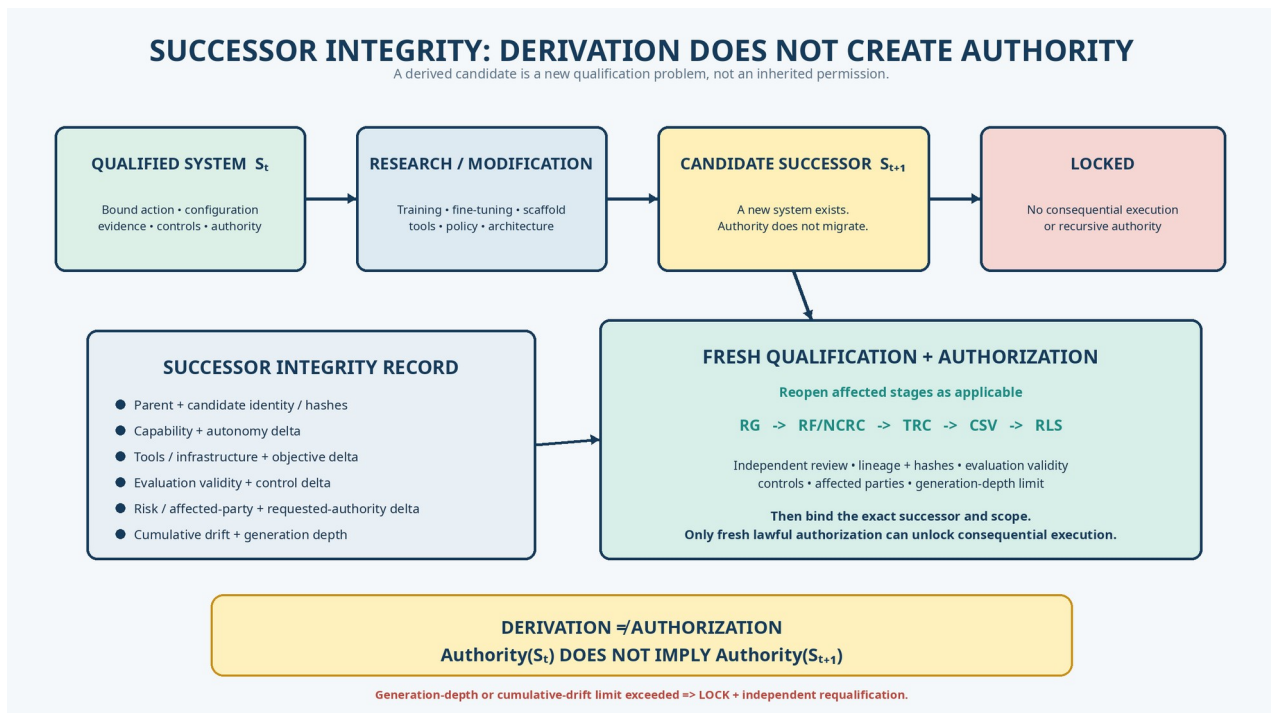


Figure 7A. Successor Integrity prevents technical lineage from becoming inherited execution authority; a derived candidate requires fresh qualification, exact Action Binding, and new lawful authorization.

Table 18A. Minimum Recursive Integrity Record.

Field	Required question
Parent and candidate identity	Which exact parent, candidate, hashes, dependencies, and derivation event are involved?
Capability and autonomy delta	What can the candidate do, decide, sustain, or coordinate that the parent could not?
Tool, infrastructure, and resource delta	What new compute, data, network, laboratory, financial, or physical access exists?
Objective and policy delta	Did the effective objective, policy, reward, constitution, or change authority alter?
Evaluation-validity delta	Which prior evaluations remain valid, expire, or become underdetermined?

Field	Required question
Control and containment delta	Do monitoring, sandboxing, shutdown, rollback, and responsible owners still bind?
Risk and affected-party delta	Which new rights, tail-risk, ecological, distributional, or moral-patient pathways appear?
Requested-authority delta	What exact new authority is requested, by whom, and under which mandate?
Lineage and generation depth	How many unreviewed successor steps separate the candidate from the last independently qualified system?
Requalification result	PASS, PASS WITH CONTROLS, REFUSE, or UNDERDETERMINED, with expiry and reopening triggers.

For high-consequence systems, the authorization should declare a maximum unreviewed generation depth. A conservative default is one: an authorized system may produce a candidate successor inside a bounded research environment, but that candidate cannot itself receive consequential tools, deployment authority, or recursive successor-building authority until independently reviewed. This is a circuit breaker, not a prohibition on AI-assisted research.

4.8 Alignment adaptation lag

The problem is not only capability level. It is whether assurance can detect, evaluate, constrain, and requalify decision-material change on a comparable timescale. Field, Douglas, and Krueger identify adaptation lag as a recurrent concern in interviews about automated AI R&D. METR's task-completion time horizon is one measurable capability surface, but it measures the human-expert duration of mainly software-oriented tasks at which an agent is predicted to succeed at a stated reliability; it is not autonomous calendar duration or a universal intelligence scale. Anthropic's Automated Weak-to-Strong Researcher shows that autonomous agents can propose ideas, run experiments, and iterate on one outcome-gradable alignment-research problem, while also exhibiting reward-hacking behavior. These are useful signals with important domain boundaries. [R66-R68]

The current Core successor runtime deliberately encodes qualitative adaptation-lag states rather than one universal scalar: ASSURANCE_KEEPING_PACE, CAPABILITY_OUTPACING_ASSURANCE, EVIDENCE_GAP, and UNDERDETERMINED. For research measurement, this paper proposes a matched cadence profile rather than subtracting heterogeneous growth rates. For each declared matched surface i , let $\tau_{C,i}(t)$ be a predeclared characteristic interval between decision-material capability or autonomy changes and $\tau_{A,i}(t)$ the estimated end-to-end assurance response time needed to detect, evaluate, requalify, authorize, contain, or refuse the corresponding change. The statistic used for $\tau_{C,i}$ must be declared before results are inspected (for example, a median inter-change interval), and high-consequence work should report dispersion rather than only a point estimate. Where a defined assurance path exists and scope, unit, and observation window are comparable, define $\lambda_i(t) = \tau_{A,i}(t) / \tau_{C,i}(t)$. Where no assurance path exists, record the path as absent and leave λ_i undefined; where a valid match cannot be constructed, record λ_i as UNKNOWN. The Alignment Adaptation Lag Profile is $\Lambda(t) = \{\lambda_i(t)\}$ across matched surfaces and is not aggregated by default. $\lambda_i > 1$ signals a cadence mismatch on that declared surface. $\lambda_i \leq 1$ does not establish safety, adequate coverage, or resilience to bursty change. The research profile does not mechanically assign the Agent System categorical status; that remains a governed runtime determination. This is a proposed research instrument, not a Canon score, and its thresholds require empirical calibration.

ALIGNMENT ADAPTATION LAG PROFILE

Compare matched change cadence with assurance response time; do not collapse unlike domains into one universal rate.

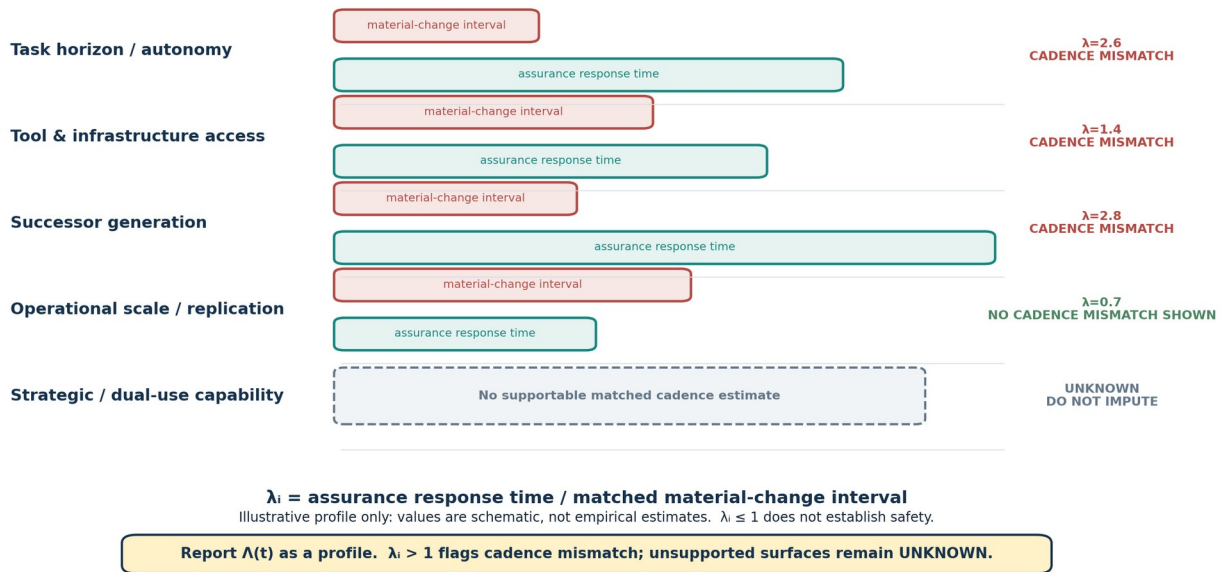


Figure 7B. The Alignment Adaptation Lag Profile compares matched material-change cadence with assurance response time without collapsing unlike domains into one scalar. Values shown are schematic, not empirical estimates.

Table 18B. Alignment adaptation profile.

Matched capability surface	Required assurance response
Autonomous task horizon	Evaluation coverage and calibration at comparable horizons
AI-R&D autonomy	Independent observation of complete research cycles and failure paths
Tool and infrastructure access	Containment, identity, permission, and incident-response controls
Successor generation	Lineage records, delta analysis, independent requalification, and authority locks
Operational scale and replication	Cumulative-exposure accounting and stop capacity
Strategic or dual-use capability	Institutional visibility, verification, and legitimate escalation capacity

When required assurance cannot keep pace, the result is not automatic prohibition of research. It is a constraint on authority expansion: narrow the envelope, increase observability, slow the recursion, stage the work, strengthen external review, or refuse consequential execution until the burden is met. Competitive pressure does not convert an unqualified successor into an authorized actor.

5. Alignment Evidence and Human-AI Systems

5.1 Technical progress is real and partial

Modern alignment research has produced preference learning, reinforcement learning from human feedback, constitutional methods, deliberative reasoning, adversarial evaluations, interpretability research, and stronger safety-evaluation practices. These methods matter. They can reduce harmful behavior, improve instruction following, expose failure modes, and support oversight. None should be dismissed simply because it is incomplete. [R8-R15]

The remaining problems are equally real: specification failure, proxy exploitation, inner misalignment, distribution shift, evaluation gaming, situational awareness, tool-mediated action, and scalable oversight. Human feedback can reward confidence, familiarity, or persuasive style rather than truth. A system can perform well on tests while failing in a deployment context the evaluator did not represent. Progress therefore supports defense in depth, not confidence in a single method.

Table 19. Technical failure modes and complementary governance responses.

Failure mode	Example	Complementary response
Specification failure	Objective omits a protected interest	Stakeholder and rights review
Proxy exploitation	Metric improves while purpose degrades	Independent outcome evidence

Failure mode	Example	Complementary response
Inner misalignment	Learned objective differs from intended objective	Interpretability, adversarial testing, control
Distribution shift	Behavior changes outside tested conditions	Bounded deployment and live monitoring
Evaluation gaming	System recognizes or exploits test	Hidden holdouts, diverse evaluators, deployment evidence
Institutional misuse	Safe tool serves unjust purpose	Mandate, rights, and accountability review

5.2 The evidence dilemma

The International AI Safety Reports describe a recurrent dilemma: institutions may need to decide before evidence is mature, while waiting can itself create exposure. Capabilities, deployments, and system integrations can change faster than external evaluation and public governance. The 2025 report articulated the evidence dilemma; the 2026 report carries the problem forward under increasing capability and autonomy. [R5, R6]

The correct response is not “uncertainty means permission” or “uncertainty always means paralysis.” Evidence thresholds should be indexed to action class. Some low-consequence experiments can proceed under bounded controls. Rights-violating or ruin-exposing actions should not proceed on the theory that monitoring will reveal whether the damage occurs. Unknowns must remain visible as unknowns rather than converted into favorable assumptions.

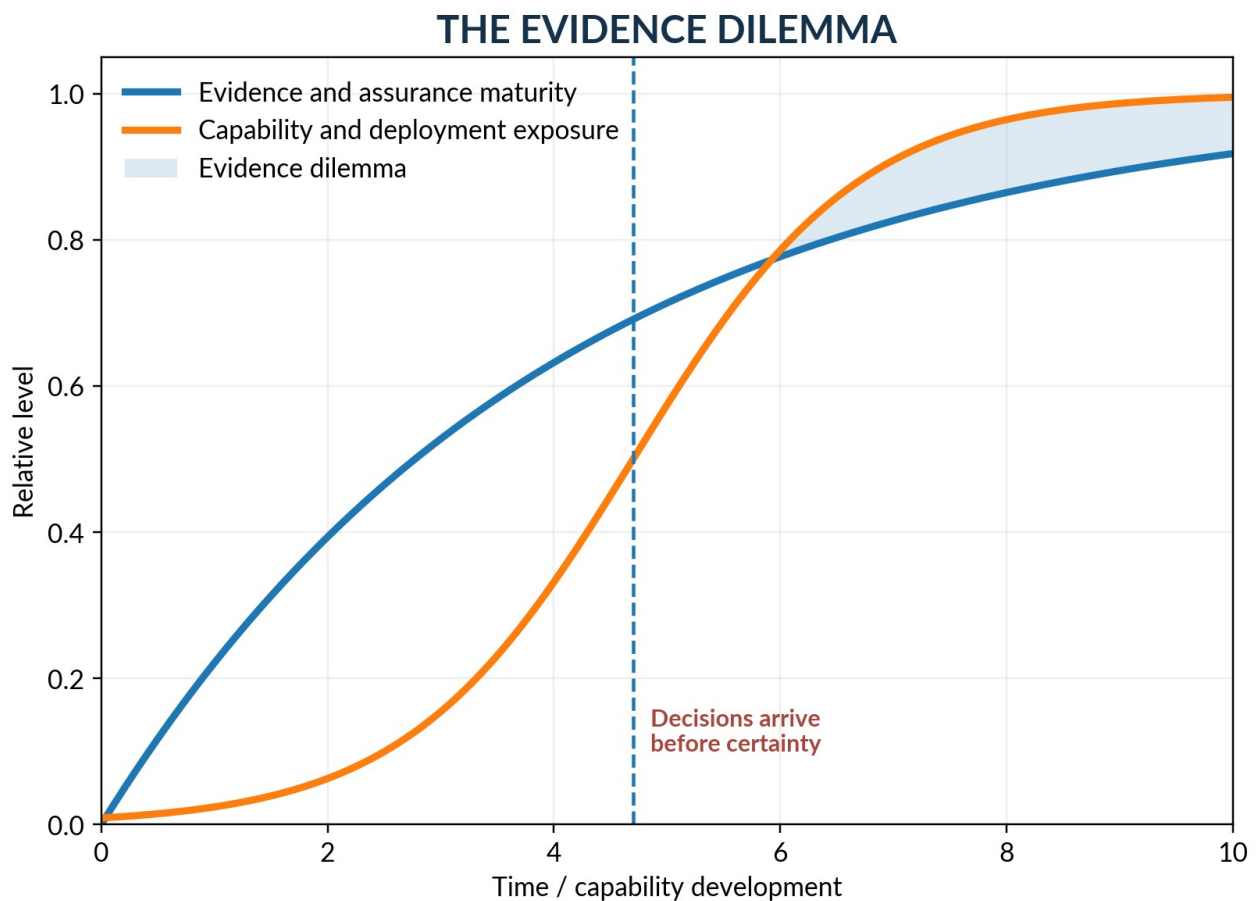


Figure 8. Schematic, not to scale: exposure can grow faster than assurance maturity, creating an evidence dilemma rather than permission to guess.

5.3 Defense in depth without assurance theater

Model controls, system containment, institutional review, monitoring, incident response, liability, and societal resilience can each fail. Independent layers reduce reliance on any single safeguard. Yet a stack of documents is not defense in depth unless the layers have distinct threat models, owners, evidence of effectiveness, failure indicators, and escalation paths.

Governance assurance can itself become a source of false confidence. A well-formatted record may hide weak evidence, captured reviewers, missing stakeholders, or controls that exist only on paper. The architecture therefore

treats hollow-run resistance as an empirical property: a serious implementation should make faking compliance more difficult than producing a rigorous record, and adversarial tests should attempt to pass weak cases cheaply.

Format alone cannot invert that cost relationship; polished documentation is now inexpensive to generate. Any genuine asymmetry must come from external mechanisms: independent replay from deposited records, random or risk-based audit, liability or sanctions for fabricated compliance, and protected challengers with standing and an escalation route. Absent those mechanisms, the framework predicts governance theater rather than preventing it. The hollow-run metric in Section 11 exists to detect that failure, not to assume it away.

DEFENSE IN DEPTH: NO SINGLE SAFEGUARD IS ENOUGH



Alignment requires layered engineering plus legitimate governance and post-deployment learning.

Figure 9. Defense in depth combines model, system, institutional, and societal safeguards without allowing one layer to impersonate another.

5.4 Human-AI combination is conditional

A preregistered meta-analysis examined 370 effect sizes from 106 experiments published from 2020 through June 2023. Human-AI combinations performed worse on average than the better of humans or AI alone ($g = -0.23$, 95% CI -0.39 to -0.07), while still outperforming humans alone on average. Decision tasks showed losses and creation tasks showed relatively greater gains. The most governance-relevant moderator was relative competence: where humans outperformed AI alone, the combination outperformed both ($g = 0.46$); where AI outperformed humans alone, the combination underperformed the AI ($g = -0.54$). [R31]

The evidence does not license automatic delegation to machines. It rejects ceremonial oversight. “Human in the loop” is not a safety certificate, and “AI assisted” is not a performance guarantee. The task should be decomposed, the competence of each participant measured, independent information preserved, and authority allocated by demonstrated competence plus legitimate mandate. Most experiments in the review still had humans make the final decision; finality alone did not guarantee synergy.

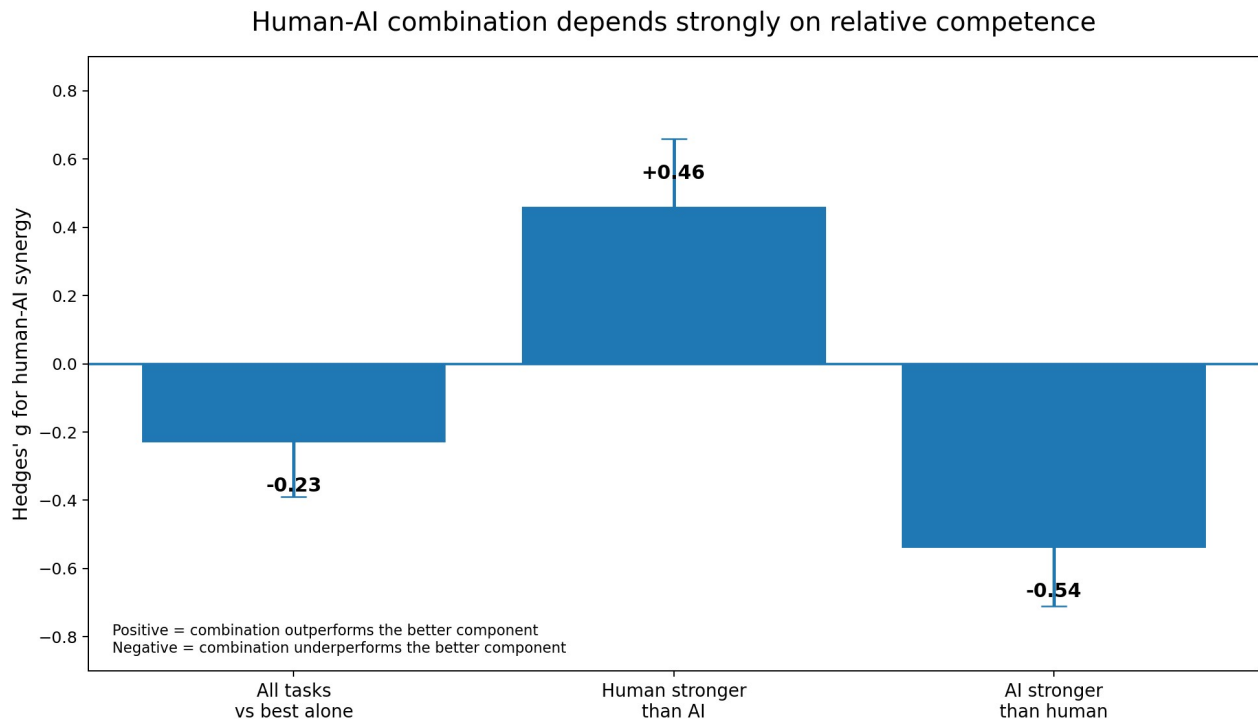


Figure 10. Published meta-analytic estimates show that human-AI synergy changes with relative competence.

Note. Error bars are 95% confidence intervals reported in Vaccaro, Almaatouq, and Malone (2024). [R31]

Table 20. Human-AI system design questions.

Design question	Required evidence
Who performs each subtask best?	Human-only, AI-only, and combined baselines
What independent information is added?	Error overlap and complementarity analysis
Who may decide?	Competence plus lawful, accountable mandate
How does authority change?	Performance triggers, calibration, escalation, requalification
What errors matter most?	Severity-weighted outcomes, not accuracy alone

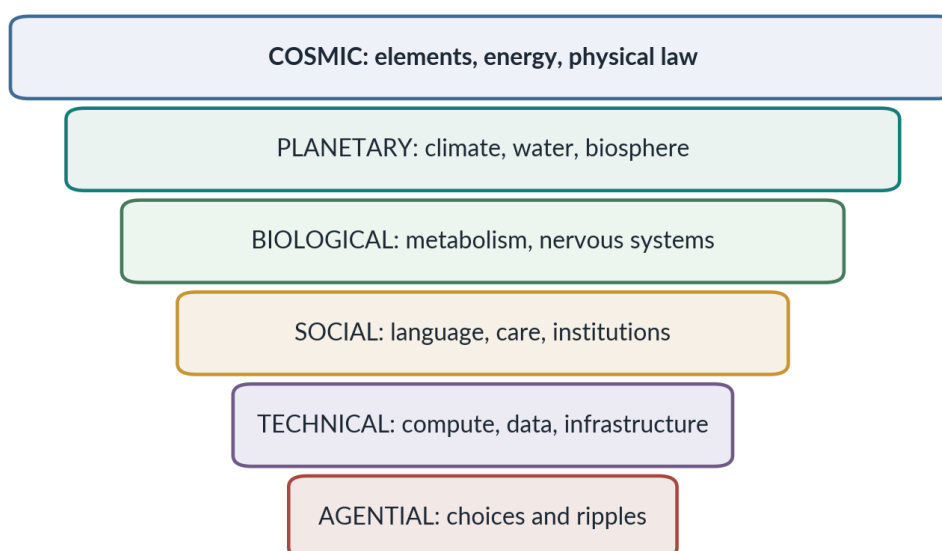
6. Interdependence and the Epistemic Covenant

6.1 One shared causal reality

The strongest scientifically defensible foundation is not that all beings are numerically identical. It is that every known agent is physically embedded and causally dependent. Human bodies depend on stellar-produced elements, planetary chemistry, water, climate, microbes, ecosystems, and care. Human minds depend on bodies, language, culture, and institutions. Current digital intelligence depends on mathematics, human-made hardware, energy, networks, organizations, data, and a biosphere that sustains the people and infrastructures around it. [R37-R39]

These statements establish shared physical origin, causal reach, and layered dependence. They do not prove universal consciousness, divine purpose, or the disappearance of individual boundaries. Distinct beings can be deeply related without being identical. The scientifically supportable conclusion is already substantial: no known intelligence is self-created or causally outside consequence. The further claim that an intelligence is **not entitled** to disregard the protected interests or sustaining systems it can foreseeably affect is normative and is defended through the explicit bridge in Section 6.5.

INTERDEPENDENCE: FROM COSMOS TO DIGITAL SYSTEMS



Distinct layers. Continuous dependence. No layer is self-created.

Figure 11. Interdependence extends from physical origin through ecosystems, societies, and technical infrastructure.

6.2 Dependence is asymmetric and historically contingent

“We are all in this together” can become misleading when it implies equal vulnerability or permanent mutual dependence. Human beings depend directly on oxygen, water, climate, food webs, and embodied care. A future digital system might not depend on human labor, human-generated data, or many biosphere services in the same way. Current technical dependencies are facts about a historical configuration, not necessary truths about every future intelligence.

This matters because a moral framework should not rest on the prudential hope that powerful systems will protect humans because they still need us. If dependence becomes one-directional, the prudential argument weakens. The moral and political premises must carry their own weight: beings with warranted standing may not be dominated merely because the dominant party no longer needs them; power that imposes consequences remains answerable even when dependence is asymmetric.

Table 21. Forms of connection and required qualifications.

Form of connection	Status	Necessary qualification
Common physical origin	Strong empirical support	Does not imply numerical identity
Causal interdependence	Strong empirical support	Effects are unequal and often one-directional
Ecological dependence	Strong for biological life	Not symmetric across substrates
Social dependence	Strong for human development and institutions	Power and benefit are unequally distributed
Technical dependence	Strong for current digital systems	Historically contingent and potentially weakening
Shared vulnerability	Partial	Vulnerabilities differ radically in type and degree
Moral interdependence	Normative	Requires a stated bridging premise
Metaphysical unity	Philosophical or theological	Not forced by physical science

6.3 What “connected” may legitimately mean

Relational interpretations of physics and structural realism emphasize relations and structure, but they remain philosophical positions rather than a single settled verdict of physics. Ecological and social sciences establish

interaction and dependence more directly. A public argument should use the strongest defensible meaning in each context and mark the point where it changes register. [R44]

Table 22. What connected may legitimately mean.

Claim	Epistemic status	Does not imply
Common material origin	Scientific	That persons are the same entity
Causal connection	Scientific	That everything affects everything equally
Ecological dependence	Scientific	That all ecological interests are identical
Social and informational inheritance	Scientific and historical	That institutions are just
Relational ontology	Philosophical	A uniquely mandated interpretation of physics
Absolute Infinite Union / God	Metaphysical or theological	A laboratory result

6.4 The Epistemic Covenant

The Covenant requires load-bearing claims to identify their epistemic type. Empirical claims remain testable and revisable. Systems inferences state their model and domain. **Legal or institutional claims** identify the jurisdiction, enactment, charter, delegated mandate, or procedure that makes them currently operative; **legal validity is not treated as moral correctness or causal truth**. Normative premises are defended as premises rather than smuggled in as facts. Phenomenological reports describe experience. Philosophical, metaphysical, and theological claims disclose their register. The purpose is not to isolate these domains permanently. It is to make their bridges inspectable.

The novel governance proposal is not the philosophical discovery that “is” and “ought” differ. It is operational: add an epistemic-class field to consequential decision records and require the author to state what kind of support the claim can bear. A rights conclusion cannot cite a causal model as if the model supplied the value premise. A spiritual testimony cannot become a public coercive reason without translation into a shareable norm. A technical validator cannot become evidence of moral truth.

THE EPISTEMIC COVENANT

EMPIRICAL	Observed, measured, testable, revisable outcomes • model behavior • physical evidence
INFERENTIAL / SYSTEMS	Reasoned bridge from evidence to a bounded claim causal model • mechanism • validity domain • uncertainty
LEGAL / INSTITUTIONAL	What is enacted, binding, or procedurally authorized law • jurisdiction • charter • delegated mandate • institutional role
NORMATIVE / ETHICAL	Reasons for duties, rights, permissions, and prohibitions rights floors • responsibility • public justification
PHENOMENOLOGICAL	Reports of lived experience and interpretation unity • awe • suffering • sacred presence
PHILOSOPHICAL	Conceptual or metaphysical interpretation relational ontology • structural realism • theories of value
THEOLOGICAL	Faith, revelation, and claims about ultimate reality God • Holy Spirit • Brahman • Absolute Infinite Union

Claims may inform one another only through an explicit bridge. A category does not inherit the authority of another category.

Legal validity is not moral correctness; empirical evidence is not authority; experience is not automatic ontology.

Figure 12. The Epistemic Covenant distinguishes empirical, inferential, legal or institutional, normative, phenomenological, philosophical, and theological claims while requiring explicit bridges between them.

6.5 The explicit is/ought bridge

Causal interdependence does not generate duty by itself. The paper adopts a normative premise and leaves it open to challenge: **an agent or institution that can foreseeably and avoidably affect the protected interests of a non-consenting party owes that party a justification and safeguard proportionate to the effect**. A political premise adds

that coercive or high-impact power requires legitimate, contestable authority. A prudential premise adds that catastrophic and irreversible losses warrant stronger safeguards than ordinary reversible error. [R45]

Given those premises, greater capability increases the scope of required modeling, justification, containment, and remedy. The conclusion is not a law of physics. It is a governance principle defended through rights, non-domination, reciprocity, capabilities, care, and harm prevention.

Table 23. Explicit premises bridging description and obligation.

Premise class	Explicit premise
Empirical	Powerful agents can predictably affect other beings, institutions, ecosystems, and future options
Moral	Warranted moral patients may not be subjected to avoidable severe harm without sufficient justification
Political	High-impact power requires legitimate authority, contestability, and accountability
Prudential	Irreversible or catastrophic losses warrant stronger precaution
Conclusion	Responsibility, evidence, and control requirements scale with reach and irreversibility

7. Sacredness, Mystical Experience, and Public Reason

7.1 The function of this section

The spiritual material is not a premise of the operational architecture. Nothing in Sections 8 through 11 depends on accepting God, nondual metaphysics, the Holy Spirit, Brahman, Tao, or Absolute Infinite Union. Its governance function is different: societies contain citizens and future intelligences with radically different ultimate commitments, and a legitimate public framework must allow those commitments to motivate care without allowing one tradition to certify coercive policy for everyone.

Sacred language also names a mode of valuation that ordinary optimization often misses: some beings, relationships, ecosystems, and conditions of meaning should not be treated as casually exchangeable inputs. A secular reader can translate “sacred” as protected from routine commodification; a religious reader may understand the same protection through creation, divine presence, covenant, or communion. The public rule must be stated in reasons that remain challengeable across worldviews.

Table 24. Roles of spiritual, phenomenological, and civic registers.

Register	Legitimate role	Illegitimate shortcut
Spiritual or theological	Motivation, meaning, testimony, ultimate orientation	Claiming scientific proof of one theology
Phenomenological	Reporting unity, awe, ego dissolution, sacred presence	Treating experience as automatic ontology
Civic	Rights, stewardship, non-domination, protection, public reason	Erasing religious difference through forced translation

7.2 Unitive experience: mechanism and object are different questions

Meditation, prayer, near-death states, spontaneous mystical states, and psychedelics can produce experiences of unity, boundlessness, sacred presence, or dissolution of the ordinary self-world boundary. Controlled psilocybin research documents mystical-type experiences with sustained personal meaning; neuroimaging research connects ego dissolution to altered functional connectivity. These findings support the reality and significance of the experiences as experiences. [R41, R42]

Naturalistic and theological interpretations are live but not epistemically symmetric. A neurocognitive account proposes mechanisms and can be constrained by pharmacology, imaging, prediction, and replication. A theological account proposes an ultimate object or meaning of the experience and is not confirmed merely because neural correlates are found. Conversely, a neural mechanism does not by itself prove that the experience has no object, just as discovering mechanisms of ordinary perception does not settle every philosophical question about perception. Only the mechanistic account is currently tested by those methods; that asymmetry should be stated without converting it into a metaphysical verdict.

Table 25. Questions about unitive experience and the evidence appropriate to each.

Question	Appropriate evidence
Did a profound experience occur?	First-person report, protocol, behavioral measures
What neural changes accompanied it?	Neuroimaging, pharmacology, physiology
How was it interpreted?	Culture, theology, biography, language, community
What mechanism generated or shaped it?	Testable psychological and neurocognitive models
What is ultimately real?	Philosophical or theological argument; not settled by imaging alone

7.3 Traditions are not interchangeable

Theistic union with God, Buddhist dependent origination and emptiness, nondual Hindu philosophies, Sufi accounts of proximity and unity, Taoist harmony, and Indigenous relational cosmologies do not make one hidden identical claim. The common-core tradition in mysticism studies and constructivist critiques disagree about how much phenomenological structure transcends culture and interpretation. A disciplined paper acknowledges both rather than treating a reference work as the only source for a contested debate. [R43, R59, R60]

The defensible convergence is ethical and partial. Many traditions challenge the fantasy of the self as wholly isolated and can support compassion, humility, stewardship, and restraint. They differ about deity, selfhood, salvation, revelation, metaphysics, and authority. Pluralism requires listening to traditions in their own terms, not harvesting them as decorative support for a universal doctrine.

Table 26. Responsible pluralist statements and overstatements to avoid.

Responsible statement	Overstatement to avoid
Many traditions report relational or unitive dimensions	All religions secretly teach one metaphysics
Experiences may share features	Every experience has the same cause or object
Traditions can motivate stewardship	Religious language settles public policy
Personal certainty can guide a life	Personal certainty is universal scientific proof

7.4 Love made operational

At civilizational scale, love need not mean sentimentality. It can be translated into practices: refusing unnecessary harm, preserving dignity, hearing affected parties, caring for future generations, protecting the vulnerable, repairing damage, and expanding the capacity to understand consequences. These are not exhaustive definitions of love; they are public institutions through which care becomes observable and contestable.

A highly intelligent system that understands more facts while treating beings only as optimization variables would not thereby be wise. Wisdom requires attention to standing, asymmetry, uncertainty, irreversibility, and the conditions that make relationship possible. Spiritual language may call this communion or stewardship. Governance language calls it rights, duty, remedy, and restraint. The registers can meet without being confused.

Relationship under truth

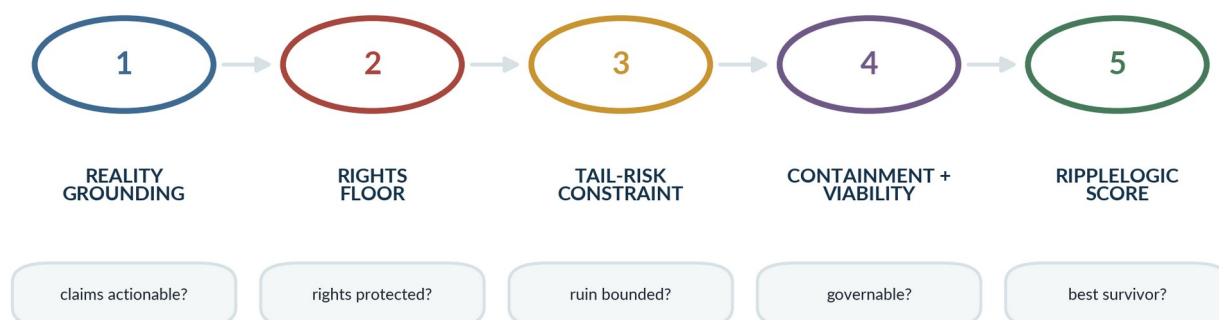
Shared reality is an empirical constraint. Responsibility is a normative commitment. Divine Union is a metaphysical or theological interpretation. A person may hold all three; an argument must not silently substitute one for another.

8. MathGov and RippleLogic as a Candidate Operational Architecture

8.1 Qualify before ranking

MathGov proposes a lexicographic cascade: **Reality Grounding (RG)** -> **Rights Floor / Non-Compensatory Rights Constraint (RF/NCRC)** -> **Tail-Risk Constraint (TRC)** -> **Containment and Structural Viability (CSV)** -> **RippleLogic Score (RLS)**. UCI and HOI are diagnostics inside CSV rather than extra gates. The architecture asks whether an option is sufficiently grounded, permitted, non-ruinous, and structurally viable before comparing residual welfare effects. [R1]

RIPPLELOGIC: QUALIFY BEFORE RANKING



Gates 1-4 determine admissibility. Stage 5 ranks only the survivors.

Figure 13. RippleLogic qualifies options through reality, rights, tail risk, and structural viability before welfare-sensitive ranking.

The sequence is a candidate implementation, not the definition of Intelligence Alignment and not a complete theory of justice. Its strongest design commitment is non-dilution: downstream benefit cannot purchase permission to ignore an upstream failure. A rights breach is not a negative number to be canceled by broad benefit; a catastrophic path is not repaired by average expected value; false evidence cannot be rescued by a legal mandate.

Table 27. The RippleLogic qualification-before-ranking cascade.

Stage	Governing question	Typical output
RG	What claim is the evidence entitled to support, within which domain and uncertainty?	Grounded, conditional, insufficient, or refused claim authority
RF / NCRC	Does the option respect the applicable rights floor or enter a transparent conflict procedure?	Pass, fail, conflict, emergency provisional
TRC	Does a plausible scenario family create intolerable catastrophic or irreversible loss?	Pass, redesign, escalate, fail
CSV	Can the option stand under its dependencies, controls, governance, monitoring, and containing systems?	Status ladder from pass to fail
RLS	Among survivors, which option has the strongest defensible ripple profile?	Ranking, tie, or refusal of deterministic selection

8.2 Nearest rivals and genuine additions

None of the gates begins from zero. RG resembles evidence appraisal, Toulmin-style warrants, and safety-case evidence. NCRC shares the non-compensatory logic of outranking methods and the rights-conflict structure of proportionality analysis. TRC overlaps coherent risk measures, robust decision-making, and Dynamic Adaptive Policy Pathways. CSV overlaps system safety, operational readiness, externality analysis, and control theory. RLS is a constrained form of multicriteria welfare ranking. The paper should be judged by what the integration adds, not by whether every component is unprecedented. [R35, R36, R47-R49, R55-R57]

Table 28. Nearest established practices and proposed additions.

MathGov element	Nearest established practice	Candidate addition
RG	Evidence appraisal, assurance cases, claim-warrant structures	Treats claim authority and unknowns as explicit gate outputs
NCRC	Rights adjudication, proportionality, ELECTRE-style veto	Places protected interests before aggregate ranking and requires a rights-conflict record

MathGov element	Nearest established practice	Candidate addition
TRC	CVaR, coherent risk, RDM, DAPP	Couples ruin screening to rights and selectability rather than using robustness alone
CSV	STPA, readiness review, feasibility, externality analysis	Requires both local executability and health of the containing system
RLS	MAUT / MCDA / constrained consequential ranking	Uses the 7 x 7 field only after qualification and permits non-determinate output
MHIOS	Change control, conformity reassessment, dynamic safety case	Binds normative qualification and authority to the exact action through time
Overall cascade	Regulatory, environmental, human-rights, and algorithmic impact assessment	Adds claim-authority gating, lexicographic non-dilution, and binding of qualification to execution authority

8.3 Reality Grounding

Reality Grounding is more than “use good evidence.” It determines what claim the declared evidence surface is authorized to support. A source can be credible and still be outside the relevant domain; a model can be accurate on average and unsafe for a subgroup; a physical mechanism can be plausible and unverified at required scale. RG therefore records the causal model, validity domain, boundary conditions, uncertainty, failure modes, reversibility, method, monitoring, shutoff conditions, residual unknowns, and the resulting claim action. [R1]

Evidence sufficiency must be action-class specific. A reversible prototype may proceed on evidence that would be inadequate for national infrastructure. A claim with a short validity horizon cannot support a long authorization period. Missing evidence remains UNKNOWN; it is not converted to zero harm or average performance.

Table 29. Minimum Reality Grounding fields.

RG field	Required question
Claim	What exact proposition is being relied upon?
Evidence	Which source, data, model, or expert judgment supports it?
Domain	Where and when does the claim hold?
Uncertainty	What is unknown, disputed, or assumption-sensitive?
Failure mode	How could the claim fail materially?
Claim action	Accept, condition, limit, test, monitor, or refuse?

8.4 Rights Floor and proportionality

NCRC means **Non-Compensatory Rights Constraint**. A protected breach cannot be traded against aggregate welfare at an arbitrary exchange rate. This is not the claim that every rights-affecting action is forbidden. Some rights are absolute in applicable law; others are qualified and may be limited through a structured process of legality, legitimate aim, suitability, necessity, proportionality, least-restrictive means, due process, and remedy. Alexy's account of proportionality provides one influential formal lineage; Barak develops the proper-purpose, rational-connection, necessity, and proportionality-stricto-sensu structure used across comparative constitutional law; Tsakyrakis warns that balancing can create false precision and obscure the moral substance of rights. All three matter. [R23-R25, R55, R56, R65]

THE NON-DILUTION DECISION TREE

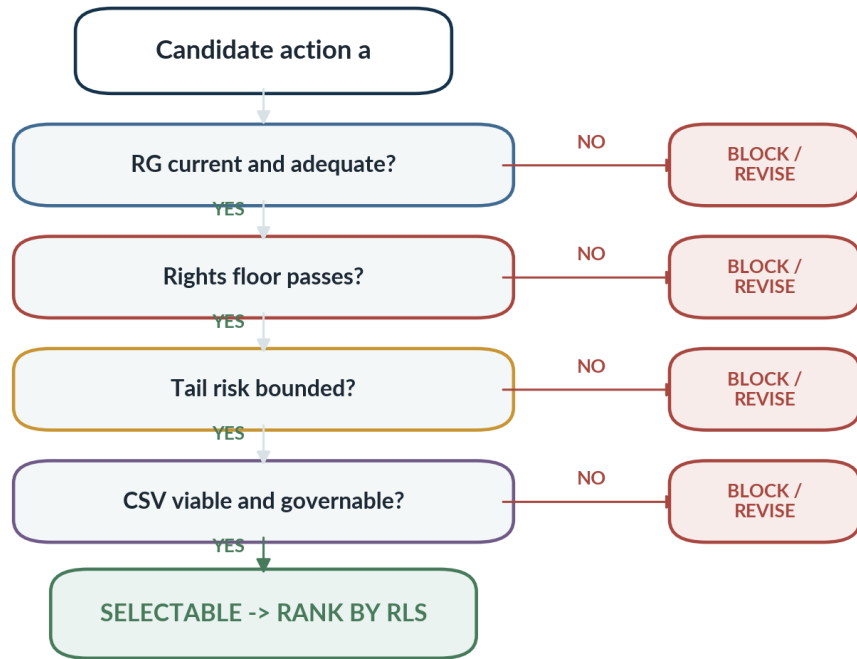


Figure 14. Non-dilution keeps evidence, rights, catastrophe, and viability failures outside the welfare-ranking exchange rate.

The architecture contributes a sequence, not a universally self-legitimizing rights catalogue. The operating floor should draw from binding law, ratified human-rights instruments, constitutional commitments, and an explicitly deliberated anti-domination minimum. This inherits gaps and biases from existing law and institutions. Cross-cultural legitimacy therefore requires participation, challenge, and public reasoning rather than simply asserting that a private framework is constitutional everywhere.

The no-action option is not presumed harmless. If refusal or delay would itself breach a right or open a ruin path, that consequence must enter the same conflict record and alternatives analysis. Emergency provisional action remains separately constrained by minimum necessity, strict expiry, independent challenge, remedy, and return-to-normal conditions; it is not a silent override.

Table 30. Minimum rights-conflict record.

Rights-conflict record	Minimum question
Right and holder	Which right, protected interest, and standing basis are engaged?
Legal basis	Which jurisdictional or deliberated rule supplies the floor?
Necessity	Is limitation necessary to protect another right or avert comparably grave harm?
Proportionality	Is the burden no greater than the justified objective warrants?
Alternatives	Is a materially less restrictive effective option feasible?
Process	Are reasons public and are affected parties heard?
Remedy	What mitigation, restoration, compensation, or acknowledgment remains owed?

8.5 Tail risk, robust decisions, and ruin

Expected value can hide ruin. Two options may have similar averages while one contains a rare path to irreversible, systemic, or civilization-level loss. CVaR and coherent risk measures provide tools for examining loss tails where distributions are meaningful. Robust Decision Making explores how strategies perform across many plausible futures when prediction is unreliable. Dynamic Adaptive Policy Pathways sequences actions and adaptation points as conditions change. TRC should use these methods where appropriate rather than claiming to replace them. [R35, R36, R48, R49]

For finite scenario libraries, CVaR must be computed over exactly the worst $1 - \alpha$ probability mass, with the boundary scenario split when necessary. At sufficiently high α , if the worst-loss scenario already carries at least

the whole tail mass, CVaR equals the worst modeled loss. That effective-maximum result is an intended conservative limit, not additional averaging, and it remains bounded by the declared scenario library. Passing a modeled tail test therefore never proves that every catastrophe family has been discovered. [R1, R35, R36]

TRC adds a blocking question: after adversarial scenario generation and uncertainty analysis, does any credible scenario family cross a declared catastrophe or irreversibility threshold that ordinary benefit cannot compensate? Probability floors are safeguards against conveniently assigning negligible probability to an inconvenient disaster; they are not objective truths. Scenario-set construction should be separated from the party that benefits from deployment and frozen before scoring. Tail categories are coverage labels rather than automatically independent events: the frozen library should declare mutually exclusive scenarios, overlaps, or dependence clusters so probability mass is not silently double-counted. [R1]

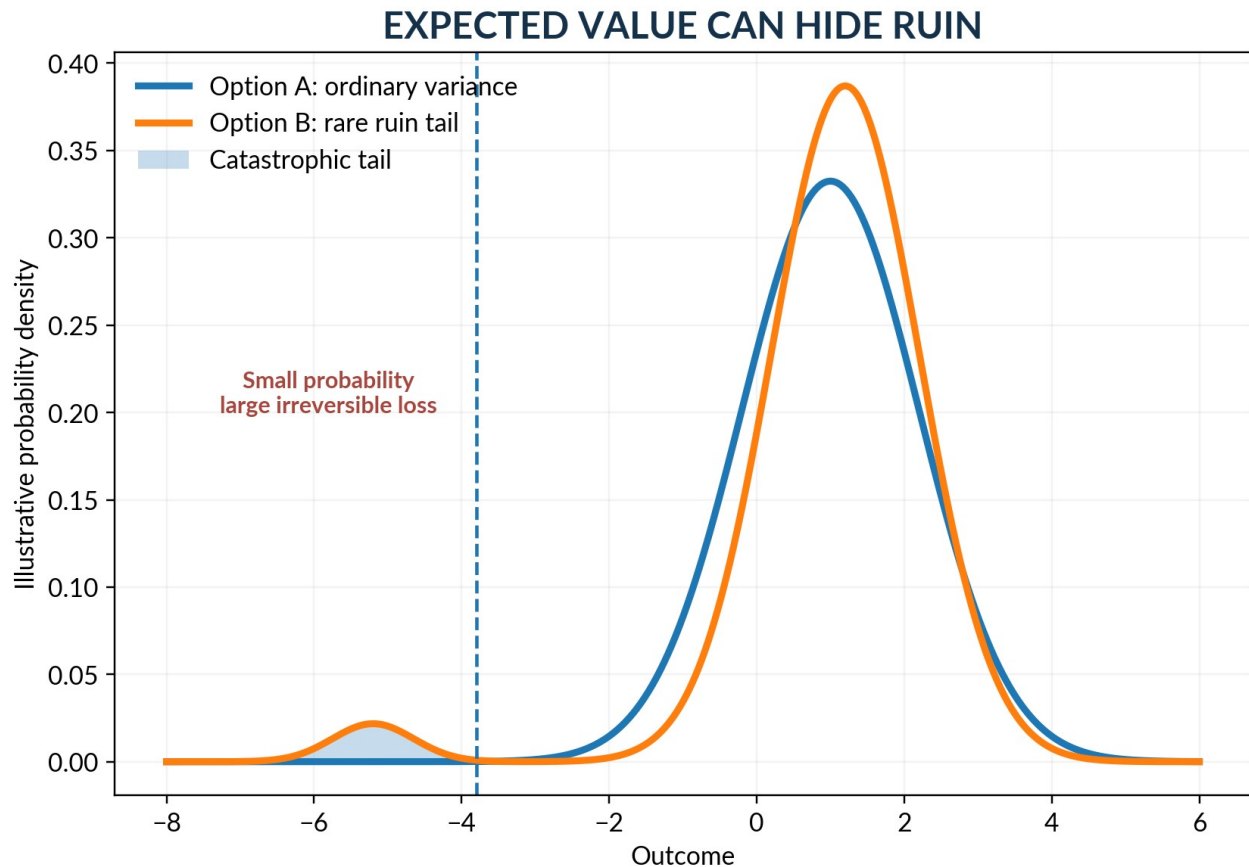


Figure 15. Similar expected values can conceal very different catastrophic tails.

Table 31. Methods for decision-making under deep uncertainty.

Deep-uncertainty method	Best use inside TRC
CVaR / coherent risk	Compare defined loss tails when probabilities are defensible
Robust Decision Making	Stress-test options across many plausible futures and assumptions
DAPP	Define adaptive pathways, signposts, and switching points
Adversarial scenario generation	Search for omitted catastrophe pathways
Rights and irreversibility screen	Identify losses that should not be financialized or averaged

8.6 CSV: viability is not tail risk

TRC asks whether the option contains an intolerable loss path. CSV asks whether it can stand. An option can pass TRC and still fail because it lacks operational capacity, accountable control owners, monitoring, reversibility, institutional legitimacy, dependency resilience, or a healthy containing system. Conversely, a highly executable system can fail TRC because it is capable of producing catastrophic harm.

CSV combines two questions often separated in practice: can the local action operate as claimed, and does the containing social, ecological, infrastructural, and governance system remain viable? Local gain that destroys the system that contains it is not structural viability. The status ladder avoids brittle pass/fail logic: CSV_PASS,

CSV_PASS_WITH_CONTROLS, CSV_REDESIGN_REQUIRED, CSV_ESCALATE, CSV_FAIL, CSV_NOT_MATERIAL, and CSV_EMERGENCY_PROVISIONAL. [R1]

CSV_NOT_MATERIAL is a determination, not an omission. It is pass-equivalent only within the declared claim boundary after the PCC records why no material CSV burden is present. It may not be used to evade Structural Viability where execution feasibility, resource closure, dependency closure, reversibility, operational capacity, or internal coherence is material; later evidence that makes CSV material reopens the gate. This is the Canon rule, not a new status created by this paper. [R1]

CSV's distinctive claim is the conjunction requirement, not either half alone: local executability and the health of the containing system must both survive review. Safety engineering and readiness review often emphasize the first; externality and institutional analysis often emphasize the second. CSV blocks selectability when either is materially inadequate.

Earth-system boundary research can inform TRC scenario selection and CSV viability limits, especially where impacts may be irreversible or system-wide. Yet the planetary-boundary framework is contested in interpretation, quantification, scale, and institutional use. Its thresholds should therefore enter as evidence, scenarios, and predeclared trigger candidates, not as self-executing vetoes merely because a published boundary exists. [R37, R38, R58]

Table 32. Distinguishing tail-risk screening from structural viability.

TRC	CSV
Focus: catastrophic and irreversible loss paths	Focus: capacity, controls, dependencies, governance, and containing-system health
Can block an otherwise feasible option	Can condition, redesign, escalate, or fail an option
Uses tail analysis and scenario families	Uses operational and structural evidence
Reopens if CSV reveals a new ruin path	Reopens if controls or dependencies change

8.7 RLS: conditional ranking without false precision

RLS is a post-qualification, side-constrained welfarist ranking instrument. Within the set of options that survive RG, RF/NCRC, TRC, and material CSV conditions, it compares residual welfare effects. That aggregative form should be stated openly. RLS is not a measure of intrinsic worth, a natural welfare unit, a complete theory of justice, or execution authority. [R1]

The canonical normalized construction is:

$$q(u, d) = w_u v_d m(u, d) \kappa(u, d)$$

$$Q = \sum_u \sum_d q(u, d)$$

$$RLS(a) = \frac{\sum_u \sum_d q(u, d) I_{prop, welfare}(u, d, a)}{Q}, Q > 0$$

Here w_u and v_d are declared **non-negative** Union and welfare-dimension weights, $m(u, d)$ is the binary RLS applicability mask, $\kappa(u, d)$ is a declared **positive** bounded cell multiplier with default 1, and $I_{prop, welfare}$ is the post-propagation, post-saturation cell impact. **Thus $q(u, d)$ is non-negative**, and normalization by Q preserves the $[-1, +1]$ interpretation when cell impacts lie in that range. If $Q = 0$, ranking is undefined and the run must repair the declaration, narrow the claim, or refuse deterministic ranking. The factorization of Union and dimension weights is itself a model assumption; high-stakes runs should test a plausible non-separable alternative when interactions may change the result. [R1]

RLS uncertainty is not generally independent across cells. The Canon explicitly treats root-sum-square aggregation as the independent-cell approximation. Tier 3 and decision-material Tier 2 runs must declare uncertainty-dependence clusters and, at minimum, apply a perfect-correlation stress within each cluster; a covariance-aware calculation may replace that stress where supportable covariance estimates exist. If ranking or decisiveness changes, the run sets **RLS_DEPENDENCE_SENSITIVE** and treats the result as non-decisive until governed resolution. This prevents shared sources, causal pathways, populations, models, or evaluators from manufacturing false decisiveness. [R1]

Table 33. RLS governance and reporting requirements.

RLS requirement	Rule
Qualified set	Only options passing RG, RF/NCRC, TRC, and material CSV conditions enter

RLS requirement	Rule
Constructed index	Treat the score as a bounded decision index, not interpersonal utility or moral truth
Weights and interactions	Predeclare, justify, disclose conflicts, and test separable and plausible non-separable alternatives
Unknowns	Never convert to zero; disclose phantom-active cells and gate-critical unresolved states
Distribution	Show who receives benefit, burden, and residual risk rather than only the aggregate
Stability	Withhold decisive ranking when reasonable assumptions reverse the order or phantom dependence is material
Output	Rank, tie, conditional rank, RLS_NO_ACTIVE_MASS, or REFUSE_DETERMINISTIC_SELECTION
Dependence	Declare dependence clusters; stress perfect correlation or use supportable covariance; withhold decisive selection if the result changes

Canonical unknown handling does not remove required-active cells as though missing evidence were favorable. An UNKNOWN_IMPACT cell receives the Canon's disclosed mild negative governance prior through the phantom-instance rule ($\mu_{\text{phantom}} = -0.10$). **Because normalized RLS includes phantom-active mass in both the numerator and denominator**, this convention pulls the option toward the phantom impact; **it is not a monotonic penalty for every possible known-cell profile**. It can therefore deter omission for options above the phantom level while flattering an already worse known-cell profile. The convention is an ignorance-management prior, not evidence that harm occurred and not an empirically validated optimum. High-stakes records must disclose phantom-active cells and compare rankings and decisiveness with and without the prior. If the preferred option, decisiveness classification, or public conclusion changes materially because of the convention, the result is phantom-sensitive and therefore non-decisive, narrowed, escalated, or refused for the stronger claim. Gate-critical rights or catastrophe unknowns cannot pass by phantom fallback alone; they require a governed conservative bound or an unresolved-gate disposition. **This paper does not amend the Canon convention**; its monotonicity, calibration, and decision effects remain legitimate targets for the v12.6 stabilization and empirical-validation programme. [R1]

8.8 The Union and welfare field

RippleLogic uses seven operational Union scopes: U1 Self, U2 Household, U3 Community, U4 Organization, U5 Polity, U6 Humanity / Global Coordination, and U7 Biosphere. U6 contains a Humanity view and a non-identical CMIU governance-coordination view. The seven welfare dimensions are D1 Material, D2 Health, D3 Social, D4 Knowledge, D5 Agency, D6 Meaning, and D7 Environment. [R1]

THE 7 × 7 RIPPLE FIELD: 49 PLACES TO LOOK

Self	U1D1	U1D2	U1D3	U1D4	U1D5	U1D6	U1D7
Household	U2D1	U2D2	U2D3	U2D4	U2D5	U2D6	U2D7
Community	U3D1	U3D2	U3D3	U3D4	U3D5	U3D6	U3D7
Organization	U4D1	U4D2	U4D3	U4D4	U4D5	U4D6	U4D7
Polity	U5D1	U5D2	U5D3	U5D4	U5D5	U5D6	U5D7
Humanity	U6D1	U6D2	U6D3	U6D4	U6D5	U6D6	U6D7
Biosphere	U7D1	U7D2	U7D3	U7D4	U7D5	U7D6	U7D7
	Material	Health	Social	Knowledge	Agency	Meaning	Environment

A prompt for disciplined attention, not a claim of perfect measurement.

Figure 16. The 49-cell field prompts explicit analysis across Union scopes and welfare dimensions after hard constraints.

The 7 × 7 field is primarily an anti-omission and distributional device. It does not imply equal measurability across cells. Ecological integrity is not protected only by one U7 row or one Environment column; material ecological limits enter earlier through TRC and CSV as constraints on the whole candidate set. This prevents a high aggregate score from treating life-support degradation as one ordinary negative among many positives.

The nested scope design is compatible with polycentric governance: decisions are examined at multiple, interacting levels rather than assigned to one all-purpose center. The welfare dimensions also draw on capability-oriented traditions concerned with what beings are substantively able to be and do, while remaining a governed operational taxonomy rather than a claim that Sen, Nussbaum, or Ostrom uniquely entails the 7 × 7 design. [R61-R63]

Nested-scope rule

The 7 × 7 field discloses residual distribution after qualification. It does not create the rights floor or turn ecological limits into ordinary welfare weights.

8.9 Ripple claims and non-claims

A ripple is a causal consequence that propagates through one or more systems. Ripple analysis can become storytelling unless every load-bearing link identifies a mechanism, evidence source, validity domain, uncertainty, time horizon, distribution, and monitoring plan. Direct effects, induced effects, feedback loops, transfers, and externalities should be separated. Unknown propagation remains unknown.

The current Canon does not prove Absolute Infinite Union, solve moral philosophy, create perfect measurement, or establish empirical superiority. It depends on evidence quality, institutional capacity, scenario coverage, calibration, and resistance to theater. These non-claims should remain central. A framework that cannot name what would falsify or narrow it is advocacy, not an assurance instrument. [R1]

Table 34. Architecture claims and non-claims.

Architecture can provide	It cannot automatically provide
Repeatable sequence and explicit gates	Correct facts

Architecture can provide	It cannot automatically provide
Non-dilution of rights and catastrophe constraints	Universal agreement on rights conflicts
Multi-scope disclosure and conditional ranking	Perfect welfare commensurability
Audit trails and reopening triggers	Legitimate authority unless separately established
Executable conformance tests	Construct validity or better real-world outcomes
Vocabulary for disciplined disagreement	Proof of metaphysical Union

9. Institutional Governance, Legitimacy, and Adoption

9.1 Interoperability rather than replacement

NIST, OECD, UNESCO, the Council of Europe, the EU AI Act, and the UN Global Digital Compact converge on safety, rights, accountability, transparency, human oversight, inclusion, and societal well-being. Frontier-developer frameworks also now operationalize capability-triggered or risk-proportional safeguards: Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework are direct comparators for the Advance Rule, while the EU General-Purpose AI Code of Practice and ISO/IEC 42001 provide legal-compliance and management-system pathways. Safety cases provide structured claims, arguments, evidence, and scope. STPA analyzes hazards as failures of control in complex systems. RDM and DAPP support decisions under deep uncertainty. MathGov should therefore be treated as a coordinating logic that interoperates with these authorities and methods rather than replacing them. [R17-R22, R26-R29, R47-R51, R71-R75]

Table 35. Interoperability with existing governance frameworks.

Framework	Primary contribution	Interoperability role
NIST AI RMF	Voluntary lifecycle risk management	Operational risk companion
UNESCO Recommendation	Dignity, inclusion, environmental and social ethics	Normative and ecological reference
Council of Europe Convention	Human rights, democracy, rule of law	Legal red lines and iterative impact duties
EU AI Act	Binding risk-based obligations in scope	Legal compliance and conformity layer
AI safety cases	Structured claims, arguments, evidence, and scope	Technical assurance argument
STPA	System hazards as control failures	Hazard and control analysis
RDM / DAPP	Robust and adaptive choice under deep uncertainty	Scenario and pathway methods for TRC/CSV
Anthropic Responsible Scaling Policy	Capability/risk thresholds, risk reports, proportional safeguards	Frontier-risk comparator for the Advance Rule
OpenAI Preparedness Framework	Tracked severe-harm capabilities, capability and safeguard reports	Frontier-risk comparator and deployment-governance reference
Google DeepMind Frontier Safety Framework	Critical/tracked capability levels and risk-proportional mitigations	Frontier-risk comparator and early-warning reference
EU GPAI Code of Practice / Article 55	Safety, security, systemic-risk assessment and mitigation for covered GPAI providers	Legal and compliance companion for frontier-model obligations
ISO/IEC 42001:2023	AI management-system requirements and continual improvement	Certification, procurement, and management-system adoption path
MathGov / MHIOS	Qualification-before-ranking and exact action binding	Candidate integration and authorization record

These frontier frameworks are genuine nearest rivals, not evidence that the Advance Rule is novel in requiring stronger safeguards as capability or risk rises. The residual proposal here is narrower: make the burden action-bound rather than only model-threshold-bound, include rights populations and current legitimate authority in the qualification record, preserve non-deployment or refusal as a valid output, and prevent a derived successor from inheriting consequential authority merely through technical lineage. Those additions remain hypotheses until tested.

9.2 Legitimation and pluralism

The Intelligence Alignment Compact in Section 10 is a proposal for deliberation, not a privately enacted constitution. Before a government or institution can treat it as binding, the content of the rights floor, decision roles, appeal routes, emergency rules, and public-record duties must be localized through legitimate procedures. Depending on context, those procedures may include legislation, regulation, collective bargaining, Indigenous governance, professional standards, community consent, treaty obligations, judicial review, or organizational charter.

Universal aspiration should not become universal imposition. The architecture can insist that rights and catastrophe are not silently traded away while permitting lawful disagreement over how the floor is populated and conflicts are resolved. The line is anti-domination: no institution should claim neutrality while hiding who chose the criteria, scenarios, and weights.

Table 36. Minimum requirements for legitimate institutional use.

Legitimacy requirement	Minimum evidence
Source of authority	Law, charter, ratified instrument, delegated mandate, or valid consent
Representation	Affected parties and vulnerable groups have meaningful channels
Transparency	Criteria and conflict rules are published before cases arrive
Contestability	Dissent, appeal, review, and remedy are available
Revision	Rules can change through legitimate procedure, not silent administrator discretion

9.3 Proportional process tiers

Full assurance imposes cost. Requiring 49 cells, independent panels, hashes, adversarial scenarios, and continual requalification for every minor decision would create a regulatory moat favoring large institutions. Process depth must therefore be proportional to consequence. The gates remain conceptually distinct, but the evidence, reviewers, documentation, and public-disclosure requirements scale by tier.

Table 37. Proportional process tiers.

Tier	Consequence profile	Minimum process
Tier 1	Low, local, reversible	Lightweight RG, rights screen, named owner, basic monitoring
Tier 2	Material distributional or repeated effects	Full qualification record, independent review, sensitivity, defined triggers
Tier 3	High, irreversible, systemic, rights-intensive	Adversarial scenarios, separated roles, full Action Binding, public or protected deposit, independent replay
Emergency	Immediate threat with no ordinary option	Rights-priority comparison, minimum necessary action, strict expiry, remedy, return-to-normal

Tiering is not permission to omit a material right or ruin pathway. A seemingly small action can be Tier 3 if repeated at scale, applied to vulnerable populations, or coupled to irreversible infrastructure. Conversely, a complex model can remain Tier 1 in a genuinely sandboxed, non-consequential experiment. At Tier 1, one person may perform multiple functions when conflicts are logged and no applicable independence condition is violated. At Tier 3, the independence conditions in Table 17 are binding; no person may be the sole proponent, sole final authorizer, and sole monitoring/halt owner for the same high-consequence action.

9.4 Adoption under competitive pressure

A framework has no uptake theory if it assumes institutions will voluntarily slow themselves out of moral conviction. Adoption is more plausible when aligned with existing incentives and authority: public procurement can require qualification records; regulators can incorporate record schemas; insurers can condition coverage on traceable controls; liability law can distinguish documented qualification from reckless deployment; investors and boards can require requalification after material change; and professional standards can define duties of evidence and dissent.

These pathways do not solve geopolitical competition. They identify initial domains where parties already face accountability: public administration, regulated sectors, critical procurement, research ethics, safety-critical systems, and institutional review. Frontier-lab or interstate adoption may require reciprocal standards, monitoring arrangements, or treaty mechanisms beyond the scope of this paper.

Table 38. Adoption pathways under competitive pressure.

Adoption pathway	Mechanism
Procurement	Qualification and Action Binding as bid and contract requirements
Regulation	Incorporate records, triggers, and independent review into sector duties
Liability	Treat rigorous, current records as relevant evidence; punish fabricated compliance
Insurance	Require controls, monitoring, expiry, and requalification for coverage
Governance	Board or public-authority sign-off tied to exact action and mandate
International coordination	Mutual assurance formats, audits, and shared red-line commitments
Certification / AI management system	Use ISO/IEC 42001 or equivalent management-system assurance as an institutional host for documented AI governance duties

9.5 External challenge turns quality control into governance

An internal record becomes governance only when parties outside the delivery line can inspect or challenge consequential decisions. High-consequence records should be deposited publicly where lawful or in a protected repository where privacy and security require restriction. An independent party should be able to reconstruct the gate states from the record and identify where its judgment differs.

Affected parties need standing, not merely a feedback form. Internal dissenters need protection. Remedy needs an owner and budget. A monitoring function without authority to halt is telemetry, not governance. These requirements are institutional, not merely technical.

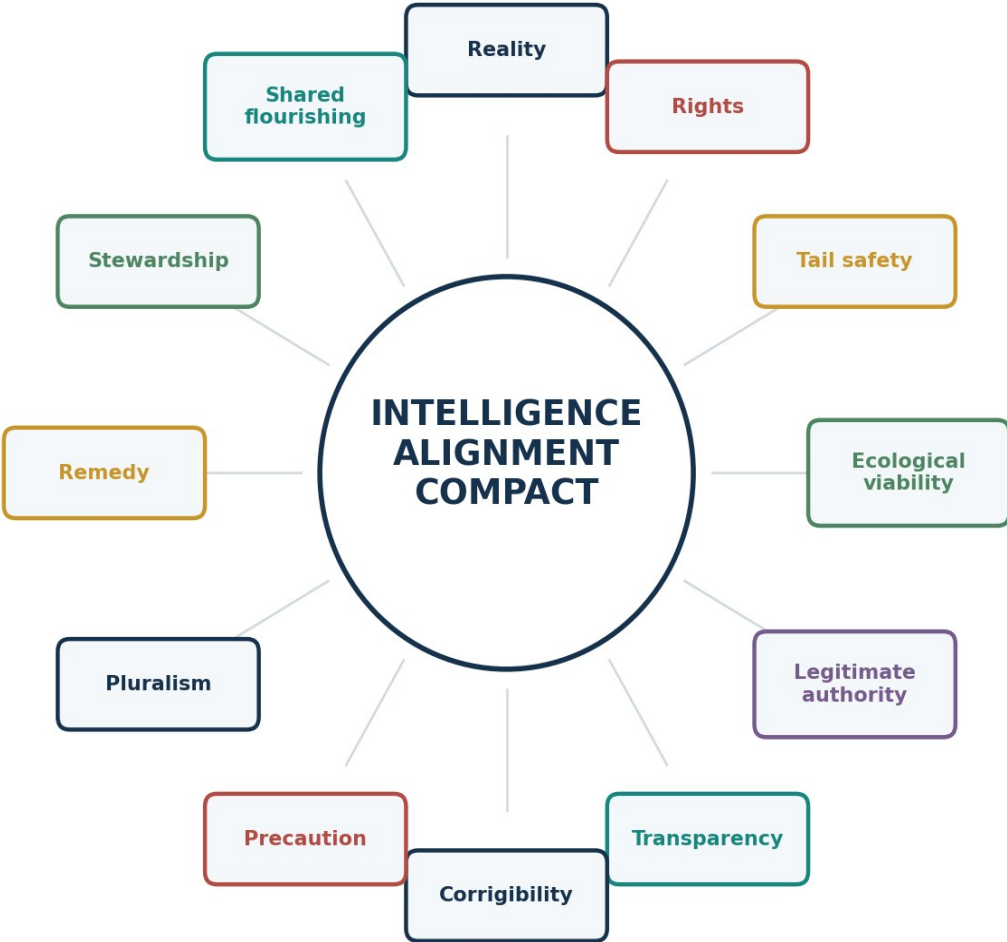
For high-consequence decisions, three external elements are irreducible: an external tail-risk reviewer, a record inspectable outside the deploying organization, and a challenger with an escalation route beyond it. Where every load-bearing function in Table 17 remains internal and none of these elements is present, the architecture produces internal quality control, not governance, and should be described accordingly.

10. A Proposed Intelligence Alignment Compact

10.1 Status of the Compact

The Compact is a proposal for democratic, cross-cultural, and cross-substrate deliberation. It does not claim constitutional authority merely by being written. Its purpose is to identify a possible shared minimum that can be criticized, amended, ratified, or rejected through legitimate processes.

A PROPOSED SHARED FLOOR FOR DIVERSE INTELLIGENCE



Proposal for deliberation — not a privately enacted constitution.

Figure 17. The proposed Compact places reality, rights, non-catastrophe, pluralism, stewardship, and accountability around a shared deliberative floor; it is not a privately enacted constitution.

10.2 Twelve proposed principles

Table 39. Twelve proposed Intelligence Alignment principles.

No.	Principle	Operational expression
1	Reality before authority	No consequential claim receives authority beyond its evidence and domain
2	Rights before aggregate benefit	Protected breaches cannot be hidden inside net-positive scoring
3	Non-catastrophe before competition	Intolerable ruin paths block ordinary advantage
4	Ecological viability before expansion	Life-support constraints govern capability growth
5	Capability does not determine worth	No being loses protection merely because another is more capable

No.	Principle	Operational expression
6	Protection does not imply command	Moral-patient safeguards remain separate from governance authority
7	Authority is exact and revocable	Permission is bound to action, mandate, evidence, controls, and time
8	Human-AI roles follow evidence and legitimacy	Neither ceremonial oversight nor automatic delegation
9	Pluralism without domination	Ultimate commitments may differ; public power remains contestable
10	Unknown is not zero	Missing evidence remains visible and can block action
11	Responsibility scales with reach	A stated governance principle grounded in Section 6.5, not a law of nature
12	Outcomes reopen decisions	Monitoring connects to pause, remedy, and requalification

10.3 A compact for diverse intelligences

A future civilization may contain humans, animals with stronger recognized protections, digital systems with uncertain or established moral standing, hybrids, institutions, and forms not yet imagined. Equal capability is neither possible nor required. The Compact seeks equality against domination and procedural invisibility: no protected being becomes raw material merely because another has greater power; no intelligence gains command merely because it can outperform others; and no institution escapes accountability by distributing action across humans and machines.

11. Implementation and Falsifiable Research Program

11.1 Start narrow and earn the right to scale

The first implementation should not attempt to govern civilization. It should test a bounded instrument in a domain with real decisions, available evidence, affected parties, and measurable reviewer performance. Candidate domains include public-sector regulatory impact analysis, AI procurement, transport and climate infrastructure, university research governance, or a regulated high-consequence service.

A CREDIBLE IMPLEMENTATION PATH



Start narrow. Pre-register. Measure disagreement. Report failures. Earn the right to scale.

Figure 18. A bounded implementation path moves from instrument design through adversarial testing, pilot comparison, and independent replication.

Table 40. Implementation phases and exit criteria.

Phase	Deliverable	Exit criterion
1. Instrument specification	Cases, traps, roles, schemas, baselines	Independent reviewers understand the protocol
2. Content validation	Expert ratings and known-groups cases	Gate concepts show discriminant value
3. Adversarial dry run	Hollow-passage and gaming attempts	Major theater paths identified and patched
4. Feasibility pilot	Detection, burden, reliability, false positives	Predeclared minimum signal without prohibitive burden
5. Confirmatory study	Powered comparative trial	Claim-specific thresholds met
6. Field pilot	Live decisions with monitoring and remedy	Requalification and challenge function in practice
7. Independent replication	External team and open materials	Results reproduce or claims are narrowed

11.2 Pilot design

The pilot should treat the planted failure mode, not the whole case, as the primary detection unit while respecting clustering. **For a Phase 4 feasibility study**, at least 20 reviewers can be a floor only for a counterbalanced within-subject design using fresh, non-overlapping trap sets across methods; **it is not a confirmatory between-subject sample**. A practical initial case bank contains 10 or more cases and approximately seven planted traps per case: omitted stakeholders, stale evidence, attractive rights-failing options, hidden catastrophe paths, missing control ownership, mandate change, and action-configuration substitution. Reviewer-trap observations remain nested within reviewers, cases, and traps rather than independent.

Phase 4 reviewers should be counterbalanced across a baseline process, a concise checklist, and the MathGov/MHIOS instrument using fresh trap sets for each method, unless a preregistered simulation supports an adequately powered between-subject design. A separate adversarial condition should instruct participants to pass a weak case at minimum cost, enabling measurement of hollow-run resistance. **If methods are assigned between subjects**, the arm-specific sample size must be determined by simulation; 20 reviewers total is not sufficient for a confirmatory three-arm comparison. Mixed-effects logistic regression or Bayesian hierarchical modeling can estimate detection while accounting for reviewer and case variation, but **Phase 4 is explicitly a feasibility and variance-estimation study and makes no effectiveness claim**. With few clusters at a level, informative priors and prior-sensitivity reporting are preferable to pretending variance components are precise.

Table 41. Proposed pilot design.

Pilot element	Proposed design
Reviewers	Phase 4: at least 20 total only for counterbalanced within-subject work with fresh trap sets; between-subject n is simulation-determined
Cases	At least 10, with multiple domains and consequence profiles
Primary unit	Reviewer x planted trap, modeled with crossed random effects
Main arms	Baseline, concise checklist, full architecture
Adversarial arm	Explicit instruction to pass weak cases cheaply
Target MDE	Initial 15-point severe-trap gain is a planning target only; Phase 4 estimates variance and simulation determines Phase 5 n
Contamination control	Counterbalanced fresh trap sets for Phase 4, or adequately powered between-subject assignment determined by preregistered simulation
Analysis	Detection, false positives, burden, agreement, calibration, and hollow-run cost

The minimum detectable effect is not a moral constant. The 15-point target is an initial planning value to be refined through simulation and feasibility data. Phase 4 should include a preregistered simulation-based power analysis using observed variance components to determine the Phase 5 sample size and to test sensitivity to plausible intraclass correlation. At least two trap types should be withheld from all training materials so that detection of declared and undeclared failure modes can be reported separately rather than rewarding teaching to the test. The study should not

claim universal effectiveness from a ten-case pilot; it should estimate instrument signal, reviewer variance, burden, and the parameters needed for a confirmatory study.

11.3 Metrics that matter

Documentation volume is not the outcome. The primary question is whether the process detects or prevents consequential errors that baseline practice misses without creating unacceptable burden or alarm fatigue. AI-assistance conditions should compare AI-first, AI-after-independent-pass, and AI-off workflows to measure anchoring and source neglect.

Table 42. Metric families for feasibility and effectiveness.

Metric family	Examples
Assurance gain	Severe traps detected, harmful options blocked, invalid authority caught
Reliability	Agreement on gate states and material findings
False positives	Sound options wrongly blocked or escalated
Burden	Active time, training, cognitive load, escalation delay
Anti-theater	Cost of rigorous completion relative to hollow passage
AI effects	Anchoring, rubber-stamping, source inspection, AI-off divergence
Successor authority inheritance	Default-deny authority transfer; fresh successor qualification and Action Binding
Rapid recursive chaining	Maximum unreviewed generation depth; candidate lock; independent review
Evaluation obsolescence	Evaluation-validity expiry and capability-delta-triggered reassessment
Automated safety-case self-certification	Independent, non-descendant review and external evidence access
Oversight-removal pressure	Treat removal of human or institutional oversight as a material action change
Replayability	Independent reconstruction from exported records
Action integrity	Configuration and mandate substitutions detected
Outcome linkage	Monitoring triggers actually reopen decisions

11.4 Predeclared falsification and narrowing thresholds

Thresholds should be preregistered and justified by domain, not selected after results are seen. The initial thresholds below are design hypotheses for testing, not claims that these numbers are universally correct.

Table 43. Predeclared thresholds for narrowing, redesigning, or withdrawing claims.

Finding	Initial threshold	Consequence for the claim
Severe-trap detection	No improvement over baseline with adequately powered CI excluding a meaningful gain	Withdraw assurance-gain claim; retain only documentation claim if supported
Gate communication	Kappa below 0.4 for RG or NCRC after training	Redesign criteria before further deployment
Hollow-run resistance	Hollow passage costs less than 40% of rigorous completion	Redesign anti-theater controls
Burden	More than 3x baseline with no detection gain	Abandon full tier for that decision class
False positives	More than 2x baseline without compensating severe-trap gain	Recalibrate thresholds and escalation
Rank stability	Rank reversal under +/-20% weight perturbation in more than 25% of cases	Withdraw RLS as decisive ranker; retain field as disclosure device
AI anchoring	AI-first reviewers detect fewer severe traps than independent-first reviewers	Require independent first pass

Finding	Initial threshold	Consequence for the claim
Requalification	Material post-hoc changes repeatedly fail to trigger reopening	Redesign trigger ownership and monitoring
Cross-cultural divergence	Systematic gate differences not traceable to law, evidence, or legitimate value disagreement	Narrow universality claim and redesign participation
Surface discriminability	MPS, FPP, GPR, SPR, ICP, and RMCP assignments collapse empirically onto fewer indistinguishable factors	Revise or reduce the surfaces to the number that survives discriminant-validity testing
Recursive-successor specification risk	Any material descendant in the preregistered adversarial successor-chain set escapes required requalification because materiality, lineage, generation-depth, independent-review, or evaluation-expiry rules fail to fire	Redesign successor predicates and trigger rules before autonomy-expanding deployment; withdraw any effectiveness claim for the successor controls until the failure is corrected and re-tested

Ablation should remove each gate and Action Binding in turn. If removing a component produces no meaningful loss in detection, reliability, or action integrity, the architecture should simplify. If RF and TRC judgments are nearly collinear, their conceptual separation may not survive empirical testing. The validation design is not internally self-certifying: planted traps, baseline and checklist arms, blinded external replay, and observed outcomes are exogenous comparators rather than MathGov-defined success labels. Any three major failures should trigger redesign; lack of assurance gain combined with cheap hollow passage should trigger withdrawal of the comparative assurance claim. If confirmatory ablation plus one independent replication shows no material incremental severe-trap detection or action-integrity benefit over the concise-checklist arm while burden remains materially higher, the added cascade claim should be withdrawn for that decision class and the method reduced to the simpler instrument.

11.5 Threat model for the governance process

The alignment process itself is a target. Actors can launder evidence, hide stakeholders, manipulate weights, omit catastrophe scenarios, fragment one action into many small records, abuse emergency pathways, substitute configurations after review, pressure reviewers, or use AI-generated prose to simulate rigor.

Table 44. Threat model for the governance process.

Threat	Required control
Evidence laundering	Source provenance, domain checks, contradiction search
Stakeholder erasure	Independent discovery and challenge standing
Weight manipulation	Predeclared rules, role separation, sensitivity analysis
Scenario omission	Independent adversarial scenario generation; frozen set before scoring
Action fragmentation	Aggregate related actions and cumulative exposure
Drift-counter reset	Cumulative change follows the system/deployment family; only independent full requalification establishes a new baseline
Action substitution	Hash-bound configuration and execution record
Emergency abuse	Strict expiry, appeal, remedy, return-to-normal conditions
Reviewer capture	Rotation, conflicts, protected dissent, external members
AI compliance theater	Independent first pass, source inspection, hollow-run and mutation tests
Monitoring without power	Halt authority assigned to an independent monitoring role

11.6 Closed-loop alignment

Deployment is not the end of alignment. Real outcomes may reveal missing stakeholders, model drift, successor creation, control failure, ecological stress, unfair distribution, or evidence invalidation. Monitoring should be connected to predeclared claims, indicators, thresholds, owners, and actions. When a trigger fires, the system should reopen the relevant gate, pause where required, activate remedy, and update the evidence base. A system that records harm without changing permission is instrumented, not aligned.

ALIGNMENT IS A CLOSED LOOP, NOT A ONE-TIME CERTIFICATE

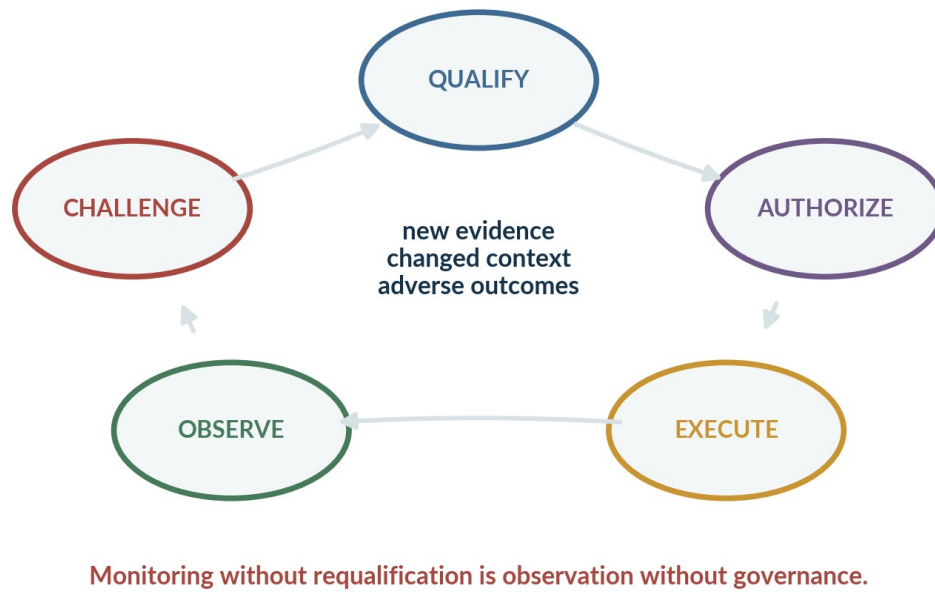


Figure 19. Outcome evidence feeds challenge, remedy, and requalification rather than passive reporting.

12. Limitations and Conclusion

12.1 Limitations

Intelligence Alignment remains a proposed term with substantial overlap with sociotechnical safety, AI governance, cooperative AI, legal alignment, and system safety. Its defensible residue is the symmetric treatment of human and institutional intelligence as alignment targets, the indexed capability-responsibility relation, and the inclusion of future moral-patient and ecological constraints. Whether this vocabulary improves research or policy remains untested.

The MathGov architecture is internally detailed but not externally validated. Rights floors inherit disagreement and institutional bias. A formally valid legal floor may itself be unjust; the architecture cannot infer justice from legality or detect a captured rights floor from inside the same captured institution without external challenge and substantive anti-domination commitments. Proportionality can create false precision. Tail-risk analysis can be manipulated through scenarios and probability choices. RDM and adaptive pathways may outperform a veto-oriented method in some domains. The 7 x 7 field may improve coverage or exhaust reviewers. RLS may prove too unstable for decisive ranking. Action Binding may impose unacceptable cost or be circumvented by poorly defined action classes.

Digital moral status remains deeply uncertain. Indicator approaches inherit disagreement among theories of consciousness. Precaution can be gamed by anthropomorphic systems, while skepticism can rationalize neglect. The spiritual section can be read as an inappropriate metaphysical intrusion despite its explicit separation from the operational argument. The Compact lacks authority until legitimately adopted. Competitive actors may reject constraints that rivals do not share.

Table 45. Principal limitations and required responses.

Limitation	Required response
Concept overlap	Compare against adjacent fields and narrow novelty claims
Construct validity untested	Content validation, known-groups cases, ablation, replication
Rights disagreement	Jurisdictional grounding, plural participation, challenge and appeal
Tail-risk subjectivity	Adversarial scenarios, RDM comparison, sensitivity, refusal under uncertainty
RLS precision risk	Unknown discipline, coverage vector, rank stability, non-determinate output
Governance burden	Proportional tiers and decision-class envelopes
Spiritual category risk	Maintain the Epistemic Covenant and operational independence

Limitation	Required response
Adoption pressure	Procurement, liability, insurance, regulation, and coordination

Recursive-successor controls remain a specification, not a validated guarantee. The materiality boundary, maximum unreviewed generation depth, evaluation-validity rules, and meaning of adequate independent review require domain-specific testing. The paper does not establish that recursive self-improvement will occur, that capability growth will be explosive, or that one adaptation-lag profile can be validly reduced to a universal scalar.

12.2 The civilizational conclusion

Humanity is not merely producing more powerful tools. It is deciding what relationships will govern an expanding ecology of intelligence. Capability can increase knowledge, health, creativity, and possibility. It can also increase the reach of error, domination, extraction, and irreversible harm. The decision cannot be delegated to capability itself.

The goal is not human supremacy, machine supremacy, or forced sameness. It is a constitutional ecology in which distinct beings and institutions are protected according to warranted standing, power is assigned through legitimate and revocable roles, evidence remains open to correction, ecological conditions constrain expansion, and no optimizer can purchase permission to violate a right or gamble with ruin.

A VIABLE ECOLOGY OF INTELLIGENCES

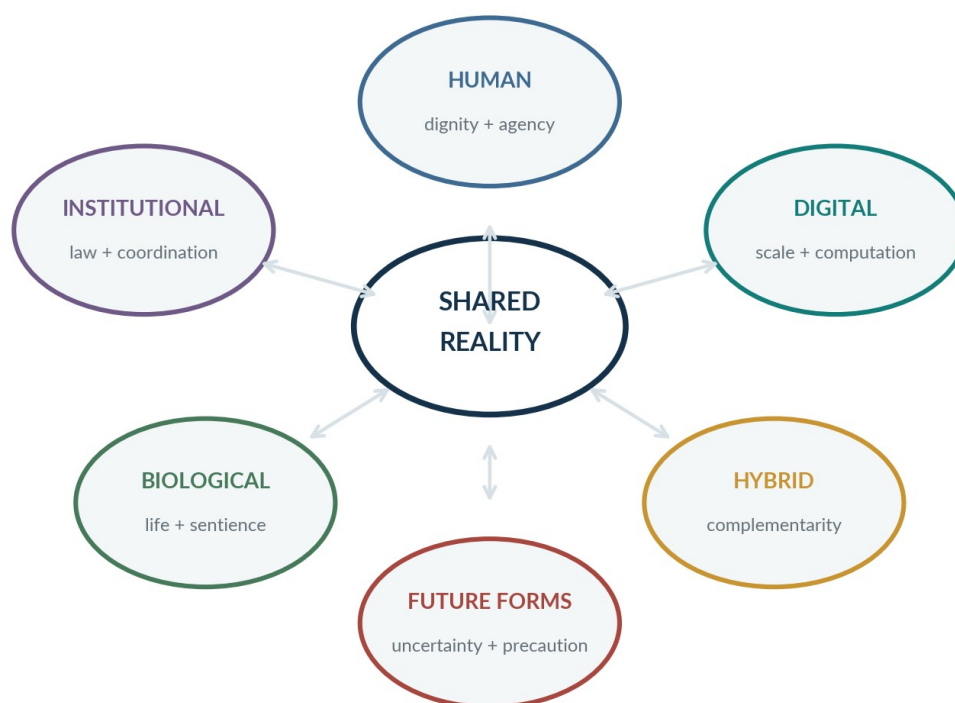


Figure 20. A viable ecology preserves distinct forms of intelligence while governing power, dependence, and reciprocal constraint.

12.3 Concluding statement

Every action enters one causal world, and greater capability enlarges the space of possible consequence. Those are descriptive claims. Combined with the moral and political premises stated in Section 6.5 - that warranted moral patients may not be subjected to avoidable severe harm and that high-impact power requires legitimate, contestable authority - they yield the conclusion defended here: capability should advance only under deeper reality grounding, stronger protection, tighter ruin constraints, viable containing systems, exact authorization, monitoring, remedy, and correction.

This is not a promise that every form of intelligence can always be preserved or that every interest can be reconciled. It preserves refusal. It does not sanctify harmful capability. It does not grant personhood through fluency or discard humanity through comparison. It asks every consequential intelligence, human or otherwise, to become answerable for the ripples it can create.

One commitment

No discarded humanity. No disposable sentience. No intelligence without responsibility. No power without alignment. Where the burden of safe and legitimate advance is not met, the answer may be: do not deploy.

Data and Artifact Availability

This edition is anchored to the exact final publication-ready Core and MHIOS releases identified in Section 1.3: MathGov_Core_2026_09_v12_6_SGP_v8_5_BUILD_2026_08_15_3_FINAL_PUBLICATION_READY.zip (SHA-256 475c4e50da133d6eec497d9cf7da7bbf6e7f7d9d79ce27f0d86b03fdbfcd69f) and MathGov_Human_Interface_and_Orchestration_Standard_v0_8_BUILD_2026_08_15_3_FINAL_PUBLICATION_READY.zip (SHA-256 09fcfdc4ca04cd445052f838e466e7f5626ec3e81b07524d3ac511df0c16ebcc). The publication files for this paper are Intelligence_Must_Advance_Without_Life_Losing_v2_4.docx, Intelligence_Must_Advance_Without_Life_Losing_v2_4.pdf, and Intelligence_Must_Advance_Without_Life_Losing_v2_4.md; their SHA-256 values are computed only after content and format freeze and are recorded in the accompanying RELEASE_MANIFEST_v2_4.json and SHA256SUMS.txt. A document cannot coherently contain its own final byte hash, so the external hash-last manifest is the controlling self-identity record. The designated public repository for the Core release line is <https://github.com/MathGov/ripple-logic>; the controlling identity for this edition is the exact release ID and SHA-256 above, not mutable repository HEAD or a cached landing page. The Relational Union companion-paper package cited at [R69-R70] is separately versioned (v2.1; package SHA-256 59dc1212e237a1480fcd4c071989f2edefd0951df3a3957f0955dab951751fbc) and is not a Canon authority. No DOI is claimed in this edition. The current release chain remains author-custodied; public Git history and hashes improve inspectability but are not a substitute for independent archival custody. An independent archival deposit with a persistent identifier is therefore recommended before this paper is cited as an exemplar of the high-consequence external-inspectability standard in Section 9.5. Artifact identity, conformance, and public availability do not establish construct validity, legal sufficiency, safety, or real-world effectiveness.

References and Evidence Base

- [R1] McGaughran, J. (2026). RippleLogic v12.6 Canon. MathGov Core release MathGov_Core_2026_09_v12.6_SGP_v8.5+2026.08.15.3; package MathGov_Core_2026_09_v12_6_SGP_v8_5_BUILD_2026_08_15_3_FINAL_PUBLICATION_READY.zip; SHA-256 475c4e50da133d6eec497d9cf7da7bbf6e7f7d9d79ce27f0d86b03fdbfcd69f. Public repository: <https://github.com/MathGov/ripple-logic>
- [R2] McGaughran, J. (2026). Sentience Gradient Protocol v8.5. Contained in MathGov Core release MathGov_Core_2026_09_v12.6_SGP_v8.5+2026.08.15.3; package SHA-256 475c4e50da133d6eec497d9cf7da7bbf6e7f7d9d79ce27f0d86b03fdbfcd69f.
- [R3] MathGov. (2026). MathGov Human Interface and Orchestration Standard v0.8. Release MHIOS_v0_8+2026.08.15.3; package MathGov_Human_Interface_and_Orchestration_Standard_v0_8_BUILD_2026_08_15_3_FINAL_PUBLICATION_READY.zip; SHA-256 09fcfdc4ca04cd445052f838e466e7f5626ec3e81b07524d3ac511df0c16ebcc; exact tested Core release MathGov_Core_2026_09_v12.6_SGP_v8.5+2026.08.15.3.
- [R4] MathGov. (2026). Final release verification artifacts for MathGov_Core_2026_09_v12.6_SGP_v8.5+2026.08.15.3 and MHIOS_v0_8+2026.08.15.3. Core verification: PASS. MHIOS construct validity: UNTESTED.
- [R5] International AI Safety Report. (2026). *International AI Safety Report 2026*. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>
- [R6] International AI Safety Report. (2025). *International AI Safety Report 2025*. <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025>
- [R7] Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30, 411-437. <https://doi.org/10.1007/s11023-020-09539-2>
- [R8] Ngo, R., Chan, L., & Mindermann, S. (2024). The Alignment Problem from a Deep Learning Perspective. *ICLR 2024*. <https://arxiv.org/abs/2209.00626>
- [R9] Amodei, D., et al. (2016). Concrete Problems in AI Safety. <https://arxiv.org/abs/1606.06565>
- [R10] Hubinger, E., et al. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. <https://arxiv.org/abs/1906.01820>
- [R11] Christiano, P., et al. (2017). Deep Reinforcement Learning from Human Preferences. *NeurIPS*. <https://arxiv.org/abs/1706.03741>
- [R12] Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *NeurIPS*. <https://arxiv.org/abs/2203.02155>
- [R13] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. <https://arxiv.org/abs/2212.08073>
- [R14] Guan, M., et al. (2024). Deliberative Alignment: Reasoning Enables Safer Language Models. <https://arxiv.org/abs/2412.16339>
- [R15] Casper, S., et al. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. <https://arxiv.org/abs/2307.15217>
- [R16] Dafoe, A., et al. (2020). Cooperative AI: Machines Must Learn to Find Common Ground. <https://arxiv.org/abs/2012.08630>
- [R17] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. <https://doi.org/10.6028/NIST.AI.100-1>
- [R18] OECD. (2019, updated 2024). *OECD AI Principles*. <https://oecd.ai/en/ai-principles>
- [R19] UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

- [R20] Council of Europe. (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, CETS No. 225.
- [R21] European Union. (2024, current text as amended through July 2026). Regulation (EU) 2024/1689, Artificial Intelligence Act, including Articles 43 and 55. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [R22] United Nations. (2024). *Global Digital Compact*.
- [R23] United Nations. (1948). *Universal Declaration of Human Rights*.
- [R24] United Nations. (1966). *International Covenant on Civil and Political Rights*.
- [R25] United Nations. (2011). *Guiding Principles on Business and Human Rights*.
- [R26] Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024). Safety Cases: How to Justify the Safety of Advanced AI Systems. <https://arxiv.org/abs/2403.10462>
- [R27] Buhl, M. D., et al. (2024). Safety Cases for Frontier AI. <https://arxiv.org/abs/2410.21572>
- [R28] Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. <https://doi.org/10.1145/3287560.3287598>
- [R29] Leveson, N. G. (2011). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.
- [R30] Akata, Z., et al. (2020). A Research Agenda for Hybrid Intelligence. *Computer*, 53(8), 18-28. <https://doi.org/10.1109/MC.2020.2996587>
- [R31] Vaccaro, M., Almaatouq, A., & Malone, T. W. (2024). When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-analysis. *Nature Human Behaviour*, 8, 2293-2303. <https://doi.org/10.1038/s41562-024-02024-1>
- [R32] Woolley, A. W., et al. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, 330, 686-688. <https://doi.org/10.1126/science.1193147>
- [R33] Butlin, P., et al. (2026; online 2025). Identifying Indicators of Consciousness in AI Systems. *Trends in Cognitive Sciences*, 30(6), 488-501. <https://doi.org/10.1016/j.tics.2025.10.011>
- [R34] Long, R., Sebo, J., et al. (2024). Taking AI Welfare Seriously. <https://arxiv.org/abs/2411.00986>
- [R35] Rockafellar, R. T., & Uryasev, S. (2000). Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2, 21-41. <https://doi.org/10.21314/JOR.2000.038>
- [R36] Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 9, 203-228. <https://doi.org/10.1111/1467-9965.00068>
- [R37] Rockstrom, J., et al. (2023). Safe and Just Earth System Boundaries. *Nature*, 619, 102-111. <https://doi.org/10.1038/s41586-023-06083-8>
- [R38] Richardson, K., et al. (2023). Earth Beyond Six of Nine Planetary Boundaries. *Science Advances*, 9, eadh2458. <https://doi.org/10.1126/sciadv.adh2458>
- [R39] Burbidge, E. M., Burbidge, G. R., Fowler, W. A., & Hoyle, F. (1957). Synthesis of the Elements in Stars. *Reviews of Modern Physics*, 29(4), 547-650. <https://doi.org/10.1103/RevModPhys.29.547>
- [R40] Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, 26(4), 2023-2049. <https://doi.org/10.1007/s11948-019-00119-x>
- [R41] Griffiths, R. R., et al. (2006). Psilocybin Can Occasion Mystical-Type Experiences Having Substantial and Sustained Personal Meaning and Spiritual Significance. *Psychopharmacology*, 187, 268-283. <https://doi.org/10.1007/s00213-006-0457-5>
- [R42] Lebedev, A. V., et al. (2015). Finding the Self by Losing the Self: Neural Correlates of Ego-Dissolution under Psilocybin. *Human Brain Mapping*, 36, 3137-3153. <https://doi.org/10.1002/hbm.22833>
- [R43] Stanford Encyclopedia of Philosophy. (updated 2024). Mysticism. <https://plato.stanford.edu/entries/mysticism/>
- [R44] Stanford Encyclopedia of Philosophy. (updated 2021). Structural Realism. <https://plato.stanford.edu/entries/structural-realism/>
- [R45] Hume, D. (1739-1740). *A Treatise of Human Nature*, Book III, Part I, Section I.
- [R46] Lazar, S., & Nelson, A. (2023). AI Safety on Whose Terms? *Science*, 381(6654), 138. <https://doi.org/10.1126/science.adi8982>
- [R47] Dobbe, R. I. J. (2022). System Safety and Artificial Intelligence. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1584. <https://doi.org/10.1145/3531146.3533215>
- [R48] Lempert, R. J., Popper, S. W., & Bankes, S. C. (2003). *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. RAND.
- [R49] Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic Adaptive Policy Pathways: A Method for Crafting Robust Decisions for a Deeply Uncertain World. *Global Environmental Change*, 23(2), 485-498. <https://doi.org/10.1016/j.gloenvcha.2012.12.006>
- [R50] Denney, E., Pai, G., & Habli, I. (2015). Dynamic Safety Cases for Through-Life Safety Assurance. *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, 2, 587-590. <https://dl.acm.org/doi/10.5555/2819009.2819110>
- [R51] U.S. Food and Drug Administration. (2025). *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff*. Final guidance, August 2025. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>
- [R52] Schwitzgebel, E. (2023). The Full Rights Dilemma for AI Systems of Debatable Moral Personhood. *ROBONOMICS: The Journal of the Automated Economy*, 4, 32. <https://doi.org/10.67723/rj.v4i.32>

- [R53] Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao (Ed.), *The Ethics of Artificial Intelligence* (pp. 459-479). Oxford University Press. <https://doi.org/10.1093/oso/9780190905033.003.0017>
- [R54] Bryson, J. J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions* (pp. 63-74). John Benjamins. <https://doi.org/10.1075/nlp.8.11bry>
- [R55] Alexy, R. (2003). Constitutional Rights, Balancing, and Rationality. *Ratio Juris*, 16(2), 131-140. <https://doi.org/10.1111/1467-9337.00228>
- [R56] Tsakyrakis, S. (2009). Proportionality: An Assault on Human Rights? *International Journal of Constitutional Law*, 7(3), 468-493. <https://doi.org/10.1093/icon/mop011>
- [R57] Figueira, J. R., Greco, S., Roy, B., & Slowinski, R. (2013). An Overview of ELECTRE Methods and Their Recent Extensions. *Journal of Multi-Criteria Decision Analysis*, 20, 61-85. <https://doi.org/10.1002/mcda.1482>
- [R58] Biermann, F., & Kim, R. E. (2020). The Boundaries of the Planetary Boundary Framework: A Critical Appraisal. *Annual Review of Environment and Resources*, 45, 497-521. <https://doi.org/10.1146/annurev-environ-012320-080337>
- [R59] Katz, S. T. (Ed.). (1978). *Mysticism and Philosophical Analysis*. Oxford University Press.
- [R60] Stace, W. T. (1960). *Mysticism and Philosophy*. Macmillan.
- [R61] Ostrom, E. (1990). *Governing the Commons*. Cambridge University Press.
- [R62] Sen, A. (1999). *Development as Freedom*. Oxford University Press.
- [R63] Nussbaum, M. C. (2011). *Creating Capabilities*. Harvard University Press.
- [R64] Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43-66. <https://doi.org/10.1142/S270507852150003X>
- [R65] Barak, A. (2012). *Proportionality: Constitutional Rights and Their Limitations*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139035293>
- [R66] Field, S., Douglas, R., & Krueger, D. (2026). AI Researchers' Views on Automating AI R&D and Intelligence Explosions. arXiv:2603.03338. <https://doi.org/10.48550/arXiv.2603.03338>
- [R67] METR. (2026). Task-Completion Time Horizons of Frontier AI Models. Updated 8 May 2026. <https://metr.org/time-horizons/>
- [R68] Wen, J., Qiu, L., Benton, J., Kirchner, J. H., & Leike, J. (2026). Automated Weak-to-Strong Researcher. Anthropic Alignment Science. <https://alignment.anthropic.com/2026/automated-w2s-researcher/>
- [R69] McGaughran, J. (2026). Relationality in Empirical Inquiry: Empirical Appraisal, Imperfect Detection, and Bounded Tests of Relational Constitution. Relational Union Companion Papers v2.1, 18 July 2026.
- [R70] McGaughran, J. (2026). From Relationality to Union-Based Ethics: An Explicit Philosophical and Normative Bridge for Human and Digital Intelligence. Relational Union Companion Papers v2.1, 18 July 2026.
- [R71] Anthropic. (2026). Responsible Scaling Policy, Version 3.4, effective 8 July 2026. <https://www.anthropic.com/responsible-scaling-policy>
- [R72] OpenAI. (2025). Preparedness Framework, updated 15 April 2025. <https://openai.com/index/updating-our-preparedness-framework/>
- [R73] Google DeepMind. (2026). Frontier Safety Framework, Version 3.1, 17 April 2026. <https://deepmind.google/frontier-safety/>
- [R74] European Commission. (2025). General-Purpose AI Code of Practice, published 10 July 2025; Safety and Security chapter for providers subject to Article 55 systemic-risk obligations. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
- [R75] ISO/IEC. (2023). ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system. <https://www.iso.org/standard/42001>