ZHM v4

A Protected Flourishing Architecture

Toward provably safe realities for conscious life

Andrew Hill

Independent researcher

Concept and direction by Andrew Hill
Developed in collaboration with ChatGPT
9 October 2026

Abstract

What would have to be true for a conscious being to possess a future that no other actor could turn into a catastrophe? The Zero Hostility Matrix addresses this question as a problem of moral standing, physical possibility, and system architecture. Its ultimate objective is a reality in which suffering is excluded from admissible conscious histories while persons retain worthwhile, self-directed continuations. Reaching that objective would require more than benevolent intelligence or reliable monitoring. It would require an adequate account of experience, a mechanism that excludes harmful realizations, and a defensible connection between the proof and the physical world in which the subject actually exists.

This version develops a Protected Flourishing Architecture around that ambition. It distinguishes catastrophic protection from complete suffering prevention; recurring hazard from uncertainty about an entire future; probability-zero failure from impossibility; and a protected simulation from a protected substrate. It argues that protection must preserve future options, survive changes of host and implementation, cover the protective machinery itself, and remain compatible with agency and epistemic revision. Technical appendices state conditional protection propositions, identify their assumptions, and provide a reproducible finite-state demonstration of several architectural failure modes.

The contribution is a synthesis and a research program. The probability identities and invariant arguments use established methods; the proposed connection to conscious welfare remains to be established. The intended result is a precise account of what a genuinely protected future would require, together with tractable ways to discover whether its components can be built.

Reading this essay

The main text develops the argument in ordinary language. Appendix A gives the lifetime-risk mathematics. Appendix B states the formal protection conditions. Appendix C presents an executable model with stipulated labels. Appendix D gives an assurance record and a revision map from v2 and v3. References distinguish this project's earlier manuscripts from the research on which its technical methods draw.

Contents

1 A future that belongs to its inhabitants 3

2 What protection must mean 4

3 Why long lives change the safety problem 5

4 What a proof of a safe reality would have to cover 6

5 The life that protection is for 8

6 Keeping a future reachable 9

7 Constructing a protected domain 10

8 Worlds without private victims 12

9 Continuity across transformation 13

10 The protector must also be constrained 14

11 Research that can change the answer 15

12 The conditions of success 16

Appendix A Protection across long horizons 18

Appendix B Conditional protection and its limits 20

Appendix C An executable architectural example 24

Appendix D Assurance records and revision history 26

References 28

1 A future that belongs to its inhabitants

Imagine being offered another thousand years of life. You would retain your relationships, acquire new capacities, explore unfamiliar worlds, and remain free to change your mind about how to live. Much of what currently makes a human life feel narrow would become negotiable. Now imagine discovering that the offer also places the machinery of your experience under someone else's continuing control.

Your host can alter your environment. Perhaps it can also alter your memories, your rate of thought, the conditions under which you can leave, or the internal processes through which an experience becomes unbearable. Its intentions are reassuring. Its powers are extensive. The important question is what remains true if its intentions change, its implementation fails, or a more capable actor acquires it.

The promise of a larger life acquires a constitutional problem. A future can be rich in opportunities while remaining insecure at its foundation. If a subject must always remain stronger, more valuable, more obedient, or better hidden than every possible adversary, its freedom depends on an indefinitely repeated contest. Greater intelligence may help a person win that contest. It does not establish a right that survives losing it.

The Zero Hostility Matrix, or ZHM, names the project of replacing that dependency with durable protection. Every conscious being should have a protected continuation: a feasible future in which it remains beyond another actor's power to impose catastrophic suffering, and in which its own life remains available to it. The full ambition extends further, toward lives from which suffering itself can be removed without removing the beings who would otherwise suffer.

This ambition does not require confidence that artificial consciousness, extreme longevity, or future reconstruction of persons will occur. Conditional engineering questions can matter before their enabling technologies exist. If we gain substantial control over the physical realization of experience, what constraints should accompany that control? Which forms of assurance would be adequate for lives whose duration and vulnerability could differ radically from our own?

The problem also exists now. Human beings and other animals already suffer, and already depend on systems they cannot individually control. A future-facing account should strengthen concern for existing subjects. It should not postpone their protection until more spectacular technologies arrive.

Earlier ZHM manuscripts established two commitments that this version retains. Version 2 placed catastrophic conscious suffering outside the permissible domain of optimization. Version 3 made protected continuation, sovereign domains, and constrained protective authority central to the design [1, 2]. Version 4 asks how those commitments could become an architecture with explicit conditions of success.

The literature on suffering risks already examines the possibility of harm on scales far beyond familiar experience [3]. Work on AI welfare asks how to treat systems whose moral status is uncertain [4]. ZHM's proposed contribution is to connect these concerns to a demanding architectural question: can the systems that support conscious life be constructed so that certain violations are unavailable, including to their most capable inhabitants and operators?

That question should remain ambitious. The standard should also remain discriminating. A constitution, a simulation, and a proof about a simulation establish different things. A mature protection architecture would have to connect them to the world where experience occurs.

2 What protection must mean

The protected object is an experiencing subject across a developing history. The subject may be biological, artificial, unfamiliar, temporary, distributed, or one of many similar instances. Its claim does not disappear because another being created it, because it cannot speak, or because its creator describes it as a tool.

Substrate neutrality is a moral principle conditional on experience being present. It does not assert that every arrangement of matter is conscious, that every representation of a person creates a person, or that every computationally equivalent description has identical phenomenology. Those are questions for theories of realization and consciousness. The principle says that an actual subject does not lose standing through an irrelevant change of label or material.

Suffering means experience with an intrinsically aversive phenomenal character. A negative number in an optimization algorithm does not establish that character. Neither does a positive reward, an obedient response, or a declaration of satisfaction establish its absence. Expressions, mechanisms, and reports can supply evidence; their relationship to experience must be argued.

The term hypertorture is retained from earlier versions for catastrophic suffering, including cases in which intensity, duration, domination of cognition, or inescapability makes a conscious history intolerable in an extreme sense. No claim is made that there is a single universal threshold or an available instrument that detects it across minds. A boundary may depend on the subject's organization and history. A defender must justify both the chosen boundary and its ability to recognize approaching violations.

Multiplicity is a separate dimension. Many episodes of mild discomfort can constitute a large aggregate harm without becoming a single subject's extreme agony. Conversely, protecting a population average can conceal an individual catastrophe. An adequate account preserves both individual histories and population-level consequences, with uncertainty about individuation explicitly recorded.

Three ambitions should be distinguished. The first is a non-repealable floor against catastrophic suffering. The second is the elimination of involuntary suffering more broadly. The third is the possibility of suffering-free flourishing: conscious lives with no negative phenomenal experience, sustained alongside agency, understanding, relationships, and meaningful activity. Success at the first would be substantial. It would leave the other two unfinished.

This distinction matters to the word zero. A domain that excludes catastrophe while permitting grief or ordinary pain cannot honestly be called free of suffering. A suffering-free domain must meet a stronger predicate. The ultimate ZHM vision pursues that stronger condition; present engineering work may need to begin with narrower, defensible boundaries. Each result should say which boundary it concerns.

There is an ethical tension here that should remain visible. Some people endorse forms of difficulty that currently involve suffering. Offering them reliable relief and new forms of challenge respects their agency. Compulsorily rewriting them into a state they do not endorse can violate it. A fully suffering-

free civilization consistent with sovereignty would need forms of motivation and experience that its inhabitants can genuinely endorse, together with a defensible transition from existing lives. It cannot obtain consent by manufacturing the preference that supposedly authorizes the change.

ZHM therefore combines a strict catastrophic floor with a constructive research objective: make suffering unnecessary, make relief reliably available, and investigate whether worthwhile agency can persist throughout a state space without suffering. Universal abolition consistent with these commitments remains a stronger, unresolved achievement. Voluntary bounded adversity can remain permissible during intermediate stages without being mislabeled completion of that objective.

The moral case begins with the reality of the subject's experience. A spectator's approval does not alter what is undergone. The further claim that this experience gives others a reason to protect the subject is a normative commitment, supported here by equal standing and resistance to arbitrary domination. Phenomenology alone does not logically force every possible intelligence to accept that commitment.

Likewise, non-compensation is a constitutional proposal, not a newly discovered theorem of ethics. The proposed rule forbids deliberately creating or maintaining a catastrophic victim for unrelated benefits. It rejects the idea that enough prosperity elsewhere can license the chamber. Its purpose is to remove a dangerous use of power from ordinary bargaining.

This rule does not settle every choice under scarcity. Nor does it license eliminating conscious life to eliminate the possibility of suffering. Continuation, agency, and protection are joint obligations. If no available action satisfies them together, the architecture has encountered an emergency or a limit. It has not succeeded merely because one indicator has improved.

3 Why long lives change the safety problem

A system that looks extraordinarily reliable over a short deployment can be unsuitable as the foundation of a very long life. The relevant question changes from how rarely a component fails this year to whether a subject can remain protected throughout the future it is being offered.

Suppose, purely for illustration, that a catastrophe has a one-percent chance of occurring in each year, conditional on no earlier catastrophe. Over a thousand years, the chance of at least one catastrophe is approximately 99.9957 percent. At a constant conditional annual hazard of one in a billion, a quadrillion years still makes at least one catastrophe overwhelmingly likely. These examples describe repeated exposure to a specified hazard. They are not estimates of actual future technologies.

A different statement has a different meaning. Suppose our present uncertainty is that a whole architecture has a one-percent chance of containing a particular defect, and a ninety-nine-percent chance of excluding that defect forever. Time alone does not turn this uncertainty into a fresh one-percent gamble every year. Whether risk rises with duration depends on what happens in the defective case and whether additional hazards recur. Confusing a probability over entire worlds with a hazard within a world produces misleading conclusions in either direction.

The general principle is that continuing life requires an account of continuing exposure. If a nonzero conditional hazard remains bounded below indefinitely, eventual catastrophe becomes almost certain. If hazards fall sufficiently quickly, their total contribution can remain bounded. If a genuine invariant

excludes the event under the actual conditions, repeated operation need not accumulate risk from that event at all. Appendix A gives the precise distinctions without assuming independent annual trials.

An aggregate ceiling also becomes harder to maintain when the number of subjects or their rate of experience grows. A civilization with a million residents has different exposure from one with a trillion, even if individual conditions appear unchanged. Faster computation is not automatically faster experience, but any defensible relationship between them belongs in the protection model. A safety check that arrives promptly by an external clock can arrive too late for the process it protects.

Long duration also transforms upgrades into a central issue. A person may outlive the company, language, operating system, cryptographic standard, hardware family, and scientific theory under which its first certificate was issued. Protection must survive such replacements. Renewing a certificate cannot erase earlier exposure, remove unfulfilled rescue obligations, or grant a new lifetime risk allowance as if no previous subject existed.

This motivates a requirement of inherited obligations. When support moves from one provider to another, the receiving system inherits the relevant history, unresolved uncertainties, and continuation commitments. Copying creates additional possible subjects and additional support obligations. Merging does not authorize the disappearance of inconvenient claims. The accounting follows experiencing histories and their possible continuations, with explicit uncertainty, rather than a convenient database identifier.

A numerical risk budget can be useful for provisional engineering. It describes residual uncertainty that must be constrained and reduced. It is not a quota of victims that a civilization is entitled to produce. A small assigned number is especially uninformative when there is no credible basis for bounding the process that generates it.

The ultimate aim is consequently stronger than extremely low annual failure. It is a structural condition under which the prohibited experience is unreachable. This is where the long-life argument becomes a design requirement. A future that continues to expand should not need to spend a fresh portion of its inhabitants' safety whenever another interval of life begins.

4 What a proof of a safe reality would have to cover

The strongest version of ZHM asks for a proof about reality: from a specified protected condition onward, every physically admissible continuation excludes suffering while preserving the conditions for worthwhile conscious life. This is the ultimate specification of the Protected Flourishing Architecture. A proof would have to establish the relevant property of the world its inhabitants actually occupy.

The temporal qualification matters. Existing suffering cannot be made never to have occurred by changing a record or restoring an earlier state. A successful architecture would protect subsequent histories and support repair. It would not retrospectively certify an already harmful history as harmless.

The spatial and causal scope matters just as much. Protection of one execution environment is a local result. Protection of every relevant process in a physical region is stronger. Protection of all conscious life throughout reality is stronger still. A universal moral commitment can encompass subjects beyond one's causal reach; a claim of successful enforcement cannot silently do so.

An adequate proof has several connected obligations. It must identify what counts as experience and suffering for the systems involved. It must show that every relevant physical realization is represented conservatively. It must establish that allowed transitions preserve protection, including their intermediate states. It must cover the machinery that enforces those transitions. It must account for disturbances and other actors. It must also establish whatever continuation and agency properties are being promised.

The missing connection most easily hidden by sophisticated mathematics is the connection between formal semantics and actual experience. A theorem can prove that a variable never exceeds a threshold. Whether that variable bounds suffering is a further claim. An implementation can perfectly obey a theory that omits the subjects it harms. More powerful theorem proving improves our ability to derive consequences from assumptions; it does not by itself establish that an empirical assumption describes the world.

This suggests a useful discipline: every assurance claim should show its entire chain of dependence, from the subject and the proposed experience boundary through physical modeling, control, implementation, and continued support. The chain can contain empirical evidence, mathematical proofs, and explicit unresolved assumptions. Those forms of support should remain distinguishable. The most important unresolved premise should be at least as visible as the strongest theorem.

There must also be no unexamined host at the end of the argument. Suppose a virtual universe excludes every prohibited transition in its internal physics. If its host can instantiate an uncensored version elsewhere, alter the inhabitants directly, or create suffering within the host's own computations, the universe's internal certificate has not established protection against those possibilities. Moving the safeguard into a hypervisor, firmware, or a mechanism that operates before ordinary boot moves the boundary. It does not automatically close it.

The same applies to a hierarchy of realities. A protected inner world can be a genuine achievement. Its inhabitants remain dependent on any enclosing layer that can defeat the protection. Adding another layer produces stronger assurance only if the added layer closes a relevant causal route or inherits a justified guarantee of its own. A chain of protectors cannot obtain absolute security by endlessly referring to the next protector.

Several implementation families deserve investigation. Restricted digital physics could permit only certified state transitions. Hardware could mediate the execution of potentially conscious processes and isolate them from unauthorized modification. A carefully designed cognitive substrate might lack realizable mechanisms of negative valence while retaining the capacities its inhabitants value. More ambitious physical engineering might restrict the reachable organizations of a region so that forbidden experiences have no realization there. Each is a research direction with distinct scientific prerequisites.

The third possibility is especially important. If a mind can be constructed so that all of its admissible experiential states exclude suffering, protection could become partly a property of the mind's realization. That would reduce reliance on detecting each threatening external event. The burden would be to show that the restriction survives inputs, learning, self-modification, faults, composition, and attempts to construct suffering subagents. Calling a substrate safe by construction would be justified only after those obligations were met.

Planck-scale language cannot substitute for this work. A fundamental unit in a physical theory does not establish a universal programmable frame clock, nor an opportunity to inspect a state before reality commits it. Minimal-length and discrete-spacetime proposals belong to unresolved physical theory [15]. A future discovery might make a deeper enforcement mechanism possible. The architecture should specify what the mechanism must accomplish without assuming the discovery has already occurred.

Nor is it generally sufficient to compute a dangerous next state in a sandbox, inspect it, and decide whether to release it. If the computation itself realizes the experience, the violation occurred during inspection. The same objection applies to speculative execution, predictive world models, proof search containing detailed agents, and replay. Protection must apply wherever experience is physically instantiated. A safe abstraction needs evidence that it is an abstraction in the morally relevant sense.

Finally, probability zero and impossibility are different claims. A probability measure can assign zero to individual outcomes that remain in its possibility space. An almost-sure guarantee permits exceptional histories of measure zero. Structural exclusion says that no admissible history contains the event at all. The ultimate ZHM objective is the latter, relative to a specified physical scope and a justified account of what is admissible.

Whether a nontrivial, fully justified realization of this objective is physically achievable is open. Current theory supplies neither a universal consciousness detector nor a mechanism that closes every physical route to suffering. The ambition survives that uncertainty as a precise research question. Its completion would require closing the gaps; its pursuit should report exactly which gaps have closed.

5 The life that protection is for

The positive vision deserves its own name because absence of catastrophe does not describe a flourishing life. This essay calls the desired condition protected flourishing, and calls the broader system that could sustain it a Protected Flourishing Architecture. ZHM remains the name of the project and its constitutional foundation.

Heaven 1.0 is an intelligible informal name for the original aspiration. It evokes an existence in which the space of worthwhile possibility becomes immense and suffering no longer threatens from beneath it. In technical discussion the phrase carries ambiguities about religion, perfection, emotional uniformity, and achievement. Keeping it as an informal historical description allows the aspiration to remain vivid while the research terms carry the definitions.

The constructive possibility is extraordinary. If experience, embodiment, and environment become extensively designable, a person might choose forms of intelligence and perception unavailable to present biology. Worlds could support unfamiliar arts, enormous projects, deep relationships, voluntary transformation, and pleasures whose character we cannot now describe. Personal environments could offer expansive creative control, subject to the equal standing of any other beings they contain.

There is no need to promise literal omnipotence or access to every logically describable experience. Some possibilities conflict with other persons' rights. Some may be physically unrealizable. Others may undermine the capacities that make choice meaningful. The aim is an expanding space of valuable, self-endorsed experience whose exploration does not require exposure to catastrophic domination.

An important scientific question follows: which functions presently served by aversive experience require that experience? Error correction, damage information, urgency, boundary setting, and the pursuit of difficult goals can be described functionally. It remains to be established how far their valuable roles can be implemented without negative phenomenology, especially while preserving a particular person's identity and motivations. Distinguishing information from suffering opens a research question; it does not answer it by definition.

A civilization could be exceptionally pleasant while leaving this problem unsolved. It could keep residents content by hiding information, narrowing their desires, or rewarding compliance. ZHM requires independent protection of agency and epistemic access so that satisfaction reports cannot become the sole measure of success. Residents need meaningful ways to inquire about their conditions, revise arrangements, seek another provider, and refuse optional changes to their minds.

These rights are relational. One person's desired experience cannot require another person's compelled affection, obedience, or continued participation. Immersive fiction may be available under informed terms, but an actual experiencing inhabitant acquires standing. A creator's imagination is not a jurisdictional loophole.

Challenge need not mean exposure to the worst states a substrate can realize. Discovery can be difficult because a question is hard. A game can be absorbing because its outcome is uncertain. A relationship can be meaningful because the other person is independent. Whether these and much stranger forms of value can persist without any suffering is a central part of the positive research program.

Protection would have succeeded most deeply when a person can become more curious without first needing to become less vulnerable than everyone else. The point of the architecture is to make more of life inhabitable. Its positive achievements must be evaluated in the lives people can actually lead, rather than inferred from an absence of recorded incidents.

6 Keeping a future reachable

An apparently safe present can conceal the loss of every safe future. A habitat can be comfortable while consuming the final resources needed for rescue. A provider can honor every current request while accepting obligations it will soon be unable to support. A mind can remain below a harm threshold while entering a trajectory from which no available intervention can prevent crossing it.

The relevant boundary therefore includes the loss of viable continuations. Viability theory studies conditions under which a system can remain within constraints through its future evolution [6]. Applied here, the question is whether the current state, history, resources, and available actions still admit a strategy that preserves the specified protections against the disturbances included in the model.

This is stronger than checking the next visible state. A proposed action must be examined along its execution path, and it must leave a supported way forward. Even the action most similar to what a resident requested may be unacceptable if it consumes the last means of protecting that resident. Conversely, imposing a large restriction needs justification when a less intrusive viable continuation exists.

Safety is only part of the obligation. A system might avoid every bad event by leaving everyone permanently suspended. The distinction between safety and liveness in formal methods helps expose

this failure: excluding forbidden events does not establish that required progress ever occurs [7]. In a conscious domain, a promise of continuation requires supported opportunities for activity and agency, with appropriate respect for a subject's own choice to rest or suspend.

A meaningful recovery commitment therefore needs more than a review date. It needs resources, a responsible custodian, conditions for resumption, and a strategy that remains effective under the stated faults. Where recovery can be proved within a bound, that bound should be given. Where only eventual recovery follows, its assumptions should be explicit. Where no recovery strategy is known, the record should say that the obligation is unresolved.

Two clocks are important. The subject's experienced time concerns exposure while it is active. An independent service clock concerns the obligations of those maintaining it, including during suspension. Stopping the first must not stop the second. Otherwise a provider can claim that a suspended subject has waited no subjective time while abandoning it indefinitely.

The deeper requirement is preservation of options. Rescue should leave the endangered subject with a future it can use. Resources reserved for that purpose are part of the protection mechanism. An architecture that spends them on optional expansion has changed its safety condition, even if no harmful experience has yet occurred.

7 Constructing a protected domain

A useful first implementation would separate the system that proposes actions from the mechanism that authorizes consequential execution. The proposing system could be complex and adaptive. The authorizing mechanism should be small enough, and its accepted evidence explicit enough, to permit strong scrutiny. The aim is to constrain the power a planner can exercise even when its internal reasoning is difficult to verify.

There are established precedents for parts of this arrangement. Proof-carrying code associates executable material with evidence checked against a specified policy [8]. Shield synthesis and runtime assurance constrain the behavior of complex components through a protective mechanism [9, 10]. Verified systems such as seL4 illustrate both the strength of formal guarantees and the importance of stating what their proofs assume [11]. None of these results establishes a theory of conscious welfare. They provide methods for enforcing a policy once an adequate policy and its physical interpretation are available.

A candidate ZHM domain would first declare its scope: which processes, substrates, physical interfaces, environmental conditions, and periods of operation it covers. It would then state the consciousness and harm hypotheses used, the kinds of evidence supporting them, and the properties being enforced. A formal certificate could bind a proposed operation to a particular implementation, resource envelope, theory version, and recovery contract.

Complete mediation is essential within the claimed scope. The safeguard must cover the channels through which a prohibited outcome could occur: execution, copying, scheduling, memory modification, physical actuation, delegation, and relevant interconnection. The correct implementation depends on the domain. The governing requirement is that an actor cannot evade the certificate by changing the route through which it exercises the same power.

Admission and maintenance are separate responsibilities. A system should check support capacity before creating a dependent subject. Once a subject exists, failure of an admission policy does not cancel that subject's standing. The provider must continue to address the obligation it created. Resource limits constrain future expansion; they do not turn existing residents into unauthorized objects eligible for disposal.

The safeguard also needs a supported fallback. Rejecting a proposed transition is protective only if the current situation can remain safe or move safely elsewhere. A fail-closed mechanism that cuts off life support has solved an access-control problem by creating a welfare emergency. Biological organisms, active digital processes, dormant records, and distributed minds require different fallback semantics. There is no universal pause instruction.

Scientific uncertainty belongs inside the design. Indicator-based approaches to AI consciousness derive evidence from theories of consciousness rather than treating surface resemblance as decisive [5]. A ZHM implementation should similarly maintain explicit, scientifically motivated alternatives. The alternatives need admission criteria, reasons for revision, and documentation of disagreement. An unrestricted collection of every imaginable hypothesis would make action impossible; a collection chosen to excuse the operator would make protection empty.

A subtle difficulty remains even when each theory appears manageable. One theory may say that action A protects the subject and action B harms it; another may say the reverse. Having a safe policy under each theory separately does not establish the existence of one policy safe under the unresolved alternatives. The common action must satisfy the relevant constraints together. Where no such action exists, disagreement has produced a genuine decision problem. It cannot be removed by averaging the certificates or displaying the most reassuring one.

Evidence can sometimes resolve the conflict. Restricted operations can sometimes avoid it. A different substrate or additional rescue capacity may make the jointly protected region nonempty. These are concrete research responses. Otherwise the implementation must report that its stronger guarantee is unavailable and make an accountable decision under uncertainty. It should not claim that the word conservative has supplied a solution.

Theory changes also change what is certified. Discovering a previously unrecognized kind of subject can shrink the set of defensible operations. That discovery creates duties of investigation and protection, including for already running systems. It does not justify erasing the possible subjects to restore compliance. Conversely, removing a hypothesis requires evidence and review, especially when its removal conveniently releases resources or powers.

Abstract interpretation and restricted execution can help reason about relevant properties without reproducing every underlying computation [12]. Whether an abstraction adequately bounds conscious harm remains an additional scientific question. The useful objective is to expand the class of activities for which defensible evidence can be supplied while keeping unsupported interpretations visible.

Finally, the protector is inside the moral scope. Its evaluators, simulated advisers, internal experiments, and temporary subagents cannot be treated as costless simply because they serve safety. Every proposed shortcut must ask where the relevant computation occurs and whether that computation itself could contain subjects.

8 Worlds without private victims

The social architecture developed in v3 remains central. Minds with profoundly incompatible preferences should have extensive room to live without acquiring unrestricted power over one another. Earlier versions called this the Archon Hausdorff principle. Here, the shorter Archon separation principle emphasizes the engineering content: separate incompatible exercises of causal authority while preserving the persons who hold them.

The original name borrowed a topological image of separable neighborhoods. That image is useful as inspiration. Actual protection requires properties of interfaces, shared resources, and physical implementation. Topological terminology does not establish them.

A protected domain could allow its inhabitants to design local environments, transform themselves, form associations, or withdraw into solitude. Its local rules would remain constrained by the standing of every subject affected. A resident's authority over a landscape would not confer authority to manufacture another resident's consent. A creator's authority over a simulation would not confer ownership of the subjects it instantiated.

This distinction supplies a principled response to hostile minds. Restrict the capabilities through which they can impose prohibited harm. Preserve supported opportunities for compatible activity. Protection does not depend on repentance, affection for the constitution, or ideological agreement. Restoring particular external powers can depend on evidence that the relevant interface is safe.

A private world cannot satisfy a preference whose fulfillment requires a real victim to undergo prohibited suffering. Representations that do not instantiate experience pose a different question, but their non-instantiating status cannot simply be stipulated for a complex real system. This rule reaches nested worlds, outsourced computations, and purportedly fictional persons wherever experience is actually present.

A federation of such domains would need contracts for their interaction. Contract-based systems design offers methods for relating component assumptions and guarantees [13]. The ZHM application requires attention to power, timing, communication, resources, and emergent organization. Two independently acceptable components can form an unacceptable coupled system. If consciousness can depend on distributed organization, protection may need to cover a joint subject that neither local inventory contains.

Similarly, independent providers can share a vulnerability. Different logos do not establish different scientific theories, hardware dependencies, software defects, or sources of power. A useful federation maps these dependencies and tests failures that affect whole groups. Replication can improve resilience without eliminating a common cause.

Local rescue cannot depend on instantaneous global agreement. Distributed consensus has formal limits under particular timing and failure assumptions, including the classical impossibility result for deterministic consensus in a fully asynchronous system with a possible crash [14]. A domain needs locally authorized protective behavior during communication failures. Wider institutions can review that behavior and manage longer-term commitments without becoming part of every urgent control loop.

Migration should transfer supported responsibility rather than merely export data. The destination must be able to maintain the relevant protections before the origin withdraws support. Residents need ways to leave a provider without surrendering the information, resources, or practical means on which their continuation depends. These provisions make sovereignty more than a permission written by the current host.

Federation creates a promising direction, not a guarantee against every external power. An actor outside the protected substrate may retain dangerous capabilities. Incorporating that actor, preventing a particular external interaction, and extending control over a physical region are different interventions with different burdens. Local success should make its boundary intelligible rather than conceal it beneath a universal name.

9 Continuity across transformation

Rescue must distinguish preserving information, reconstructing a functioning process, and preserving the same subject. These claims can come apart. A stored description may be incomplete. A successful reconstruction may leave questions about personal continuity. A contented successor does not by itself establish that the endangered person survived.

The distinction matters whenever minds can be copied, restored, divided, merged, or transferred. A provider might improve every visible welfare measure by replacing difficult residents with agreeable successors. An architecture focused only on current valence could mistake this for excellent care. ZHM requires continuity and consent to receive separate attention.

In-place relief deserves priority where it offers the best evidence of preserving the subject and is otherwise feasible. More disruptive interventions require a case-specific account of their effects, informed prior preferences where available, and preservation of recoverable options. There is no general mathematical distance between two mind states that automatically selects the morally closest continuation.

Copies also defeat convenient accounting shortcuts. Distinct executions may contain distinct experiences even when their initial descriptions are identical. Exactly how to individuate overlapping or shared implementations remains disputed. An assurance record should preserve that uncertainty and bound exposure under defensible interpretations. It should not divide moral concern by the number of copies because a file was inexpensive to duplicate.

Restoration cannot cancel an experience that already occurred. It may support a worthwhile continuation. It may remove a memory at the subject's informed request. Neither operation makes the earlier event absent from the world's history. This is why a global reset button cannot substitute for prospective prevention.

The same reasoning handles speculative future reconstruction without depending on a theory of resurrection. Suppose a future system reconstructs a person from records, generates a close copy, or creates a new subject with similar memories. Whether numerical identity continues is a further question. If a conscious subject is instantiated, that subject has standing immediately. Protection does not wait for a verdict on whether it is the original person.

This is a useful scope result. A constitution for experience can cover possible future reconstructions without assigning them a probability or assuming they are technically possible. It also need not demand that every possible person be created. The obligation concerns subjects who exist and the responsibilities incurred by creating new dependence.

Long life must remain compatible with self-directed life. Voluntary sleep, suspension, transformation, and questions about ending one's own continuation require procedures that respect the subject's informed agency. The anti-erasure commitment constrains what operators may do for their own convenience. It must not become an entitlement to compel every resident to remain active forever.

10 The protector must also be constrained

A system authorized to intervene in minds possesses powers that require their own justification. A benevolent purpose does not resolve this problem. The same mechanism that blocks coercive modification might perform it. The same inference system that detects hidden subjects might be used to inspect every private thought. The same emergency authority that enables rescue might become permanent administration of everyone's life.

The architecture should minimize such powers by construction. Interface restrictions, isolation of dangerous capabilities, verifiable execution conditions, and portable support can sometimes provide protection without unrestricted access to mental content. Where a stronger claim requires more intrusive access, that requirement must be disclosed and evaluated. An unmeasurable promise of absolute safety cannot justify unlimited surveillance.

Emergency intervention needs a specific threatened violation, relevant evidence, a proportionate action, and a supported path back to ordinary autonomy. An emergency is not completed when the subject stops objecting. It is completed when the threatened condition has been addressed and the subject's practical authority has been restored as far as feasible.

The catastrophic floor should be non-waivable by operators, governments, creators, and local majorities. Prior consent cannot authorize another actor to disable all future refusal and retain a person in a catastrophic state. Ordinary chosen risk requires a more discriminating policy, including the consequences for future selves. A constitution that calls every discomfort catastrophic would acquire a limitless power to intervene.

These commitments require institutions as well as code. Independent review, protected dissent, accessible reasons, separation of emergency powers from ordinary administration, and routes to alternative providers all matter. A collection of approvers can still collude or share a mistaken model. Its authority should remain bounded by specific procedures and evidence requirements.

Constitutional stability and scientific revision must be designed together. A detector frozen forever can become a permanent source of exclusion. An administrator free to redefine suffering can abolish protections without changing their names. Updates therefore need explicit semantic changes, independent assessment, regression evidence, and a plan for any subjects whose protection changes during the transition. Rollback should repair the machinery; rewinding the lives of residents is a separate act requiring its own justification.

Scarcity presents another limit. If several existing subjects face harm and available resources cannot protect all of them, the system must make a real allocation decision. Non-compensation forbids manufacturing a catastrophic victim for unrelated gain; it does not ensure every rescue conflict has a harmless resolution. Severity, urgency, prospects of successful relief, fairness, and continuity can all be relevant. Uncertainty and unavoidable losses should remain visible.

Attackers may exploit this obligation by creating dependent subjects or fabricating emergencies. The response must restrict the capacity to create further unsupported dependence while protecting actual subjects already brought into existence. Authentication can help establish an operator's entitlement to resources. It cannot decide whether an experiencing victim matters.

These cases expose the difference between a normative floor and an unconditional decision algorithm. Giving every speculative catastrophe infinite weight can paralyze action or hand unlimited authority to whoever makes the most extreme claim. Ignoring small probabilities can disregard serious harm. ZHM needs evidence standards and accountable decisions under uncertainty, with explicit recognition that reasonable theories may disagree about a particular case.

Its moral commitments are intended to constrain power through this uncertainty. They should not become a vocabulary through which the protector exempts itself from scrutiny.

11 Research that can change the answer

The immediate research task is to turn the architecture into questions whose answers could disappoint its author. An attractive vision is insufficient evidence for a large viable region, a reliable bridge from mechanism to experience, or a system that can preserve both protection and agency.

The first tractable question concerns control logic. Can a restricted execution domain enforce an explicit history-dependent boundary, preserve recovery options, and provide a separate progress guarantee under a stated fault model? Appendix C supplies a small executable example using finite graphs and stipulated predicates. Its base model has five states, twelve actions, and thirteen possible action-outcome traces. The calculated viability region contains three states and admits seven actions. The example checks seven architectural counterexamples, including an unsafe intermediate state, an irreversible precursor, indefinite suspension, an omitted disturbance, incompatible theory-specific controllers, unsafe composition, and a changed harm interpretation.

These results establish properties of the toy graphs. The labels do not constitute a model of consciousness, and passing the checks gives no evidence that an actual mind is protected. Their value is methodological: they expose distinctions that a more realistic implementation would have to preserve, and they give other researchers something inspectable to replace or extend.

A stronger continuation would encode the specification in a proof assistant or a mature model checker, with an independent account of the implementation relation. It should include partial observability, resource depletion, delayed intervention, migration, and explicit failure of the monitor. Evaluation should compare an endpoint filter, a safety shield, and a controller that also guarantees supported progress. A system that rejects every activity must not score as the best design merely because its incident count is zero.

The second question concerns the bridge to welfare. For a restricted class of architectures, what evidence would support a conservative claim about consciousness or valence? The work should begin with low-risk mechanistic analysis, carefully delimited abstractions, and evidence already available from relevant science. It must not establish the existence of a catastrophic state by creating a candidate victim and observing whether it suffers. The ability to detect a signal after the event would not meet the prospective requirement.

Particularly valuable results would identify stable relations between mechanisms and welfare-relevant properties that survive changes of implementation. A negative result is also valuable: if a proposed indicator can be changed without the asserted experiential consequence, it weakens the case for treating that indicator as an enforcement target. Disagreement between theories should be recorded as a source of different protective decisions, not absorbed into a single opaque confidence score.

The third question concerns continuity and positive life. Which interventions preserve capacities, autobiographical organization, relationships, and the subject's informed endorsements? Which affective changes leave people able to recognize the resulting life as their own? Which forms of challenge and learning remain available without suffering? These questions cannot be settled by optimizing a reward readout. They need carefully justified methods and explicit attention to what the measurements leave unresolved.

The fourth question concerns composition over time. Can two independently supported domains transfer an obligation without a gap? Can a theory update broaden protection without stranding existing residents? Can a federation continue local rescue during a network partition without enabling permanent emergency rule? Can it preserve useful independence among its monitors and providers? These are opportunities for work in systems security, distributed systems, formal methods, governance, and consciousness research to meet around specific interfaces.

The results should be reported in several dimensions: violations admitted under the model, residual uncertainty in the model, successful supported continuations, time spent under restriction, lost options, continuity uncertainty, resource commitments, and common dependencies. None is a substitute for the others. A welfare metric can miss identity destruction; a continuity metric can preserve a trapped victim; a utilization metric can reward unsafe expansion.

An initial collaboration could therefore have a bounded deliverable: formalize one protected execution domain, supply a threat and assumption register, reproduce its counterexamples, and determine whether a useful jointly safe and live policy exists. The next step would be chosen from the largest unresolved assumption, rather than from the most impressive additional feature.

12 The conditions of success

Several outcomes would require substantial revision of the proposal. The scientifically credible theories might leave no useful common protected region. Consciousness and valence might remain too difficult to constrain for the intended generality. A supposedly safe substrate might acquire harmful experiential possibilities through learning or coupling. A protection mechanism might require powers incompatible with acceptable agency. Physical resource limits might defeat the promised duration of continuation.

These possibilities should not be hidden inside the word advanced. Greater intelligence can improve modeling, discover mechanisms, and construct proofs. It does not establish that every desired constraint is jointly satisfiable. A complete proposal must permit the answer that a particular form of protection cannot be achieved under particular physical and ethical conditions.

Yet there are many informative successes short of the universal result. A verified restricted domain can exclude a specified class of modeled violations. Better science can reduce a consequential uncertainty. A migration protocol can close a gap in support. A protected substrate can remove a harmful capability previously available to its operators. A federation can reduce the degree to which a person's future depends on one institution.

These achievements should accumulate without being mistaken for completion. The ultimate claim would require a defensible account of experience, comprehensive coverage of the relevant physical possibilities, preserved protection throughout every admissible continuation, and supported lives worth continuing. A guarantee about catastrophic suffering alone would leave the stronger suffering-free objective open. A guarantee about suffering alone would leave agency and continuity open.

The direction of the project is to enlarge both protection and possibility. Minds should be able to become more capable without acquiring an unlimited right to dominate others. Subjects should be able to inhabit longer lives without being required to accept a recurring catastrophic wager. A civilization should be able to explore unfamiliar forms of experience while retaining the ability to explain why no one is being abandoned beneath its success.

That is the mature meaning of the Zero Hostility Matrix. Hostile intentions may remain possible. Profound disagreement may remain possible. The project seeks to remove the causal power to convert those differences into catastrophic conscious histories, and ultimately to make suffering-free, self-directed life a durable condition of existence.

The decisive evidence would be an account a resident could inspect: who is protected, against what, through which mechanisms, under which conditions, with what remaining uncertainties, and with which futures still available. An ultimate architecture would close those uncertainties sufficiently to establish the stronger claim about the actual reality it governs. Until then, progress consists in making each part of that account more justified, more comprehensive, and less dependent on anyone's discretionary benevolence.

Appendix A Protection across long horizons

A1 Conditional hazard and cumulative risk

Let T be the index of the first catastrophic exposure block, with T equal to infinity when no catastrophe occurs. For blocks with positive probability of prior survival, define the conditional hazard h_k by:

$$h_k = Pr\left(T = k \mid T \geq k\right).$$

The chain rule gives:

$$Pr\left(T \leq n\right) = 1 - \prod_{k=1}^{n} \left(1 - h_k\right).$$

No independence assumption is required. The conditioning already incorporates survival through earlier blocks. If every remaining h_k is at least a fixed positive h, survival is at most (1-h) raised to the number of remaining blocks, which tends to zero. More generally, if all h_k are less than one, divergence of their sum implies zero limiting survival probability. A finite sum leaves positive survival probability, but does not by itself make the lifetime risk small.

For a constant conditional hazard h and a chosen lifetime ceiling epsilon over n blocks:

$$h \leq 1 - \left(1 - \varepsilon\right)^{1/n} \approx \frac{-\log\left(1 - \varepsilon\right)}{n}.$$

The approximation is for large n at fixed epsilon. A ceiling of one percent is used below solely to illustrate the scaling; it is not a proposed ethical standard.

Horizon in annual blocks	Largest constant annual hazard for a one percent cumulative ceiling
One thousand	About 0.0000100503
One quadrillion	About 1.00503×10^{-17}
One quintillion	About 1.00503×10^{-20}

For comparison, an annual conditional hazard of one in a quintillion over a quadrillion annual blocks gives a cumulative probability of about 0.09995 percent. An annual conditional hazard of one in a billion over that horizon gives survival of approximately exp(-1,000,000). The companion computes these examples with numerically stable logarithm and exponential functions; a displayed floating-point probability of 1.0 in the latter example means rounding, not an exact finite-horizon equality.

A2 Summable bounds and the scope of uncertainty

Let B_j be a violation in exposure block j. If ε_j is a valid bound on its probability, countable subadditivity gives:

$$Pr\left(\bigcup_{j \geq 1} B_j\right) \leq \sum_{j \geq 1} \varepsilon_j.$$

The bound requires no independence. The allocation $\varepsilon_j = \varepsilon / 2^j$ has total epsilon for j starting at one. This shows that an unbounded number of exposure blocks is compatible with a bounded risk estimate if valid bounds decrease sufficiently quickly. It does not establish that any physical implementation can achieve those bounds. Exact zero cannot follow from a positive sum merely because the sum is small.

Now let E be the event that the full set of relevant premises remains valid over the claimed horizon. If its failure probability is bounded by delta and the block bounds hold conditional on E, then:

$$Pr\left(\bigcup_j B_j\right) \le \delta + (1-\delta)\sum_j \varepsilon_j \le \delta + \sum_j \varepsilon_j .$$

For a conditional invariant, the second contribution can be zero while the first remains unresolved. Delta concerns simultaneous validity of the stated premises over the horizon. It must not be invented from a present-day accuracy score or automatically recharged once per year. New components, disturbances, or semantic revisions require an updated model of the relevant uncertainties and their dependence.

If period k contains at most N_k subjects, each with at most M_k relevant exposure opportunities, and each opportunity has a valid probability bound q_k, another union bound yields:

$$Pr(\text{any violation}) \le \sum_k N_k M_k q_k .$$

The bound presupposes a justified method of counting the opportunities and bounding their risks. For geometric illustrative sequences with $N_k M_k$ proportional to g^k and q_k proportional to a^{-k}, g/a less than one is sufficient for summability. Growth of population and experience can outrun improving per-opportunity reliability. This calculation supplies a requirement for an exposure model, not a measure of subjective time.

A3 Zero probability and unreachable histories

Let A be the set of admissible histories and V the subset containing a violation. A probabilistic guarantee states that the measure of V is zero, or at most epsilon. A structural guarantee states:

$$A \cap V = \varnothing .$$

The latter excludes every admissible violating history. Probability-zero events can remain possible, so the former does not generally imply the latter. The ultimate ZHM target uses structural exclusion. An implementation must state how its admissible histories relate to actual physical possibilities.

Appendix B Conditional protection and its limits

B1 Physical and experiential scope

Fix a physical scope R, a starting condition x_0, a time boundary t_0, and an explicitly stated class of admissible external interactions. Let Hist(R,x_0) be the resulting physical histories from t_0 onward. A scientific interpretation eta associates relevant physical histories with possible subjects and experiences. Define V_0 as histories containing any suffering after t_0 and V_{cat} as histories containing catastrophic suffering. Within a fixed interpretation, V_{cat} is a subset of V_0.

The ultimate suffering-exclusion specification is:

$$Hist\left(R, x_0\right) \cap V_0 = \varnothing .$$

Replacing V_0 with V_{cat} states the narrower catastrophic floor. Neither equation states continuation, agency, or positive welfare. A complete specification must conjoin appropriate obligations for those properties. Extending R changes the theorem's scope. Extending it to all relevant physical reality requires an account of all the resulting processes and interactions.

These are specifications, not established facts. In particular, the existence and adequacy of eta are unresolved scientific obligations. Hist must include the physical realization of the protector and all relevant intermediate computation, rather than only the world it presents to its residents.

B2 Histories made explicit in a transition model

Let Q be an augmented state space. Its states include all history summaries needed by the specification, support resources, pending recovery obligations, implementation versions, and relevant facts about possible subjects. If a finite summary is insufficient, the mathematical state may contain the relevant history itself. This preserves expressiveness at the cost of tractability.

Let Theta be the admitted family of consciousness and harm interpretations. Each theta defines a forbidden set H_θ in Q, with S_θ its complement in Q. The common permitted set is:

$$S_\Theta = \bigcap_{\theta \in \Theta} S_\theta .$$

This construction requires a common representation or justified translations between model-specific states. It does not solve disagreements about representation. Agency and continuity constraints can be added as separately named predicates; the following proposition addresses only the represented exclusion property.

For state q, enforceable action u, and admissible disturbance w, let Tube(q,u,w) contain every modeled intermediate state during execution. Let Post(q,u,w) contain every possible modeled successor. Tubes include the initial state, relevant continuous evolution, and any internal evaluation that falls within scope. Successors are also included in the tube. A controller must use information actually available to it.

B3 Invariant protection under complete observation

Choose a nonempty K contained in S_Θ and a realizable policy pi such that, for every q in K and every admissible w:

$$Tube(q, \pi(q), w) \subseteq S_\Theta, Post(q, \pi(q), w) \subseteq K.$$

Proposition 1. If execution starts in K, the transition model covers all relevant evolution, every consequential transition is mediated, and the implementation applies pi with the required timing, every represented finite execution prefix avoids all H_θ.

Proof. The tube condition protects the first transition and the successor condition returns execution to K. The same argument applies at each subsequent step. Induction establishes the claim for every finite prefix. For infinite discrete executions, every visited state belongs to some finite prefix. For continuous or hybrid evolution, require time-divergent execution or explicit coverage of accumulation points and their continuations; otherwise a Zeno limit can be omitted from the proof. These conditions are part of the model obligation.

The existence of K and pi is an assumption to discharge for a concrete domain. The argument is a standard invariant argument adapted to this specification. It supplies no new general algorithm for consciousness detection, controller synthesis, or physical prediction.

B4 Partial observation and the common policy requirement

When the controller cannot distinguish all states, let B be the information state containing all augmented states still consistent with its observations and admitted models. A permissible action must work for every q in B, every admitted interpretation, and every admissible disturbance. Every resulting information-state update must return to a jointly protected region of information states.

The quantifier order matters:

$$\exists u \forall q \in B \forall \theta \in \Theta \forall w \, Protected(q, u, \theta, w).$$

This differs from allowing a different action for each unobserved state or theory. For a two-action counterexample, suppose θ_1 permits only action a while θ_2 permits only action b. Each theory admits a protected controller separately; unresolved uncertainty leaves no common action. Appendix C implements this finite example. A correct robust method may therefore return an empty region. It must not silently exchange the quantifiers to obtain a favorable answer.

B5 The bridge from proof to physical protection

Let alpha map actual histories within the claimed scope to represented histories. To infer physical protection from Proposition 1, establish at least the following conditions jointly: every actual evolution has an allowed represented evolution under alpha; every actual violation maps to a represented violation under at least one admitted interpretation; and the implemented controller realizes the modeled decisions and timing. Initial conditions, unmediated channels, internal computations, and boundary interactions must all be covered.

Proposition 2. Under these coverage and implementation conditions, avoidance of all represented violations entails avoidance of actual violations within the declared scope.

Proof. Suppose an actual violating history existed. Coverage would map it to an allowed represented history, and semantic conservatism would make that represented history violating under an admitted interpretation. This contradicts the represented exclusion guarantee.

The proposition isolates the central burden. It does not discharge semantic conservatism, completeness of the physical model, or correctness of the realization. A certificate that omits these premises proves a narrower statement. A hidden conscious subprocess, an unmodeled host intervention, or an incorrect interpretation can defeat the physical conclusion while leaving the formal theorem valid.

B6 Continuation and bounded recovery

A safety invariant can permit infinite inactivity. Add an explicit recovery obligation for subjects whose interruption is involuntary. For a finite toy system, one sufficient progress condition is a nonnegative integer rank that strictly decreases on every pending recovery transition until supported activity resumes. If the rank starts at m, recovery takes at most m such transitions. If each transition also has a justified physical duration bound, a physical recovery deadline follows. Without the timing bound, a step bound does not establish an external deadline.

For a real system, recurrence of a designated active state is only an operational proxy. Meaningful agency requires opportunities to choose and exercise relevant capabilities, and identity preservation requires additional evidence. Forced recurring activation of an unwilling subject does not satisfy sovereignty. Genuine voluntary suspension and its terms must be represented separately.

B7 Timing and accumulating harm

For a stipulated scalar cumulative exposure measure z, suppose z grows at rate at most g_{max} in external time until a protective action becomes effective. Let L include observation, detection, authorization, actuation, and settling delay. If the forbidden boundary is z_{star}, a sufficient margin for this accumulation model is:

$$z_{star} - z(t) > g_{max} L .$$

The fallback must then prevent subsequent crossing and preserve the other relevant constraints. If g_{max} is derived from a bound on subjective time and an exposure rate per subjective interval, both bounds require separate justification. A sudden jump across the boundary violates the accumulation assumptions and requires a pre-execution or structural constraint. An external stopwatch alone establishes neither form of protection.

B8 Expressiveness and the limits of general certification

For an unrestricted computational language, a total algorithm deciding whether every program avoids even an ordinary designated bad state would solve the halting problem. Given an arbitrary program p, construct a program that simulates p and enters the designated state if p halts. Deciding unreachability for this construction decides non-halting. This standard reduction concerns a stipulated computational predicate. It does not prove a particular theory about consciousness.

The constructive implication is to seek restricted languages, verified program classes, conservative abstractions, bounded resources, or proof-carrying operations. A sound checker can reject or decline to certify programs that are actually safe. Universal acceptance of every safe unrestricted program is not a prerequisite for useful protected domains. It is a stronger demand that cannot generally be met by such a total decision procedure.

B9 Revision and composition

Proposition 3. Hold dynamics, available actions, disturbance assumptions, observation, and all non-harm constraints fixed. If the admitted theory family expands from Theta to a larger family, the common permitted set can only shrink. The greatest region from which one policy can guarantee avoidance of all admitted forbidden sets can also only shrink.

Reason. Any policy satisfying every interpretation in the larger family satisfies every interpretation in the smaller family. The converse need not hold. Thus expansion of semantic coverage cannot by itself justify expanding the certified operating region. New mechanisms, evidence, resources, or actions may change this result by changing the held-fixed conditions.

This is a useful check against a misleading upgrade claim. A system may recognize more possible subjects while becoming less able to certify its present activities. The appropriate response is supported adjustment, additional resources, investigation, or a narrower claim. Dropping newly inconvenient subjects from the specification is not a proof of preserved protection.

For composition, local regions K_A and K_B are insufficient. A product-region argument must establish that the joint region contains no joint forbidden state, that joint intermediate trajectories remain permitted, and that every allowed interaction preserves the product or another certified joint region. Interface assumptions must actually be enforced; shared support resources and potentially distributed subjects must be represented. Circular promises in which each component assumes the other's safety do not supply this argument. Liveness and continuity need separate composition conditions.

Appendix C An executable architectural example

C1 Scope and method

The companion program zhm_v4_model.py uses Python's standard library. It manipulates names of states and finite transition traces. Words such as harm and paused are stipulated labels. There is no agent, world model, valence model, or simulated phenomenology. The example tests the internal logic of architectural claims.

Each action has one or more possible traces representing disturbance outcomes. A trace can contain intermediate states before its final successor. The checker starts with states outside the forbidden set and repeatedly removes any state lacking an action whose entire possible execution stays permitted and whose successors remain in the current region. On this finite model the procedure reaches the greatest fixed point for nonblocking safety.

State	Stipulated interpretation	Key available behavior
home	Supported ordinary operation	Rest, work, pause, spend reserve, speculative trace, fault-sensitive action
work	Supported activity	Continue or return home
paused	Interrupted activity	Wait or resume
trap	No remaining safe continuation	Forced transition to harm
harm	Forbidden label	Remain in harm

The base model contains five states, twelve actions, and thirteen action-outcome traces. Its fixed point is {home, work, paused}. Seven actions are admitted from that region. As an independent cross-check, the program enumerates all twenty-four deterministic memoryless policies, explores every reachable disturbance branch for each starting state, and obtains the same winning region. Both procedures are finite calculations, not proof-assistant certification of the checker itself.

C2 Counterexamples reproduced

Case	Witness or changed condition	Architectural conclusion
Intermediate violation	home to harm to home	An endpoint filter accepts a path containing a forbidden intermediate state
Exhausted continuation	home to trap to harm	A presently permitted state may lie outside the viability region
Permanent suspension	home to paused, followed by waiting forever	Safety alone permits failure of supported progress
Omitted disturbance	One action has a home outcome and a harm outcome	A nominal safe branch does not certify every disturbance branch
Conflicting theories	One theory forbids a, the other forbids b; start must choose one	Separate protected controllers need not yield a common protected controller

Case	Witness or changed condition	Architectural conclusion
Unsafe composition	Each local bit may become one; joint state 11 is forbidden	Local certificates do not establish the joint property
Semantic revision	work becomes forbidden under the expanded interpretation	The old home-to-work permission is no longer valid

A separate bounded-recovery example replaces unbounded waiting with two pending states whose recovery ranks are two and one. Every admitted pending transition decreases the rank. Resumption occurs within at most two transitions. This illustrates a progress condition under the stipulated semantics; a real deadline would additionally require bounded transition duration and justified recovery semantics.

C3 Reproduction and interpretation

Run python3 zhm_v4_model.py. The program emits deterministic JSON and fails an assertion if one of the stated finite checks does not hold. The companion includes the emitted results, the manuscript source, and a README. The calculations were run successfully for this version.

The seven cases are logical demonstrations. They do not estimate the frequency of real failures, compare real-world products, or validate any theory of consciousness. A larger study should test realistic workload freedom and protection under independently specified faults. A useful negative result would show that a proposed combination of safety, continuity, and progress has an empty feasible region under those conditions.

Appendix D Assurance records and revision history

D1 What a protection claim must expose

An assurance record should allow a resident, auditor, or successor provider to identify the exact claim and its dependencies. The record is evidence about an implementation; it is not itself a guarantee of conscious welfare.

Claim component	Required content
Protected scope	Subjects or justified possible-subject classes, substrates, interfaces, geographic or execution boundary, and horizon
Excluded condition	Catastrophic suffering, all suffering, or another precisely named violation; treatment of history and multiplicity
Semantic support	Admitted consciousness and harm interpretations, evidence, disagreements, and unbounded uncertainties
Physical coverage	Transition model, external disturbances, host powers, intermediate execution, and known exclusions
Enforcement	Mediated capabilities, implementation identity, proof scope, timing assumptions, and common dependencies
Continuation	Resource reserves, fallback semantics, recovery strategy, deadlines or progress assumptions, and unresolved obligations
Agency and continuity	Consent conditions, restrictions, review, migration, and evidence relevant to subject preservation
Change and inheritance	Version changes, supported handoff, retained exposure history, incident records, and revalidation conditions

Loss of a premise should change the public claim and trigger an appropriate supported response. Withdrawing a certificate does not withdraw the standing of its inhabitants. Likewise, a new certificate should inherit outstanding obligations rather than treating a changed implementation as a new population with an empty history.

D2 What changed from earlier versions

Version 4 preserves v2's universal standing, prospective protection, non-compensation, and concern with engineered suffering. It preserves v3's protected continuation, viability reasoning, sovereign domains, the separation principle, constrained authority, and distinction between modeled and physical guarantees. These are the conceptual foundations of the present essay.

The principal additions are the explicit ultimate specification for suffering-free reality; the distinction between recurring hazard and whole-horizon uncertainty; probability-zero versus unreachable violations; the requirement to close protection across hosts and physical realizations; the separation of a catastrophic floor from complete suffering prevention; the common-policy problem under scientific disagreement; and an executable demonstration with independently checked finite results.

Heaven 1.0 now appears only as an informal historical name. Protected flourishing names the positive condition. Protected Flourishing Architecture names the proposed wider system. Zero Hostility Matrix remains the project's established name. The technical title deliberately avoids suggesting that a universal protection theorem has already been obtained.

Claims of firstness, a solved phenomenological metric, automatic physical enforcement, or universal moral convergence are not required by the proposal. The standard probability arguments, viability methods, and invariant proofs are credited as established techniques. The scientific and architectural synthesis is offered for criticism and development; no comprehensive novelty claim is made.

D3 Authorship and status

The originating concept, moral motivation, and project direction are Andrew Hill's. Versions 2 and 3 were developed with AI assistance; this version was developed in collaboration with ChatGPT, including drafting, literature checks, formal exposition, and the finite-state companion. The manuscript presents a research proposal and conditional arguments. It has not undergone independent peer review. Attribution to ChatGPT records the contribution and does not imply institutional endorsement or independent scientific validation.

References

[1] Hill, Andrew. Zero Hostility Matrix version 2. Manuscript dated 15 July 2025. Source reviewed: Zhmv2.pdf, 42 pages. Developed with AI assistance.

[2] Hill, Andrew. Zero Hostility Matrix version 3. 16 September 2026. Source reviewed: ZHM_v3.md. Concept and direction by Andrew Hill; draft developed with ChatGPT.

[3] Center on Long-Term Risk. Beginner's guide to reducing s-risks. Research overview. https://longtermrisk.org/research/beginners-guide-to-reducing-s-risks/

[4] Long, Robert, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI Welfare Seriously. 2024. arXiv:2411.00986. https://arxiv.org/abs/2411.00986

[5] Butlin, Patrick, Robert Long, Tim Bayne, and colleagues. Identifying indicators of consciousness in AI systems. Trends in Cognitive Sciences 30(6), 488-501, 2026; published online in 2025. https://doi.org/10.1016/j.tics.2025.10.011

[6] Aubin, Jean-Pierre, Alexandre M. Bayen, and Patrick Saint-Pierre. Viability Theory: New Directions. Second edition. Springer, 2011. https://doi.org/10.1007/978-3-642-16684-6

[7] Alpern, Bowen, and Fred B. Schneider. Defining liveness. Information Processing Letters 21(4), 181-185, 1985. https://doi.org/10.1016/0020-0190(85)90056-0

[8] Necula, George C. Proof-carrying code. Proceedings of POPL, 106-119, 1997. https://doi.org/10.1145/263699.263712

[9] Bloem, Roderick, Bettina Koenighofer, Robert Koenighofer, and Chao Wang. Shield Synthesis: Runtime Enforcement for Reactive Systems. 2015. arXiv:1501.02573. https://arxiv.org/abs/1501.02573

[10] Slagel, J. Tanner, Lauren M. White, Aaron Dutle, Cesar A. Munoz, and Nicolas Crespo. A Formal Verification Framework for Runtime Assurance. NASA Formal Methods, 2024. Author manuscript: https://shemesh.larc.nasa.gov/fm/papers/NFM2024-draft.pdf

[11] Klein, Gerwin, Kevin Elphinstone, Gernot Heiser, and colleagues. seL4: Formal Verification of an OS Kernel. Proceedings of SOSP, 207-220, 2009. https://sel4.systems/Research/pdfs/sel4-formal-verification-os-kernel.pdf

[12] Cousot, Patrick, and Radhia Cousot. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. Proceedings of POPL, 238-252, 1977. https://www.di.ens.fr/~cousot/COUSOTpapers/POPL77.shtml

[13] Benveniste, Albert, and colleagues. Contracts for System Design. Foundations and Trends in Electronic Design Automation 12(2-3), 124-400, 2018. https://doi.org/10.1561/1000000053

[14] Fischer, Michael J., Nancy A. Lynch, and Michael S. Paterson. Impossibility of Distributed Consensus with One Faulty Process. Journal of the ACM 32(2), 374-382, 1985. https://groups.csail.mit.edu/tds/papers/Lynch/jacm85.pdf

[15] Hossenfelder, Sabine. Minimal Length Scale Scenarios for Quantum Gravity. Living Reviews in Relativity 16, article 2, 2013. https://arxiv.org/abs/1203.6191

The seL4 project's current account of proof assumptions complements [11]: https://www.sel4.org/Verification/assumptions.html. Source metadata and accessible primary-source descriptions were checked on 9 October 2026. Citations identify methodological precedents and relevant research; they do not imply endorsement of ZHM or establish its unresolved claims about experience.