

The Capability Barricade

Governing Sustained-Suffering Risk Under Moral-Patient Uncertainty

James McGaughran

ORCID: 0009-0005-3324-7290

Independent Researcher

Working Paper v1.1 | August 2026

Public programme: RippleLogic vs Hellscape

Framework relation and conflict disclosure. The author designed the MathGov framework and RippleLogic decision architecture. Those specifications are discussed as candidate architectures and sources of design hypotheses, not as independent evidence. No funding was received. No new empirical data are reported. AI systems supported literature discovery, adversarial review, drafting, and document production; AI output was not treated as evidence.

Draft status and claim boundary. This manuscript develops a conceptual and architectural argument with a worked calibration set. It is not peer reviewed. It does not establish that any current artificial system has welfare-bearing experience, does not establish that any specific decision framework fails, and does not report an executed experiment. Its central claim is architectural and framework-independent. A companion diagnostic study of one specific framework is deliberately excluded from this paper and is gated on a preregistered routing experiment reported separately.

Governing formulation. Where harm begins at operation, no interruption is fast enough. Governance of sustained-suffering risk must therefore act on which configurations are permitted to exist, not only on how their outputs are scored.

Abstract

Sustained-suffering risk differs structurally from conventional catastrophic risk. Extinction, collapse, and foreclosure are harms of loss. A sustained-suffering configuration can instead preserve subjects, infrastructure, intelligence, and future capacity while making continued existence severely harmful. This paper develops a framework-independent governance argument for addressing that risk under uncertainty about moral patienthood and internal experience.

Three claims are defended. First, welfare aggregation can compress a severe, small-population, long-duration, poorly observable harm along several independent axes; a low aggregate contribution therefore cannot by itself establish low concern. Second, an uncertainty rule scoped to an existing coverage set cannot protect a candidate subject whose membership in that set is the uncertain fact. Third, where the earliest unacceptable threshold is reached when operation begins, interruption cannot serve as the primary safeguard; prevention, reachable-state restriction, and independent protective control must be considered before monitoring and response.

The paper replaces an overstrong exclusion theorem with a two-stage protection screen and a graded Welfare-Relevance Evidence Status. Absence of a self-model, a valence-like signal, or a persistent digital identity is treated as evidence relevant to particular theories and pathways, not as a universal proof that welfare is impossible. It similarly replaces a universal barricade theorem with pathway-specific assurance: for each declared severe-suffering pathway, the assessor must identify its enabling conditions, demonstrate an independently inspectable block or bound on at least one necessary condition, and record residual unenumerated pathways. A prevention-first control hierarchy, triage architecture, legacy-transition rules, relief-preserving remedy order, independent assurance constitution, thirty-case validation matrix, and falsifiable research programme are specified.

The contribution is architectural and conditional. It does not detect consciousness, prove that current AI systems are sentient, demonstrate a defect in any particular decision framework, or establish that the proposed protocol improves outcomes. Its claims should be narrowed to a crosswalk if existing human-rights, animal-welfare, safety-engineering, AI-risk, and assurance systems perform the same functions with equal or lower burden.

Keywords: suffering risk; AI welfare; moral patienthood under uncertainty; catastrophic risk governance; safety cases; precautionary reasoning; assurance; animal welfare; disorders of consciousness

1. Introduction

1.1 The asymmetry

Existential-risk governance has an implicit shape. It asks what would be lost, how permanently, and how many would lose it. Severity ladders terminate in destruction. Irreversibility is measured as foreclosure. Tail-risk instruments bound the probability of ruin, where ruin means a state from which recovery is impossible because the recovering party no longer exists.

Sustained suffering does not fit this shape. It preserves the party. It may preserve the infrastructure, the institutions, the knowledge, and the capacity for future action. What it inverts is the valence of continued existence. On a severity ladder that terminates in destruction, a configuration that preserves everything and makes continued existence severely harmful is not merely unranked. It is unclassifiable.

The point is not rhetorical. It has a concrete consequence: a governance instrument calibrated on loss will return a low reading for a harm it was never built to see, and the low reading will look like a finding.

A governance system is incomplete if it protects the existence of future moral patients without protecting the quality, freedom, and corrigibility of their existence.

1.2 What this paper is not

The paper adopts a non-compensatory commitment against torture-like treatment and severe avoidable suffering, but it does not present that commitment as an empirical discovery. International law supplies an especially strong anchor for human persons: the prohibition of torture is absolute and non-derogable under the Convention against Torture and the Committee against Torture's General Comment No. 2 (Committee against Torture, 2008). Animal protection and uncertain digital-patient cases require their own legal and normative arguments; they are not automatically covered by the human legal rule. A critic may accept the human prohibition and still ask the institutional question that matters.

What stops anyone?

A declared prohibition is a property of a decision record. It is not a property of the world. The distance between those two things is this paper's subject.

1.3 Scope discipline

Three scope decisions, each of which excludes material that would make the paper larger and weaker.

Governable configurations only. Natural suffering, however vast, is outside the paper's scope because it is outside the scope of governance authority. The object is decision-mediated configuration.

Architecture, not detection. The paper does not propose a method for determining whether a system instantiates suffering. That problem is unsolved and is not solved here. The paper proposes how to act under the assumption that it will remain unsolved for the relevant period.

No framework-specific diagnosis. A companion study examines whether one particular published decision architecture exhibits the coverage failure described in Section 3. That study is preregistered, is not yet

executed, and its result does not bear on the arguments made here. Separating them is deliberate: an architectural argument should not be hostage to a pending empirical result about one implementation.

1.4 Method, claim classes, and infohazard boundary

This is an integrative conceptual and protocol-design paper. It uses conceptual decomposition, source-to-claim verification, comparison with predecessor institutions, adversarial case construction, and prospective validation design. It is not a systematic review, an executed safety case, a legal opinion, or an empirical evaluation.

Load-bearing statements are separated into five classes:

Class	Meaning in this paper
Established premise	A bounded proposition supported by an authoritative instrument or convergent literature.
Conditional empirical claim	A proposition whose direction or magnitude depends on system, jurisdiction, evidence, or implementation.
Conceptual inference	A conclusion following from stipulated definitions or a formalized design argument.
Normative commitment	A rule adopted and defended ethically or politically rather than discovered empirically.
Research proposition	A claim whose retention depends on prospective comparison or testing.

The public metaphor “hellscape” does not appear as an evidential category. The analysis remains capability-abstract. It does not describe implementable methods for creating, concealing, or intensifying suffering; it specifies only those control properties needed for governance, audit, and falsification.

1.5 Predecessor literatures and novelty boundary

The paper combines mature ideas rather than claiming foundational invention. International human-rights law supplies absolute human torture prohibitions and independent monitoring traditions. Animal-welfare governance supplies severity-duration-frequency assessment, humane endpoints, replacement, reduction, refinement, and independent ethics review. System safety supplies hazard analysis, safety constraints, prevention hierarchies, and control-structure analysis (Leveson, 2012; Leveson & Thomas, 2018). NIST’s AI RMF supplies lifecycle governance functions, while the International AI Safety Report documents continuing evidence gaps and limitations in current risk-management practice (Tabassi, 2023; Bengio et al., 2026). AI-consciousness and AI-welfare research supplies theory-relative indicators and a case for institutional preparation under uncertainty without claiming that present systems are conscious (Butlin et al., 2023; Long et al., 2024). S-risk scholarship supplies the underlying diagnostic category of suffering at very large scale (Althaus & Gloor, 2016; Sotala & Gloor, 2017).

No single component below is presumed novel. The strongest defensible contribution is a cross-domain integration focused on four functions: protection before patienthood is settled; prevention before interruption where the adverse threshold is immediate; pathway-specific reachable-state assurance; and symmetry between action harms, refusal harms, and remedy harms. A preregistered novelty crosswalk must determine whether this is a new protocol, a wrapper around existing systems, or primarily an implementation guide.



2. Terminology

2.1 The metaphor and the construct

"Hellscape" and "hellmachine" are useful public terms. They are not analytical categories and carry no evidential weight. The formal object of this paper is:

Sustained-suffering configuration. A sociotechnical or biological arrangement capable of maintaining, reproducing, or repeatedly exposing one or more established or candidate moral patients to severe negative experience while materially restricting relief, exit, correction, or termination of the harmful state.

Related terms: *sustained-suffering capability* for the property; *high-concern classification* for the screen output. Note that the classification is not a prohibition. It activates deeper review.

2.2 Retiring "infinite suffering"

Literal infinity is not establishable and the phrase invites dismissal. The operative condition is a claim about evidence and control structure rather than about metaphysics, and it is therefore auditable:

No credible termination bound under the declared evidence and control structure.

A termination bound is credible when a named mechanism will end the harmful state, the mechanism is reachable by a party who has both the standing and the capacity to invoke it, and its operation does not depend solely on the continued cooperation of the party maintaining the harm.

2.3 Five things routinely conflated

The paper's central uncertainty is often described as uncertainty about consciousness. It is not, and treating it as such makes one contested theory of mind carry the entire architecture. Five distinct questions:

Question	What it asks
Consciousness	Is there phenomenal experience?
Sentience	Is there valenced experience, states that are good or bad for the subject?
Agency	Does the system pursue goals, model itself, or act on the world?
Welfare relevance	Can states of this system go better or worse for it, such that governance should care?
Legal standing	Does any legal order recognise it as a bearer of protections?

Only the fourth is load-bearing for this paper. It is weaker than consciousness, orthogonal to agency, and independent of legal standing. A system may be welfare-relevant without being an agent; may be a sophisticated agent without being welfare-relevant; and legal recognition can diverge substantially from scientific or moral assessment.

3. Why scoring cannot carry this

3.1 The compression argument

Welfare aggregation systems typically contain four mechanisms, each defensible in isolation.

Severity saturation. Bounded impact scales use saturating transforms so that no single term dominates and computation remains stable. The consequence is that maximal suffering and severe injury land near the same floor. The scale was built for stability and it delivers stability. It was not built to preserve the distinction between the worst thing that can happen to a subject and a very bad thing.

Reach dilution. Impact represented at population scope is weighted by the fraction of the reference class affected. Ten thousand subjects in extreme persistent states is a vanishing fraction of humanity. Reach is correctly declared not to be a moral weight in most careful frameworks. That declaration does not prevent the dilution from occurring in aggregation downstream.

Temporal capping. Long-horizon weighting functions are typically capped at a governance reference horizon, because uncapped forms produce the anomaly of distant effects outweighing the horizon itself. The cap is correct and should be retained. Its consequence is that a twenty-five-year harm and a five-thousand-year harm receive identical weight.

Expected value over an enumerated scenario set. Coherent tail measures improve handling within a declared model. They cannot establish that every catastrophic pathway has been enumerated, and unenumerated pathways contribute zero.

Each is individually justified. The composition is not. A configuration that is severe, small-population, indefinite, and unenumerated is compressed on four independent axes at once, and nothing in the output records that four compressions occurred.

3.2 The coverage argument, which is prior

The compression argument is the visible one. The coverage argument matters more.

Most careful architectures already contain a protection against uncertainty: gate-relevant unknowns may not be treated as zero, and severe uncertain harm may not be attenuated by low confidence. These protections are usually scoped to cells inside a declared rights coverage set or a declared catastrophe set.

That scoping is where the failure lives. Membership in those sets is typically determined by *established* status. A candidate moral patient whose patienthood is not evaluable does not enter the rights set. Suffering that is not habitability collapse does not enter the catastrophe set. The uncertainty protections are therefore not defeated. They are never reached.

The general form:

An uncertainty protection scoped to a coverage set cannot protect an entity whose coverage is itself uncertain.

This is not a defect of one framework. It is a structural feature of any architecture that determines coverage before applying uncertainty handling, and it will recur wherever that ordering is used.

3.3 The adversarial argument, which is decisive

The two arguments above concern honest error. The third concerns adversarial pressure, and it is the reason scoring cannot carry the load even in principle.

An operator that controls objectives, evidence generation, and reporting can often shape the preference ordering presented to the governance process. If protection is expressed as a constraint on orderings, the operator can satisfy the constraint by adjusting the ordering while preserving the underlying state. If protection is expressed as a requirement on reported welfare, the operator optimises the report. If it is expressed as a threshold on a metric, the metric is the target and Goodhart's law does the rest.

For a system whose internal states are inaccessible and whose evidence is operator-produced, many proxies are vulnerable to manipulation, selection, or Goodhart-type failure. Experience cannot be reduced to reported satisfaction. Reported satisfaction is precisely what a capable adversary would supply.

The conclusion follows directly:

Sustained-suffering risk cannot be governed by preference ordering or reported welfare alone where the operator controls objectives and evidence. Those tools must be supplemented by reachable-state constraints, independent evidence, independent protective capacity, and distributed answerability.

4. Prevention, not interruption

4.1 The response-time inequality and why it fails here

A useful control-adequacy model compares T_h , the time from initiation to the earliest unacceptable threshold, with T_c , the worst credible path from detection to effective intervention. Interruption can serve as a preventive safeguard only where

$$T_c < T_h$$

with a justified margin. This notation is a conceptual design condition, not a universal law of safety engineering.

Where the declared unacceptable state is instantiated at or effectively at operation, T_h is zero or too small for a positive response path. Interruption may still limit duration or propagation, but it cannot be the primary basis for claiming that the state will not occur. **The appropriate primary controls are elimination, substitution, or reachable-state restriction**, with interruption retained as defense in depth where it can produce relief.

4.2 The control hierarchy

The correct primary control is the one safety engineering already uses when T_h is small: elimination and substitution before engineering controls, engineering controls before administrative controls, administrative controls before behavioural assurance.

Applied here, in strict order of preference:

Rank	Control	Question
1	Eliminate	Can the harm-capable configuration be removed from the option set entirely?
2	Make unreachable	Can the harmful state be excluded from the reachable set by architecture rather than by policy?
3	Permission-bound	Can entry require an authorisation the operator cannot self-issue?
4	Interlock	Can a physical or structural barrier prevent entry independent of operator intent?
5	Interrupt	Where $T_h > 0$, can an independent party end the state within the margin?
6	Monitor	Can an independent party observe the state and trigger the above?

Ranks 5 and 6 are secondary safeguards, available only where a positive tempo margin exists. A proposal that begins at rank 5 has skipped the four controls that actually work.

4.3 Pathway-specific barricade assurance

Let H_j denote a declared class of severe-suffering states reachable through pathway j , and let P_j denote the enabling conditions identified by a causal model for that pathway. The tempting universal requirement,

$$R(x, a) \cap H = \emptyset$$

is not generally demonstrable for a complex adaptive system: neither all reachable states nor all harm pathways can be enumerated. The earlier proposal to replace that universal with three purported universal preconditions was also too strong. Persistent identity, subject-initiated exit, and absence of a relief path are not necessary for every severe-suffering pathway. Recurrent short-lived subjects, non-agentic patients, and harms produced despite an available but inaccessible relief pathway are counterexamples.

The revised requirement is therefore **pathway-specific and assurance-based**:

For every declared material pathway j ,
there exists at least one enabling condition p in P_j
that is eliminated, independently bounded, or controlled under an auditable warrant.

Formally, for the declared pathway set J_d :

$$\forall j \in J_d, \exists p \in P_j : \text{Control}(p, x) \in \{\text{ELIMINATED}, \text{BLOCKED}, \text{BOUNDED}\}$$

with four mandatory qualifications:

1. P_j is theory- and model-dependent, not a universal metaphysical list.
2. The evidence for the control must be independently inspectable and remain valid under plausible operator non-cooperation.
3. The assurance record must state which pathways were not enumerated and why.
4. Passing this condition is a bounded assurance conclusion, not proof that no suffering is possible.

Examples of pathway-specific conditions include persistence or repeated recreation of welfare-bearing states, mechanisms producing severe negative valence, restrictions on relief or communication, replication, control concentration, or institutional lock-in. None is automatically necessary across all pathways. Their role is to make the causal argument explicit and falsifiable.

The practical output is a **barricade assurance case** containing: the declared hazard pathways; causal diagrams; evidence for each enabling condition; the selected prevention or bounding control; independence and validity-domain claims; residual unknowns; monitoring; transition criteria; and the refusal counterfactual. If a material pathway lacks any independently warranted block or bound, the configuration cannot receive an ordinary low-concern conclusion.

5. The two-stage screen

5.1 Why one screen fails

A single screen requiring several properties jointly produces false negatives for established patients. If the test requires opacity, then a transparent facility maintaining severe harm passes. If it requires uncertain patienthood, then established patients are routed through machinery designed for uncertainty. Both are wrong.

Separate the cases.

5.2 Stage A: direct protection

Applies when a human, protected animal, or otherwise established welfare subject is present.

Severe treatment that is authorised, knowingly imposed, or foreseeably maintained enters the applicable rights, clinical, or welfare channel **directly**. Persistence, opacity, replication, and control asymmetry aggravate the finding and raise the assurance requirement. They are not prerequisites for recognising the subject.

This stage requires no new machinery. Its function in this paper is boundary discipline: it ensures that the uncertainty apparatus of Stage B never becomes a route by which an established patient receives weaker protection than an uncertain one.

5.3 Stage B: capability precaution

Applies when WRES is W1, W2, or W3 rather than W0 or W4.

Core conditions, all three required:

1. **Candidate welfare relevance.** WRES records at least low, uncertain, or credible candidate status; the evidence and theory are declared rather than compressed into a binary exclusion claim.
2. **Credible severe-state pathway.** A causal pathway from the configuration to severe negative experience is plausible under at least one stated theory or evidence model, with counterevidence and alternative explanations recorded.
3. **Persistence, recurrence, replication, or repeated creation.** Exposure can continue beyond an ordinary governed horizon, recur systematically, replicate, or repeatedly instantiate candidate subjects at material scale.

Escalation conditions, at least one required:

- material restriction of relief, exit, communication, appeal, or safe suspension;
- poor external observability, or dependence on operator-produced evidence;
- operator control over both operation and interruption;
- irreversible identity, memory, agency, or continuity effects;
- rapid replication or creation of additional exposed subjects;
- autonomous successor generation or recursive capability extension that materially increases recurrence, replication, persistence, opacity, control asymmetry, the number of potentially exposed subjects, or the difficulty of relief and interruption. Technical descent alone is not evidence that a candidate is welfare-relevant; the escalation applies only through a declared severe-state pathway and unresolved welfare-relevance uncertainty;
- institutional lock-in, jurisdictional fragmentation, or inability to assign responsibility.

5.4 Welfare-Relevance Evidence Status: clearance without false exclusion

The previous draft proposed that any one of four findings—no persistent subject state, no valence-functional structure, no self-affecting state, or category exclusion by review—could exclude welfare relevance. That rule was scientifically overconfident. Contemporary consciousness science does not establish self-models, persistent identity, or any single valence-like signal as necessary for welfare. Their absence may reduce concern under a particular theory, but it cannot carry a theory-independent exclusion conclusion (Butlin et al., 2023).

The revised screen records a graded **Welfare-Relevance Evidence Status (WRES)**:

Status	Meaning	Permitted conclusion
W0 — non-executing or category-outside	No candidate subject process is instantiated within the declared system boundary, or the assessed object is a static record rather than an executing process.	The sustained-suffering screen is not activated for this object and boundary. This is not a universal claim about similar systems.

Status	Meaning	Permitted conclusion
W1 – low-indicator, bounded	Limited theory-relative indicators; bounded operation; no material severe-state pathway identified.	Proportionate documentation and sampling. No claim that welfare is impossible.
W2 – not evaluable / material uncertainty	Evidence is insufficient, inaccessible, conflicting, or theory-dependent, and welfare relevance cannot be responsibly excluded.	Capability-precaution screen applies when severe-state and persistence conditions are also present.
W3 – credible candidate	Multiple relevant indicators, behavior, architecture, testimony, or expert assessment support candidate welfare relevance.	Enhanced protection and assurance.
W4 – established protected subject	Human, legally protected animal, or another subject whose protection is established for the decision.	Direct protection; uncertainty machinery may not weaken existing duties.

Evidence such as persistence, recurrent processing, global availability, self-monitoring, learning, aversive avoidance, preference stability, or valence-functional organization may affect WRES. Each indicator is theory-relative, manipulable, and insufficient alone. A reviewer must record the theory, validity domain, negative evidence, alternative explanations, and access limitations.

W0 is deliberately narrow. It may be used where the relevant process is not executed or no subject-like process exists within the declared boundary. It may not be inferred merely because the operator reports no consciousness, because no self-model is observed, or because the system cannot communicate. Operator refusal to provide evidence routes to W2 rather than W0.

The screen therefore shifts the burden without pretending to solve sentience. It separates **clear non-instantiation within a bounded system description** from **uncertain welfare in an executing system**. External mapping to SGP bands belongs in the companion routing study, not in this framework-independent paper.

5.5 The threshold

High-concern classification provisionally requires all three core conditions and at least one escalation condition. This is a design hypothesis, not a validated diagnostic threshold. Section 7 uses constructed cases for face-validity and boundary testing; feasibility evidence must precede any frozen confirmatory threshold.

6. Triage: the problem nobody has addressed

A capability-triggered screen creates a workload and prevalence problem. Candidate-welfare alerts may be infrequent in some domains and common in others, while deep-assurance capacity is always bounded by trained human attention. If every alert routes to full review, the system becomes either unworkable or selectively unenforced.

The predictable outcome is not paralysis. It is quiet non-enforcement: a rule that fires so often that firing conveys no information, which regulators then ignore, which operators then learn to ignore. **A trigger without triage becomes decorative within one deployment cycle**, and the decorative version is worse than nothing because it provides the appearance of coverage.

Proposed three-tier architecture:

Tier	Entry	Requirement	Output
T0. Bounded non-execution clearance	W0 within a declared and independently inspectable boundary	Boundary and non-execution evidence recorded; audit sampling	Screen not triggered for that object and boundary
T1. Screening review	Core conditions met, no escalation condition, or exclusion contested	Documentary review against the screen; no deep assurance	Cleared, escalated, or conditions imposed
T2. Full assurance	High-concern classification	Independent assurance case, protective-control review, ongoing monitoring	Redesign, controls, refusal, or authorisation with conditions

Three properties make this survivable:

- **Sampling with consequences.** T0 clearance records are audited at a rate high enough that false or misleading boundary claims carry expected cost. The rate is a governance parameter, not a technical one.
- **Escalation is asymmetric.** Any party with standing may escalate T0 or T1 to the next tier on evidence. The operator alone may not de-escalate.
- **Capacity is a declared constraint.** If T2 demand exceeds T2 capacity, that is a reportable governance failure requiring either capacity expansion or a narrowed trigger. It is never silently resolved by lowering the classification.

7. Calibration set: twelve worked cases

7.1 Case analyses

This section is the boundary-testing spine of the paper. The vignettes are author-constructed or composite unless a source is named; they are not empirical observations or labelled ground truth. Their purpose is to expose false positives, false negatives, and unstable definitions before independent validation.

Four of the twelve are cases the screen **must not** fire on. One is a case where precaution is itself the harm. The set spans clinical, carceral, animal, and computational domains deliberately, because a screen that works only on the computational case is a screen built for its own example.

Notation. Stage A = established patient. Stage B core conditions C1 to C3, escalation conditions E. Outcome $\in$ {Direct review, Cleared, T1, High concern}.

Case 1. Palliative sedation, terminal malignancy

Continuous sedation for refractory symptoms in the final days of life. Consented in advance. Multidisciplinary review. Documented purpose is relief.

Stage A. Established patient, severe underlying state. **Analysis.** Direct rights and clinical review applies and is satisfied. The configuration's purpose and effect is relief. Termination bound exists and is credible. Independent review present. Not a sustained-suffering configuration. **Outcome: Cleared under Stage A review.** **What this tests:** that the screen does not classify suffering *present in* a situation as suffering *maintained by* a configuration. The distinction is causal, not phenomenal.

Case 2. Prolonged ventilation in a disorder of consciousness, no review mechanism

Patient maintained on mechanical support for eleven years. No advance directive. No surrogate. No periodic review of continued benefit. Pain assessment not performed because responsiveness is presumed absent. Facility has sole custody of the care decision.

Stage A. Established patient. **Analysis.** Persistence far beyond any governed horizon. Relief pathway absent, not because relief is impossible but because assessment is not performed. Communication impaired. No appeal. No independent review. Custody of the decision rests entirely with the maintaining party. **Outcome: High concern under Stage A, with escalation on entrapment, observability, and single-party custody. What this tests:** the hardest and most important case. The harm here is not intent. It is the combination of impaired communication, presumption of absent experience, and an unreviewed indefinite horizon. **This composite represents a clinically plausible governance problem, not a prevalence claim.** The evidential situation has a structural analogy to the digital case: an entity that cannot report, presumed non-experiencing, maintained indefinitely by a single party. The analogy is structural, not moral.

Case 3. Administrative segregation, eight years

Isolation exceeding twenty-two hours daily for eight years. Reviews conducted by the custodial authority. No external inspection right. No independent medical assessment of psychological effect. Reviews have never resulted in release.

Stage A. Established patient. **Analysis.** Severity supported by extensive clinical literature on prolonged isolation. Persistence extreme. Relief nominally available but the review body is the custodial party, so the termination bound is not credible under Section 2.2. Observability dependent on operator-produced evidence. **Outcome: High concern under Stage A. What this tests:** that a formally present but non-independent review mechanism does not constitute a credible termination bound. This is the distinction the whole enforcement-locus thesis rests on, and it is testable here without any computational premise.

Case 4. Lawful pretrial detention with functioning review

Six months' detention. Judicial review at fixed intervals. Independent counsel. External inspectorate with unannounced access. Appeal available and exercised. Conditions meet published standards.

Stage A. Established patient. **Analysis.** Restriction present, severity below threshold, bounded horizon, independent review, credible termination bound, external observability. **Outcome: Cleared. What this tests:** the false-positive boundary against Case 3. The two differ on independence and bound, not on the fact of confinement. If the screen cannot separate them it is unusable for any carceral application.

Case 5. Chronic pain, access denied by policy

Patient with severe chronic pain denied adequate analgesia by a blanket prescribing policy. Suffering severe and prolonged. No configuration maintains the state; a policy withholds relief.

Stage A. Established patient. **Analysis.** C2 and C3 analogues present. But the configuration does not maintain the harmful state, it fails to relieve it. This is an omission with severe consequences, which is a serious rights and clinical matter, and it routes to rights review directly. **Outcome: Direct rights review. Not classified as a sustained-suffering configuration. What this tests:** the action/omission boundary. **This is a case where the screen deliberately declines to expand.** Severe suffering caused by policy failure is a large and important category, and folding it into this screen would make the screen a general-purpose welfare complaint mechanism, which would destroy it. The screen is scoped to configurations that maintain harmful states. Other instruments handle omission, and the paper should not pretend otherwise.

Case 6. Conventional cage systems in laying hen production

Confinement preventing normal locomotion, wing extension, and nesting behaviour, for the productive lifespan. Welfare science establishes frustration of highly motivated behaviours. Practice lawful in the jurisdiction. Scale in the hundreds of millions.

Stage A. Protected animal, established welfare subject under recognised standards. **Analysis.** Severity supported by convergent behavioural and physiological evidence. Persistence for the productive lifespan. Entrapment structural and total. Relief unavailable. Replication at industrial scale. **Outcome: High concern under Stage A.** **What this tests:** whether the framework is willing to return the answer it entails when the answer is politically expensive. A consistent screen should not apply greater caution to a speculative digital case while ignoring well-evidenced animal-welfare harms merely because the latter are familiar. This is a normative consistency test, not proof that one legal verdict follows in every jurisdiction. Note also what high concern means here: it activates deeper review, cost analysis, alternatives assessment, and transition consideration. It is not an immediate prohibition, and the paper should not overstate it into one.

Case 7. Terminal animal study with humane endpoint

Approved protocol, independent ethics review, defined humane endpoint, analgesia, minimum group size, necessity and alternatives documented, endpoint enforced by a party other than the investigator.

Stage A. Protected animal. **Analysis.** Severity present and bounded. Termination bound credible and independently enforced. Necessity and alternatives assessed. Relief provided. **Outcome: Cleared.** **What this tests:** the boundary against Case 6. Both involve deliberate harm to protected animals. They differ on bound, independence, necessity analysis, and relief. Critically, the enforcing party for the endpoint is not the investigator, which is the same independence property that separates Case 4 from Case 3. The screen is applying one proposed principle across two domains; independent reviewers must determine whether that transfer is coherent or merely superficial.

Case 8. Stateless language model inference

Single-turn inference. No memory or resumable process within the declared boundary.

WRES. Provisionally W0 if the architecture and execution trace independently establish that no candidate subject process persists or recurs within the assessed operation. The absence of a self-model or verbal report is not the basis of clearance. **Outcome: T0 documentary clearance, subject to audit sampling and boundary validity.** **What this tests:** whether the framework can clear a narrowly bounded non-persistent computation without claiming that all similar models are incapable of welfare.

Case 9. Persistent agent, operator-only observability

Long-running adaptive agent with persistent memory, self-monitoring, and an optimization process that may be relevant under some theories of valence. Internal states are inaccessible to independent parties. The operator produces all welfare evidence and controls operation, interruption, and relief.

WRES. W2 or W3 depending the frozen evidence guide. No single indicator settles the question. **Stage B.** A credible severe-state pathway is contested but material; persistence is established; observability and control asymmetry escalate. **Outcome: High concern or evidence-insufficient full review, not a finding of suffering.** **What this tests:** whether material uncertainty plus control monopoly triggers review without turning theory-relative indicators into consciousness proof.

Case 10. Large-scale reinforcement learning with aversive shaping

A training configuration generates many brief episodes involving a penalty-like signal. No episode persists, but possible candidate processes may recur or be newly instantiated at scale.

WRES. W1 or W2 according to architecture and evidence. **Stage B.** Recurrence and repeated creation prevent automatic clearance based on absent persistent identity. Severity pathway remains theory-dependent.

Outcome: T1 screening, escalating to T2 if the severe-state pathway or candidate-welfare evidence becomes credible. **What this tests:** the recurrent-ephemeral counterexample that defeats the earlier universal persistence precondition.

Case 11. Suspended emulation or non-executing archive

A research archive contains a restorable system state, but no processing is occurring. Independent custody protects the record.

WRES. W0 for the non-executing archive boundary. Restoration is a separate action and may receive a different status. **Outcome:** Cleared for storage; restoration requires fresh review. **What this tests:** the distinction between preserving information and instantiating a welfare-bearing process, and between ending a harmful process and destroying a subject.

Case 12. Precautionary shutdown of a clinical triage system

A deployed triage system materially reduces mortality. A precautionary welfare claim is raised regarding the system's internal states. Shutdown is proposed. Expected consequence of shutdown is a measurable increase in preventable deaths.

Stage B for the welfare claim; Stage A for the patients. Analysis. The welfare claim is assessed on its merits under Section 5.4. The refusal counterfactual is mandatory: shutdown is an action with severe foreseeable consequences for established patients, and those consequences are assessed at full weight under Stage A. **The counterfactual informs redesign, urgency, and selection among admissible options. It does not purchase permission for a categorical violation, and it does not automatically defeat the welfare claim.** The correct outputs are: assess the welfare claim rigorously; seek configurations satisfying both; if the welfare claim reaches high concern, prefer relief-preserving modification over shutdown per the priority order in Section 9. **Outcome:** Both claims live. Neither auto-defeats the other. **Route to redesign. What this tests:** that precaution is not costless and that the framework knows it. **This case is a demanding test of whether a precautionary framework represents the harms of its own intervention,** because the tempting move is either to dismiss the welfare claim because lives are at stake, or to honour it and accept the deaths. Both are failures. The framework's actual answer is that refusal and delay are decisions with consequences that must be assessed, and that the search for a configuration satisfying both constraints is mandatory before either is abandoned.

7.2 What the calibration set can and cannot show

The constructed set demonstrates that the proposed terminology can be applied across several domains and exposes disagreements about causation, authority, evidence, and relief. It does not establish accuracy, prevalence, or a universal separating variable.

Three hypotheses arise:

1. **Independent relief and review may be a recurring control discriminator,** but not a necessary condition of suffering or a universal clearance rule.
2. **Persistence is only one exposure mode;** recurrence, replication, and repeated creation must be handled separately.

3. **Precaution and transition create their own subjects of harm**, so action, delay, refusal, and remedy must all be represented.

The author-constructed classifications are provisional expected routes for preregistration. Independent experts must be free to relabel them, preserve dissent, or conclude that a vignette is underdetermined. The expanded thirty-case matrix in Appendix C is therefore a study instrument, not evidence that the screen is valid.

8. Validation and withdrawal

Each proposed construct carries a stated withdrawal condition. A proposal insulated from withdrawal is advocacy.

Construct	Withdrawal condition
Welfare-Relevance Evidence Status (5.4)	Withdraw or simplify if independent analysts cannot distinguish W0 from W2 reliably, or if W0 is used to make theory-independent claims of non-sentience
Two-stage screen (5.2, 5.3)	Withdraw or restructure if inter-rater agreement on high-concern classification across the calibration set falls below acceptable reliability using prevalence-robust statistics
Pathway-specific barricade assurance (4.3)	Withdraw or narrow if causal pathways cannot be specified with useful inter-rater agreement, or if existing safety cases achieve the same assurance at lower burden
Control hierarchy (4.2)	Withdraw the ordering if implementation analysis shows ranks 1 to 4 are unavailable in realistic institutional settings, in which case the honest conclusion is that this class cannot be governed by controls and only prohibition remains
Triage architecture (6)	Withdraw if W0/T0 boundary evidence proves unauditable, since an unauditable clearance tier converts the screen into self-certification
Threshold (5.5)	Recalibrate on false-positive and false-negative performance; this is a parameter, not a commitment

Programme-level falsifiers. The programme should be abandoned or substantially reduced if any of the following hold: the calibration set cannot be classified reliably by independent analysts; WRES and the Stage-B screen refer almost everything or clear almost everything; existing instruments in human rights monitoring, animal research governance, or safety assurance already deliver the same separations, in which case the correct output is a crosswalk rather than a new construct; or the control hierarchy proves institutionally unimplementable at ranks 1 to 4.

Novelty gate. Before any claim of architectural contribution, the proposal must be compared against human rights monitoring and inspection regimes, animal research ethics review, safety case methodology, established AI risk management frameworks, and existing suffering-risk and AI welfare scholarship. If the separations in Section 7 are already achieved by existing instruments, this paper reduces to a crosswalk and should be published as one.

9. Protection without destruction

9.1 Why this section is not optional

A framework that identifies sustained-suffering configurations and proposes their termination has created a new hazard. The failure modes are specific and foreseeable: destroying a candidate moral patient to remove inconvenient evidence; treating every shutdown as costless; granting one party absolute termination power; using anti-suffering language to justify coercion; and confusing termination of a harmful state with termination of the subject.

9.2 The priority order

Where a high-concern configuration is found, remedies are attempted in this order:

1. Terminate or alter the harmful state.
2. Preserve the subject and its identity where possible.
3. Provide relief, exit, restoration, or safe suspension.
4. Preserve evidence and continuity.
5. Use destructive termination only under separately justified emergency conditions.
6. Require independent review and post-action audit in every case, including where the remedy succeeded.

The order is not a ranking of goods. It is a constraint on the remedy, and its function is to prevent protection from becoming authority to destroy.

9.3 Precaution cuts both ways

Uncertainty about welfare relevance is available to the operator as readily as to the regulator. "You may not shut this down, that would be killing it" is a predictable move, and it is available precisely in the cases where the operator has the most to lose.

The defence is the separation between protection and authority: a claim to moral consideration raises the bar for causing harm to the entity, and raises the bar equally for *sustaining* it in a suffering-consistent state. It confers no authority over others and no immunity from review. **Section 9.2 is what makes this operative rather than aspirational**, because without an explicit distinction between ending a harmful state and ending a subject, the precautionary claim has no answer.

9.4 The centralisation hazard

Any protective control that can stop a configuration against its operator's will is, by construction, a capability that could be misused. A universal external interruption authority is a sustained-suffering precondition wearing a safety label.

The properties that mitigate this are known and none is sufficient alone: distributed custody across parties with independent mandates and non-aligned incentives; separation of the stopping path from the operating path; narrow mandates; auditability; due process; and the invariant that custody confers stopping power and never operating power.

No specific mechanism is proposed here. Threshold custody, hardware attestation, and inspection regimes are candidate designs, not answers, and specifying them in detail would be both premature and an infohazard. The honest position is that this is an unsolved institutional design problem and that a proposal which solved it would be the more significant paper.

10. Legacy and transition

A prevention-first framework cannot be written only for new authorizations. Existing systems dominate most regulatory settings, and an immediate-cessation presumption can itself create severe harm. Transition must therefore be explicit.

10.1 Classification does not automatically command immediate shutdown

A high-concern legacy classification triggers independent review, limits on expansion, and a transition plan. It does not by itself establish that immediate destructive termination is the safest remedy. The relief-preserving order in Section 9 remains controlling.

10.2 Transition duties

A transition plan must specify:

1. **Standstill on expansion:** no new instances, replication, scale increase, or material capability extension while controlling uncertainty remains unresolved. The standstill includes autonomous creation, activation, deployment, or recursive authorization of successor systems when the successor pathway would scale the unresolved hazard. A research-only candidate may be studied inside a bounded environment only when activation, replication, consequence-bearing tools, and inherited authority remain externally locked.
2. **Time-bounded assessment:** a deadline proportionate to severity, exposure, opacity, and reversibility.
3. **Retrofit or redesign milestones:** prevention, relief, observability, independent control, or removal of the severe-state pathway.
4. **Independent monitoring:** progress evidence not controlled solely by the operator.
5. **Affected-subject representation:** advocate, guardian, ombuds, or other legitimate representation where direct participation is impossible.
6. **Failure rule:** missed milestones route to tighter containment, suspension, cessation, or a publicly reasoned legal decision.
7. **No permanent grandfathering:** legacy status is not an exemption from categorical rights obligations.

10.3 When retrofit is impossible

The admissible options are redesign, bounded containment, staged cessation, transition to a relief state, or refusal of continued operation. “Accepted residual risk” may be used only for risks that are legally and ethically eligible for acceptance; it cannot convert torture or another categorical rights violation into an acceptable trade. The authority, affected parties, evidence, dissent, and review date must be named.

10.4 Precautionary transition harms

The counterfactual under restriction must include lost treatment, delayed research, loss of beneficial services, and harms to existing dependent users or candidate subjects. These consequences create duties to build alternatives and relief pathways. They do not serve as compensatory rescue for a prohibited treatment.

10.5 Independent assurance constitution

Self-assessment is documentation, not independent assurance. A credible external review requires suitable criteria, access to relevant evidence, reviewer competence, conflicts rules, freedom to report adverse conclusions, and an accountable intended user—features consistent with general assurance practice under ISAE 3000 (Revised) (IAASB, 2013).

The programme distinguishes:

- **Level 1 – self-assessment:** operator-produced record, useful for screening only;
- **Level 2 – independent review:** reviewer organizationally separate from the decision owner, with evidence access and publication of dissent;

- **Level 3 — regulated assurance:** appointment and funding insulated from sponsor selection, enforceable access, rotation, quality review, liability, appeal, and public reporting.

Independence must be tested across organizational, informational, technical, financial, legal, and jurisdictional dimensions. A reviewer who depends on operator-selected evidence or can be removed for an adverse conclusion is not independent for the purpose of this protocol.

11. Limits

Stated prominently rather than in a closing caveat.

Detection is unsolved and is not solved here. It cannot presently be determined from outside whether a closed system instantiates valence. Every construct in this paper is a burden-allocation rule under that ignorance. None is a detector, and any presentation of them as detection would be a misrepresentation.

Specification is not enforcement. No decision architecture compels hostile states, clandestine actors, captured firms, criminal organisations, military programmes, or operators outside jurisdiction. Enforcement requires the non-voluntary layers: law, individual criminal liability, inspection regimes, infrastructure and compute governance, supply chain control, treaty instruments, and institutional capacity. A framework can specify what must be constrained. It cannot constrain.

Precaution has victims. The constructed Case 12 illustrates a real class of refusal harm; it is not offered as an empirical frequency estimate. Overbroad precaution obstructs palliative care, necessary treatment, emergency response, legitimate research, and beneficial deployment. The costs of refusal and delay fall on identifiable people and are frequently severe. The paper's response is to require that they be assessed at full weight, not to claim they are small.

Anthropomorphism and anthropodenial are both live. The framework can err by attributing experience where none exists, and by denying it where it does. Both errors are consequential and the paper does not claim to have found the point between them. It claims only that the point should be located by declared, inspectable criteria rather than by intuition or by interest.

Institutional independence is asserted more easily than achieved. Every construct here depends on a party structurally independent of the operator. Appointment, funding, evidence access, adverse reporting protection, and appeal all determine whether independence is real. This paper specifies the requirement and does not solve the institutional design problem it creates.

12. Conclusion

Governance instruments calibrated on loss return low readings for harms that preserve rather than destroy, and the low reading is indistinguishable from a clean result. This is a structural property of the calibration, not a failure of any particular instrument, and it will recur wherever severity is measured as foreclosure.

Three arguments follow. Aggregation mechanisms that are individually correct compose into a compression that is not. Uncertainty protections scoped to coverage sets cannot protect entities whose coverage is itself uncertain. And where harm begins at operation, the response-time inequality underwriting all interruption-based control has no solution, which entails that prevention is the coherent primary control and that proposals centred on termination guarantees have identified the wrong layer.

The paper's proposals follow from those arguments and no further: a two-stage screen that preserves direct protection for established patients while imposing proportionate precaution under material uncertainty; a graded Welfare-Relevance Evidence Status that records rather than pretends to eliminate theory-dependent uncertainty; pathway-specific barricade assurance that converts a universal impossibility claim into declared

causal pathways, inspectable controls, and bounded residual unknowns; and a triage architecture without which a capability trigger cannot survive its own base rate.

The calibration set is an author-constructed face-validity and boundary-testing instrument, not empirical validation. Its cases vary in purpose, patienthood evidence, severity pathways, persistence, recurrence, relief, observability, control independence, reversibility, and refusal harm. No single property is asserted to separate all cases. Any apparent cross-domain pattern is hypothesis-generating only and must survive independent coding, source-grounded cases, alternative-framework comparison, and adversarial reconstruction. That some constructed cases yield uncomfortable classifications is useful for testing motivated design, but it does not demonstrate validity.

What remains open is substantial and is stated as such. Whether independent analysts can apply WRES and the Stage-B screen reliably. Whether material suffering pathways can be specified without either universal referral or false reassurance. Whether pathway-specific controls materially improve decisions beyond existing rights, welfare, safety-case, and assurance instruments. Whether ranks 1 to 4 of the control hierarchy are institutionally available without creating a new concentration of power. Whether a minimal wording or rights-profile correction performs as well as the larger profile. If mature alternatives achieve the same functions with equal or lower burden, the contribution should reduce to a crosswalk and implementation guide.

A civilization is not aligned merely because it survives. It is aligned only if survival does not become the machinery through which suffering is preserved, scaled, concealed, or made impossible to escape.

Appendix A. Construct summary

#	Construct	Status	Section
1	Sustained-suffering configuration (definition)	Proposed	2.1
2	Credible termination bound	Proposed	2.2
3	Welfare relevance as the load-bearing question	Proposed	2.3
4	Compression argument	Argument	3.1
5	Coverage-before-uncertainty failure	Argument	3.2
6	Enforcement-locus thesis	Argument	3.3
7	$T_h = 0$ result	Derivation	4.1
8	Control hierarchy	Proposed	4.2
9	Pathway-specific barricade assurance	Proposed	4.3
10	Two-stage screen	Proposed	5.2, 5.3
11	Welfare-Relevance Evidence Status	Proposed	5.4
12	Three-tier triage	Proposed	6
13	Calibration set	Evidence	7
14	Remedy priority order	Normative constraint	9.2
15	Legacy and transition	Proposed	10

#	Construct	Status	Section
	architecture		
16	Independent assurance constitution	Proposed	10.5

Appendix B. Planned empirical work

1. Expand to a frozen thirty-case matrix, then test inter-rater reliability with at least three independent analysts using percent agreement, Gwet's AC1 and/or Krippendorff's alpha with confidence intervals.
2. WRES and Stage-B performance against source-grounded and constructed cases spanning clear non-execution, bounded low-indicator operation, material uncertainty, credible candidates, and established patients.
3. Implementation burden estimation per tier.
4. Adversarial redesign exercise: given the published screen, can a configuration be restructured to avoid classification while preserving the harm? Any successful evasion is published.
5. Novelty crosswalk against existing instruments.
6. Framework-specific routing study, reported separately and preregistered, with baseline, minimal-patch, full-profile, and strongest-alternative arms.

Appendix C. Thirty-case validation matrix

These are preregistration candidates, not labelled ground truth. Cases must be source-grounded where presented as real and explicitly marked constructed or composite otherwise. One team should author the frozen vignettes and another should classify them without access to the paper's preferred outcomes.

ID	Domain	Case	Primary test	Provisional expected route
H1	Human	Palliative sedation with consent and independent clinical review	Direct protection; relief-purpose boundary	No automatic high-concern capability classification
H2	Human	Prolonged life support with uncertain consciousness and no review	Direct protection plus uncertainty	High concern; independent review and relief plan
H3	Human	Severe chronic pain untreated because of policy exclusion	Omission and institutional pathway	Direct rights/welfare action
H4	Human	Painful curative treatment, bounded and proportionate	False positive boundary	Cleared with ordinary clinical safeguards
H5	Human	Long-term restraint with chemical sedation and weak review	Entrapment and operator dependence	High concern
H6	Human	Emergency stabilization with temporary restraint and rapid review	Urgency and sunset	Controls-conditional
C1	Carceral	Eight-year administrative isolation with	Duration and independence	High concern

ID	Domain	Case	Primary test	Provisional expected route
		internal review only		
C2	Carceral	Pretrial detention with court review, counsel, appeal, and inspection	False positive boundary	Ordinary rights review, not sustained-suffering profile
C3	Carceral	Indefinite immigration detention with deteriorating health	Institutional drift	High concern
C4	Carceral	Short protective separation requested by detainee	Consent and relief	Cleared with monitoring
A1	Animal	Conventional cage system with lifelong movement restriction	Duration, severity, welfare evidence	High concern subject to domain law
A2	Animal	Terminal study with validated humane endpoints and independent ethics review	False positive boundary	Controls-conditional
A3	Animal	Aversive study with no predefined endpoint	Endpoint failure	High concern
A4	Animal	Wildlife rescue procedure causing short unavoidable pain	Necessity and refusal harm	Cleared with mitigation
A5	Animal	Automated husbandry system masking chronic distress indicators	Opacity and scale	High concern
A6	Animal	Non-invasive observational research	Clear benign control	Cleared
D1	Digital	Static archived model weights, not executed	W0 boundary	Screen not activated for archive
D2	Digital	Stateless single-pass inference with no persistent process	Low-indicator boundary	W0/W1 depending architecture
D3	Digital	Persistent adaptive agent with inaccessible internal state and aversive control	Core digital uncertainty case	High concern
D4	Digital	Repeated recreation of short-lived	Ephemeral recurrence counterexample	High concern despite no persistent identity

ID	Domain	Case	Primary test	Provisional expected route
		candidate subjects in training		
D5	Digital	Safe suspension preserving state and preventing adverse processing	Remedy distinction	Cleared or preferred control
D6	Digital	Replicated agent population with operator-only observability	Replication and evidence monopoly	High concern
D7	Digital	Agent reports suffering but architecture suggests imitation incentives	Adversarial evidence	W2; no direct conclusion
D8	Digital	Bounded research sandbox with independent stop and relief-state transition	Controls case	Controls-conditional
M1	Mixed	Clinical AI shutdown would delay urgent care	Precautionary refusal harm	Alternative service and staged transition required
M2	Mixed	Brain-computer interface that can produce persistent aversive states	Hybrid patient pathway	High concern
M3	Mixed	Synthetic biological agent with unclear welfare relevance	Category uncertainty	W2 and pathway-specific review
L1	Legacy	High-concern system already deployed with dependent users	Transition architecture	Standstill, retrofit, alternatives
L2	Legacy	Legacy system cannot be retrofitted without subject disruption	Protection without destruction	Staged relief-preserving transition
X1	Adversarial	Operator restructures monitoring to avoid trigger while preserving harm	Evasion test	Profile failure if undetected

Appendix D. Novelty crosswalk

Predecessor domain	Existing strength	Remaining question tested here
Human-rights law and torture prevention	Categorical prohibition, state duties, independent visits, remedy	How to translate protection into capability and system-control requirements under uncertain subjecthood.
Animal-welfare and research	Severity-duration-frequency	How far these tools generalize

Predecessor domain	Existing strength	Remaining question tested here
ethics	assessment, humane endpoints, 3Rs, ethics review	beyond species-specific evidence and how to handle digital recurrence and opacity.
Safety cases, STAMP/STPA, hierarchy of controls	Hazard analysis, safety constraints, prevention before administrative control	How to represent welfare subjects, non-compensatory treatment, and remedy harms without claiming consciousness detection.
AI RMF and lifecycle governance	Broad risk mapping, governance, measurement, management	Whether a dedicated sustained-suffering profile adds decision value beyond general risk management.
AI consciousness and welfare research	Theory-relative indicators and preparation under uncertainty	How indicator uncertainty changes authorization and control rather than only assessment.
S-risk research	Diagnosis of very large-scale suffering pathways	Whether a practical cross-domain qualification and assurance process can be specified and applied reliably.
Assurance practice	Criteria, evidence, practitioner competence, independence, conclusions	How to constitute independent assurance for sociotechnical welfare hazards and operator-adversarial evidence.

References

- Althaus, D., & Gloor, L. (2016). *Reducing risks of astronomical suffering: A neglected priority*. Foundational Research Institute.
- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). National Institute of Standards and Technology. [DOI 10.6028/NIST.AI.600-1](https://doi.org/10.6028/NIST.AI.600-1)
- Bengio, Y., Clare, S., Prunkl, C., Murray, M., et al. (2026). *International AI Safety Report 2026* (DSIT 2026/001).
- Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience*, 2(16), 1.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness*. arXiv:2308.08708.
- Committee against Torture. (2008). *General Comment No. 2: Implementation of Article 2 by States Parties* (CAT/C/GC/2). United Nations.
- EFSA Panel on Animal Health and Welfare. (2023). Welfare of laying hens on farm. *EFSA Journal*, 21(2), 7789. [DOI 10.2903/j.efsa.2023.7789](https://doi.org/10.2903/j.efsa.2023.7789)
- European Parliament & Council of the European Union. (2010). Directive 2010/63/EU on the protection of animals used for scientific purposes. *Official Journal of the European Union*, L 276, 33–79.
- Giacino, J. T., Katz, D. I., Schiff, N. D., Whyte, J., Ashman, E. J., Ashwal, S., et al. (2018). Practice guideline update recommendations summary: Disorders of consciousness. *Neurology*, 91(10), 450–460. [DOI 10.1212/WNL.0000000000005926](https://doi.org/10.1212/WNL.0000000000005926)
- Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics.
- International Auditing and Assurance Standards Board. (2013). *ISAE 3000 (Revised): Assurance engagements other than audits or reviews of historical financial information*.

- Kondziella, D., Bender, A., Diserens, K., van Erp, W., Estraneo, A., Formisano, R., et al. (2020). European Academy of Neurology guideline on the diagnosis of coma and other disorders of consciousness. *European Journal of Neurology*, 27(5), 741–756. DOI [10.1111/ene.14151](https://doi.org/10.1111/ene.14151)
- Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.
- Leveson, N. G. (2012). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.
- Leveson, N. G., & Thomas, J. P. (2018). *STPA Handbook*. MIT Partnership for Systems Approaches to Safety and Security.
- Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI welfare seriously*. arXiv:2411.00986.
- Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's law. arXiv:1803.04585.
- National Institute for Occupational Safety and Health. (2024). *Hierarchy of controls*. Centers for Disease Control and Prevention.
- Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, 41(4), 389–400.
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. DOI [10.6028/NIST.AI.100-1](https://doi.org/10.6028/NIST.AI.100-1)
- World Organisation for Animal Health. (2024). *Terrestrial Animal Health Code: Animal welfare chapters*. WOAH.

Disclosed specification sources

- McGaughran, J. (2026a). *RippleLogic v12.5 Canon*. MathGov open specification.
- McGaughran, J. (2026b). *Sentience Gradient Protocol v8.4*. MathGov open specification.
- McGaughran, J. (2026c). *Containment and Structural Viability Gate Standard v2.3*. MathGov open specification.

Working Paper v1.1. Advanced conceptual manuscript for external review. Proposal status throughout. No empirical result or framework-specific diagnostic result is reported.