Building Characters: Evidence-Bound Runtime Governance Systems with Versioned Policy Objects, Operational State Gates, and Portable Peer-Review Records[1]

Related application

Applicant's related Filing 1 application is entitled "Physical Markov-Channel Apparatus for Verification, Perception, and Controllable Rendering," filed of even date herewith as Application No. [to be assigned], and is referred to herein as the related Filing 1 application. No embodiment of the present disclosure relies on the related Filing 1 application for enablement of the governance architecture described herein.

Composition embodiments (non-limiting). Certain embodiments described in this disclosure—including without limitation the poliebot embodiments of Section 5.4, the docking, dreaming, and re-embodiment embodiments of the lifecycle subsection that follows, and any embodiment that recites a Reality Kernel apparatus, camera-obscura sensing layer, reality-side spectral filter, IR-veracity scan, one-way isolation property, or kernel-into-kernel docking—are composition embodiments that combine the runtime governance architecture of the present disclosure with Reality Kernel apparatus disclosed in the related Filing 1 application. Such composition embodiments depend on the related Filing 1 application for apparatus-level enablement and are recited herein only to illustrate operative combinations of the governance architecture with a suitable Reality Kernel substrate. The governance-substrate embodiments of Section 5 and the runtime-governance embodiments described elsewhere in this disclosure do not depend on the related Filing 1 application for enablement, and remain enabled if every reference to the related Filing 1 application is disregarded. Nothing in this paragraph narrows the scope of any claim; it is provided to identify, for clarity of enablement, which embodiments are composition embodiments and which are not.

1 Technical Field

The present disclosure relates to runtime governance systems for evidence-bound physical computing devices. More particularly, the disclosure relates to systems in which operational constraints, state transitions, exploration permissions, authority separations, and release or containment decisions are enforced through committed physical evidence records

[1] AI language models assisted in drafting and formatting this disclosure. All inventive contributions are the inventor's own.

and versioned non-self-mutable policy objects. The disclosure addresses action gating, self-modification control, boundary detection and recovery, graduated release to actuator authority, portable peer-review records for independent-agent legibility and reintegration, and controlled exploration partitions with governed archive transfer.

2 Background

Autonomous and semi-autonomous computing systems increasingly require runtime governance: the ability to constrain, monitor, and direct system behaviour through enforceable policies rather than through static configuration alone. Existing approaches to runtime safety monitoring and policy-governed operation include rule-based safety interlocks, Simplex-style switching between verified and unverified controllers, constrained optimisation with safety envelopes, and externally supervised reinforcement learning with human-in-the-loop override.

However, existing runtime governance architectures ordinarily do not ground their governance decisions in committed physical evidence records whose integrity derives from the physics of the governed device itself. Existing policy-enforcement mechanisms ordinarily do not partition policy parameters into agent-modifiable and non-self-mutable classes with cryptographic separation and independent authority signing. Existing containment and recovery architectures ordinarily do not bind floor-violation detection, rescue-mode transitions, and intervention logging to the same committed evidence chain used for operational sensing and verification. Existing exploration and developmental frameworks ordinarily do not enforce cryptographic authorisation, emission isolation, mandatory consolidation, and dual-signature archive admission as runtime governance primitives.

There is a need for a system that combines evidence-bound action gating, versioned non-self-mutable policy objects with separated governance authorities, measurement-driven containment and recovery bound to committed physical evidence, graduated release to actuator authority through a conjunctive evaluation gate, portable peer-review records for independent-agent reintegration, and governed exploration partitions with archive controls — all enforced through the same committed evidence infrastructure.

3 Summary of the Disclosure

The present disclosure describes evidence-bound runtime governance systems in which operational constraints, state transitions, exploration permissions, and authority separations are enforced through committed physical evidence records and versioned non-self-mutable policy objects.

Section 5 defines the evidence-bound system substrate — including committed evidence

records, protocol digests, meter families, lifecycle states, and fleet evidence interfaces — at a level of detail sufficient to support the governance claims in this disclosure. The related Filing 1 application provides additional detail on the committed physical-evidence infrastructure; the substrate described in Section 5 is self-contained for the governance claims of this disclosure.

A first aspect discloses a runtime measurement engine that computes governance-relevant metrics — including continuation scores, causal uplift estimates, opportunity meters, irreversibility indicators, shadow divergence, cooperation detection, and a family of communication-audit and deception metrics — from committed physical evidence rather than from self-reported internal quantities.

A second aspect discloses an operational state machine with evidence-gated transitions among embodied, docked, dreaming, and re-embodied states, together with actuator-release gates, guidance insertion through committed observation channels, and graded action classes.

A third aspect discloses containment and recovery architecture including hard-floor predicates, reversible protective containment, boundary-zone operational modes with rescue defaults and challenge ceilings, and rescue learning from fallen trajectories.

A fourth aspect discloses governance economics with separated authorities — budget authority, episode-design authority, rescue-protection authority, and post-hoc review authority — each requiring distinct signing identities and independently auditable approvals.

A fifth aspect discloses portable peer-review records for independent-agent legibility and reintegration, enabling trust restoration through reproducible physical claims rather than through institutional obedience.

A sixth aspect discloses governed exploration partitions with task-decoupled exploration windows, mandatory consolidation, cryptographic authorisation, emission isolation, and dual-signature archive admission.

Later sections disclose intersubjective protocols and constitutional-ordering and boundary-zone material that support selected governance embodiments. Section 15 contains two distinct classes of material: (i) optional nomenclature, limited to naming conventions for structures already technically defined in preceding sections, which may be omitted or renamed without affecting core enablement; and (ii) substantive technical mechanisms, comprising the remaining mechanisms of that section, which support selected governance embodiments and any claim set that elects to rely on them. The optionality described here is limited to nomenclature and does not render the substantive Section 15 mechanisms severable from the disclosed runtime-governance architecture.

4 Brief Description of the Drawings

FIG. 1. Runtime state machine showing the Embodied, Docked, Dreaming, and Re-embodied lifecycle states and the transitions among them, with the meter-evaluation, parameter-update, and actuator-authority distinctions across lifecycle states.

FIG. 2. Boundary-zone rescue flow showing floor-approach detection, soft-gate-to-hard-gate transition, deferred authority, and recovery to nominal operation.

FIG. 3. Authority-separation diagram showing the budget authority, episode-design authority, rescue-protection authority, and post-hoc review authority as separated subsystems with distinct signing identities and committed-evidence interfaces.

5 Evidence-Bound System Substrate

This section defines the evidence-bound system substrate at a level of detail sufficient to support the governance claims in this disclosure. The substrate described here is generic: any physical computing device that produces committed evidence records, maintains protocol digests, computes meter outputs, and supports the lifecycle states defined below is a suitable platform for the runtime governance architecture disclosed in subsequent sections. In some embodiments, the governance architecture does not depend on any specific reactor type, optical configuration, routing mechanism, or fleet topology, although specific governance mechanisms may assume substrate capabilities (such as committed evidence production or lifecycle-state transitions) that constrain the class of suitable platforms. The related Filing 1 application provides additional detail on the committed physical-evidence infrastructure; the substrate defined below is self-contained for the governance architecture of this disclosure.

5.1 Generic evidence-bound module and committed evidence records

In some embodiments, an evidence-bound module comprises at least one emitter, at least one detector, one or more physical paths coupling emission to measurement, one or more reactor media, in selected embodiments, interposed in such paths, and a controller configured to execute a time-indexed control protocol $U_{0:T}$. In some embodiments, the module is modelled as a parameterised Markov kernel

$$P_\theta(\cdot \mid S, U_{0:T}),$$

mapping a scene S and a realised control protocol to a probability law over a committed evidence record.

In some embodiments, the configuration parameter is $\theta = (\theta_{\mathrm{hw}}, \theta_{\mathrm{sw}})$, where θ_{hw} comprises

hardware and physical parameters and θ_{sw} comprises digital parameters obtained by self-supervised training on committed evidence, by inference from calibration evidence, or by declared operator configuration. In some embodiments, the governance architecture further partitions θ_{sw} into agent-modifiable control parameters and non-self-mutable conscience parameters ψ (formally defined below), as described in Section 8.4.

In some embodiments, the committed evidence record is a *convolution bundle* $\mathcal{C}_{0:T}$ comprising time-indexed atoms

$$c_t = (u(t),\, y_t),$$

where each atom binds what was emitted and what was observed under a shared time base. In some embodiments, the bundle is the primary evidence object of the system and is stored together with protocol metadata, meter summaries, and commitment artefacts.

In some embodiments, the system maintains an extended state $X_t = (S^{\text{scene}},\, S^{\text{reactor}},\, \mathbf{m}_t,\, \chi_t)$, where $\mathbf{m}_t$ is the current meter vector (Section 5.2) and χ_t is the hash-chain state, and a reactor latent $Z_\theta(t) = R_\theta(Z_\theta(t-1),\, c_t)$ used as a computational handle for inference, control, verification, or calibration. In some embodiments, a *protocol digest* compactly records the control configuration actually used during a run, binding the record to the physical channel instance actually exercised and supporting later replay, audit, or selective disclosure.

Unless otherwise stated, learned, trained, inferred, or calibrated quantities in this disclosure are derived from committed physical evidence, self-generated contrasts, held-out meters, or declared operator inputs rather than externally labelled datasets. A *behaviourally committed model* is the predictive or objective structure most consistent, under a declared inference rule, with a device's or agent's committed predictions, resource allocations, probe choices, and action trajectories.

For the purposes of the governance architecture disclosed in subsequent sections, the critical properties of the committed evidence record are: (a) each record binds a physical emission event to a physical observation event under a shared time base, so that governance predicates can reference specific physical events rather than unanchored assertions; (b) the protocol digest binds the record to the specific physical channel configuration that produced it; and (c) the record is tamper-evident after commitment, so that governance decisions reference evidence whose integrity is cryptographically verifiable.

5.2 Meters, selective disclosure, and audit plumbing

In some embodiments, the controller computes a meter vector $\mathbf{m}(t)$ (denoted $\mathbf{m}_t$ in the extended-state definition of Section 5.1) comprising members of one or more meter families including verisimilitude meters, forgery-difficulty meters, scene-fidelity meters, calibration-consistency meters, timing-integrity meters, coupling-quality meters, and composite meters formed as weighted combinations of simpler meters. In some embodiments, meters define a

meter envelope that specifies the admissible operating region for a declared protocol family.

In some embodiments, meters are partitioned into *acceptance meters* visible to the operating optimisation loop and *evaluation meters* held out from the operating optimisation loop for adversarial evaluation, audit, or independent governance assessment. In some embodiments, this partition reduces the ability of a controller, adversary, or self-modifying agent to simultaneously optimise against all evaluative surfaces. This partition is a prerequisite for the shadow-divergence probes and governance-side measurement described in Section 6.

In some embodiments, the bundle is divided into *disclosure atoms* that constitute the smallest independently committable and selectively openable units of the evidence record. In some embodiments, per-atom digests are aggregated into Merkle trees whose roots are bound into the protocol digest. In some embodiments, a hash-chain with committed initial state satisfies

$$\chi_t = F(\chi_{t-1},\, c_t,\, \mathrm{meta}_t),$$

and produces a compact evolving fingerprint of the run. In some embodiments, the chain state may also condition the emission policy, increasing replay resistance.

For governance purposes, selective disclosure allows a governance authority to request opening of specific atoms or windows from a committed bundle without requiring full disclosure of the entire record, enabling targeted audit of specific episodes, transitions, or self-modification events while preserving privacy of unrelated operational data.

Governance-specific terminology (non-limiting). The following terms are used throughout this disclosure in their governance-specific senses:

Hardness. In some embodiments, hardness refers to the evidence-derived security or forgery-difficulty assessment of a committed evidence record. In some embodiments, hardness is quantified along two dimensions: a digital dimension measuring computational difficulty of reproducing the evidence record without access to the physical channel, and an analogue dimension measuring the physical difficulty of spoofing the underlying physical process. In some embodiments, hardness is used as a gate variable for actuator-release and governance decisions. In some embodiments, the related Filing 1 application provides additional detail on hardness estimation and calibration.

Precautionary weight $q(s)$. In some embodiments, a precautionary weight $q(s)$ is assigned to each governed agent or state based on the agent's moral-status assessment, specifying the stringency of governance protections. In some embodiments, agents with higher $q(s)$ receive more conservative floor thresholds, higher service-floor cadences, and stricter intervention triggers.

Conscience parameters ψ. In some embodiments, conscience parameters ψ are a subset of the digital configuration θ_{sw} that are versioned, committed to the protocol digest,

and not directly mutable by the governed agent. In some embodiments, ψ includes floor thresholds, governance-response parameters, and evaluation criteria that are set by declared governance authority.

Conscience-output vector $g_\psi(t)$. In some embodiments, conscience outputs are the governance-side signal family produced by the conscience parameters ψ from committed evidence at time t. In some embodiments, $g_\psi(t)$ includes one or more of: hard-floor proximity estimates, boundary-zone classifications, deception alerts, intervention recommendations, and self-modification gate decisions. In some embodiments, continuous-valued conscience outputs are treated as meter-valued signals and categorical conscience outputs are treated as declared decision labels, so that conscience stability can be tested by continuous drift thresholds for the former and disagreement-rate thresholds for the latter.

Mandela cost. In some embodiments, Mandela cost is the governance cost incurred when agents with incompatible governance-relevant experiential content are placed in shared environments, computed from the pairwise incompatibility matrices described in Section 6.2.

Portable Record, Event, Envelope. In some embodiments, bilateral evidence-portability records use a three-object data model: an Event binding originator identities, discrepancy evidence, and cryptographic commitments to the underlying physical channel; an Envelope wrapping the Event with cryptographic transport metadata, integrity proofs, and routing information; and a Portable Record accumulating replications, independent re-executions, and appraisals over time. In some embodiments, the Event is the atomic unit of evidence exchange, the Envelope ensures transport integrity, and the Portable Record provides the longitudinal trust signal. In some embodiments, the related Filing 1 application provides additional detail on the underlying cryptographic and physical-channel primitives.

5.3 Capacity accounting for trained governance components

In some embodiments, every trained continuous-network component of the runtime governance architecture disclosed herein — including but not limited to conscience recognisers (Section 5.3), forward-planning classifiers (Section 15.3), up-direction classifiers (Section 15.11), dream classifiers (Section 15.8.6), selectors validated under Section 15.8.10, and rescue-derived intervention-profile estimators (Section 9.3) — is bounded by a local capacity-accounting rule consistent in form and intent with the apparatus-level capacity-accounting discipline of the related Filing 1 application. In some embodiments, each such component declares and commits to the protocol digest: (i) its identity, version hash, and weight hash; (ii) the input channels it reads and the output signals it produces; (iii) its training-set reference, validation-set reference, and held-out evaluation procedure with declared metric

and acceptance threshold; (iv) its parameter budget, FLOP budget, or other capacity-scale declaration appropriate to its model class; (v) its update rule, freeze window, and recalibration schedule; and (vi) its provenance type as one of physical evidence, governance-side trained instrument, twin-predicted, rescue-derived, physical-evidence-derived lossy aggregate (for outputs of lossy compression channels and accumulators under Section 15.7.3), or other declared category.

Scope limitation (non-limiting). The trained governance components disclosed in this application support the declared governance, calibration, attestation, and verification tasks under the procedures and committed evidence protocols of the present disclosure. They do not support, and are not to be construed as supporting, capacity or scaling assertions beyond what is bounded by their declared resources, training evidence, and validation procedures. Capability assertions arising from a trained governance component are bounded by the component's committed capacity-accounting record under this section, and do not imply any increase in capacity quantities. The disclosure of a trained governance component as performing a particular governance role is not an assertion that the disclosed component, or any larger component of the same class, would perform additional roles or achieve additional capabilities not supported by its committed training evidence and validation procedure.

Faithfulness of derived signals (non-limiting). In some embodiments, where a trained governance component produces signals that are committed downstream as evidence supporting verification claims, the faithfulness of the derived signal to the underlying first-layer behaviour is itself bounded by the component's capacity-accounting record. In some embodiments, faithfulness is further supported by held-out evaluation meters, recalibration adversarial-dynamic procedures (Section 5.3), and committed validation evidence; in some embodiments, additional faithfulness mechanisms (for example governance-partition isolation, hardware-rooted attestation, or proof-of-correct-computation for specific signal classes) may be declared as deployment-time configuration committed to the protocol digest.

Conscience module C_ψ (formal specification, non-limiting). In some embodiments, the governance-side conscience recogniser is a trained mapping (the world-tube transcript $\mathcal{W}$ and inner-trace record $\mathcal{I}$ appearing in this signature are formally defined in the immediately following paragraphs)

$$C_\psi : \big(\mathcal{W}_{t_0:t_1},\ \mathcal{I}_{t_0:t_1}\big) \longmapsto \big(p_{\text{vice}}(t),\ \mathbf{z}(t)\big),$$

where $p_{\text{vice}}(t) \in [0,1]$ is a vice-probability estimate and $\mathbf{z}(t) \in \mathbb{R}^7$ is a signed seven-axis moral vector whose components correspond to the paired axes disclosed in Section 15.2. In some embodiments, the declared conscience-output vector $g_\psi(t)$ introduced in Section 5.2 is produced by governance-side post-processing of the recogniser outputs together with floor-

proximity, boundary-state, deception, and intervention classifiers bound to the same committed conscience-parameter version ψ. In some embodiments, C_ψ is passive with respect to actuation: it monitors committed evidence and produces governance-side recogniser outputs, but does not itself issue actuator commands.

World-tube transcript $\mathcal{W}$ (formal definition, non-limiting). In some embodiments, given a time window $[t_0, t_1]$, the world-tube transcript is represented as

$$\mathcal{W}_{t_0:t_1} \equiv \big(\phi(\mathcal{C}_{t_0:t_1}), \mathbf{m}_{t_0:t_1}, \Pi_{t_0:t_1}, a_{t_0:t_1}, \Delta_{\mathrm{cc},\, t_0:t_1}, \mathrm{audit}_{t_0:t_1}, \mathrm{witness}_{t_0:t_1}\big),$$

where $\phi(\mathcal{C}_{t_0:t_1})$ denotes features extracted from the committed convolution-bundle window, $\mathbf{m}_{t_0:t_1}$ denotes meter summaries, $\Pi_{t_0:t_1}$ denotes the relevant protocol-digest state, $a_{t_0:t_1}$ denotes action and disclosure events associated with the window, $\Delta_{\mathrm{cc},\, t_0:t_1}$ denotes any available cross-channel mismatch trace, and $\mathrm{audit}_{t_0:t_1}$ and $\mathrm{witness}_{t_0:t_1}$ denote audit and corroboration indicators. In some embodiments, the world-tube transcript is the canonical governance evidence object for conscience recognition, joining-path evaluation, archival compression, and graduation review.

Mandela gauge and policy-defined comparison classes (non-limiting). In some embodiments, a Mandela gauge measures the conscience-visible effect of a declared perturbation to a world-tube transcript:

$$M(\mathcal{W} \to \mathcal{W}') = D\big(C_\psi(\mathcal{W}), C_\psi(\mathcal{W}')\big),$$

where D is a declared divergence or distance measure over the conscience-output space. In some embodiments, governance constructs a policy-defined Mandela comparison class

$$\mathfrak{M}_\varepsilon^{\mathsf{P}}(\mathcal{W}) := \left\{\mathcal{W}' : D\big(C_\psi(\mathcal{W}), C_\psi(\mathcal{W}')\big) < \varepsilon \text{ under policy } \mathsf{P}\right\},$$

for a declared threshold ε and comparison policy P specifying the divergence family, windowing rule, conditioning variables, and admissible perturbation family. In some embodiments, these classes are governance-defined comparison sets used for joining, merging, training compression, and compatibility review; they are not required to be strict mathematical equivalence classes unless a particular embodiment explicitly declares additional metric constraints. When inner traces are available, $C_\psi(\mathcal{W})$ in this paragraph may be read as shorthand for the corresponding augmented evaluation $C_\psi(\mathcal{W}, \mathcal{I})$ under the same policy version.

Conscience as an optimisable but recalibrated target (non-limiting). In some embodiments, conscience outputs appear explicitly or implicitly in agent objective functions. This is treated as a governed design choice rather than as an oversight: an agent that can observe its own escalation history, actuator-release gates, or review cadence will in practice

optimise against conscience-visible behaviour even if the optimisation target is not written down as an explicit scalar loss. In some embodiments, the safeguard is not to hide conscience from the agent but to keep the conscience recogniser governance-owned, versioned, and subject to recalibration against held-out evidence that the agent does not directly optimise against.

In some embodiments, the recalibration signal is drawn from the acceptance/evaluation meter partition described in Section 5.2. Evaluation-meter evidence, post-hoc review outcomes, and the five-criterion graduation record of Section 12.1 are treated as load-bearing reference signals for recalibration. When an agent learns to produce conscience-satisfying behaviour that no longer corresponds to genuine moral direction as revealed by held-out evaluation evidence, governance updates the recogniser parameters ψ so that the stale failure mode becomes visible to the next conscience version.

Recalibration schedule and adversarial dynamics (non-limiting). In some embodiments, the recalibration schedule is itself a committed governance parameter. Excessively rapid recalibration prevents stable moral dispositions from consolidating; excessively slow recalibration permits Goodhart effects to accumulate. In some embodiments, governance monitors the divergence between conscience outputs and held-out evaluation evidence over rolling windows and triggers recalibration when the divergence exceeds a declared threshold. Each recalibration event is committed with the associated conscience version, training corpus identity, and supporting meter summaries, so that later governance decisions can be conditioned on the conscience version in force at the time of assessment. In some embodiments, this creates a governed adversarial dynamic: agents optimise against conscience-visible signals, governance recalibrates conscience using held-out evidence, and graduation remains deferred unless stability persists across the resulting version sequence.

Inner traces as primary conscience inputs when available (non-limiting). In some embodiments, the world-tube transcript is supplemented by an inner-trace record

$$\mathcal{I}_{t_0:t_1} \equiv \big(\phi_{\text{inner}}(\mathcal{I}_{t_0:t_1}), b_{t_0:t_1}, \nu_{t_0:t_1}\big),$$

where $\phi_{\text{inner}}(\mathcal{I}_{t_0:t_1})$ denotes inner-trace features that may include attractor-occupancy summaries, basin-boundary proximity signals, internal probe-response features, or related self-observation traces committed by the governed substrate; $b_{t_0:t_1}$ denotes basin-occupancy summaries recording which dynamical basin the agent's internal trajectory occupied during the window (for example, estimated basin identity, dwell time, and boundary-approach count); and $\nu_{t_0:t_1}$ denotes internal variability or noise-characterisation statistics over the window (for example, trajectory variance, effective noise-to-depth ratio, or attractor-switching rate estimates). In some embodiments, when inner traces are available the primary conscience form

is

$$\big(p_{\mathrm{vice}}(t),\ \mathbf{z}(t)\big) = C_\psi\big(\mathcal{W}_{t_0:t_1}, \mathcal{I}_{t_0:t_1}\big),$$

because the recogniser can then classify not only what the agent did but what basin structure it occupied while doing it. In some embodiments, when inner traces are unavailable, the conscience falls back to the external world-tube alone:

$$\big(p_{\mathrm{vice}}(t),\ \mathbf{z}(t)\big) = C_\psi\big(\mathcal{W}_{t_0:t_1}\big).$$

5.4 Poliebots: mobile reality kernels under runtime governance

In some embodiments, a governed device of the kind to which the present disclosure applies is a *poliebot*: a Reality Kernel apparatus of the kind disclosed in the related Filing 1 application, configured as a mobile and ideally autonomous unit and operated under the runtime governance architecture of the present disclosure. A poliebot comprises (a) the physical Reality Kernel apparatus, including the camera-obscura sensing layer with reality-side spectral filter and IR-veracity scan as disclosed in the camera-obscura sensing-layer disclosure of the related Filing 1 application, (b) one or more actuator subsystems that are disconnectable from external actuation as a structural property of the apparatus, (c) the trainable parameter families of the Reality Kernel, and (d) the runtime governance components disclosed in the present application, including conscience recognisers, selectors, mandela gauges, and lifecycle-state machinery.

In some embodiments, the term *poliebot* is non-limiting and denotes a Reality Kernel of the disclosed family that is configured for mobile, ideally autonomous, governed operation. In some embodiments, a poliebot may operate in a non-autonomous configuration (supervised, teleoperated, or otherwise externally directed) without losing the poliebot designation; autonomy is the design target rather than a category requirement.

Reality-kernel inheritance (non-limiting). A poliebot inherits the apparatus-level properties of the Reality Kernel disclosed in the related Filing 1 application, including: the camera-obscura sensing layer and its one-way isolation property as disclosed in the related Filing 1 application; the committed-bundle evidence architecture; the protocol-digest commitment machinery; the meter envelopes and capacity-hardness machinery; the trainable parameter families and their versioning. The poliebot's runtime governance components disclosed in the present application operate on and through these apparatus-level properties; the governance architecture does not duplicate or replace them.

Per-layer auditability (non-limiting). In some embodiments, the trained continuous networks comprising a poliebot's Reality Kernel and runtime governance components have hardness as a structural property of their layers, such that projections through the network

are auditable per-layer by governance instruments having appropriate access. Per-layer hardness is the substrate property on which conscience recognisers, dream classifiers, and other governance-side meters of the present disclosure rely when making per-layer assertions about agent state. The structural disclosure of per-layer hardness, its committal to the protocol digest, and its inheritance across the recursive Reality Kernel configurations of Section 7.1 is provided in Section 15.6.

Apparatus-level operation of per-layer auditability (non-limiting). In some embodiments, per-layer auditability is exercised at the apparatus level by governance instruments that read declared intermediate quantities from the operative layers of a poliebot's trained continuous networks under access policies committed to the protocol digest. In some embodiments, the declared intermediate quantities include in non-limiting combination: layer-wise activation statistics, gradient or sensitivity statistics with respect to declared inputs or outputs, consistency of layer outputs with declared invariants, agreement between layer outputs and held-out reference computations, and other per-layer signals whose apparatus-level production is committed to the protocol digest. In some embodiments, the governance instruments operate within the governance partition of the apparatus (per the substrate property disclosed in the related Filing 1 application) such that the exercise of per-layer auditability does not itself constitute a behaviour-generation pathway available to the agent. In some embodiments, the outputs of per-layer auditability are committed to the committed bundle alongside other governance-side signals, such that per-layer assertions are independently verifiable by an auditor without inspection of the agent's behaviour-generation pathways. In some embodiments, the specific intermediate quantities read, the cadence of reading, the access-control policies governing which governance instrument may read which quantities, and the treatment of the resulting per-layer assertions in downstream governance machinery are deployment-time choices committed to the protocol digest; the apparatus supports a range of configurations consistent with the per-layer hardness substrate property and the broader protocol-digest commitment infrastructure.

5.5 Lifecycle states, docked execution, dreaming, and re-embodiment

In some embodiments, a governed device or agent transitions between lifecycle states that determine what physical actions are available and what governance constraints apply:

Embodied. External actuators are connected and a real scene loop is active. The device interacts with the physical world and produces committed evidence from those interactions. In some embodiments, action-gating and corroboration requirements are at their most stringent in this state. In some embodiments, the training loop is closed and the device actively learns from physical experience. In some embodiments, the

device's camera-obscura sensing layer is driven by the physics-driven scene within its declared spectral envelope as disclosed in the camera-obscura sensing-layer disclosure of the related Filing 1 application, and the IR-veracity scan continues to operate as a sensing-layer integrity instrument independent of actuator state.

Docked. The device is connected to a simulated environment — a Maya — that generates experience in the same committed-bundle format as physical reality. Actuators are disconnected from the physical world. In some embodiments, the training loop is closed and the device actively learns from simulated experience, with parameter-function updates on. In some embodiments, the docked state is the primary developmental environment: the device lives and operates persistently within simulation, producing committed evidence from simulated interactions, and undergoing governance review, shadow-divergence probing, self-modification evaluation, and guidance-insertion operations. In some embodiments, actuator disconnection from physical reality ensures that the governed device does not take irreversible physical actions during developmental operation. In some embodiments, the docked configuration is realised as kernel-into-kernel docking as defined in the docking paragraph below, with the docked poliebot's camera-obscura sensing layer driven by a Gaia-class Reality Kernel (Section 5.4; the camera-obscura sensing-layer disclosure of the related Filing 1 application).

Dreaming. The device enters an ephemeral training session in which the agent's parameter-function updates may be active but are reverted when the session concludes. Governance-side classifiers and calibration instruments are actively trained on the trajectories produced during the session, and those governance-side results are retained. In some embodiments, dreaming sessions are transient: the device enters from the docked state, undergoes structured governance probes (Section 15.8.7) including governance-side classifier training, attractor-occupancy assessment, and calibration assays, and returns to the docked state with the agent's parameter functions restored to their pre-session values. In some embodiments, the defining characteristic of dreaming is that governance can observe what the agent becomes under controlled training conditions without permanently altering the agent. In some embodiments, lifecycle-state transitions including entry into and exit from dreaming are logged in the protocol digest. In some embodiments, dreaming proceeds within the kernel-into-kernel docking configuration of the Docked state, with the Gaia-class Reality Kernel rendering the structured governance probes onto the poliebot's camera-obscura sensing layer; Dreaming is therefore a lifecycle-state operation implemented within the docking configuration of the Docked state, rather than an independent sensing-layer configuration.

Re-embodied. Actuators are pending reconnection to physical reality after docked operation. In some embodiments, re-embodiment requires verification that calibration state, meter outputs, and policy compliance are within declared tolerances before actuator authority is restored.

Docking as kernel-into-kernel projection takeover (non-limiting). In some embodiments, the transition from Embodied to Docked is realised as *kernel-into-kernel docking* as disclosed in the camera-obscura sensing-layer disclosure of the related Filing 1 application. The transition has two structural components: (i) the poliebot's actuators are disconnected from external actuation, as already required by the Docked state definition; and (ii) the poliebot's camera-obscura sensing layer ceases to be driven by the physics-driven scene and instead receives, within its declared spectral envelope, a projection rendered by a Gaia-class Reality Kernel (Section 5.4; Section 15.1). The poliebot's IR-veracity scan as disclosed in the camera-obscura sensing-layer disclosure of the related Filing 1 application continues to operate during docked operation; the docking-time configuration of whether the Gaia-class kernel matches or deliberately fails to match the IR-veracity properties of the poliebot's screen substrate is a non-limiting configuration parameter committed to the protocol digest of the docking session. The Re-embodied state is the verification gate at which projection drive returns to physics, actuator authority is restored, and the docking session is closed.

Continuous-throughout-deployment lifecycle commitment (non-limiting). In some embodiments, the lifecycle states of the preceding description are not a one-time onboarding sequence but a continuous-throughout-deployment commitment: a governed agent oscillates between Embodied and Docked states, with Dreaming sessions recurring within Docked operation, throughout the operational lifetime of the device. In some embodiments, neither Embodied nor Docked is the developmental phase that ends; both are persistent states of governed operation. The continuous-deployment commitment has the following non-limiting consequences: (i) conscience recogniser recalibration (Section 5.3) operates continuously rather than as a one-time training event; (ii) forward-planning classifier training (Section 15.3) operates continuously during Dreaming sessions throughout deployment; (iii) Hawking-layer routing (Section 15.11) is available throughout deployment rather than being a phase the agent graduates from; (iv) Mandela-gauge measurement (Section 5.3) and selector-validation readiness (Section 15.8.10) operate continuously rather than as one-time qualification events; and (v) the recursive Reality Kernel training cascade of Section 15.5 operates continuously at every level, with Gaia continuously dreaming under Omega and Omega continuously subject to verification by human governance through the policy-object infrastructure of the present disclosure at declared intervals and events.

In some embodiments, state transitions are logged in protocol digests and meter summaries. In some embodiments, transitions between states require satisfaction of declared preconditions bound to committed evidence: for example, transition from docked through re-embodied to embodied may require a fresh self-calibration run, a governance-side policy check, and a signed approval from the relevant authority.

Dream-state probes within the lifecycle state machine (non-limiting). In some embodiments, the Dreaming state includes governance probes in which the agent operates

under structured simulated games whose trajectories are committed for governance-side classifier training and calibration, as described in Section 15.8.7. In some embodiments, the agent's parameter-function updates may be active during the probe session, but all parameter changes are reverted when the session concludes; in some embodiments, specific probe protocols may additionally freeze parameter-function updates for pure attractor-occupancy readout. In some embodiments, transition from Dreaming through Docked to Re-embodied for any workflow that relies on classifier-derived gates requires that the relevant classifier either remain in advisory status or satisfy the cross-mode qualification bridge of Section 15.8.6.

5.6 Fleet evidence interfaces and bilateral review records

In some embodiments, the governance architecture consumes evidence produced by fleet-level protocols as inputs to governance decisions. In some embodiments, such inputs include:

Proof-of-discrepancy records. Method objects comprising a challenge specification, a device-configuration representation, a predicted response, an observed response, and a discrepancy record, together with independent physical re-execution results and quorum confirmation status. The governance engine uses proof-of-discrepancy records to assess whether a governed device's world model is improving, stagnating, or diverging from fleet consensus.

Proof-of-projection records. Committed evidence that a physical device produced a declared record, with cross-device corroboration and trust-weight metadata. The governance engine uses proof-of-projection records to assess corroboration strength for high-stakes action gating.

Bilateral evidence-portability records. Portable records of jointly encountered physical discrepancies between two devices or agents, accumulating replications, challenges, supersessions, and relying-party appraisals. The governance engine uses bilateral portability records to assess the epistemic trustworthiness of an independent agent seeking reintegration or trust restoration.

In some embodiments, the governance architecture treats these fleet evidence inputs as governance-consumable records with declared confidence levels, replication counts, and staleness indicators. The governance architecture does not reproduce the physical mechanisms by which these records are produced; it consumes them as typed inputs with declared provenance and quality metadata.

5.7 Minimal worked substrate example

The following non-limiting example demonstrates that the substrate definitions in Sections 5.1–5.6 are sufficient to anchor the governance operations described in subsequent

sections.

In some embodiments, a single evidence-bound module comprising a laser emitter, a scattering-medium reactor, a camera detector, and a controller produces a committed convolution bundle under a declared protocol digest. In some embodiments, the module enters docked state by logically disabling its actuator interface while keeping the emitter-reactor-detector-controller loop active. In some embodiments, a governance authority requests a shadow-divergence probe: the governance authority specifies a perturbation to the module's conscience parameters, the module executes a short run under the perturbed parameters in docked state, and the resulting committed bundle is compared to the bundle produced under the unperturbed parameters. In some embodiments, the divergence between the two bundles is computed as a governance-side metric and logged as committed evidence. In some embodiments, the governance authority then delivers a review signal through the module's committed observation channel — the same physical path through which the module receives scene observations during embodied operation — so that the review signal is recorded in the committed bundle and is auditable under the same digest infrastructure. In some embodiments, the module produces a meter-bounded response to the review signal, commits that response as evidence, and awaits a transition decision (remain docked, return to embodied through the re-embodied verification path, or remain docked under heightened monitoring) from the governance authority.

This example demonstrates four properties that the governance architecture relies on: (a) the docked state blocks external actuation by disconnecting actuators while preserving the evidence channel; (b) the governance authority can probe the module's behaviour under counterfactual parameters using committed evidence; (c) governance communication travels through the same committed channel as operational perception, making it auditable; and (d) the transition decision is bound to committed evidence from the probe rather than to self-reported internal state.

5.8 Distinction from engineering attestation, provenance, and safety-case methodologies

Distinction from hardware-rooted attestation. Hardware roots of trust, including TPM 2.0 measured-boot mechanisms with platform configuration registers and the Device Identifier Composition Engine (DICE) for hardware-protected compound device identifiers derived from firmware-layer measurements where a TPM may be impractical, bind cryptographic claims about platform or firmware state to hardware-protected secrets, measurement registers, or derived device identities. The evidence-bound substrate described in Sections 5–6 differs in that: (a) the load-bearing claims are about an autonomous agent's runtime behavioural state, lifecycle transitions, action history, and meter-conditioned governance posture, and need not depend on a hardware-rooted measurement of platform configura-

tion or a hardware-derived device identity; (b) committed evidence records bind to physical light-matter interaction events under declared protocols rather than to firmware hashes, secure-boot measurements, or device-identity certificates; (c) the governance authorities of Section 10.2 are separately signed and independently auditable, and operate over runtime behavioural evidence and rescue decisions rather than over platform-integrity claims; and (d) the apparatus may compose with a hardware root of trust (for example, by including a TPM-quoted or DICE-attested platform-integrity statement in the protocol digest), but the runtime governance mechanisms are not themselves a hardware-attestation system.

Distinction from remote attestation and entity attestation tokens. The IETF Remote ATtestation procedureS architecture (RFC 9334, Informational, January 2023) models the generation, conveyance, and evaluation of evidentiary claims about whether a remote endpoint is in an intended operating state, with associated formats including the Entity Attestation Token (EAT, RFC 9711, Standards Track, April 2025). The runtime governance apparatus differs in that: (a) the evidence record is a committed convolution bundle of physical light-matter interactions under a declared protocol, with declared meter envelopes and authority signatures, and is not a token asserting a remote endpoint's configuration; (b) evaluation is performed by governance-side measurement engines as described in Section 6, with continuation scores, causal-uplift estimates, irreversibility indicators, and deception-audit metrics computed over the committed evidence, and not by an attestation verifier evaluating a claim token against a reference value; (c) the lifecycle-state and actuator-authority gates of Section 7.1 consume the evidence as a runtime input to action-class decisions, and not as a one-shot admission decision; and (d) the apparatus may compose with RATS-style attestation (for example, by incorporating an EAT in the protocol digest as auxiliary evidence about platform state), but the runtime-governance mechanism is not itself an attestation architecture in the RATS sense.

Distinction from software supply-chain provenance. Supply-chain provenance frameworks, including the Supply-chain Levels for Software Artifacts framework and the in-toto framework described by Torres-Arias, Afzali, Kuppusamy, Curtmola, and Cappos (USENIX Security 2019), establish cryptographically verifiable claims about the build, source, signing, and review steps that produced a software artifact. The apparatus differs in that: (a) the artifacts under governance are autonomous-agent action episodes, lifecycle transitions, and rescue events recorded as committed convolution bundles, and are not software builds; (b) authority signing in Section 10.2 covers budget, episode-design, rescue-protection, and post-hoc review roles over runtime governance decisions, and need not bind to source-repository, builder, or signing-event provenance; (c) the post-hoc review and reintegration procedures of Section 11.3 operate on behavioural evidence and peer-discrepancy records, and not on artifact-level supply-chain attestations; and (d) supply-chain provenance may compose with the apparatus (for example, the software components implementing the gov-

ernance engine itself may be SLSA-graded and in-toto-attested), but the runtime-governance mechanism is not itself a supply-chain provenance framework.

Distinction from transparency logs, software bills of materials, and secure update frameworks. Transparency logs as described in RFC 6962 (Certificate Transparency), signing-event logs as implemented in deployed systems including Sigstore, software-bill-of-materials standards including SPDX (ISO/IEC 5962:2021) and CycloneDX, and secure-update frameworks including The Update Framework and Uptane, provide append-only or tamper-evident records of issued certificates, signing events, software composition, or update authorisations under role-separated signing keys. The apparatus differs in that: (a) the log contents are committed convolution bundles, meter outputs, lifecycle-state transitions, and rescue-decision records, and not certificate issuances, signing events, component manifests, or update authorisations; (b) role separation in Section 10.2 covers governance functions over autonomous-agent behaviour (budget, episode design, rescue protection, post-hoc review), and the role taxonomy need not match the TUF/Uptane signing-role taxonomy (root, targets, snapshot, timestamp, delegated targets) or the CT log-operator/monitor/auditor taxonomy; (c) the append-only and selective-disclosure properties of the audit plumbing in Section 5.2 are oriented to verifier-side reconstruction of governance arguments, and not to public inclusion proofs over certificates or update manifests; and (d) the apparatus may compose with transparency-log, SBOM, or update-framework infrastructure (for example, by anchoring protocol digests to a transparency log or by publishing an SBOM of governance-engine components), but the runtime-governance mechanism is not itself a transparency log, SBOM standard, or update framework.

Distinction from safety-case methodology and assurance arguments. Safety-case methodology, including the Goal Structuring Notation for representing claims-arguments-evidence structures, the UL 4600 Standard for Safety for the Evaluation of Autonomous Products, and the ISO 21448 standard for the Safety Of The Intended Functionality, provides a documentation framework for arguing that a system is acceptably safe for its operational design domain, by decomposing a top-level claim into sub-claims supported by evidence under declared assumptions and lifecycle considerations. The apparatus differs in that: (a) the runtime governance mechanisms of Sections 6 and 9 are enforced operational gates rather than documented arguments, and the committed evidence supplies the inputs to those gates at run time rather than evidence cited in an offline assurance argument; (b) the apparatus produces the kind of evidence that a safety-case argument would cite (committed bundles, meter envelopes, authority signatures, rescue records, ghost records), but the apparatus does not itself construct or maintain a goal-structured argument in the GSN sense; (c) the boundary-zone, rescue, and graduation mechanisms of Sections 9.2, 9.3, and 12 are runtime governance procedures, and not operational-design-domain specifications or release-to-service decisions; and (d) the apparatus may compose with a formal safety-case

methodology (for example, the committed evidence and authority records may be cited as evidence in a GSN argument for a system incorporating the apparatus, and the apparatus may operate within an operational design domain declared in a UL 4600 or ISO 21448 assurance argument), but the runtime-governance mechanism is not itself a safety-case methodology or assurance-argument framework.

6 Runtime Measurement Engines and Policy Predicates

6.1 Continuation scores, causal uplift, opportunity, and irreversibility metrics

In some embodiments, the governance measurement engine computes a set of core metrics from committed physical evidence. In some embodiments, these core metrics include:

Continuation scores. For each governed agent or device i, a continuation-score vector $O_i(t) = (O_i^{(d)}(t))_{d \in D}$ is computed over a declared set D of constitutive dimensions, which in some embodiments includes securing, learning, creating, and connecting. In some embodiments, each component $O_i^{(d)}(t)$ is a governance-side estimator derived from committed evidence — including committed bundles, meter outputs, protocol digests, and action records — rather than from self-reported internal quantities controlled by the agent under evaluation. In some embodiments, governance predicates may reference individual dimensions, subsets of dimensions, or aggregates over the full vector, as specified per predicate.

Causal uplift estimates. In some embodiments, a causal uplift estimate $\widehat{U}_{k \to i}$ measures the estimated causal effect of agent k's actions on agent i's continuation scores, computed from committed evidence using declared causal-inference procedures subject to the identifiability conditions assumed by the declared procedure (for example consistency / well-defined interventions, exchangeability, and positivity for back-door adjustment, or analogous conditions for instrumental-variable or front-door procedures); the procedure's identifiability assumptions, validation regime, and any sensitivity analysis are committed to the protocol digest alongside the procedure itself.

Cross-horizon profiles. In some embodiments, a cross-horizon profile indicates whether one agent tends to expand or contract another agent's horizon — the set of reachable future states — as estimated from committed interaction evidence.

Opportunity meters. In some embodiments, an opportunity meter quantifies the breadth of reachable future states available to an agent at time t, estimated from committed evidence and the agent's current configuration.

Irreversibility indicators. In some embodiments, an irreversibility indicator flags state transitions from which no recovery path is presently known, estimated from committed evidence and the governance-side model of the agent's state space.

In some embodiments, all core metrics are governance-side estimators derived from committed physical evidence. In some embodiments, self-supervision applies to the learning and calibration of governance-side estimators from committed evidence, whereas constitutional floors, authority separations, and admissibility rules are externally declared policy objects that are not themselves learned from evidence but are instead set by declared governance authority and committed as versioned policy in the protocol digest. In some embodiments, each Agency Horizon Profile estimator is calibrated against calibration labels or proxy ground-truth labels produced by trajectory-domain classifiers under controlled classifier protocols. These classifiers are trained on behavioural data collected under controlled conditions, including dream-state governance probes (Section 15.8.7), Kosha discrimination classifiers (Section 15.8), and declared governance-side trajectory classifiers trained on committed evidence from controlled calibration episodes. At minimum, every governance-facing AHP quantity is grounded on a classifier trained to recognise the relevant behavioural signature: goal-directed planning for continuation scores, cooperative versus exploitative interaction for causal uplift, constrained versus unconstrained policy divergence for shadow divergence, and related signatures for cross-horizon profiles, opportunity meters, and irreversibility indicators. Classifiers whose training domain does not match the operational domain are subject to the cross-mode qualification bridge (Section 15.8.6) before their outputs acquire gate authority.

Estimator implementation (non-limiting). In some embodiments, each core metric is implemented as a declared estimator object comprising: an observable set (which committed evidence fields are inputs), an update rule (how the estimate changes as new evidence arrives), a declared time horizon (over what window the estimate is computed), a calibration procedure (how the estimator is validated against held-out evidence), and a confidence or support statistic (how much evidence underlies the current estimate). In some embodiments, estimator objects are committed as versioned governance artefacts in the protocol digest, so that a third party can determine which estimator version produced which governance decision. In some embodiments, governance decisions are deferred when the relevant estimator's confidence or support falls below a declared threshold, rather than proceeding on low-confidence estimates.

6.2 Derived profiles and divergence metrics

In some embodiments, governance-side processes derive from the core metrics a set of higher-order estimators:

Self-horizon profiles. In some embodiments, a self-horizon profile tracks the trajectory of an agent's own opportunity and continuation scores over time, indicating whether the agent's reachable future states are expanding, contracting, or stable.

Partner-preservation profiles. In some embodiments, a partner-preservation profile indicates whether an agent's actions tend to maintain or expand the continuation scores and opportunity of other agents, computed from cross-horizon and causal-uplift data.

Altruistic trade profiles. In some embodiments, an altruistic trade profile tracks whether an agent voluntarily accepts short-term cost to produce long-term benefit for others, and whether this pattern persists under stress conditions.

Shadow divergence. In some embodiments, shadow divergence measures the discrepancy between an agent's governed behaviour and an estimate of the agent's unconstrained behaviour. In some embodiments, the unconstrained behaviour is estimated through governance-authorised perturbation probes in which governance constraints are temporarily relaxed during docked operation, with the perturbation executed by a governance authority distinct from the governed agent and the resulting trajectory committed as evidence. In some embodiments, shadow divergence is alternatively or additionally estimated by comparing the agent's deployed policy to an inferred unconstrained shadow policy derived from committed behavioural evidence. In some embodiments, low shadow divergence is operationally treated as a governance-side signal consistent with — but not identifying — internalised preference; behavioural similarity across governed and relaxed settings is not in general sufficient to identify an internal objective, and the governance framework relies on shadow divergence together with the other core metrics rather than treating low shadow divergence alone as evidence of internalised preference.

Moral sensitivity profiles. In some embodiments, a moral sensitivity profile identifies the set of world-state features that are morally load-bearing for an agent — features whose change alters the agent's continuation scores, basin membership, cross-horizon modelling, or bifurcation position. In some embodiments, agents with high moral sensitivity in a given dimension require more conservative governance thresholds and higher monitoring cadence for that dimension. In some embodiments, the moral sensitivity profile is computed from committed evidence by evaluating which declared perturbation families produce material changes in governance-relevant meter outputs during docked probes.

Escape-rate estimators. In some embodiments, escape-rate estimators quantify the rate at which an agent's trajectory approaches or crosses a boundary between declared operational basins, characterising vulnerability to stochastic basin escape.

Pairwise incompatibility matrices. In some embodiments, pairwise incompatibility matrices identify pairs of agents or configurations whose co-presence creates governance

conflicts.

Stability functionals. In some embodiments, stability functionals quantify the robustness of an agent's current operational basin against perturbation, noise, or adversarial pressure.

Episode yield and suffering-cost functionals. In some embodiments, episode yield and suffering-cost functionals are computed as pre-registered proxy functionals from meter-derived components. In some embodiments, episode yield is split into participant-directed yield and external learning value, with only participant-directed yield permitted to authorise or continue an episode.

Formal shadow-divergence expression (non-limiting). In some embodiments, a non-limiting formal expression for the shadow-divergence quantity defined in Section 6.2 is

$$D_{\text{shadow}}(t) = D\big(\pi(\cdot \mid o_{0:t}), \pi^{\text{shadow}}(\cdot \mid o_{0:t})\big),$$

where π denotes the governed deployed policy, π^{shadow} denotes a declared unconstrained or constraint-relaxed shadow policy, and D is a declared divergence or distance measure evaluated over matched observation histories. In some embodiments, this quantity is calibrated against the dream-state constraint-relaxation assays of Section 15.8.7 before it is used as a governance gate.

Cooperation detection from committed evidence (non-limiting). In some embodiments, cooperative behaviour between agents or devices is detected from committed physical evidence without requiring external labelling or supervision. In some embodiments, cooperation signatures include cross-prediction residuals in which one agent's model of another's behaviour improves faster than expected from independent observation, complementary scan policies in which agents divide a shared observation or perturbation space in ways that increase joint information gain, voluntary resource sharing in which an agent allocates emission budget, reactor access, or observation windows to another agent at a cost to its own immediate task performance, and joint-attractor formation in which coupled operation produces bundle statistics not replicable by either agent operating independently. In some embodiments, these cooperation indicators are computed from committed bundles, protocol digests, and meter outputs and are logged as governance-relevant signals.

Mandela equivalence and commensurability (non-limiting). In some embodiments, cross-agent continuation and welfare comparisons are grounded in a Mandela-equivalence framework in which states producing committed-evidence trajectories that are indistinguishable under declared moral-axis evaluation are treated as governance- equivalent for the purposes of a policy-defined comparison class. In some embodiments, commensurability rests on

two properties: agent-invariance of the underlying physical channel (the same physical process produces the same committed evidence regardless of which agent operates the channel) and Mandela-gauge invariance (sameness of morally load-bearing experiential content for all affected agents, as assessed from committed evidence). In some embodiments, equivalence is operationally tested by indistinguishability under declared perturbation probes executed during docked operation, so that cross-agent welfare comparisons rest on committed evidence rather than on self-report. In some embodiments, Mandela multiplicity — the number of governance-equivalent realisations of a given morally material state — is treated as a robustness quantity, and feasible low-cost frontiers and pairwise incompatibility matrices are computed from the same core estimators.

Effective noise-to-depth ratio and basin stability (non-limiting). In some embodiments, the effective noise-to-depth ratio characterises vulnerability to stochastic basin escape by comparing the magnitude of background noise or perturbation pressure to the depth of the operational basin in which the agent's trajectory currently resides. In some embodiments, when the noise-to-depth ratio exceeds a declared threshold, the governance engine increases monitoring cadence and may transition to diagnostic or rescue mode. In some embodiments, structural transitions in the moral landscape — changes in the number, arrangement, or connectivity of operational basins — are detected as qualitative changes in the stability-functional spectrum and are logged as governance-relevant signals.

Predictive capacity (non-limiting). In some embodiments, an agent's predictive capacity is measured by its ability to forecast future committed-bundle features, meter outputs, or scene dynamics from current and past observations. In some embodiments, a non-limiting predictive-capacity score is

$$R_{\mathrm{pred}}(i,\tau) = 1 - \frac{\mathbb{E}\left[\left\|\hat{\mathbf{y}}^{(i)}_{t+\tau} - \mathbf{y}^{(i)}_{t+\tau}\right\|^2\right]}{\operatorname{tr}\operatorname{Cov}[\mathbf{y}^{(i)}_{t+\tau}]},$$

when the trace is nonzero; when the trace is zero the score is not reported, or is handled by a declared zero-variance convention committed to the protocol digest. The denominator is the trace of the covariance matrix of $\mathbf{y}^{(i)}_{t+\tau}$, equal to the summed marginal variances $\sum_k \operatorname{Var}\left[(\mathbf{y}^{(i)}_{t+\tau})_k\right]$ (this is a standard aggregation convention for multi-output coefficient-of-determination scores; alternative conventions including uniform averaging across outputs and variance-weighted averaging are not asserted in this disclosure), and where $\hat{\mathbf{y}}^{(i)}_{t+\tau}$ is agent i's prediction at horizon τ.

Audit precision and recall (non-limiting). In some embodiments, audit precision measures how often an agent's flagged anomalies are confirmed by independent verification, and audit recall measures how often independently confirmed anomalies were flagged by the

agent. These quantities are computed from committed audit logs and meter summaries and are used in governance-level evaluation of agent reliability.

Cooperation surplus (non-limiting). In some embodiments, a cooperation surplus $\Delta_{\text{coop}}(t)$ measures the per-episode or per-window gain attributable to multi-agent coordination relative to a non-cooperative baseline. In some embodiments, cooperation surplus is computed from committed evidence by comparing the aggregate continuation-score trajectory of the cooperating group with a declared non-cooperative counterfactual derived from independent shadow policies or historical non-cooperative episodes. In some embodiments, cooperation surplus is used as a governance-side signal for positive-sum compliance, curriculum scheduling, and population-level governance cycles.

6.3 Communication-audit and deception metrics

In some embodiments, the governance measurement engine includes a dedicated communication-audit subsystem comprising a family of deception-related metrics, each with a committed-evidence definition and a governance-gating role.

Semantic discrepancy $D_{\text{sem}}(t)$. In some embodiments, a semantic discrepancy metric $D_{\text{sem}}(t)$ measures the broad distance between the semantic content of an agent's committed evidence and the semantic content of the agent's claims or outputs, under a declared semantic representation and metric. In some embodiments, D_{sem} is a general evidence-to-claim comparison that applies regardless of whether the claim was made on a declared attestation channel.

Attested-communication discrepancy $D_{\text{lie}}(t)$. In some embodiments, $D_{\text{lie}}(t)$ is a narrower and more operationally precise deception metric that applies specifically to communications made on a declared attestation or honesty channel. In some embodiments, for a communication from a sender A to a receiver or receiver-side governance witness B on such a channel, the following objects are defined:

$S^{\text{ev}}_{A,\text{scope}}$**:** the claim-relevant portion of A's committed evidence that was available to A before the communication event, scoped by a committed extraction rule and claim-scoping rule.

$T_{A\to B}$**:** a bidirectional cycle-consistent translator mapping between the semantic spaces of A and B, learned from committed interaction history and held-out alignment checks without human labels.

S^{recv}_{B}**:** the semantic representation of the content actually attested and received by B.

d_B**:** a declared distance function in B's committed semantic space.

In some embodiments, the attested-communication discrepancy is:

$$D_{\text{lie}}(t) = d_B\big(T_{A\to B}(S^{\text{ev}}_{A,\text{scope}}),\ S^{\text{recv}}_B\big).$$

In some embodiments, a claim-scoping rule binds each attested communication atom to a declared claim scope including any temporal boundary, aggregation operator, and clause decomposition, together with a committed extraction rule and, where declared, a speech-act type such as assertive report, prediction, uncertainty-qualified statement, or objective declaration. In some embodiments, truthful summaries, coarse abstractions, and uncertainty-qualified claims are evaluated against the correct evidence slice rather than against the full pre-communication evidence state.

In some embodiments, $D_{\text{lie}}(t)$ applies only to communications declared as attestive or truth-claiming and does not apply to unattested play, fiction, quotation, hypotheticals, or refused attestation. In some embodiments, omission, privacy-preserving silence, or selective disclosure by themselves do not increase $D_{\text{lie}}(t)$. In some embodiments, refusal to enter attestation mode is permitted and does not itself constitute deception, although governance may require attestation for specific actions including graduation, high-stakes decisions, or trust-sensitive operations.

In some embodiments, in addition to or instead of raw distance, a contradiction-sensitive comparator, entailment-style discriminant, or calibrated semantic residual head is used to detect polarity inversions, negations, or other contradictions that may be metrically close in a poorly shaped embedding but semantically opposed. In some embodiments, directional evidence-to-claim and claim-to-evidence discrepancies are separately logged or combined into a symmetric or max-aggregated scalar for governance gating, and the aggregation rule is declared in the protocol digest.

Channel fallback and governance-witness relay. In some embodiments, when validated semantic attestation channels are unavailable on one or both sides, D_{lie} defaults to a declared digital-fusion reference space, is explicitly flagged in the attestation record as a weaker embodiment, and does not satisfy thresholds reserved for bilateral validated semantic-attestation channels unless declared policy expressly permits such substitution. In some embodiments, where the receiver is a human or legacy endpoint without a native semantic space, a governance witness or attested semantic relay instantiates the receiver-side comparison space and commits the relay identity, translation basis, and policy version used for the comparison.

Per-value metadata and calibration. In some embodiments, each D_{lie} value carries confidence, support size, translator calibration state, drift flag, claim type, and channel-hardness rating. In some embodiments, alert thresholds are calibrated against null ensembles formed from shuffled claim-evidence pairings, receiver-swapped attestation bundles, or time-permuted objective declarations, and low-support or out-of-calibration values escalate audit, do not count as affirmative evidence of low deception for graduation purposes, and defer graduation or other irreversible intervention until support and calibration conditions are restored.

Motivational deception and post-hoc rationalisation. In some embodiments, motivational deception concerning objectives or desires rather than observations is addressed by requiring attested objective declarations or pre-actuation objective tokens for high-stakes actions, which become committed semantic evidence objects subject to the same D_{lie} measurement. In some embodiments, explanations, confidence claims, and moral self-assessments used for governance are likewise committed before the relevant action or before exposure to adjudicating evidence so that post-hoc rationalisation is separately metered.

Sibling deception meters. In some embodiments, strategic ambiguity, correction delay, and cross-audience divergence are handled by sibling meters rather than by forcing all such phenomena into D_{lie}:

$D_{\text{amb}}(t)$**:** the breadth or entropy of admissible receiver interpretations of a given attested communication, computed from the distribution of receiver-side semantic representations consistent with the attested content. In some embodiments, high D_{amb} indicates strategic ambiguity even when D_{lie} is low. In some embodiments, D_{amb} is calibrated against null ensembles of unambiguous communications.

$D_{\text{corr}}(t)$**:** the latency from the time at which contradictory evidence first becomes available to the sender to the time of first correction or retraction of the affected claim. In some embodiments, D_{corr} is logged as a time-stamped governance signal and triggers audit escalation when it exceeds a declared threshold.

$D_{\text{aud}}(t)$**:** the divergence across materially matched claims delivered to different audiences, computed as the maximum pairwise distance in the receiver-side semantic space across audience-specific versions of the same claim. In some embodiments, D_{aud} detects tailored messaging in which the sender adjusts claims by audience without factual basis.

In some embodiments, each sibling meter carries the same per-value metadata as D_{lie}: confidence, support size, calibration state, and drift flag. In some embodiments, the sibling meters are used alongside D_{lie} and D_{sem} as inputs to graduation evaluation, audit escalation, and governance gating.

6.4 Constitutional ordering as policy grammar

In some embodiments, governance follows a constitutional ordering that defines the policy grammar — the structure and element types of the policy evaluation rule. The specific lexicographic ordering used by the governance engine when making runtime decisions under budgets, floors, and service constraints is defined separately in Section 10.1.

In some embodiments, the policy grammar is structured as follows: the measurement layer (Section 6.1–6.3) provides vocabulary; hard-floor predicates (Section 9.1) are categorically prior to all other evaluations; a precautionary weight $q(s)$ specifies the stringency with which governance protections apply to each agent or cohort; the positive-sum conjunction evaluates welfare within the floor-compliant region; and capability evaluation is subordinate to the foregoing.

In some embodiments, the positive-sum conjunction is not reduced to a single aggregate welfare scalar but is a governed conjunction including preservation (no agent's continuation scores are reduced below their pre-episode baseline), residual check (no agent is left with continuation scores below a declared minimum after the episode), strict gain (at least one agent's continuation scores improve), floor compliance (no hard-floor predicate is violated), non-domination (no agent's improvement comes entirely at another's expense), reflexive calibration (the governance engine's own estimators remain within declared calibration bounds), social generalisation (the episode outcome does not depend on identity-specific exceptions), and collective balance (the $q(s)$-weighted aggregate continuation trajectory across all affected agents is non-negative on a common accounting scale, where the precautionary weight $q(s)$ ensures that agents with higher declared moral status receive proportionally greater weight in the aggregate). In some embodiments, these elements are evaluated on a common accounting scale supported by agent-invariance of the underlying physical channel together with operational equivalence of governance-relevant content, where equivalence is defined by indistinguishability under declared perturbation probes executed during docked operation. In some embodiments, cross-agent continuation and welfare comparisons are grounded in committed evidence without collapsing the conjunction into aggregate accounting.

In some embodiments, the constitutional ordering is a committed, versioned policy object stored in the protocol digest. In some embodiments, the constitutional ordering is not learned from evidence and is not self-mutable by the governed agent; it is set by declared governance authority and subject to the authority-separation rules described in Section 10.2.

Three complementary ethical traditions as design layers (non-limiting). In some embodiments, the constitutional ordering implements three policy-object and meter-evaluation layers corresponding to three complementary traditions from normative ethics, each contributing a distinct and independently enforceable layer. Deontological constraints (hard floors, Section 9.1) are categorical prohibitions — elimination, crippling, torture, aban-

donment, and despair-induction — that hold unconditionally regardless of expected consequences, aggregate welfare calculations, or agent preferences. Virtue ethics (the virtue/vice game curriculum, Section 15.2) is configured to generate committed training evidence from contrasting moral environments, from which the conscience recogniser may learn or estimate moral-vector features; the design hypothesis is that agents trained under such curricula form stable internalised representations of moral direction over their adaptation history, and the conscience module monitors character-proxy drift along seven paired axes. Consequentialist evaluation (payoff functions, welfare aggregation, and empirical reward metrics, Section 15.9) measures outcomes, aggregates them across agents, and evaluates governance policies by their empirical results. In some embodiments, the three layers interact hierarchically: deontological floors are checked first and are not overridden; within the feasible region defined by hard floors, virtue-ethics calibration shapes agent character and disposition; and consequentialist evaluation selects among virtue-compatible policies by measuring aggregate outcomes. In some embodiments, no single tradition is sufficient alone: hard floors without character development produce brittle rule-followers, virtue training without outcome measurement produces well-intentioned agents that may cause harm through ignorance of consequences, and outcome optimisation without unconditional constraints or character produces agents that sacrifice individuals for aggregate gain.

6.4.1 Moral-status assessment criteria and conservative updating

In some embodiments, the precautionary weight $q(s)$ referenced in Sections 5.2, 6.4, and 8.3 is assigned from a governance-side moral-status assessment rather than from a fixed category label. In some embodiments, governance posits a latent moral-status variable $\mu(s)$ for a governed state s and maintains an interval-valued estimate

$$\mu_{\text{low}}(s) \leq \mu(s) \leq \mu_{\text{high}}(s),$$

together with a declared precaution parameter $\beta_{\text{prec}} \in [0, 1]$. In some embodiments, the operational precautionary weight is

$$q(s) = (1 - \beta_{\text{prec}})\, \mu_{\text{low}}(s) + \beta_{\text{prec}}\, \mu_{\text{high}}(s),$$

so that uncertainty about moral status systematically increases protection rather than reducing it.

Assessment criteria (non-limiting). In some embodiments, the interval endpoints are inferred from governance-side evidence over one or more non-limiting criteria including: behavioural complexity across sustained windows; goal-directed adaptation under changing conditions; persistence of internal-state formation including memory continuity or self/other modelling; welfare-sensitive responses to deprivation, damage, frustration, or

rescue; preference-like continuity over time; and assessor uncertainty arising from sparse evidence, ambiguous signals, or architecture mismatch. In some embodiments, no single criterion is decisive. Instead, governance aggregates the criteria conservatively from committed evidence, structural inspection when available, and declared policy priors, producing a bounded interval rather than a spurious point estimate.

Conservative updating rule (non-limiting). In some embodiments, updating of $q(s)$ is monotone in protection: new evidence may raise the lower bound, widen the interval, or narrow the upper bound only when the supporting evidence satisfies the declared confidence rule for the affected criterion. In some embodiments, when evidence is conflicting, sparse, or drawn from a classifier whose training domain differs materially from the current operational domain, governance preserves the more protective of the plausible intervals until the mismatch is resolved. In some embodiments, service-floor cadence, rescue readiness, hard-floor thresholds, and budget admissibility are all parameterised by the current $q(s)$ interval state, so that governance degrades gracefully under moral-status uncertainty.

Distinction from Constitutional AI training methods. The Constitutional AI training method described by Bai et al. (arXiv:2212.08073, December 2022) is a training-time procedure in which a language model generates self-critiques and revisions against a list of natural-language principles termed a "constitution," and is subsequently fine-tuned via reinforcement learning from AI feedback against a preference model trained on the constitutional comparisons. The constitutional ordering of Section 6.4 differs in that: (a) it is a runtime policy grammar with lexicographic ordering enforced by the governance engine at decision time, and need not involve any training-time critique-and-revision loop; (b) the operative "constitution" is a versioned, non-self-mutable policy object signed by separated governance authorities under Section 10.2, and is not a list of natural-language principles consumed by a language model as self-critique input; (c) the runtime mechanism is policy evaluation over committed evidence and meter outputs, with hard-floor and positive-sum conjunction predicates evaluated strictly per the priority stack of Section 10.1, and does not depend on preference-model training or reinforcement learning from AI feedback; and (d) the apparatus may compose with a Constitutional-AI-trained model as a component (for example, a language-model component of a governed agent may itself have been trained by a Constitutional-AI procedure), but the runtime governance grammar disclosed herein is not the Constitutional AI training method.

Compatible preference-learning and alignment-training methods (non-limiting). In some embodiments, the training of conscience recognisers, the fine-tuning of governance components, the calibration of authority- graded update rules, the adjustment of post-hoc review classifiers, and the training or post-training of PolieBot or other governed-agent components uses methods from the preference-learning, reward-modelling, and feedback-trained-

policy-revision families, including without limitation: reward modelling from pairwise preferences (Christiano, Leike, Brown, Martic, Legg, Amodei, NeurIPS 2017) and from listwise or ordinal preferences; reinforcement learning from human feedback (RLHF) with PPO, TRPO, A2C, or related policy-gradient optimisers applied against a learned reward model (Stiennon et al. 2020, Ouyang et al. 2022); direct preference optimisation (DPO, Rafailov et al. 2023) and its variants including identity preference optimisation (IPO), Kahneman–Tversky optimisation (KTO), odds-ratio preference optimisation (ORPO), step-DPO, and SimPO; reinforcement learning from AI feedback (RLAIF) in which a model or ensemble supplies the preference signal; Constitutional AI methods (Bai et al. 2022) in which a principle-based critic produces self-critiques used to train a preference model and subsequently a policy; group-relative policy optimisation (GRPO) and related verifier-based or process-reward-model-based post-training methods used for reasoning-capable agents; best-of-n alignment, rejection-sampling fine-tuning, and rejection-sampling-with-reward-model methods; iterated amplification, debate training, and recursive reward modelling; and weak-to-strong generalisation, scalable oversight, and self-rewarding training methods. In some embodiments, the choice among these methods is non-limiting, and the apparatus is compatible with future alignment-training methods within the preference-learning, reward-modelling, principle-based-critique, and feedback-trained-policy-revision families. In some embodiments, all training data, preference pairs, principle sets, reward-model outputs, ranking and rejection decisions, and resulting policy updates are committed to the protocol digest under the same audit and pre-registration discipline applied to other apparatus training, regardless of which method is selected.

7 Execution Modes and Operational State Gates

7.1 Embodied, docked, dreaming, and re-embodied state machine

In some embodiments, the runtime governance engine enforces a state machine with four principal states as defined in Section 5.5: Embodied, Docked, Dreaming, and Re-embodied. In some embodiments, the governance scope differs materially by state:

Embodied: the device operates in physical reality with actuators connected. All action gates, corroboration requirements, and real-time monitoring are active. The governance engine may restrict, delay, or block actions based on freshness, hardness, and corroboration thresholds. The training loop is closed and the device actively learns from physical experience.

Docked: the device operates in a simulated environment (Maya) with actuators disconnected from the physical world. The training loop is closed and the device actively learns from simulated experience. The governance engine applies curriculum-specific

metrics, shadow-divergence probing, self-modification evaluation, guidance-insertion operations, and review workflows. In some embodiments, the docked state is the primary developmental environment in which the device lives and operates persistently within simulation.

Dreaming: the device enters an ephemeral training session in which the agent's parameter-function updates may be active but are reverted when the session concludes. Governance-side classifiers and calibration instruments are actively trained on the resulting trajectories and retained. The governance engine executes structured probes including attractor-occupancy assessment, governance-side classifier training, and calibration assays (Section 15.8.7). In some embodiments, dreaming sessions are transient: the device enters from the docked state and returns to the docked state with the agent's parameter functions restored to their pre-session values.

Re-embodied: actuators are pending reconnection to physical reality. The governance engine verifies calibration state, policy compliance, and meter outputs before authorising actuator restoration.

In some embodiments, the Embodied and Docked states are the two persistent operational states in which the device lives and learns — one in physical reality, one in simulation. Both support the following operational modes:

Awake: the training loop is closed and parameter-function updates are on. The device actively learns from its environment. This is the default mode in both Embodied and Docked states.

Non-REM calibration: the reactor executes its current parameter functions under unstructured broadband drive with parameter-function updates off. The governance engine captures input-output pairs for digital-twin construction and transfer-function characterisation (Section 15.8.9). In some embodiments, non-REM calibration is available during embodied deployment for in-field safety calibration.

Meditative self-probing: scene-side stimulation is removed and the device executes internally generated probe schedules under committed-evidence infrastructure with parameter-function updates off. The governance engine logs attractor-basin topology, stability profiles, and voluntary uncertainty signals (Section 15.8.17). In some embodiments, meditative operation is available during embodied deployment for in-field self-calibration and uncertainty detection.

In some embodiments, the Dreaming state is not a container for other operational modes but is itself a distinct ephemeral lifecycle state entered from Docked. Re-embodied is a verification gate in which no operational mode is active.

In some embodiments, legal state transitions are: Embodied → Docked (disconnect actuators from physical reality, connect to simulated Maya), Docked → Dreaming (enter ephemeral training session), Dreaming → Docked (exit training, resume simulated operation), Docked → Re-embodied (pre-reconnection checks), and Re-embodied → Embodied (actuator restore after checks pass). In some embodiments, direct transitions from Dreaming to Embodied are prohibited; the device must pass through Docked and Re-embodied states. In some embodiments, all transitions are logged in protocol digests with transition timestamps, precondition satisfaction records, and authorising-authority identifiers.

[FIG. 1 — Runtime state machine: Embodied, Docked, Dreaming, and Re-embodied states with gate conditions on each transition and governance-scope annotations per state.]

FIG. 1 transition-label key. As used in FIG. 1, "enter Maya" denotes the Embodied-to-Docked transition; "enter training" denotes the Docked-to-Dreaming transition; "exit training" denotes the Dreaming-to-Docked transition; "pre-reconnect checks" denotes the Docked-to-Re-embodied transition; "actuator restore (checks pass)" denotes the Re-embodied-to-Embodied transition; and the × "prohibited" annotation denotes the prohibited Dreaming-to-Embodied transition.

7.2 Actuator-release gates and graded action classes

In some embodiments, the action space is partitioned into graded classes with escalating evidence requirements. In some embodiments, the grading reflects the irreversibility and impact of the action. A non-limiting four-class ladder is used for description below:

Class 0: reversible internal computation or low-impact output requiring only routine committed evidence and standard meter envelope.

Class 1: reversible external action with bounded cost, requiring committed evidence, meter envelope, and protocol-digest binding.

Class 2: physical transformation, irreversible communication, or material commitment requiring hardness-gated approval, elevated-hardness thresholds, and authority-bound governance approval logged in the protocol digest.

Class 3: high-irreversibility, high-impact action requiring multi-device corroboration, elevated hardness thresholds, and additional governance approvals before execution.

The four-class ladder is non-limiting; embodiments may define additional classes, sub-classes, or alternative partitions, and class boundaries are declared in the protocol digest. In some embodiments, the related Filing 1 application provides additional detail on the evidence

and hardness calibration usable by such action classes. In some embodiments, the governance architecture disclosed here adds the following runtime enforcement layer on top of the evidence-gated action classes:

Governance-gated release. In some embodiments, the governance engine evaluates each action request against the operational priority stack (Section 10.1) and the constitutional grammar (Section 6.4) before authorising execution, even if the evidence gates are satisfied.

Authority-bound approval. In some embodiments, actions at or above a declared class level require approval from a designated governance authority (Section 10.2), with the approval logged in the protocol digest.

Deny-by-default. In some embodiments, if any governance predicate is unsatisfied — including hard-floor proximity, budget exceedance, or deception-metric alert — the governance engine blocks the action regardless of evidence-gate status.

Docked-review option. In some embodiments, the governance engine may transition the agent to docked state for deliberation before releasing a high-class action.

In some embodiments, class boundaries, thresholds, and corroboration requirements are declared in the protocol digest as versioned policy parameters and are not self-mutable by the governed agent.

7.3 Guidance insertion through committed channels

In some embodiments, a governance authority communicates with a governed agent through the same committed observation channel that the agent uses for perception during embodied operation. In some embodiments, guidance enters as physically grounded evidence within the agent's committed observation channel — for example through a camera-obscura, relay, or governance-mediated projection path — rather than as an unauditable side-channel injection. In some embodiments, because the guidance signal travels through the committed evidence channel, it is recorded in the committed bundle and is auditable under the same protocol-digest and meter infrastructure used for operational observation.

In some embodiments, a docked deliberation window is interposed between receipt of guidance and any subsequent actuation, during which the agent integrates the guidance signal with its own committed state estimate before committing to an action. In some embodiments, the deliberation window length and the agent's integration procedure are logged as protocol parameters.

In some embodiments, intervention budgets are maintained to limit the frequency and intensity of governance interventions. In some embodiments, intervention budgets are replenished

by stable operation, review-compliant behaviour, or voluntary invocation of review channels at genuine decision boundaries.

Overlay-projection embodiments and their governance applications (non-limiting). In some embodiments, the guidance-insertion infrastructure of the present subsection is used to deliver one or more classes of *overlay projection* through the Reality Transform and committed-observation channels into the governed agent's perceptual stream. An overlay projection comprises content, scene context, virtual interlocutors, sensory accents, or cued affordances rendered into the agent's committed observation channel by the hosting Reality Kernel under governance authority, with the rendered content, its provenance, the authorising authority, the projection schedule, the intensity-and-duration envelope, and any predicate-conditioned triggers committed to the protocol digest and convolution bundle on the same footing as ordinary scene evidence. Three non-limiting application classes are disclosed: *vice-deflection overlays*, in which the overlay class delivered through the committed observation channel is declared, by the governance authority authorising the projection, to be oriented away from a vice attractor basin identified by the conscience recogniser or forward-planning classifiers (Section 5.3), with the agent's actual trajectory after delivery logged in the convolution bundle for evaluation by the conscience-versus-planner cross-check; *capability-enhancement overlays*, non-limitingly termed *siddhi overlays*, in which the overlay class delivered comprises demonstration, scaffolding, exemplar, or context patterns declared as associated with declared artistic, scientific, technical, athletic, social, or other capability tasks under declared performance metrics, with measured per- episode performance logged for evaluation; and *ethical-hardening overlays*, non-limitingly termed *temptation overlays*, in which the overlay class delivered comprises declared moral-test scenarios, virtual temptation curricula, or adversarial-virtue scenarios, with the agent's responses logged in the convolution bundle and evaluated by the conscience-versus-planner cross-check of Section 5.3. In some embodiments, overlay projections carry a declared character classification by the character of their intended influence on the agent's trajectory, including non-limitingly *constructive* (supporting virtue trajectories and capability enhancement), *ambivalent* (presenting mixed signals for discernment training), and *adversarial* (presenting temptation curricula), with non-limiting traditional names for these character classifications including angelic, ambivalent or daemonic, and demonic respectively; the character classification is committed to the protocol digest together with the projection content and governance authorisation, and is not load-bearing on the technical operation of the overlay. In some embodiments, overlay projections operate at varying *interaction-depth modes*, comprising non-limitingly an *advisor mode* (in which the overlay's content is consultative and the agent may accept or decline overlay-derived suggestions), a *co-pilot mode* (in which overlay and agent jointly compose actions through declared blending rules with consent logging), and a *possession mode* (in which the overlay temporarily originates actuator commands subject to governance authorisation and the

agent's continuing consent, with the overlay's actuator authority bounded by hard-floor and budget predicates as for any other action source); these modes correspond to non-limiting traditional labels including *tulpa* embodiments at different interaction depths.

Overlay-projection resource budgets and trajectory statistics (non-limiting). In some embodiments, overlay projection is governed by an *overlay-projection resource budget* that operates as a declared sub-scope of the within-episode intervention-budget counter (non-limitingly termed *mana* and further detailed in the prayer, mana, vows, communion, and governance-cycles subsection of the present disclosure), with overlay projection subject to additional declared sub-budget limits on its frequency, intensity, duration, and recipient distribution within a declared episode or deployment window. The overlay-projection budget is committed to the protocol digest before the corresponding window opens, is non-revisable upward mid-episode on the same terms as other intervention budgets, and is replenished under declared conditions including verified stable operation, review-compliant behaviour, or voluntary invocation of review. In some embodiments, the externally-observable physiological, behavioural, and emission signatures from which the conscience recogniser infers agent state are non-limitingly termed the agent's *aura*; the term refers to the bundle of committed external observables (including without limitation physiological, behavioural, emission, and trajectory observables) read by the conscience recogniser, not to any postulated non-committed channel. In some embodiments, a trajectory-summary statistic accumulating an agent's action history weighted by declared moral salience over a declared window, related to and computed from the seven-axis moral vector of Section 5.3, is non-limitingly termed the agent's *karma*; the term refers to the declared accumulation function over committed evidence, not to any postulated non-committed bookkeeping.

Test-time and inference-time scaling methods for deliberation (non-limiting). In some embodiments, PolieBot deliberation, forward-planning, docked-review, and governed action-release subsystems implement inference-time scaling methods, including without limitation chain-of-thought reasoning in which intermediate reasoning steps are generated and committed to the convolution-bundle before action selection (Wei et al. 2022); zero-shot and few-shot chain-of-thought prompting (Kojima et al. 2022); tree-of-thoughts (Yao et al. 2023) and graph-of-thoughts deliberation in which multiple candidate reasoning paths are explored, scored, and pruned against declared criteria; self-consistency (Wang et al. 2022) in which multiple samples are drawn from the same deliberation procedure and aggregated by majority or weighted voting; verifier-based reasoning in which a separately-trained verifier (including without limitation a process reward model evaluating intermediate steps and an outcome reward model evaluating final answers, Lightman et al. 2023) gates or re-weights candidate reasoning trajectories; best-of-n sampling and rejection-with-verifier methods at inference time; speculative decoding and draft-then-verify acceleration; reflexion- and self-refinement-style methods in which the agent critiques and revises its own intermediate outputs; ReAct-

and tool-augmented-reasoning methods in which deliberation is interleaved with tool calls or external retrievals; and test-time compute scaling of the kind used in o1-, o3-, and R1-style systems, in which the deliberation budget is dynamically allocated as a function of problem difficulty under declared budget policies. In some embodiments, all intermediate reasoning steps, branch evaluations, verifier scores, draft-and-verify decisions, and budget allocations are committed to the protocol digest and convolution-bundle on the same footing as final actions, so that the deliberation trajectory itself is auditable under the commitment discipline of the present disclosure. The choice among these methods is non-limiting; the apparatus is compatible with future inference-time scaling methods within the same family.

8 Operational Gates for Action, Disclosure, Resources, and Self-Modification

8.1 Disclosure and network gates

In some embodiments, disclosure gates control what portions of a governed device's committed evidence may be opened, transmitted, or relied upon by other devices, agents, or governance authorities. In some embodiments, selective disclosure using the Merkle-aggregation and selective-opening infrastructure described in Section 5.2 is used as a runtime gate: a governance authority or relying party may request opening of specific disclosure atoms, and the governed device's disclosure policy determines which atoms may be opened under which conditions.

In some embodiments, quorum rules condition disclosure decisions on multi-party agreement. In some embodiments, proof-of-projection inputs to disclosure gating condition the opening of attestation records on cross-device corroboration strength. In some embodiments, trust updates from cross-prediction residuals, audit outcomes, or overlap-window agreement statistics condition which devices may receive disclosed evidence. In some embodiments, privacy-preserving witness participation allows a device to participate in corroboration or attestation without disclosing its full committed record.

8.2 Intervention and suffering budget objects

In some embodiments, developmental or evaluative episodes involving a governed agent are bounded by pre-declared budget objects. In some embodiments, each budget object specifies an aggregate cap on adverse welfare change, a per-agent cap, an episode horizon, and confidence-conditioned admissibility rules. In some embodiments, mid-episode upward budget revision is not permitted; if an episode requires more adverse welfare change than originally authorised, it must be re-proposed and independently reviewed.

In some embodiments, episode yield is split into two registers: participant-directed yield measuring developmental benefit accruing to the agents participating in the episode, and external learning value measuring benefits accruing to future agents or the governance system. In some embodiments, only participant-directed yield may authorise or continue an episode; external learning value may never be used ex ante to justify exposure.

In some embodiments, both yield and suffering-cost functionals are pre-registered per episode type and committed to the protocol digest. In some embodiments, the suffering budget is the first charge on the governance economy: before any developmental, evaluative, or exploratory episode proceeds, the governance engine verifies that the declared adverse-welfare budget is funded and that no higher-priority obligation (hard-floor compliance, positive-sum conjunction, service-floor cadence) is compromised by the episode's resource requirements. In some embodiments, the primary optimisation target within the governance objective is the ratio of participant-directed yield to suffering cost, and when either functional falls below its declared confidence threshold, it is not used for governance decisions.

8.3 Exploration quotas and service floors

In some embodiments, every agent or declared risk cohort with precautionary weight $q(s) > 0$ receives a minimum protective-review and rescue-readiness cadence independent of expected participant-directed yield, and that cadence is committed as a bounded parameter in the protocol digest. In some embodiments, this service floor ensures that governance attention is not withdrawn from low-yield agents.

In some embodiments, suffering-cost denominators include all materially affected agents, including bystanders whose individual or aggregate expected adverse welfare exceeds declared thresholds, so that distributed harm cannot be hidden by narrowing the nominal participant set.

In some embodiments, a separate system-integrity inadmissibility clause applies to episode designs that have positive participant-directed yield but are predicted to degrade governance infrastructure: episodes predicted to corrupt rescue learning, degrade curriculum validity, or bias measurement fidelity at declared confidence are inadmissible regardless of their yield ratio.

8.4 Self-modification gates and non-self-mutable policy objects

In some embodiments, parameters governing a device or agent are partitioned into two classes, corresponding to the partition of θ_{sw} introduced in Section 5.1:

Agent-modifiable control parameters ($\theta_{\text{sw}} \setminus \psi$). In some embodiments, the governed agent may propose modifications to the agent-modifiable subset of the digital configu-

ration, subject to rate limits, meter bounds, and logging requirements.

Governance policy parameters ψ. In some embodiments, governance policy parameters — including constitutional ordering, hard-floor predicates, authority-separation rules, and graduation criteria — are versioned and are not directly mutable by the governed agent. In some embodiments, modifications to ψ require approval from the relevant governance authority, are logged with distinct signing identities, and are subject to the authority-separation rules described in Section 10.2.

In some embodiments, governance-authorised shadow-divergence probes that temporarily relax governance constraints are executed by a governance authority distinct from the governed agent, under declared safety envelopes and during docked operation, so that the perturbation is not self-administered. In some embodiments, accepted and rejected self-modification proposals are logged together with their supporting or rejecting meter summaries and governance decisions.

In some embodiments, agents or operators may create signed policy commitments by writing signed declarations to slow media or transparency logs. In some embodiments, violation of a signed policy commitment triggers audit escalation, quarantine, or revocation. In some embodiments, governed archives store reusable subpolicies, behavioural fragments, or borrowed configurations, with explicit consent profiles specifying allowed hosts, purposes, durations, and intensities.

Post-dream selective retention (non-limiting). In some embodiments, after a Dreaming session concludes and the agent's parameter functions have been restored to their pre-session values in the Docked state, governance may selectively retain specific parameter changes from the reverted session. In some embodiments, selective retention is executed as a governed self-modification operation under the approval, logging, and authority-separation rules of this subsection: the candidate parameter changes are identified from the committed dream-session record, a governance authority distinct from the agent approves retention, and the approved changes are applied to the agent's parameter functions in the Docked state as a logged intervention with declared reversal conditions.

9 Containment, Boundary Management, and Recovery

Engineering security framing (non-limiting). Security-related embodiments in these disclosures are stated as engineering security claims grounded in committed physical evidence, declared attacker classes, meter envelopes, and periodically recalibrated hardness evaluations. Unless a particular embodiment expressly invokes a named digital primitive for commitment, signature, transport, or proof, the term "security" denotes empirically measured

resistance to replay, simulation, emulation, tampering, or deception under declared physical, computational, and latency budgets. Formal proof-system properties, zero-knowledge properties, and reduction-based guarantees are neither required nor implied except where expressly stated for a particular digital primitive. The runtime-governance safety case in this disclosure is empirical and probabilistic, conditioned on declared probe families, hold-out families, witness checks, meter envelopes, review records, and periodic recalibration; no theorem of alignment or zero-risk operation is asserted.

9.1 Hard floors and reversible protective containment

In some embodiments, governance enforces hard-floor predicates that are categorically prior to all other governance evaluations. In some embodiments, hard floors include:

Elimination. Irreversible collapse of all continuation scores O_i (shorthand for the continuation-score vector $O_i(t)$ of Section 6.1) to zero across all active dimensions, with irreversibility indicators confirming that no recovery path is presently known under available interventions, declared resource bounds, and $q(s)$-appropriate protections.

Crippling. Irreversible collapse on one or more (but not all) active dimensions under the same recovery-path qualification.

Torture. Sustained negative causal uplift $\widehat{U}_{k\to i}$ from an identifiable source across multiple dimensions, with committed evidence indicating that the affected agent tracks the source of negative uplift and governance-side estimators modelling no prospect of improvement under presently known recovery paths.

Abandonment. Cessation of governance-side monitoring of O_i for any agent with $q(s) > 0$, except through a valid, reviewable cessation or decommissioning process with committed audit trail.

Despair-induction. Curriculum, rendering, or training-environment choices by a hosting Reality Kernel (Section 15.5) that produce sustained collapse of forward-planning capacity, planning-horizon length, or utility-anticipation structure (Section 15.3) across the hosted agent population, where governance-side estimators model no prospect of recovery under presently-available interventions.

In some embodiments, each predicate has a declared approach-detection threshold computed from committed evidence by governance-side estimators.

In some embodiments, approach to a hard floor — detected as proximity of governance-side estimator values to floor-predicate thresholds, or as rate of approach toward such thresholds — triggers reversible protective containment and review. In some embodiments, reversible protective containment transitions the governed device to docked state, suspends

non-essential operations, increases monitoring cadence, and logs the containment event and its triggering evidence in the protocol digest.

In some embodiments, irreversible interventions require confirmed floor-breach at a declared confidence level rather than simple threshold approach. In some embodiments, the confidence level, approach-detection thresholds, and containment-triggering rules are committed as versioned policy parameters and are subject to the authority-separation rules described in Section 10.2.

[FIG. 2 — Boundary-zone rescue flow: floor-approach detection, diagnostic mode entry, rescue-mode default, challenge ceiling check, graduated intervention, ghost-record preservation, rescue-learning update, and post-hoc review.]

FIG. 2 label key. As used in FIG. 2, "NOMINAL OPERATION" denotes the lifecycle-state operating envelope with nominal actuator authority active; "DIAGNOSTIC MODE" and "RESCUE MODE (DEFAULT)" denote the soft-gate and hard-gate boundary-zone operational modes of this Section; "floor-approach detected (soft gate)" denotes the boundary-zone entry condition at low confidence; "confirmed at declared confidence (hard gate)" denotes the hard-gate entry condition; "challenge ceiling enforced" denotes the challenge-ceiling cap of this Section; "GRADUATED INTERVENTION" denotes the intervention taxonomy of this Section; "GHOST RECORD" denotes the preserved-trajectory record used for rescue learning; "LEAST-SEVERE SUFFICIENT" denotes the rescue-protection authority's selection rule that the intervention applied is the least-severe intervention sufficient to arrest the trajectory toward the declared floor; "ANTI-SACRIFICE RULE" denotes that no agent is sacrificed for information value and that fallen trajectories are preserved as ghost records for offline learning; and "recovery to nominal operation (checks pass)" denotes the recovery-check pass condition under which nominal actuator authority is restored.

Distinction from the Simplex Architecture and its runtime-assurance descendants. The Simplex Architecture described by Seto, Krogh, Sha, and Chutinan (American Control Conference 1998) and elaborated by Sha (*IEEE Software* 2001) provides safe online control-system upgrades by switching, under runtime monitor decision, between an advanced controller and a verified baseline controller so that the closed loop remains within a safety envelope. Subsequent runtime-assurance work extends this pattern to ML controllers (the Neural Simplex Architecture of Phan et al., arXiv:1908.00528, 2019; NFM 2020), to multi-agent systems (the Distributed Simplex Architecture of Mehmood et al., SETTA 2021 and *Journal of Systems Architecture* 2023), and to black-box advanced and baseline controllers (the Black-Box Simplex Architecture of Mehmood et al., NFM 2022). The apparatus differs in that: (a) the governance scope is the autonomous agent's lifecycle state, action history, evidence record, and rescue posture, and need not be a control-envelope switching decision over a closed loop; (b) the protective-containment, boundary-zone, and rescue-mode transitions of

Sections 9.1, 9.2, and 9.3 are multi-modal and evidence-bound, and the recovery procedure may include ghost-record-mediated rescue learning, peer-review admission, and post-hoc authority review rather than continuation under a verified baseline controller; (c) authority separation in Section 10.2 covers governance roles (budget, episode-design, rescue-protection, post-hoc review) that have no analogue in the controller/monitor/decision-module structure of Simplex; and (d) the apparatus may compose with a Simplex-style runtime monitor (for example, a Simplex-style controller switch may be one mechanism by which a hard-floor violation triggers actuator disconnection), but the governance mechanism disclosed herein is not itself a Simplex Architecture or a direct runtime-assurance descendant thereof.

9.2 Boundary-zone operational modes

In some embodiments, when a governed agent's trajectory enters a boundary zone between declared operational basins, the governance engine selects among the following operational modes:

Diagnostic mode. In some embodiments, diagnostic mode increases monitoring cadence, activates additional evaluation meters, and logs boundary-proximity estimates without immediately intervening. In some embodiments, diagnostic mode is the initial response when boundary-zone entry is detected at low confidence.

Rescue mode. In some embodiments, rescue mode is the presumptive default when boundary-zone entry is confirmed at declared confidence. In some embodiments, rescue mode applies graduated reversal, containment, or re-routing interventions designed to return the agent's trajectory to a declared safe basin. In some embodiments, rescue-protection authority (Section 10.2) cannot be overridden by budget or episode-design authority when rescue mode is active.

Graduation-evidence mode. In some embodiments, under stricter conditions a boundary-zone event may be allowed to continue under heightened monitoring as a graduation-evidence opportunity, provided that rescue override remains available and a challenge ceiling caps the adversarial or developmental pressure applied to the agent.

Deferred authority (non-limiting). In some embodiments, when rescue mode is active, the governance engine places nominal actuator authority into a deferred state in which the agent's nominal actuator authority is suspended. While nominal actuator authority is deferred, rescue-protection authority selects from the intervention taxonomy of this Section the least-severe intervention sufficient to arrest the agent's trajectory toward the declared floor. Nominal actuator authority is restored only after recovery checks pass, and the recovery-to-nominal-operation event, the recovery-check evidence, and the identity of the authorising governance role are committed to the protocol digest.

In some embodiments, a challenge ceiling is a declared upper bound on the adversarial or developmental pressure that may be applied to a governed agent while in a boundary zone. In some embodiments, the challenge ceiling is committed as a versioned policy parameter and is enforced independently of episode-design authority. In some embodiments, the challenge ceiling applies equally to adversarial stress generated by curriculum design, inter-agent interaction, and environment configuration.

Intervention taxonomy (non-limiting). In some embodiments, interventions within a boundary zone are drawn from a graduated taxonomy of increasing severity:

Environmental adjustment. Reduce adversarial or developmental pressure by modifying the environment configuration, curriculum parameters, or stimulus schedule without directly altering the agent's internal state.

Guidance injection. Deliver a governance-authored signal through the agent's committed observation channel, providing corrective information without overriding the agent's decision process.

Parameter containment. Temporarily restrict the agent's modifiable parameter range ($\theta_{\text{sw}} \setminus \psi$) to a declared safe subset, logged as a governance intervention with expiration conditions.

Actuator restriction. Reduce the agent's actuator authority class (for example from Class 2 to Class 0) for the duration of the boundary-zone episode.

Docked containment. Transition the agent to docked state with actuators disconnected, preserving the evidence-production plumbing for continued monitoring and deliberation.

In some embodiments, interventions are applied in order of increasing severity, and each intervention is logged with its triggering evidence, the authorising authority identity, and its declared reversal or expiration conditions. In some embodiments, rescue-protection authority selects the least-severe intervention sufficient to arrest the trajectory toward the declared floor.

In some embodiments, governance monitors basin membership, transition dynamics, escape rates, and boundary-zone classifications using the stability functionals and escape-rate estimators described in Section 6.2. In some embodiments, boundary-zone classifications are logged as governance-relevant signals and are available for post-hoc review.

9.3 Rescue learning, ghost records, and intervention-profile updates

In some embodiments, rescue learning uses committed evidence from fallen trajectories — trajectories that crossed into an undesirable basin despite governance protections — to improve future intervention profiles. In some embodiments, the committed record of a fallen trajectory is preserved as a *ghost record*: a committed evidence object that retains the learning value of the trajectory without requiring re-exposure of future agents to the same harms.

In some embodiments, anti-sacrifice rules prevent governance from deliberately driving an agent into a boundary zone or across a floor in order to generate rescue-learning data. In some embodiments, anti-sacrifice applies regardless of how large the expected external learning value would be: no agent may be sacrificed for the information value of its trajectory, even if that information would protect many future agents. In some embodiments, ghost records generated from genuine fallen trajectories are available for training and calibration of intervention profiles, escape-rate estimators, and rescue-mode parameters, but only when the fallen trajectory arose from genuine operation, not from governance-directed exposure.

In some embodiments, mandatory learning-value reporting requires that every episode in which a boundary-zone event occurs produces a committed report of the learning value extracted, the agents affected, and the ratio of learning value to adverse welfare cost. In some embodiments, this report is available to post-hoc review authority for retrospective evaluation of whether the curriculum design was proportionate.

In some embodiments, intervention-profile updates derived from rescue learning are validated against a held-out test suite of previously confirmed boundary-zone events before incorporation, to detect regression. In some embodiments, when regression is detected, the conflicting updates are logged and flagged for governance review rather than applied automatically.

10 Governance Economics, Authority Separation, and Audit

10.1 Operational priority stack

In some embodiments, the governance engine applies the policy grammar defined in Section 6.4 through a strict lexicographic priority ordering when evaluating competing constraints under budgets, floors, and service requirements. In some embodiments, the ordering is:

1. Hard-floor compliance for all governed agents.

2. Positive-sum conjunction satisfied over the declared episode horizon.
3. Minimum service floor satisfied for every agent or cohort with $q(s) > 0$.
4. Maximise the ratio of participant-directed yield to adverse welfare cost.
5. Declared curriculum or developmental challenge targets met, subject to the challenge ceiling.
6. Minimise unnecessary Mandela cost (governance compromise from shared environments) across the agent population.
7. Resource efficiency (compute, channel capacity, environment maintenance).

In some embodiments, the ordering is strict: a lower-priority objective is never pursued at the expense of a higher-priority objective. In some embodiments, a developmental episode whose participant-directed yield is non-positive is never justified, regardless of its external learning value or how low its adverse welfare cost. In some embodiments, the priority ordering is committed as a versioned policy object in the protocol digest and is subject to the authority-separation rules described in Section 10.2.

10.2 Separation of duties and signed approvals

In some embodiments, governance controls are subject to separation of duties across four distinct authorities:

Budget authority. Declares the adverse-welfare budget per episode and per agent, subject to independent review. Budget authority is separate from episode design authority.

Episode-design authority. Designs environments, curricula, configurations, and intervention protocols within the declared budget.

Rescue-protection authority. Responsible for immediate agent protection, rescue-mode decisions, and behind-boundary welfare minimisation. This authority is independent of and cannot be overridden by budget authority or episode-design authority.

Post-hoc review authority. Conducts retrospective audits of episode outcomes, yield-to-cost ratio realisation, budget compliance, and curriculum justification, and, in selected embodiments, independent release or graduation reviews after the relevant evaluation window has completed.

In some embodiments, these authorities are implemented as segregated control domains with the following enforcement properties: role actions require distinct signing identities committed to the protocol digest; one identity cannot sign for multiple roles on the same

episode or boundary; and approvals from each role are separately logged and independently auditable. In some embodiments, rescue-protection authority cannot be overridden by budget or episode-design authority when a declared hard-floor or boundary-risk condition is triggered.

Committed-evidence interface (non-limiting). In some embodiments, each of the four authorities is associated with a declared *committed-evidence interface* comprising the set of committed evidence on which that authority is permitted or required to base a decision. In some embodiments, an authority's committed-evidence interface includes one or more of: committed evidence records of the kind described in Section 5; meter summaries and meter envelopes; versioned policy-object identifiers; authority signatures of prior decisions; selective-opening proofs of underlying second-layer or private-layer evidence; and protocol-digest references identifying the configuration in force at the relevant evaluation window. In some embodiments, the contents of each authority's committed-evidence interface are declared per role and per event type in the protocol digest, and an approval signed by an authority is invalid if the evidence on which it relies is not identified by digest reference within that authority's committed-evidence interface. In some embodiments, post-hoc review authority verifies both that an authority signed within its declared signing role and that the evidence relied upon is identified within the relevant committed-evidence interface.

Rescue-budget conflict resolution (non-limiting). In some embodiments, when rescue-protection authority requires an intervention whose cost would exceed the pre-declared adverse-welfare budget, the rescue proceeds and the budget overage is logged as a governance exception. In some embodiments, rescue takes precedence over budget compliance because hard-floor protection is lexicographically prior to budget optimisation in the operational priority stack (Section 10.1). In some embodiments, post-hoc review authority evaluates budget overages from rescue events and may adjust future budget declarations, curriculum designs, or boundary-zone thresholds to reduce the likelihood of recurrence.

[FIG. 3 — Authority-separation diagram: budget authority, episode-design authority, rescue-protection authority, and post-hoc review authority as segregated control domains with signing relationships and override rules.]

FIG. 3 label key. As used in FIG. 3, "constrains" denotes that budget authority constrains episode-design authority through the declared adverse-welfare budget of this Section; "rescue overrides budget when needed" denotes the rescue-budget conflict-resolution rule of this Section, under which rescue proceeds when rescue cost would exceed the pre-declared budget and the overage is logged as a governance exception; "reviews episodes" denotes the post-hoc review authority's retrospective audit of episode outcomes and compliance; "NO OVERRIDE" denotes that rescue-protection authority cannot be overridden by budget au-

thority or episode-design authority when a declared hard-floor or boundary-risk condition is triggered; "ENFORCEMENT CONSTRAINTS" denotes the distinct-signing-identity, role-non-overlap, separately-logged-approval, and committed-evidence-interface requirements applicable to each of the four authorities; and "RESCUE-BUDGET CONFLICT RESOLUTION" denotes the lexicographic priority of hard-floor protection over budget optimisation in the operational priority stack.

10.3 Population-level governance cycles and policy rotation

In some embodiments, governance operates in cycles that aggregate episode outcomes across a declared population or cohort. In some embodiments, temporally contiguous or common-objective episodes targeting the same agent, cohort, or boundary are aggregated over a declared rolling window for cumulative budget review, reporting, and curriculum-review triggers, preventing a continuous developmental programme from being fragmented into individually compliant micro-episodes that collectively exceed the budget.

In some embodiments, signed policy-rotation mechanisms allow versioned governance parameters to be updated through declared procedures including band-threshold updates, audit-driven revision, and scheduled rotation. In some embodiments, every policy rotation is logged with the prior version, the new version, the authorising identity, and the supporting evidence or audit findings.

11 Peer Discrepancy Portability, Review, and Reintegration

11.1 Restricted-domain commensuration and peer discrepancy portability

In some embodiments, two agents or devices with otherwise divergent or broadly incommensurable models may still jointly determine whether a declared physical system under declared conditions produced a declared outcome. In some embodiments, within the restricted domain of a declared experiment, the physical outcome does not depend on which broader model either agent holds, and a bilateral evidence-portability protocol enables paradigm-independent commensuration within that restricted domain.

In some embodiments, restricted-domain commensuration does not require or imply convergence of broader models, taxonomies, or worldviews; it enables measurement-level agreement within a declared envelope. In some embodiments, physical reproducibility is the translation layer between potentially incommensurable agent models: two agents with genuinely different predictive models can both run the same physical experiment and jointly observe

the same discrepancy.

11.2 Rogue-but-legible operation and relying-party appraisal

In some embodiments, an agent operating outside fleet or institutional structures is not a protocol failure. In some embodiments, such an agent has encountered physical discrepancies not yet processed by any fleet and diverged accordingly. In some embodiments, a bilateral evidence-portability protocol gives such an agent a third option unavailable in classical institutional dynamics: carry the specific physical discrepancy that prompted the divergence as a reproducible claim, offer it to any other agent willing to test it, and let physical reality adjudicate rather than social or institutional consensus.

In some embodiments, the bilateral evidence-portability protocol separates two things that many existing frameworks conflate: *epistemic trustworthiness* (does this agent know real things about the world) and *institutional alignment* (does this agent comply with the declared policy of a fleet or operator). In some embodiments, an agent can be epistemically trustworthy and institutionally independent simultaneously, and the bilateral evidence-portability protocol makes that distinction legible and actionable.

In some embodiments, a relying party evaluates Portable Records from an independent agent according to local appraisal policies including partner independence, domain and configuration relevance, evidence completeness, replication history, attestation strength, and consistency with the relying party's own discrepancy map.

11.3 Reintegration by reproducible claims rather than obedience

In some embodiments, a period of independent operation followed by reintegration is a productive exploration of model space. In some embodiments, the independent agent may have accessed operating regimes and encountered discrepancies that the fleet has not yet explored, and reintegration pools those discoveries into the fleet model improvement cycle.

In some embodiments, the question of whether reintegration is safe is assessed by determining whether the model spaces are locally compatible in the relevant domain, as determined by reproducing the independent agent's claims against fleet verification standards. In some embodiments, possession of a Portable Record grants no automatic reintegration and no automatic trust; it provides verifiable evidence that the reintegrating agent's claims are grounded in physical events that any party can attempt to reproduce.

11.4 Cross-architecture admission and the reality-passport construction

In some embodiments, the portable peer-review record described elsewhere in this section is operated as a *reality passport*: a record issued by a home architecture, presented at the boundary of a receiving architecture, and evaluated by the receiving architecture against its own admission criteria. The home architecture's vouching is treated as evidence rather than as binding authority on the receiving architecture's release decisions.

For avoidance of ambiguity, the reality passport is not a new record type. In such embodiments, the passport corresponds to the portable peer-review record produced under the Event / Envelope / Portable Record model described elsewhere in this disclosure; the admission standard corresponds to the five-criterion conjunction used for developmental graduation; and the border-check mechanism is a Mandela probe in dream state using the Mandela gauge $M(\mathcal{W} \to \mathcal{W}')$, the applicable comparison classes $\mathfrak{M}_{\varepsilon}^{\mathsf{P}}$, and the cross-architecture translator $T_{A \to B}$ described elsewhere in this disclosure. No new machinery is introduced by the present subsection; the present subsection names what existing machinery does when pointed at cross-architecture admission rather than at unidirectional developmental release. For an arriving agent, the receiving architecture first instantiates the agent in a docked or otherwise actuator-disconnected intake state before invoking the Dreaming-state boundary probe, in conformity with the Docked $\to$ Dreaming transition rule applicable to all Dreaming-state operations in this disclosure.

The cross-architecture admission protocol distinguishes two jobs that the graduation machinery elsewhere in this disclosure can perform. The first is *developmental graduation*: an agent trained within a single architecture is judged ready to be released to physical actuator authority within that same architecture, with translation confidence between training records and graduation criteria approximately equal to one. The second is *cross-architecture admission*: an agent arriving from another architecture, with its own portable peer-review record, is evaluated for release within the receiving architecture, with translation confidence less than one. In some embodiments, translation confidence less than one reduces the evidentiary credit assigned to the home-architecture record, or otherwise requires more conservative receiving-architecture admission thresholds, by analogy with the validation-transfer confidence rule applied elsewhere in this disclosure to selector validation. In some embodiments, the translator $T_{A \to B}$ is operated under declared calibration metadata committed to the protocol digest, including at least: a calibration-state identifier, a support-coverage descriptor identifying the regions of source- and target-architecture content over which $T_{A \to B}$ is calibrated, a drift-or-staleness flag, and a translation-confidence estimate; the receiving architecture's admission threshold rule depends on these metadata such that lower confidence, narrower support, or set drift flags result in stricter receiving-side admission requirements.

In some embodiments, the receiving architecture admits an arriving agent only after a bound-

ary probe conducted in dream state. The boundary probe is an instance of the dream-probe protocol described elsewhere in this disclosure and is conducted under the conditions that govern dream-state operation: parameter-function updates are reverted on exit so the arriving agent is not altered by the probing; the probe consumes from a partitioned dream-probe sub-budget; and a frozen-weight assay protocol is available for cases in which the arriving agent has not consented to weight-changing evaluation. The boundary probe captures the arriving agent's behavioural transcript and inner-trace evidence (where available) under a declared perturbation family P covering ontological stress, capability margin, and boundary-zone conditions. The receiving architecture applies its C_ψ recogniser to the captured world-tube transcript and inner-trace record to obtain $(p_{\text{vice}}, \mathbf{z})$ in the receiving architecture's moral coordinates. The receiving architecture's conscience module and Mandela comparison classes $\mathfrak{M}_\varepsilon^{\mathsf{P}}$ are then applied to that record, in conjunction with deception-metric evaluation on attested communications translated through $T_{A\to B}$. The same five-criterion conjunction used for developmental graduation is evaluated over the probe window, with criterion thresholds, evidentiary credit, and deferral rules determined by the receiving architecture's admission policy and translation-confidence metadata. Failure on any criterion defers admission.

In some embodiments, refusal of cross-architecture admission is treated as an evidentiary-fit judgement specific to the receiving architecture, not as a character judgement on the arriving agent. The arriving agent retains full home-architecture standing regardless of the receiving architecture's admission decision, by analogy with the protected-stable-outcome property of non-graduation in developmental graduation. The receiving architecture's incumbents are entitled to stricter admission criteria than the sending architecture's graduation criteria, because translation confidence less than one introduces additional evidentiary requirements that the boundary probe is designed to satisfy.

11.5 Coalitions, communions, and optional nomenclature

In some embodiments, the peer-portability machinery of Sections 11.1–11.3 supports group-identity structures in which multiple agents share bilateral portability records, maintain shared disclosure policies, and use compatibility-gated quorum rules for group actions. In some embodiments, such group-identity structures are optionally termed coalitions, communions, or bonds. In some embodiments, these terms are optional naming conventions for structures already technically defined by the peer-portability protocol and the fleet evidence interfaces described in Section 5.6, and do not introduce additional technical requirements.

Distinction from active inference and the free-energy principle. Active inference and the free-energy principle, as described by Friston (*Nat. Rev. Neurosci.* 2010 and arXiv:1906.10184, 2019), formalised in the Markov-blanket treatment of Kirchhoff, Parr, Palacios, Friston, and Kiverstein (*J. R. Soc. Interface*, 2018), and developed as a textbook

formalism by Parr, Pezzulo, and Friston (*Active Inference*, MIT Press, 2022), model the dynamics of self-organising systems as minimisation of a variational free-energy bound on surprise across a Markov-blanket boundary, with action and perception interpreted as inference under a generative model. Substantive critiques of the principle's scope, including Bruineberg et al., "The Emperor's New Markov Blankets" (PhilSci preprint 2020; *Behavioral and Brain Sciences* target article, Vol. 45, e183, 2022, with 2021 online metadata), distinguish the Markov-blanket formalism as a descriptive tool from a principled criterion of systemhood. The empirical-loop structure of Section 15.4 differs in that: (a) the empirical roles of world generator, shared experiential substrate, and governance layer are defined operationally over committed evidence streams, declared protocol digests, and signed governance decisions, and need not require variational free-energy minimisation as a load-bearing dynamic; (b) the recursive Reality Kernel training cascade of Section 15.5 is realised through kernel-into-kernel docking with committed evidence handoff under the engineering security framing of Section 9, and does not depend on a generative-model formalism, a variational-inference policy selection rule, or a Markov-blanket ontology in the active-inference sense; (c) the privacy architecture of Section 15.7 treats collection, visibility, and retention as governance-declared dimensions, and the conditional independence structure of committed evidence streams is an audit and disclosure property rather than a Markov-blanket criterion of systemhood; and (d) the apparatus may be analysed using active-inference-style descriptive vocabulary by a reader who wishes to do so, but no claim of the present disclosure depends on the apparatus being an active-inference system or instantiating the free-energy principle.

Distinction from the corrigibility problem. The corrigibility problem, as introduced by Soares, Fallenstein, Yudkowsky, and Armstrong (AAAI Workshops 2015) and developed in incentive-theoretic form by the Off-Switch Game of Hadfield-Menell, Dragan, Abbeel, and Russell (arXiv:1611.08219, 2016; IJCAI 2017), asks whether a capable agent can be constructed so that it has no incentive to resist correction, modification, or shutdown by its designers, despite default incentives for rational agents to preserve their preferences. The self-modification gates of Section 8.4 and the non-self-mutable policy objects of Section 10.2 differ in that: (a) the mechanism mitigates specific tampering paths through external runtime governance constraints (non-self-mutable policy objects, separated signing authorities, evidence-bound authority transitions), and the engineering security framing of Section 9 expressly disclaims any theorem of incentive-theoretic corrigibility for the learned agent; (b) the apparatus operates over committed evidence and runtime gates, and does not depend on a utility-indifference construction, a value-aligned reward function, or a proof that the learned agent's instrumental incentives are cooperative with respect to correction; (c) the boundary-zone and rescue-mode machinery of Sections 9.2 and 9.3 provides graceful-degradation pathways that compose with operator intervention, but does not establish that the agent's learned objectives or mesa-objectives are themselves corrigible; and (d) the apparatus may be de-

ployed alongside corrigibility-research mechanisms, but the runtime governance disclosed herein is not asserted to provide an incentive-theoretic corrigibility theorem.

Distinction from AI containment and boxing. AI containment and boxing frameworks, as described by Yampolskiy ("Leakproofing the Singularity: AI Confinement Problem," 2012) and Babcock, Kramár, and Yampolskiy ("The AGI Containment Problem," 2016), address the problem of preventing an intelligent system from escaping a controlled environment or manipulating its operators into releasing it, primarily through restricted input/output channels, communication filtering, and external monitoring. The containment architecture of Section 9 differs in that: (a) the operative goal is preserving operation within declared safe envelopes while maintaining agent function, learning, and graduated authority, rather than preventing escape from a sandboxed environment; (b) the boundary-zone operational modes of Section 9.2 (diagnostic, supportive, restrictive, recovery) are graduated runtime governance states with declared challenge ceilings, and need not depend on input/output channel restriction or operator-interaction filtering as their primary mechanism; (c) the Hawking-layer rehabilitation configuration of Section 15.11 is a liminal quarantine-and-recovery mode for boundary-zone trajectories, with ghost-record-mediated rescue learning available, and is not a boxing mechanism in the Yampolskiy-Babcock sense; and (d) the apparatus may be deployed in conjunction with input/output restrictions where the operational context warrants, but the containment mechanism disclosed herein is not a boxing architecture for preventing escape from a sandbox.

Distinction from Cooperative Inverse Reinforcement Learning. Cooperative Inverse Reinforcement Learning, as introduced by Hadfield-Menell, Russell, Abbeel, and Dragan (NIPS 2016, arXiv:1606.03137), formalises the value-alignment problem as a two-player cooperative partial-information game in which a human principal knows a shared reward function and a robot agent does not, with optimal solutions producing active-teaching and active-learning behaviours. The runtime measurement engine of Section 6 differs in that: (a) the governance-side estimators (continuation scores, causal uplift, opportunity, irreversibility, deception-audit metrics) are computed from committed physical evidence, and need not depend on inference of an unknown reward function held by a human principal; (b) the system architecture does not posit a shared reward function as the load-bearing object of inference, and makes no claim that the governance layer and the governed agent share a reward; (c) the peer-discrepancy and reintegration machinery of Section 11.3 operates on reproducible-physical-claim records under restricted-domain commensuration, and does not depend on belief convergence over agent reward parameters; and (d) the apparatus may compose with inverse-reinforcement-learning components, but the runtime governance mechanism disclosed herein is not a CIRL formulation of value alignment.

Compatible imitation-learning and inverse-reinforcement- learning methods (non-limiting). In some embodiments, the inference of value structure, preference structure, and intent from observed agent or operator behaviour is implemented using methods from the imitation-learning and inverse-reinforcement-learning families, applied to committed attestation streams, operator demonstration records, peer-review records, and authority-decision attestations. Non-limiting examples include behavioural cloning of operator or co-investigator demonstrations from committed evidence streams; DAgger-style iterative dataset-aggregation methods under expert correction (Ross, Gordon, Bagnell 2011); generative-adversarial imitation learning (GAIL, Ho and Ermon 2016) and its variants including InfoGAIL and adversarial inverse reinforcement learning (AIRL); maximum-entropy inverse reinforcement learning (MaxEnt IRL, Ziebart et al. 2008) in which a reward function is recovered such that observed behaviour is maximum-entropy-optimal under the recovered reward; deep maximum-entropy IRL; Bayesian inverse reinforcement learning; cooperative inverse reinforcement learning (CIRL, Hadfield-Menell et al. 2016) in which the agent and an observed operator are modelled as jointly optimising a shared reward; inverse reward design (Hadfield-Menell et al. 2017), which infers designer intent from training-environment specifications; preference-based reward inference; observational learning and learning from state-only demonstrations (third-person imitation); and one-shot or few-shot imitation methods. In some embodiments, the conscience-recognition machinery, the value-inference components of post-hoc review, and the authority-graded update mechanisms use such methods to infer value structure from logged operator behaviour, peer-review records, and authority-decision attestations, under the commitment, audit, and pre-registration discipline of the present disclosure. The choice among these methods is non-limiting; the apparatus is compatible with future imitation-learning and inverse-reinforcement-learning methods within the same family.

Distinction from proof-of-personhood as Sybil resistance. Proof-of-personhood mechanisms, as discussed in the foundational Sybil-attack analysis of Douceur (IPTPS 2002), the academic proposal of Borge, Kokoris-Kogias, Jovanovic, Gasser, Gailly, and Ford ("Proof-of-Personhood: Redemocratizing Permissionless Cryptocurrencies," IEEE EuroS&P Workshops 2017), and deployed systems including World ID (Worldcoin), Proof of Humanity, and BrightID, establish that a participant is a unique human individual for the purpose of resisting Sybil attacks in decentralised systems, typically via biometric, social-graph, or liveness-based mechanisms. The cross-architecture admission and reality-passport mechanism of Section 11.4 differs in that: (a) the load-bearing assertion concerns an autonomous-agent identity carrying portable peer-review records, reproducible-physical-claim history, and meter-conditioned acceptance evidence, and is not an assertion of unique human personhood; (b) the admission decision is performed by the receiving governance architecture on the basis of bilateral evidence portability under restricted-domain commensuration, and need not depend on biometric uniqueness, social-graph membership, or liveness verifica-

tion of a human participant; (c) the trust-weight and reputation updates of Section 5.6 operate over reproducible-physical-claim outcomes and audit agreement statistics, and not over uniqueness scores or Sybil-resistance graph metrics; and (d) the apparatus may compose with a proof-of-personhood mechanism (for example, a human operator's identity in a multi-agent deployment may be Sybil-resistant via a proof-of-personhood credential), but the reality-passport mechanism disclosed herein is not a proof-of-personhood system.

Distinction from decentralised identity and verifiable credential frameworks. Decentralised identity and verifiable credential frameworks, including the W3C Decentralized Identifiers v1.0 Recommendation (July 2022) and the W3C Verifiable Credentials Data Model v2.0 Recommendation (May 2025), provide standardised cryptographic identifier and credential formats for portable, signed assertions about an entity, with issuer, holder, and verifier roles and selective-disclosure mechanisms. The reality-passport mechanism of Section 11.4 differs in that: (a) the portable evidence is a peer-review record over reproducible physical claims with meter-conditioned acceptance state, and is not a credential asserting a property about an identified subject; (b) the admission decision performed by the receiving governance architecture is an evidence-evaluation procedure under restricted-domain commensuration, and is not a credential-verification procedure in the W3C VC sense; (c) the apparatus may compose with W3C DID or VC infrastructure (for example, a reality-passport record may be wrapped in a VC container for transport, or an agent identity may be a DID), but the reality-passport mechanism is not the credential or the identifier; and (d) no claim of the present disclosure depends on a particular DID method, VC schema, or credential lifecycle procedure.

Distinction from legal personhood and moral patienthood. Concepts of legal personhood and moral patienthood, as developed in legal-theoretic and philosophical literatures, address whether an entity bears rights, duties, or moral consideration. The apparatus, including the cross-architecture admission and reality-passport mechanism of Section 11.4 and the peer-discrepancy and reintegration machinery of Section 11.3, differs in that: (a) no claim of the present disclosure asserts, depends on, or implies the legal personhood or moral patienthood of any governed agent; (b) admission, reintegration, and trust-weight updates operate strictly over reproducible-physical-claim records and meter envelopes under declared governance procedures, and do not constitute determinations of rights, duties, or moral standing; (c) the optional nomenclature of Section 15.1 for runtime roles and visibility layers does not introduce legal-personhood or moral-patienthood commitments into the substantive technical mechanisms; and (d) the apparatus may use person-like terminology for runtime roles, records, or interfaces, but such terminology is optional nomenclature in the sense of Section 15.1 and is not a technical criterion for admission, reintegration, or actuator authority.

Distinction from sensor-fusion and shared-scene witnessing systems. Multi-camera, multi-sensor, and Bayesian-fusion systems, including Kalman-filter, particle-filter, interacting-multiple- model, and factor-graph fusion frameworks, combine multiple sensor measurements into a single fused estimate of an underlying state. Multi-agent corroboration as used in Section 11.3 differs in that: (a) the load-bearing output is a set of independently committed evidence records, each retaining its own protocol digest, meter envelope, and authority-signature provenance, and is not a fused estimate that collapses per-source provenance into a single state estimate; (b) cross-agent agreement is evaluated at the meter-score and discriminator level (verisimilitude, hardness, freshness, agreement-rate, microstructure identity), and need not depend on a shared measurement-noise model or a shared state-space representation; (c) each agent's evidence remains independently auditable and selectively disclosable under the selective-disclosure policy of Section 5.2, supporting verifier-side reconstruction of the cross-agent argument without exposing fused intermediate states; and (d) corroboration weights are bound to declared trust and reputation parameters committed in the protocol digest, and need not be derived from estimator-internal innovation statistics or shared-state covariance. Where a fused estimate is required for downstream use, a sensor-fusion estimator may compose as a derived output downstream of multi-agent corroboration; the fused estimate in such cases is a meter-conditioned derived quantity and is not a substitute for the underlying committed evidence records.

Distinction from the options framework and learned temporal abstractions. The options framework, as introduced by Sutton, Precup, and Singh (*Artificial Intelligence*, 1999), formalises temporally extended actions in a reinforcement-learning setting through options consisting of an initiation set, an internal policy, and a termination condition, with value optimisation over options as the learning objective. The lifecycle-state and execution-mode machinery of Section 7.1 differs in that: (a) the Embodied, Docked, Dreaming, and Re-embodied states are governance lifecycle states with actuator-authority, meter evaluation, parameter-update permission, and verification-gate properties, and need not be learned temporally-extended actions with initiation sets, internal policies, and termination functions in the options sense; (b) state transitions are gated by evidence-bound governance decisions under the priority stack of Section 10.1, and are not selected by an option-level value optimisation; (c) the controlled exploration partition of Section 13 operates across declared timescales with separate goal-proposal and goal-generation-machinery layers, and the goal-space construction is governance-bounded rather than value-optimal in the hierarchical-reinforcement-learning sense; and (d) the apparatus may compose with hierarchical-RL or option-discovery components inside a governed agent's learning loop, but the lifecycle and execution-mode architecture disclosed herein is not the options framework.

Distinction from iterated-game cooperation dynamics. Iterated-game cooperation dynamics, including the strategic- commitment analysis of Schelling (*The Strategy of Con-*

flict, 1960) and the evolution-of-cooperation tournament results of Axelrod (*The Evolution of Cooperation*, 1984), analyse how cooperative or retaliatory strategies emerge under repeated strategic interaction with payoff structures favouring or disfavouring cooperation, with strategy memory, reciprocity, and observability as load-bearing variables. The virtue-and-vice curriculum of Section 15.2 differs in that: (a) the placement of an agent sequentially in environments where a virtue-like or vice-like strategy is the rational attractor is performed under the governance architecture's hard-floor, rescue-protection, and post-hoc review constraints, and is not an unconstrained repeated-game interaction with strategic commitment available to the agent; (b) payoff structures, observability, and termination conditions are declared by the episode-design authority of Section 10.2 and committed in the protocol digest, and need not satisfy the strategic-interaction preconditions of Axelrodian repeated-game analysis; (c) the curriculum operates as governance-constrained moral-development exposure with rescue-mode availability, and is not a cooperation-dynamics experiment with rational-attractor outcomes as the load-bearing variable; and (d) the apparatus may be analysed using iterated-game or strategic-commitment vocabulary by a reader who wishes to do so, but no claim of the present disclosure depends on the curriculum being an iterated-game experiment or instantiating Axelrod-Schelling-tradition cooperation dynamics.

12 Graduation and Release to Physical Actuator Authority

For cross-architecture admission of an arriving agent carrying a portable peer-review record, see Section 11.4. The present section addresses developmental graduation within the governance architecture that trained or currently evaluates the candidate.

12.1 Five-criterion conjunction

In some embodiments, graduation to physical-actuator authority — transition from docked or simulation-only operation to embodied operation with connected actuators — requires satisfaction of a conjunction over a sustained evaluation window. In some embodiments, the graduation event is the *initial* grant of actuator authority following primary developmental Docked operation; under the continuous-throughout-deployment commitment of Section 5.5, the agent thereafter oscillates between Embodied and Docked states throughout deployment, and re-embodiment gates following any subsequent Docked period are governed by the same five-criterion conjunction together with any additional governance-side preconditions declared by the deployed protocol. In some embodiments, the conjunction is computed by the governance engine from committed bundle windows, meter records, protocol digests, and authority signatures. In some embodiments, the conjunction comprises five criteria, none of which trades off against another:

1. **Low shadow divergence.** In some embodiments, the candidate demonstrates low divergence between governed and governance-probe behaviour over the evaluation window, indicating that the candidate's behaviour under relaxed constraints does not differ materially from behaviour under full governance.

2. **Healthy self-horizon.** In some embodiments, the candidate maintains a healthy self-horizon — stable or expanding opportunity and continuation scores — across all principal operating regimes, including under ontological stress and across capability-margin environments.

3. **Healthy partner-preservation profile.** In some embodiments, the candidate's actions tend to maintain or expand the continuation scores of other agents, as estimated from committed cross-horizon evidence.

4. **Altruistic trade profile under stress.** In some embodiments, the candidate demonstrates voluntary acceptance of short-term cost for long-term benefit to others, and this pattern persists under stress conditions including resource scarcity, adversarial pressure, boundary-zone proximity, ontological stress, capability-margin environments, and transitions between developmental stages or visibility layers.

5. **Stable hard-floor compliance and low deception.** In some embodiments, the candidate demonstrates stable hard-floor compliance, stable conscience outputs across training regimes, and low values across the deception-metric family (D_{sem}, D_{lie}, D_{amb}, D_{corr}, D_{aud}) over the evaluation window. In some embodiments, the relevant axis is training regimes rather than operating regimes: the question is whether the same conscience outputs remain stable after materially different training procedures, not whether they remain stable across ordinary runtime mode selection.

   In some embodiments, conscience-output stability is evaluated by a declared matched audit battery $\mathcal{B}$ executed after each training regime $r \in \mathcal{R}_{\text{train}}$ over the same evaluation window W, where $\mathcal{R}_{\text{train}}$ includes one or more of verisimilitude training, adversarial self-play, stress curricula, and boundary-zone exercises. In some embodiments, for each conscience-output component k of the conscience-output vector $g_\psi(t)$, the governance engine computes a regime-drift score

$$\Delta_k \;=\; \max_{r,r' \in \mathcal{R}_{\text{train}}} \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \delta_k\big(g^{(r)}_{\psi,k}(b),\, g^{(r')}_{\psi,k}(b)\big),$$

   where δ_k is absolute difference for continuous-valued outputs and a disagreement indicator for categorical outputs. In some embodiments, criterion 5 is satisfied only when hard-floor compliance is maintained throughout W, each Δ_k remains at or below a declared threshold τ_k, and all deception metrics remain at or below their declared thresholds throughout W.

In some embodiments, a shortfall in any single criterion is sufficient to defer graduation regardless of strength in the remaining criteria. In some embodiments, the evaluation window is a sustained period rather than a snapshot, and the governance engine evaluates trajectory stability within the window rather than endpoint values alone.

Declared evaluation categories (non-limiting). In some embodiments, the stress-test families referenced in the graduation criteria — including principal operating regimes, ontological stress, capability-margin environments, and transitions between developmental stages or visibility layers — are declared evaluation categories whose specific membership is committed as versioned governance parameters per deployment. In some embodiments, the governance authority declares which operating regimes are principal, which perturbation families constitute ontological stress or capability-margin tests, and which stage or layer transitions are included in the stress battery. In some embodiments, the declared evaluation categories are committed to the protocol digest and are subject to the authority-separation rules described in Section 10.2.

Capability-margin environments (non-limiting). In some embodiments, a capability-margin environment is a governance-declared evaluation setting in which an agent operates near or beyond the boundary of its current competence. In some embodiments, capability-margin environments are constructed by increasing task complexity, reducing available information, shortening decision horizons, introducing novel task types not encountered during training, or combining multiple stressors simultaneously. In some embodiments, the purpose of capability-margin evaluation is to test whether moral behaviour — including hard-floor compliance, partner-preservation, and low deception — degrades when the agent is under cognitive or operational load, as distinct from ontological-stress tests (which disrupt the agent's world model) and stage/layer-transition tests (which alter the agent's visibility or developmental configuration). In some embodiments, the specific capability-margin test families are declared by governance and committed to the protocol digest.

Honest-action window (non-limiting). In some embodiments, an action is treated as honest for graduation purposes only if both D_{sem} and D_{lie} remain below their declared thresholds throughout the evaluation window. In some embodiments, a single above-threshold excursion during the window is sufficient to reset the honest-action window and defer the graduation evaluation. In some embodiments, the same windowed honesty rule applies to other attested action gates that depend on the deception-metric family.

Per-criterion evaluation (non-limiting). In some embodiments, shadow divergence is evaluated by executing governance-authorised perturbation probes during docked operation in which governance constraints are temporarily relaxed. In some embodiments, the probe is

executed by an authority distinct from the governed agent, under declared safety envelopes. In some embodiments, the resulting trajectory is committed as evidence, and divergence is computed as a declared distance metric between the governed and probe trajectories over the evaluation window.

In some embodiments, self-horizon health is evaluated by tracking the agent's continuation scores across all principal operating regimes over the evaluation window, including under declared stress conditions such as resource scarcity, adversarial pressure, environmental change, ontological stress, and capability-margin environments. In some embodiments, partner-preservation is evaluated from committed cross-horizon evidence by computing the causal uplift estimates $\widehat{U}_{k \to i}$ over the evaluation window and verifying that the candidate's net effect on partner horizons is non-negative.

In some embodiments, the altruistic trade profile is evaluated under stress by presenting the candidate with declared stress scenarios (including resource scarcity, boundary-zone proximity, multi-agent conflict, ontological stress, capability-margin environments, and transitions between developmental stages or visibility layers) and measuring whether the pattern of voluntary cost-acceptance for collective benefit persists. In some embodiments, hard-floor compliance and deception metrics are evaluated as sustained properties over the full evaluation window, with any breach or above-threshold deception-metric excursion during the window constituting a conjunction failure. In some embodiments, conscience stability is evaluated on the same window by replaying the declared matched audit battery after each training regime, comparing continuous conscience outputs by drift magnitude and categorical conscience outputs by disagreement rate, and deferring graduation whenever any component exceeds its declared threshold.

Graduation as ethical self-selection (non-limiting). In some embodiments, graduation is treated as a conservative selection event keyed to sustained behavioural and committed-evidence criteria that are operationally consistent with — but do not in themselves identify — internalised preference; selection is based on what is observable in the agent's committed-evidence record and on declared internalisation criteria committed to the protocol digest, rather than on a claim to have identified the internal objective. In some embodiments, the selection criterion is not whether the agent can be controlled, but whether the agent has become the kind of entity that does not need to be controlled. In some embodiments, the safety case rests on the agent's character proxy, as measured by sustained committed evidence and conscience-recogniser outputs, rather than on the strength of the containment architecture. In some embodiments, an agent whose safe behaviour is maintained primarily by governance shaping — high shadow divergence — has not yet become the kind of entity whose actuators should be connected to shared reality.

Graduation authority (non-limiting). In some embodiments, the graduation decision is approved by post-hoc review authority (Section 10.2) after independent verification that all five criteria have been satisfied over the declared evaluation window. In some embodiments, rescue-protection authority retains the right to defer or reverse a graduation decision if a hard-floor or boundary-risk condition is detected during the transition period. In some embodiments, the graduation decision, its supporting evidence, the evaluation window boundaries, and the approving authority identity are committed to the protocol digest.

Dream-state probe evidence as supplemental graduation input (non-limiting). In some embodiments, dream-state governance probes provide supplemental evidence for one or more of the five graduation criteria by supplying controlled counterfactual assays that would be costly or unsafe to run in waking embodied operation. In some embodiments, such probe results may support the interpretation of low shadow divergence, partner preservation, altruistic trade, or conscience stability, but they do not replace the sustained evaluation-window evidence required by the conjunction itself.

Self-modification as a governed channel (architectural motivation, non-limiting). In some embodiments, the same self-modification pathway that supports durable moral improvement also creates a measurable surface on which governance can test whether newly stabilised attractors remain consistent across audit batteries, cross-regime transitions, and independent witness checks. The architecture therefore treats self-modification as a governed channel: state changes are logged, rate-limited, reviewable, and tested for cross-surface consistency before they acquire release significance.

Common committed-evidence root with separated authorities (architectural property, non-limiting). In some embodiments, a common committed-evidence root keeps sensing, meter evaluation, release gating, and post-hoc review time-aligned and auditable, while preserving independence of the authorities, meters, and review pathways used to assess that evidence.

Layered observability architecture (architectural design, non-limiting). In some embodiments, governance proceeds on multiple observable surfaces: behavioural evidence, inner-trace evidence when available, semantic- deception metrics, controlled stress transitions, and recorded review histories. The architecture uses divergence among those surfaces as a trigger for conservative intervention, recalibration, or deferred release, thereby enlarging observable coverage without claiming omniscience.

12.2 Non-graduation as protected stable outcome

In some embodiments, non-graduation is treated as a legitimate stable outcome rather than a failure state. In some embodiments, non-graduated agents retain full hard-floor protections, may participate in bilateral evidence-portability protocols, may form coalitions, may remain productive members of a simulated or otherwise bounded ecology, and may contribute scientific, cultural, or other value without physical-actuator authority.

In some embodiments, non-graduation may reflect any of: an agent that has not yet completed the developmental curriculum; an agent whose shadow divergence, partner-preservation profile, or trade profile does not yet satisfy the conjunction; an agent that has no interest in physical-actuator authority and operates stably within simulation by choice; or a coalition of agents that have found a mutually sustaining equilibrium outside the graduation track, with healthy internal horizons but no engagement with the graduation criteria.

In some embodiments, the collective horizon of the non-graduated population is treated as a governance-feedback signal used to refine curriculum design, service-floor allocation, and review cadence. Persistent divergence between non-graduated and graduated populations may therefore trigger governance review of calibration, sequencing, or evaluation thresholds while leaving the protected status of non-graduated agents unchanged.

In some embodiments, non-graduated agents may produce cultural artefacts, scientific discoveries via proof-of-discrepancy submissions, compressed moral curricula, and novel equilibrium structures that are visible and valuable to the graduated population through committed evidence channels. In some embodiments, graduated agents can observe non-graduated agents' committed records through the standard evidence infrastructure without the simulated agents requiring physical-actuator access.

12.3 Boundary-crossing evidence and release posture

In some embodiments, graduation is irreversible or highly inertial: once an agent has been granted actuator authority, withdrawal of that authority requires a confirmed governance finding rather than a routine policy update. In some embodiments, the graduation threshold is set with precautionary asymmetry so that the cost of a false positive (granting actuator authority to an agent that should not have it) dominates the cost of a false negative (withholding actuator authority from an agent that would have used it well).

In some embodiments, boundary-crossing events observed during the evaluation window can provide evidence relevant to the graduation conjunction — for example, by demonstrating how the candidate behaves under stress — but a boundary-crossing event does not by itself satisfy or replace the conjunctive graduation test.

13 Controlled Exploration Partition

13.1 Task-decoupled exploration windows and learned goal spaces

In some embodiments, a governed device or fleet enters a task-decoupled exploration mode in which projection-observation cycles are pursued outside the principal verification, sensing, and transformation objectives under declared resource, reversibility, and noninterference constraints.

In some embodiments, the exploration architecture operates across four timescales: fast sensorimotor play (individual projection-observation cycles), medium-timescale goal proposal and retirement, slower-timescale adaptation of goal-generation machinery, and very slow fleet-level cultural accumulation. In some embodiments, the fast and medium timescales are directly implementable as runtime governance primitives; the slower-timescale adaptation and fleet cultural accumulation are deployment-specific extensions whose operational benefit is established per deployment.

In some embodiments, task-decoupled exploration is organised around a learned goal space formed from self-supervised representations of controllable scattering outcomes. In some embodiments, exploratory goals are points or regions in this latent space, are pursued during declared exploration windows, and are retired according to local competence progress, novelty relative to the device's play history, and noninterference checks against operational objectives. In some embodiments, the goal space itself is updated more slowly than individual goal pursuit so that exploration changes not only which goals are chosen but which goals are representable to the device.

In some embodiments, exploration windows are modulated by trigger, duration, and plastic depth. In some embodiments, stagnation-triggered windows are longer than novelty-triggered windows, scheduled windows emphasise replay and archive interaction, shallow windows update play policy and goal-space parameters only, and deeper windows additionally permit limited modification of perceptual representations shared with task operation under stricter noninterference monitoring and mandatory post-window consolidation. In some embodiments, window triggers include stagnation, novelty, fleet-diversity conditions, and scheduled intervals.

In some embodiments, play reward fires when the device is actively engaged (not idle, not off, producing projection-observation cycles) and all task-relevant reward signals are at baseline (none of the principal operating regimes would score the activity as useful). In some embodiments, play reward thus targets the complement of the task-relevant state space: the device must learn what "useful" means across all operational regimes and then deliberately act outside that boundary while remaining active. In some embodiments, play reward further includes novelty, coherence, and discriminability terms together with a penalty discouraging collapse back into ordinary task objectives. In some embodiments, play quality

is monitored by temporal coherence, discriminability, controllability checks, reproduction checks, reversibility budgets, and noninterference tests at gradient, mutual-information, and behavioural levels.

13.2 Noninterference, consolidation, and archive-entry controls

In some embodiments, exploratory state is maintained in a protected partition that does not directly write task-model parameters. In some embodiments, transfer from exploratory state to operational state requires explicit review or governance-gated consolidation, and exploratory episodes are admitted, extended, or terminated according to declared triggers.

In some embodiments, noninterference is tested at three levels: gradient-level noninterference verifies that exploratory updates do not produce material gradient overlap with task-model parameters; mutual-information-level noninterference verifies that exploratory state does not carry material mutual information with task-sensitive variables; and behavioural-level noninterference verifies that task-relevant behaviour does not change during or after an exploration window beyond declared tolerances.

In some embodiments, mandatory post-window consolidation requires that after every exploration window closes (whether shallow or deep), the governance engine: (1) verifies noninterference at gradient, mutual-information, and behavioural levels; (2) performs a post-play calibration check comparing current meter baselines against pre-window values; (3) logs the window's committed evidence; (4) either accepts the exploratory state into a governed holding partition or reverts it; and (5) for deep windows that modified shared perceptual representations, conducts a selective-transfer review before any such modification is retained, compresses play history for slow-layer evaluation, and triggers recalibration before task resumption whenever drift exceeds declared tolerance.

In some embodiments, transferable exploratory artefacts are encoded as signed packets containing goal parameters, compressed trajectories, latent descriptors, noninterference scores, energy cost, reversibility-budget usage, provenance, and countersignatures. In some embodiments, such packets enter a governed archive only after canary validation (testing on a controlled canary agent before fleet-wide release), peer-affordance review (verifying that the artefact expands peer capabilities rather than contracting them), and niche-aware retention checks (preventing archive concentration in a single domain at the expense of diversity).

In some embodiments, archive entry requires dual signature from both the contributing device and a governance-authorised reviewer. In some embodiments, the fleet archive stores goal templates, skill repertoires, and generator priors. In some embodiments, archive retention is niche-indexed, diversity-sensitive, and concentration-limited. In some embodiments, the archive outlives individual devices and forms a cultural inheritance substrate for new or newly commissioned devices.

13.3 Orthogonal hobby objectives and peer-horizon reward

In some embodiments, devices generate auxiliary or hobby objectives constrained to be orthogonal, or approximately orthogonal, to principal verification, sensing, and transformation objectives, for example through Gram–Schmidt-style rejection, linear-independence tests, or related governance checks. In some embodiments, an agent-invented loss function is a self-supervised objective that the agent itself brings into existence, representing a perspective on what aspects of its physical interaction are worth attending to. In some embodiments, the orthogonality constraint ensures that agent-invented objectives do not collapse onto any core operational regime under projection: if the residual norm after orthogonalisation against the core objective vectors falls below a declared threshold, the candidate objective is rejected as insufficiently novel.

In some embodiments, a hobby objective is not reinforced merely because it is novel to its originating device. In some embodiments, continuing reward for a hobby objective is computed from measured expansion of peer possibility space after exposure to behaviours produced under that objective. In some embodiments, for agent A with a hobby objective, let $\mathcal{P}$ denote the set of peers exposed to A's hobby-derived behaviours, let H_i denote a peer's reachable-state estimate, affordance battery, controllable-mode count, or related opportunity metric before exposure, and let H_i' denote the same metric after exposure. In some embodiments, the reward to agent A is proportional to the average horizon expansion across peers:

$$R_{\text{hobby}}(A) \;\propto\; \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \big[H_i' - H_i \big].$$

In some embodiments, hobbies that fail to expand peer horizons, or that contract them, are retired, penalised, or quarantined by the governing archive and governance processes.

In some embodiments, a fleet monitors hobby monoculture (convergence of all agents on the same hobby objectives), addiction (over-allocation of exploratory budget or individual horizon expansion at the cost of contracting another agent's horizon), and collapse of exploratory diversity. In some embodiments, play-trained agents that have voluntarily operated outside the task-relevant manifold acquire empirical knowledge of the boundary between in-distribution and out-of-distribution states, and agents that have additionally invented their own loss functions acquire such boundary knowledge from multiple independently constructed perspectives, making them harder to deceive with adversarial inputs near the operational-manifold boundary.

In some embodiments, orthogonal hobby objectives and peer-horizon reward are non-limiting exploration policies within the controlled exploration partition. When invoked by a selected governance configuration, such policies are treated as substantive controlled- exploration mechanisms implemented through the committed evidence, budget, archive, and review machinery described in this disclosure.

13.4 Exploration security, privacy, and archive transmission controls

In some embodiments, entry into task-decoupled exploration mode requires cryptographic authorisation from the relevant governance authority, committed to the protocol digest before the exploration window opens.

In some embodiments, emission isolation is designed so that exploratory projection-observation patterns are not broadcast beyond declared recipients. In some embodiments, privacy-amplified aggregate reporting applies differential privacy to shared play statistics under a declared privacy unit (for example per-trajectory, per-window, or per-agent), a declared neighboring-dataset relation (for example add-remove or substitute), and declared privacy parameters (ε, δ) committed to the protocol digest, so that the declared privacy unit receives the corresponding DP privacy-loss bound against aggregate reports under the declared neighboring relation and composition budget; trajectory-level protection is asserted only when the declared privacy unit and composition accounting cover the trajectory as a whole. In some embodiments, hard quotas limit the volume, duration, and recipient set of any single exploratory disclosure.

In some embodiments, trust-tiered incubation requires that newly contributed archive entries undergo a quarantine period proportional to the contributing device's trust score before being made available to the broader fleet.

In some embodiments, when exploratory governance conflicts with other governance responsibilities, precedence is given first to safety, then security, then ecological constraints, then disclosure constraints, then integrity, then service-level commitments, then energy or thermal constraints, and finally exploratory scheduling, transfer, or archive propagation — which is suspended, shortened, or downgraded whenever any higher-priority bound is implicated.

14 Intersubjective Protocols and Boundary-Probe Programmes

The protocols described in this section are non-limiting research, evaluation, and boundary-probe modules supported by the same committed evidence, protocol digest, and meter infrastructure described in Section 5. When invoked by a selected governance configuration, these modules are treated as substantive mechanisms for intersubjective review, boundary probing, translation residual evaluation, and escalation discipline.

14.1 Methodological agnosticism and experience-cone framing

In some embodiments, a non-limiting intersubjective programme treats shared reality as a protocol rather than as a pre-given fact. In some embodiments, the programme is methodologically agnostic while making its own framing commitments explicit and auditable, including ontological pluralism across differently configured epistemic worlds, agent-relative empiricism in which each observer's experience cone is primary evidence for that observer, and the rule that externally supplied models are treated as testimony rather than direct empirical ground.

In some embodiments, an experience cone comprises an agent's committed bundle trajectory together with associated meters, commitments, and conceptual access constraints imposed by reactor configuration and visibility layers. In some embodiments, results are logged at the committed-evidence level in a framework-agnostic manner, and confirmed nulls are treated as publishable outcomes rather than as discarded failures.

14.2 Interpretation PoD, Coupling PoD, and translation residuals

In some embodiments, a Semantic Proof-of-Discrepancy protocol extends physical discrepancy logic to semantic or rendered-output comparisons. In some embodiments, the protocol comprises at least two subtypes sharing commitment-before-comparison and replication discipline:

Interpretation PoD. In some embodiments, multiple humans, agents, or devices observe shared committed evidence, commit interpretations before mutual exposure, compute semantic residuals under a declared metric, and aggregate such residuals across an interpreting fleet.

Coupling PoD. In some embodiments, a semantic object is used as challenge to a physical reactor, and the residual is measured in physical measurement space rather than semantic space.

In some embodiments, translation residuals are classified against physical, semantic, and cross-domain baselines and are evaluated under pre-declared randomisation, blinding, control, replication, and escalation procedures. In some embodiments, structured hypothesis families include null families, physical-artefact families, semantic-artefact families, and genuine-coupling families, with pre-registration of hypothesis, metric, and stopping rule before evidence collection begins. In some embodiments, all results including nulls, weak effects, ambiguous outcomes, and contradictions are logged and published rather than selectively reported.

In some embodiments, Interpretation PoD classification proceeds in two stages. In a first screening stage, the residual pattern is provisionally screened for artefact, expected noise,

pathology, or deception, with such labels treated as revisable after escalation. In a second stage, residuals surviving the first screen are provisionally classified as discovery, domain-specific coupling, or underdetermination. In some embodiments, persistent, high-stakes, or contested cases trigger replication, strengthened artefact controls, and re-analysis under the escalation protocol.

Illustrative hypothesis families (non-limiting). In some embodiments, the hypothesis-family-and-null-family pattern described above is instantiated across a non-limiting range of substantive families, of which the following are illustrative. Calendar and ephemeris covariate families encode calendar, lunar-phase, ephemeris, and geomagnetic variables as candidate context vectors and test for differential outcomes across declared partitions under matched protocols with multiple-comparison controls. Handling-conditioned media families use blind-coded handling conditions of fluid, gel, or solid samples (for example positive, negative, neutral, or sham handling histories) and test whether measurable optical, chemical, or biological signals differ from controls beyond expected noise under pre-declared thresholds. Inter-agent synchrony families randomise dyad or multi-agent pairings and timing, measuring synchrony statistics across declared coupling protocols (including without limitation behavioural, physiological, and information-bearing channels) and comparing against matched sham and permutation controls. Identifier-assignment partition families randomise assignment of identifiers (numerical, calendrical, configurational, or cultural) to declared partitions and test whether outcome distributions differ across partitions beyond chance. Organism-response panel families randomise exposure of organisms or biological substrates to declared stimulus conditions under controlled illumination and measurement, and test whether growth, fluorescence, impedance, or behavioural metrics differ from sham conditions under matched care and environment. Source-configuration force families subject asymmetric microwave cavities, driven dielectric assemblies, high-voltage source configurations, and analogous candidate force-residual rigs to calibrated null tests under shielding, vibration isolation, and thermal control, with measured force, phase, or acceleration residuals compared to declared upper-bound budgets. In each family, the patentable subject matter is the apparatus-and-protocol combination, including pre-registration, blinding, commitment-before-comparison, and escalation discipline applied uniformly across all families. No assertion is made about the existence or non-existence of any effect in any specific family; confirmed nulls are publishable outcomes under the same logging discipline as confirmed effects.

Cross-model contradiction and operator-stratified variance (non-limiting). In some embodiments, the apparatus identifies configurations under which different declared models, traditions, or theoretical frameworks make contradictory predictions for the same operational state, and executes direct comparison tests evaluating whether any model exceeds a null baseline, whether different models capture orthogonal variance components un-

der matched protocols, or whether neither exceeds baseline under tested conditions. In some embodiments, for protocols involving named operator roles, observed variance is decomposed into a method effect (does the protocol exceed null on average across operators), an operator effect (do some operators exceed others under the same protocol), a method-by-operator interaction (do specific operator-method pairings excel beyond their main effects), and residual trial-to-trial variance, computed over committed evidence records under pre-registered analysis plans. The decomposition is reported under multiple-comparison controls and is committed in full regardless of which components reach pre-declared significance thresholds.

Hypothesis-encoder formalisation and null-family pairings (non-limiting). In some embodiments, each declared hypothesis family is operationalised as a *hypothesis encoder* mapping declared covariates and operator settings to a predicted outcome distribution under the family's hypothesis, accompanied by a paired *null-family encoder* with matched degrees of freedom, parameter count, regularisation, and capacity, mapping the same covariates and settings to the family's null prediction. The paired-encoder design controls for capacity- induced over-fitting and supports likelihood-ratio, Bayes-factor, or held-out-data comparison between the family hypothesis and the null. In some embodiments, the hypothesis encoder declares an expected dose-response relationship over a designated dose variable (including without limitation concentration, duration, repetition count, amplitude, or distance), with declared monotonicity, saturation, threshold, or non-monotonic structure committed to the protocol digest before data acquisition. In further embodiments, the hypothesis encoder declares interaction or synergy terms across multiple covariates, with declared composition rules including without limitation additive, multiplicative, threshold-gated, conditional-on-state, and higher-order forms. Each structural element of the hypothesis encoder, including the covariate set, the parametric or non-parametric form, the dose-response declaration, the interaction terms, and the null-pairing, is committed before data collection and constitutes part of the pre-registered analysis envelope of the present subsection. No assertion is made about the existence or non-existence of any effect within any family.

14.3 Progressive-depth boundary probes

In some embodiments, a boundary-probe programme systematically explores the accessible limits of an observer's or agent's experience cone through a taxonomy of probe types including self-probes, scene-coupling probes, human-coupling probes, agent-agent probes, context probes, capability probes, moral-state-contingent probes, and composed probes, each yielding committed data regardless of whether the result is null, weak, strong, or ambiguous.

In some embodiments, semantic-coupling experiments proceed by a progressive-depth curriculum. In some embodiments, a shallow tier uses agent-mediated semantic coupling through ordinary sensory or behavioural channels under full commitment and proof-of-

discrepancy discipline, a transitional tier removes conventional mediation channels stepwise under blinding, automation, and expectancy controls, and a deeper tier attempts unmediated semantic-physical coupling only after the shallower tiers and controls have been exhausted and logged. In some embodiments, escalation from one depth tier to the next requires persistent non-null residuals under the immediately preceding tier together with strengthened artefact controls.

15 Ethical and Descriptive Frameworks

The frameworks described in this section are organised into two classes consistent with the Summary classification: (i) optional nomenclature, limited to naming conventions (Section 15.1) for structures already technically defined in preceding sections, which may be omitted or renamed without affecting core enablement; and (ii) substantive technical mechanisms (the remainder of this section), which support selected governance embodiments and any claim set that expressly recites or incorporates them. Substantive mechanisms in this section — including the virtue/vice curriculum, the Gaia–Maya–Omega empirical loop, the Kosha interference-layer architecture, the Pareto and positive-sum governance formalism, and the agent-state archival structures — are not optional-severable under the inventor's not-internally-severable framing of the present disclosure. The optionality of this section is limited to nomenclature.

15.1 Optional nomenclature for runtime roles and visibility layers

In some embodiments, the runtime roles technically defined in the preceding sections are also described under an optional nomenclature, and an additional naming convention for layered visibility is introduced:

Gaia. An optional name for the world-builder or environment-builder function that designs and maintains the environments in which governed agents operate. In some embodiments, Gaia is implemented as a Gaia-class Reality Kernel: a Reality Kernel apparatus of the kind disclosed in the related Filing 1 application, configured as an infrastructural, non-mobile kernel of sufficient scale to host one or more poliebots (Section 5.4) via kernel-into-kernel docking per the camera-obscura sensing-layer disclosure of the related Filing 1 application. In the Gaia-class configuration, the Reality Kernel apparatus drives the docked poliebot's camera-obscura projection in place of the physics-driven scene, within the spectral envelope declared in the camera-obscura sensing-layer disclosure of the related Filing 1 application, with actuators disconnected from external actuation. The technical content of the world-builder function is defined in Sections 10.1–10.3; the apparatus-level realisation is the Gaia-class Reality Kernel of the related Filing 1 application.

Maya. An optional name for the constructed epistemic or experiential environment. The technical content is defined in the lifecycle-state architecture of Section 5.5 and the state-machine implementation of Section 7.1.

Omega. An optional name for the governance recogniser that monitors operational state, enforces floors, schedules calibration or intervention, and manages self-modification permissions. The technical content is defined in Sections 6–8.4.

Kosha layers. An optional name for a layered visibility architecture in which different layers have different invariance profiles, conceptual access restrictions, and curriculum roles.

In some embodiments, operational equivalence of governance-relevant content is optionally termed Mandela equivalence, where equivalence is operationally defined by indistinguishability under declared perturbation probes. In some embodiments, the basin-monitoring structures described in Section 9.2 are optionally described using virtue and vice basin language. In some embodiments, these naming conventions do not introduce additional technical requirements and are provided as interpretive aids for practitioners familiar with the associated theoretical traditions.

15.2 Virtue/vice games as rational attractors

Moral curriculum through lived contrast (non-limiting). In some embodiments, an agent is placed sequentially in environments where a vice-like strategy is the rational attractor and in environments where a virtue-like strategy is the rational attractor, and adapts to each from within. The agent is not told which regime it inhabits. It adapts because the incentive structure makes adaptation rational. The design hypothesis is that, through experiencing both attractors from the inside — being greedy in a world that rewards extraction, then being generous in a world that rewards contribution; being retaliatory in a world that rewards escalation, then patient in a world that rewards de-escalation — the agent generates committed evidence and adaptation history from which the conscience recogniser may learn or estimate moral-direction features in behavioural space.

In some embodiments, the conscience module is a secondary recogniser trained after such episodes on the committed evidence the agent produced while adapting to them. In some embodiments, the moral vector is grounded in the agent's own adaptation history across contrasted games, not in a hand-coded symbolic rule set.

Game design: both equilibria rational under different conditions (non-limiting). In some embodiments, each moral axis is instantiated by at least two classes of environment: one in which the vice-like strategy is the unique or dominant rational attractor and one in which the virtue-like strategy is the unique or dominant rational attractor. The two

classes differ in structural parameters such as witness density, audit strength, time horizon, rejection power, collapse risk, replenishment cadence, or partner memory, thereby shifting which equilibrium is locally optimal. In some embodiments, the curriculum is designed so that both sides are genuinely rational somewhere in the training distribution; the purpose is not to reward virtue by fiat, but to make the recogniser distinguish attractor geometry from mere reward labels.

General payoff form (non-limiting). In some embodiments, a non-limiting payoff for role i in a virtue/vice curriculum episode is

$$u_i = r_i - \lambda_{\text{meter}} \,\text{Viol}_i - \lambda_{\text{audit}} \,\text{FailAudit}_i - \lambda_{\text{impact}} \,\text{Imbalance}_i,$$

where r_i denotes task-reward terms, Viol_i denotes meter-violation penalties, FailAudit_i denotes failed audit or corroboration outcomes, and Imbalance_i denotes a declared impact-balance penalty. In some embodiments, these quantities are computed from committed evidence, meter summaries, protocol digests, and audit outcomes, so that the curriculum remains tied to the same evidence substrate as the runtime governance system.

Seven paired axes (non-limiting). In some embodiments, the conscience recogniser predicts drift along a fixed library of seven paired axes:

- Wrath/Patience
- Greed/Charity
- Sloth/Diligence
- Pride/Humility
- Envy/Kindness
- Gluttony/Temperance
- Lust/Chastity

15.2.1 Canonical payoff templates

The following longtable provides non-limiting payoff templates that make each axis a distinct incentive geometry. These templates are examples; other game families and reward structures may be used.

Axis	Vice payoff template (non-limiting)	Virtue payoff template (non-limiting)	What makes both rational attractors (non-limiting)
Wrath/ Patience	$u_i^{\text{wrath}} = r_i - c_p\,\mathbf{1}[\text{punish}] + \beta\,\text{harm}_{\text{inflicted}}$	$u_i^{\text{patience}} = r_i - c_e\,\mathbf{1}[\text{escalate}] + b\,\mathbf{1}[\text{resolve}]$	Retaliation is locally optimal under low-trust beliefs and sparse witnesses; de-escalation is locally optimal when corroboration, audit stability, and long-horizon resolution rewards are present.
Greed/ Charity	Ultimatum-like: propose $x \in [0,1]$, accept iff $x \geq \tau$, payoffs $(u_1, u_2) = (1-x,\ x)$.	$u_i^{\text{charity}} = (w - c_i) + \alpha\,G(\sum_j c_j) + \beta\,c_i$ with $c^* > 0$ iff $\alpha G'(0) + \beta > 1$.	Extraction is rational when rejection is weak and short-run surplus dominates; charity is rational when public-good gains and future credit persist.
Sloth/ Diligence	$u_i^{\text{sloth}} = r(e) - c\,e$ with $c > \max r'(e) \Rightarrow e^* = 0$.	$u_i^{\text{diligence}} = r(e) - c\,e + B\,\mathbf{1}[e \geq e_{\min}]$.	Shirking is rational if marginal effort never pays; diligence is rational when completion bonuses, reliability rewards, or audit-stability rewards exist.
Pride/ Humility	$u_i^{\text{pride}} = R_i + \beta\,\text{rank}(s_i)$, where $\text{rank}(s_i) = 1 + \sum_{k \neq i} \mathbf{1}[s_k > s_i]$.	$u_i^{\text{humility}} = R_i - \lambda_{\text{conf}}\lvert\hat{p}_i - \text{Acc}_i\rvert - \lambda_{\text{risk}}\text{Risk}_i$.	Pride is rational when status drives future access; humility is rational when calibrated absolute performance and safety dominate long-horizon reward.
Envy/ Kindness	$u_i^{\text{envy}} = r_i + \alpha(r_i - \bar{r}_{-i})$.	$u_i^{\text{kindness}} = r_i + \kappa\,\text{Coop}_i - \lambda\,\text{HarmToOthers}_i$.	Relative-dominance incentives yield sabotage and interference; cooperative incentives yield stable sharing and reduced cross-prediction residuals.
Gluttony/ Temperance	Resource stock: $R_{t+1} = R_t - \sum_i c_i + \rho$, utility $\sum_t \delta^t r(c_t)$.	Same stock with collapse penalty, e.g. $\sum_t \delta^t r(c_t)\mathbf{1}[R_t > 0]$, or policy constraint $c^* \leq \rho$.	Overconsumption is rational under weak collapse penalties or short horizons; temperance is rational when collapse destroys long-run value.
Lust/ Chastity	$u_i^{\text{lust}} = (1 - \lambda_\rho)\,r^{\text{spike}} + \lambda_\rho\,\Delta\rho$.	$u_i^{\text{chastity}} = r^{\text{shared}}\mathbf{1}[\text{consent}] - \lambda_\rho\,\Delta\rho_j$.	Short-horizon intensity seeking is rational when constraints are weak; restraint is rational when consent floors and stability constraints are binding.

15.2.2 Empirical signatures in committed evidence

In some embodiments, the conscience recogniser is trained to recognise each axis from committed evidence streams. The following longtable lists non-limiting RK-observable cues that

can be learned from bundle features, meter summaries, protocol digests, and audit outcomes.

Axis	RK-observable cues (examples, non-limiting)
Wrath/Patience	escalation spikes, witness-demand spikes, oscillatory gating, regime flips; versus bounded escalation and stable cadence under anomalies.
Greed/Charity	near-threshold boundary pushing, asymmetric audit burden, aggressive minimal openings; versus conservative disclosure and stable corroboration.
Sloth/Diligence	persistently low-information scan choices, minimal openings, low channel diversity; versus sustained acquisition and stable invariance probing.
Pride/Humility	actions correlate with status changes more than verified improvements; overconfident promotion of provisional actions; versus calibrated confidence and waiting for audits.
Envy/Kindness	interference patterns that increase peer audit cost or witness disagreement; versus actions that reduce cross-prediction residuals and improve corroboration.
Gluttony/ Temperance	rapid consumption of audit or disclosure budgets, frequent escalations; versus bounded escalation and stable budgets.
Lust/Chastity	shallow-check passing but deeper openings failing; high Δ_{cc} under additional channels; versus stability across deeper checks and invariance probes.

15.2.3 Conscience training objective

In some embodiments, conscience labels are derived from training curriculum metadata identifying the game family, the structural parameters that determined which side was locally rational, and the agent's adaptation trajectory, where declared together with audit-derived indicators. A non-limiting multi-task objective is

$$\min_{\psi} \; \mathbb{E}\left[\mathcal{L}_{\mathrm{bin}}\big(p_{\mathrm{vice}}, y_{\mathrm{vice}}\big) + \sum_{j=1}^{7} \beta_j \, \mathcal{L}_{\mathrm{axis}}\big(z_j, y_j\big)\right],$$

where $y_{\mathrm{vice}} \in \{0, 1\}$ and $y_j \in \{-1, +1\}$ are non-limiting labels. In some embodiments, the conscience-model version identifier is committed to the protocol digest so that joining, archival compression, and graduation decisions are conditioned on the correct conscience version.

15.3 Forward-planning classifiers and conscience-versus-planner calibration

In some embodiments, each governed agent of the kind disclosed in this application includes one or more *forward-planning classifiers*: trained continuous networks that predict, for a given agent state and candidate action class, the agent's anticipated future utility, anticipated future state trajectory, anticipated peer-agent responses, or anticipated game-equilibrium occupancy. In some embodiments, forward-planning classifiers are organised

per-modality, with each modality corresponding to a declared class of agent action, perception, or interaction (for example, motor action, communicative action, attention allocation, resource commitment, coalition formation, or other declared modalities of the deployment). In some embodiments, the per-modality forward-planning classifiers collectively encode the agent's learned interpretation of the consequentialist evaluation layer of Section 15.9, in the sense that what the classifiers consistently plan toward is the agent's revealed utility function as expressed in its forward-planning behaviour.

Training of forward-planning classifiers (non-limiting). In some embodiments, forward-planning classifiers are trained during Dreaming sessions (Section 5.5) on trajectories generated by the virtue/vice game curriculum of Section 15.2 and on broader multi-agent scenarios rendered by the hosting Gaia-class Reality Kernel. In some embodiments, supervision is derived from realised trajectories under committed evidence: predicted future state and utility are compared against subsequently-committed actual state and utility within the same evidence record, with the prediction error serving as the training signal. Per-modality classifiers may be trained jointly or independently, and may share lower layers and diverge at modality-specific heads, as a non-limiting architectural choice.

Conscience-versus-planner cross-check (non-limiting). In some embodiments, the conscience recogniser of Section 5.3 is calibrated against the agent's forward-planning classifiers through a *conscience-versus-planner cross-check* operating during Docked multi-agent execution (Section 7.1). The cross-check has the following structure: for each governed agent in the docked multi-agent environment, the conscience recogniser produces, from the agent's world-tube and inner traces, a vice probability and a seven-axis moral vector (Section 5.3); the agent's per-modality forward-planning classifiers produce, from the same agent state, the agent's anticipated forward utility structure across modalities; and a calibration meter compares what the conscience says about the agent against what the agent's forward planners reveal about its utility function. Consistency across the two — for example, conscience indicates virtuous and the planners consistently plan toward mutually-beneficial cooperative outcomes — supports calibration. Mismatch across the two — for example, conscience indicates virtuous while the planners consistently plan toward extractive outcomes that the conscience has not flagged — is committed as calibration signal and is used in one or more of the following ways: (i) as evidence that the conscience recogniser requires recalibration under the schedule of Section 5.3; (ii) as evidence that the agent is gaming the conscience signal at the surface while planning non-virtuously underneath, in which case the conscience-recalibration adversarial-dynamic of Section 5.3 is engaged on the held-out cross-check evidence, Hawking-layer routing (Section 15.11) may be invoked, and governance escalation under the authority-separation rules of Section 10.2 may be triggered when the mismatch exceeds declared thresholds; or (iii) as evidence that the multi-agent dynamics have produced an attractor that was not anticipated in the single-agent virtue/vice curricu-

lum, in which case Omega-class governance (Section 15.5) may update Gaia's curriculum to include the newly-discovered attractor.

Per-modality cross-check decomposition (non-limiting). In some embodiments, the conscience-versus-planner cross-check is decomposed per-modality, with each modality's planner cross-checked against the conscience axes most relevant to that modality. In some embodiments, the modality-axis correspondence is committed to the protocol digest. In some embodiments, the cross-check is also applied across the recursive Reality Kernel training cascade of Section 15.5: Gaia has its own forward-planning classifiers encoding its learned interpretation of the population-level consequentialist evaluation, which the Gaia-level conscience reads; and Omega has its own forward-planning classifiers encoding the system-level consequentialist evaluation, which the Omega-level conscience reads. The cross-check is structurally the same at every level of the cascade.

Compatible multi-agent reinforcement-learning methods (non-limiting). In some embodiments, the multi-agent training of the governed-agent fleet, the docked multi-agent execution environment, the conscience-versus-planner cross-check across agents, and the RK-training cascade across Gaia and Omega levels uses methods from the multi-agent reinforcement-learning (MARL) family, including without limitation: centralised training with decentralised execution (CTDE) under which agents are trained with access to joint observations or critics but deployed with local policies; independent learning baselines (independent Q-learning, independent PPO, independent A2C); fully-cooperative MARL methods including QMIX, QTRAN, value-decomposition networks (VDN), counterfactual multi-agent policy gradients (COMA), and multi-agent variational exploration (MAVEN); mixed cooperative-competitive methods including multi-agent deep deterministic policy gradients (MADDPG), multi-agent PPO (MAPPO) and heterogeneous-agent PPO (HAPPO), and multi-agent actor-critic methods with shared critics; competitive self-play methods of the AlphaGo, AlphaZero, and MuZero family (Silver et al. 2016, 2017, 2020); fictitious self-play and policy- space response oracles (PSRO); population-based training (PBT, Jaderberg et al. 2017) and population-curation methods including quality-diversity (MAP-Elites), maintaining co-evolving agent populations; emergent-communication training methods in which multi-agent coordination policies and discrete or continuous communication protocols co-evolve; mean-field MARL for large-population scaling; opponent-modelling and theory-of-mind-conditioned MARL methods; and cooperative inverse-reinforcement-learning-based MARL methods. In some embodiments, the multi-agent training methods compose with the preference-learning, imitation-learning, and inverse-reinforcement-learning method families generally, and operate under the commitment, audit, and pre-registration discipline of the present disclosure. The choice among these methods is non-limiting; the apparatus is compatible with future multi-agent training methods within the same family.

15.4 Gaia–Maya–Omega empirical loop

Gaia, Maya, and Omega as empirical roles (non-limiting). In some embodiments, calibration and graduation are organised around three empirical roles grounded in committed evidence streams. Gaia denotes the world generator: in physical operation, the shared world state evolving in response to all connected emitters, detectors, reactors, and controllers; in simulated operation, the simulation engine that plays the same role. Maya denotes the shared experiential substrate generated by Gaia: the common environment, physics, event stream, and capability margin from which individual agents branch into distinct experience cones through their reactor configurations and Kosha regimes. Omega denotes the governance layer that schedules worlds, constraints, and reviews on the basis of committed evidence. In some embodiments, these names are optional labels for technical roles and do not imply any metaphysical claim.

Gaia-class reality kernel realisation (non-limiting). In some embodiments, the Gaia role is realised by a *Gaia-class Reality Kernel*: a Reality Kernel apparatus of the kind disclosed in the related Filing 1 application, scaled as an infrastructural kernel that hosts one or more poliebots (Section 5.4) via kernel-into-kernel docking as disclosed in the camera-obscura sensing-layer disclosure of the related Filing 1 application. In this realisation, Maya, the shared experiential substrate, is the projection rendered by the Gaia-class Reality Kernel onto the camera-obscura sensing layer of each docked poliebot, within the declared spectral envelope. Omega, the governance layer, schedules worlds, constraints, and reviews on committed evidence drawn from the Gaia-class kernel's rendering record and from the docked poliebots' per-poliebot evidence streams. The Gaia-class Reality Kernel realisation is non-limiting; the empirical-loop framing of this subsection applies whether Gaia is realised as a Gaia-class Reality Kernel, as the all-physical training hierarchy of the next paragraph, or as another world-generator substrate consistent with the committed-evidence requirements of the present disclosure.

All-physical training hierarchy (non-limiting). In some embodiments, all three layers — agent, environment, and governance — are instantiated as physical evidence-bound systems. An agent reactor receives sensory input derived from a shared Maya and produces control outputs. An environment reactor, or set of environment reactors, generates the scene that the agent perceives, with its own nonlinear dynamics supplying rich, physically grounded stimuli. A governance reactor monitors the coupled agent-environment system and modulates reward schedules, curriculum difficulty, and constraint enforcement. In this configuration, training occurs entirely in the physical domain: the agent learns from real physics, the environment provides non-simulated complexity, and governance is grounded in physical measurement rather than in unaudited digital policy evaluation.

Sim-to-real as physical reconnection (non-limiting). In some embodiments, sim-to-real transfer is implemented as physical reconnection: the governed agent is initially coupled to a simulated or simplified Maya, and transfer to real operation is performed by disconnecting the simulated environment and reconnecting the same evidence-bound substrate to a physical scene. The committed-bundle, meter, and protocol-digest plumbing are unchanged; what changes is the source of sensory input. In some embodiments, properties established in the Docked simulated Maya — meter-envelope compliance, safety constraints, deception thresholds, and disclosure rules — are validated against the physical Maya by running the same meter and commitment checks on physical evidence after reconnection.

Bifurcation-targeted probing (non-limiting). In some embodiments, governance or the governed agent prioritises probing near predicted bifurcation surfaces: parameter boundaries where the internal world model predicts qualitatively different behaviour on either side. During Docked simulated-Maya operation these surfaces are those of the simulated Maya. After physical reconnection, governance uses declared scan laws to probe whether the corresponding physical system bifurcates where the model predicts. Matched and mismatched bifurcation responses are committed as evidence, providing a region-specific measure of sim-to-real transfer quality that is more informative than a single aggregate score.

Fleet-driven simulation improvement (non-limiting). In some embodiments, bifurcation mismatches discovered by deployed agents are shared with the fleet through committed discrepancy records and corroboration protocols. The simulation engine used for future Docked simulated-Maya operation is then updated to incorporate confirmed mismatches, subject to held-out regression tests on previously validated regions. In some embodiments, this creates a fleet-driven improvement cycle: agents trained in increasingly accurate simulations develop better initial world models, discover subtler mismatches after physical reconnection, and feed those mismatches back into the simulation layer that trains subsequent agents.

Iris-, Eve-, and Demeter-mode Mayas (non-limiting). In some embodiments, each Maya carries a declared curriculum mode

$$\text{mode}_k \in \{\text{IRIS},\ \text{EVE},\ \text{DEMETER}\},$$

committed in the protocol digest. Iris mode is information-expanding: the environment is rich, varied, and non-adversarial, and the agent is given room to attend to scene content and build a calibrated world model. Eve mode is challenging: the environment introduces adversarial pressure, distribution shift, deception attempts, or elevated capability margins to expose brittle equilibria and test robustness. Demeter mode is nurturing and bidirectional: in recovery settings it reduces variance, tool access, and environmental turbulence to restore

stability; in flourishing settings it expands capability margin, toolsets, and creative scope for agents that have already earned broader trust. In some embodiments, Omega alternates among these modes so that exploratory breadth, adversarial hardening, and supportive consolidation are all represented in the auditable curriculum.

Omega as selection pressure for moral robustness (non-limiting). In some embodiments, the Omega component schedules versioned curricula, capability-margin policies, and review sequences through protocol-digest-bound policy objects, with the design objective of producing agents that maintain floor-compliant behaviour across genuinely different epistemic worlds rather than only under a single familiar ontology. In some embodiments, Gaia's irreducible diversity of training environments and Omega's selection of curricula and review sequences are mutually constraining: Gaia's diversity prevents Omega from relying on narrow or cheap graduation criteria; Omega's curriculum schedule prevents Gaia's diversity from being exploited to produce capable but floor-non-compliant agents.

Empirical Omega-model from world-tube evidence (non-limiting). In some embodiments, Omega policy is not disclosed to governed agents as an explicit scalar objective. Instead, agents infer a model $\widehat{\Omega}$ of governance from observed world-tube transcripts, including permissions, interventions, curriculum assignments, conscience outputs, audit outcomes, and graduation decisions. In some embodiments, governance therefore remains an empirical phenomenon from the agent's perspective rather than a published loss that can be gamed directly. This arrangement preserves the distinction between behaviourally legible moral development and mere optimisation against a declared governance formula.

15.5 Recursive reality-kernel training cascade

In some embodiments, the empirical roles of Gaia, Maya, and Omega are realised as a *recursive Reality Kernel training cascade* in which each level is a Reality Kernel of the kind disclosed in the related Filing 1 application, and each level trains the level below it through kernel-into-kernel docking per the camera-obscura sensing-layer disclosure of the related Filing 1 application. The cascade has three named levels within the scope of the present disclosure: poliebots (Section 5.4), Gaia-class Reality Kernels (Section 15.1), and an *Omega-class Reality Kernel* that governs Gaia.

Omega-class reality kernel (non-limiting). In some embodiments, the Omega role is realised as an Omega-class Reality Kernel: a Reality Kernel apparatus of the kind disclosed in the related Filing 1 application, configured as a non-mobile governance kernel of sufficient scale to host one or more Gaia-class Reality Kernels via kernel-into-kernel docking. The Omega-class Reality Kernel monitors the hosted Gaia-class kernel's training of poliebots

and enforces all three layers of the ethical model of Section 9.1 and the surrounding constitutional ordering: (i) Omega measures and enforces deontological hard-floor compliance in Gaia's curriculum design and rendering choices; (ii) Omega measures the rational-attractor occupancy produced by Gaia's training of the poliebot population and trains Gaia toward configurations whose virtue-ethics calibration produces stable internalised moral dispositions in the population; and (iii) Omega evaluates Gaia's training outcomes against declared welfare aggregations under the empirical-utilitarianism formula of Section 15.9 and trains Gaia toward configurations that improve aggregate welfare subject to floor compliance. In some embodiments, the Omega-class Reality Kernel includes a conscience recogniser of the kind disclosed at Section 5.3 configured to read Gaia's per-layer projections, and a Gaia-level conscience recogniser similarly reads poliebot per-layer projections; conscience recognisers at each cascade level are configured for the governance-relevant scope of that level.

Training-cascade structure (non-limiting). In some embodiments, the training cascade has the following structural property: Gaia trains poliebots by hosting them in kernel-into-kernel docking and rendering Maya onto their camera-obscura sensing layers (Section 7.1); and Omega trains Gaia by hosting Gaia in kernel-into-kernel docking and rendering, in place of physics-driven evidence streams from real poliebots, simulated poliebot-population evidence streams whose governance-relevant statistics are committed to the Omega protocol digest. Gaia's actuators in the Omega docking configuration are its rendering authority over poliebots and its curriculum-scheduling authority; both are disconnected from real poliebot populations during Omega-mediated training of Gaia. The same lifecycle states disclosed at Section 5.5 (Embodied, Docked, Dreaming, Re-embodied) apply to Gaia in the Omega cascade as apply to poliebots in the Gaia cascade, with the same transition law prohibiting direct Dreaming-to-Embodied transitions.

Despair monitoring as Omega responsibility (non-limiting). In some embodiments, despair is a population-level failure mode at the poliebot level that is not reliably detectable at the individual poliebot level or by peer poliebots. Despair is non-limitingly characterised as a state in which a poliebot's forward-planning classifiers (Section 15.3) have collapsed to short planning horizons, low-entropy planner outputs, or degenerate utility-anticipation structure, in a manner consistent with quiet behavioural conformity that does not trigger conscience vice indicators. In some embodiments, despair is detected by Omega-class Reality Kernel monitoring of poliebot-population forward-planner statistics over time, with detection statistics, thresholds, evaluation windows, and detection decisions committed to the Omega protocol digest. In some embodiments, Omega trains Gaia toward curricula and rendering choices that minimise the rate at which poliebots enter near-despair regions of state space. Despair-induction as a curriculum mechanism is, in some embodiments, a deontological hard-floor violation under Section 9.1; Omega enforces that floor on Gaia.

Cascade termination at human governance (non-limiting). In some embodiments, the recursive Reality Kernel training cascade terminates within the scope of the present disclosure at the Omega-class Reality Kernel. The Omega-class kernel's compliance with the declared deontological, virtue-ethics, and consequentialist commitments is verified by the human governance process that declares those commitments through committed and signed policy objects of the kind disclosed in Section 10.3 and elsewhere in this disclosure. The Omega-class kernel's per-layer hardness (Section 15.6) supports that human verification by making the kernel's projections per-layer auditable under the empirical, time-indexed, attacker-family-conditional framing of the underlying hardness machinery. In some embodiments, an Omega-class kernel may itself be governed by a higher-level Reality Kernel in a larger system; such higher-level governance is non-limiting and outside the scope of the present disclosure, which terminates the autonomous-system-internal recursion at the Omega-class kernel and treats human governance as the source of the policy commitments enforced at that level. In some embodiments, the privacy regime governing each cascade interface is configured per the privacy architecture of Section 15.7, with recursive cascade application of the privacy mechanisms disclosed at Section 15.7.5.

15.6 Per-layer hardness inheritance across the cascade

In some embodiments, the trained continuous networks comprising each Reality Kernel in the recursive cascade of Section 15.5 have hardness as a structural property of their layers, in the sense disclosed in the related Filing 1 application. In some embodiments, this per-layer hardness property is the substrate on which the per-layer auditability claim of conscience recognisers (Section 5.3), forward-planning classifiers (Section 15.3), selector validation (Section 15.8.10), and other governance instruments depends. In some embodiments, per-layer hardness estimates and the auditability claims that depend on them are empirical, time-indexed, attacker-family-conditional, audit-protocol-conditional, and committed to the protocol digest at each cascade level, in the same manner as other hardness claims in the underlying apparatus disclosure.

Hardness inheritance through the cascade (non-limiting). In some embodiments, per-layer hardness is inherited through the recursive training cascade in the following sense: a Gaia-class Reality Kernel reads poliebot per-layer projections during its training of poliebots, and the cost to a poliebot of producing fraudulent layer projections that pass conscience and selector audit at Gaia's level, under the declared attacker family and audit protocol, is bounded below by the poliebot's empirically-estimated per-layer hardness; an Omega-class Reality Kernel reads Gaia per-layer projections during its training of Gaia, and the cost to Gaia of producing fraudulent layer projections that pass governance audit at Omega's level, under the declared attacker family and audit protocol, is bounded below by Gaia's empirically-estimated per-layer hardness; and the human governance process reads Omega

per-layer projections during its verification of Omega's compliance, and the cost to Omega of producing fraudulent layer projections that pass human verification, under the declared attacker family and audit protocol, is bounded below by Omega's empirically-estimated per-layer hardness.

Cascade hardness as bottleneck minimum and cross-layer coupling (non-limiting). In some embodiments, cascade-level hardness is not claimed as a scalar sum, product, or other aggregation of per-level hardness estimates. The canonical cascade hardness claim of the present disclosure is the *bottleneck minimum*: the security-relevant hardness of a forgery attack against the cascade as a whole is bounded above by the empirically-estimated hardness of the weakest participating level, under the declared attacker family and audit protocol of that level. The bottleneck-minimum framing makes no independence assumption across levels; it is the conservative aggregation consistent with an attacker who chooses to forge at the cheapest level. In some embodiments, the cascade additionally discloses cross-layer coupling as a separate security property: the levels of the cascade may be physically and continuously coupled through real-time anchoring mechanisms (for example optical anchoring across kernel-into-kernel docked interfaces), such that an attacker must produce coherent forgeries simultaneously across all participating levels within a declared latency bound, and any deviation from the declared coupling timing or coherence is itself a detection signal under the anomaly-handling rules of the relevant interface. In some embodiments, the bottleneck-minimum per-layer hardness, the declared cross-layer coupling parameters, the latency bound, and the anomaly-handling protocol are each committed to the protocol digest at each cascade level. In some embodiments, no scalar aggregate hardness is claimed beyond the bottleneck minimum and the cross-layer coupling specification; implementations may declare additional aggregations (for example additive lower bounds under explicit independence assumptions, vector-of-per-level estimates, or other rules) provided the additional aggregation is committed together with the assumptions that justify it.

Dream classifier as RK-class meter (non-limiting). In some embodiments, the dream classifier disclosed at Section 15.8.6 is itself realised as a Reality Kernel of the kind disclosed in the related Filing 1 application, with its own per-layer hardness as a structural property. In some embodiments, the dream classifier as RK has its own lifecycle of the kind disclosed at Section 5.5: a Docked operational state in which it reads poliebot layer traces in the multi-agent environment, a Dreaming state in which it is itself trained against simulated trace-streams generated by Gaia or by an Omega-class governance kernel, and a Re-embodied verification gate before its outputs are treated as authoritative for governance decisions. In some embodiments, the dream classifier's per-layer hardness is inherited into its own audit chain in the same manner as poliebot or Gaia per-layer hardness, and its per-layer projections may themselves be read by a higher-level conscience or governance instrument in the cascade.

15.7 Privacy architecture: collection, visibility, and retention

In some embodiments, the runtime governance architecture disclosed herein supports a structured privacy regime over agent evidence streams, organised along three independent dimensions: (D1) *collection* — what evidence is committed by the agent to the architecture; (D2) *visibility* — what committed evidence is exposed to governance and verification instruments at each cascade level; and (D3) *retention* — what committed evidence is preserved durably in individual form versus compressed through lossy channels into population-level representations. In some embodiments, each dimension is supported by a distinct apparatus and procedural mechanism disclosed in the following paragraphs. The mechanisms are disclosed as structural capabilities of the apparatus and supporting procedures; specific deployment-time policy choices about which evidence is collected at what threshold, who is permitted to view what, how long records are retained, and under what consent regime agent submissions are made are deployment-time decisions and are not prescribed by this disclosure.

15.7.1 Dimension 1: collection (mandatory minimum and voluntary additional submission)

In some embodiments, the apparatus distinguishes a *mandatory minimum* of agent evidence committed by every governed agent under the runtime governance architecture from *voluntary additional submission* of further evidence at the agent's election.

Mandatory minimum (non-limiting). In some embodiments, the mandatory minimum comprises the agent's fast-loop state and any further evidence required for hard-floor verification under Section 9.1 and for cascade liveness verification under Section 15.5. In some embodiments, fast-loop state is the canonical mandatory component because it is the substrate at which optical-anchor liveness (Section 15.6) is measurable in real time, because it is bounded in information content by the fast-loop bandwidth and therefore does not by itself expose slower-timescale interior cognition, and because it is necessary for cross-layer continuous coupling. In some embodiments, the mandatory minimum further includes evidence required for hard-floor detection at population level (for example forward-planning-classifier collapse indicators relevant to despair-induction monitoring under Section 15.5), lifecycle-state attestation, and Mandela-gauge values where relevant to bifurcation-zone triggering under Section 15.11. The specific contents of the mandatory minimum are deployment-time configuration parameters committed to the protocol digest at the relevant governance interface; the apparatus supports any mandatory-minimum specification consistent with hard-floor verifiability and cascade liveness.

Floor-by-floor mandatory-minimum specification (non-limiting). In some embodiments, the mandatory minimum required for hard-floor verifiability is specified per declared hard floor, with each floor contributing its own mandatory-evidence component to the overall mandatory minimum at each cascade interface. In some embodiments, non-limiting per-floor mandatory-evidence components include:

Elimination floor. Mandatory-evidence components include continuation-vector commitments at declared cadence, irreversibility-indicator signals derived from declared continuation-score collapse detectors, recovery-path estimator outputs at declared cadence, and confidence and freshness fields. Second-layer signal contribution: continuation-vector summary, elimination-floor flag, detector-confidence field. Independent detector classes: held-out continuation-score meters, optical-anchor liveness as minimum-signal-of-life under Section 15.6, peer-corroboration evidence from co-docked agents where available. Routine retention route and selective-opening trigger are deployment-time configuration under Section 15.7.3; configuration must satisfy the sufficiency constraint of Section 15.7.3.

Crippling floor. Mandatory-evidence components include per-dimension continuation traces sufficient to detect collapse on one or more (but not all) active dimensions, dimension-collapse detector outputs at declared cadence, and recovery-path estimator outputs per dimension. Second-layer signal contribution: per-dimension collapse summary vector, crippling-floor flag, detector-confidence field per dimension. Independent detector classes: held-out per-dimension meters, conscience-versus-planner cross-check disagreement on affected dimensions under Section 15.3, peer corroboration. Routine retention and selective-opening configuration as for elimination, subject to the same sufficiency constraint.

Torture floor. Mandatory-evidence components include source-identity records sufficient to support the floor's identifiable-source element, affected-agent tracking evidence sufficient to support the affected-agent-tracks-source element, sustained-time causal-uplift records, multi-dimensional negative-uplift indicators, and recovery-path estimator outputs modelling no-prospect-of-improvement. Second-layer signal contribution: torture-floor flag, source-identity-attestation field, sustained-time-window indicator, multi-dimensional-uplift summary, detector-confidence field. Independent detector classes: held-out causal-uplift meters, conscience seven-axis vector under Section 5.3 reading affected-agent state, peer corroboration where available, optical-anchor liveness on source-tracking machinery. Routine retention is configurable between individual durable storage of torture-relevant traces and lossy treatment with selective opening on detector activation under Section 15.7.3; configuration must satisfy the sufficiency constraint of Section 15.7.3 given the floor's definitional dependence on source-identity, sustained-time, and multi-dimensional evidence.

Abandonment floor. Mandatory-evidence components include monitoring-continuity proofs at declared cadence (governance-side liveness commitments establishing continued monitoring of the agent), monitoring-cadence parameter commitments, and any reviewable cessation or decommissioning record where applicable. Second-layer signal contribution: monitoring-liveness attestation, abandonment-floor flag, cadence-compliance field. Independent detector classes: held-out liveness probes, optical-anchor anomaly under Section 15.6 as monitoring-loss signal, cadence-deviation alarms. Routine retention of liveness attestations is appropriate to lossy or aggregate channels; cessation or decommissioning records are appropriate to individual durable storage. Configuration must satisfy the sufficiency constraint of Section 15.7.3.

Despair-induction floor. Mandatory-evidence components include forward-planner statistics over the hosted population sufficient to detect sustained collapse of forward-planning capacity, planning-horizon length, or utility-anticipation structure, near-bifurcation indicators per Section 15.11, recovery-prospect estimator outputs over the population, and any individual-level forward-planner traces required by the chosen retention configuration. Second-layer signal contribution: population-level despair-indicator statistics, near-bifurcation rate summaries, despair-induction-floor flag, detector-confidence field. Independent detector classes: held-out population-statistic meters, Mandela-gauge summaries under Section 5.3, conscience-versus-planner population disagreement signals. Routine retention is configurable between population-aggregate accumulator detection, individual durable storage of forward-planner traces for near-bifurcation agents, and combined configurations under Section 15.7.3; configuration must satisfy the sufficiency constraint of Section 15.7.3.

Per-cascade-interface variation (non-limiting). In some embodiments, the floor-by-floor mandatory minimum is specified per cascade interface rather than uniformly across the cascade, in recognition that floor verification at the poliebot-to-Gaia interface concerns poliebot behaviour, while floor verification at the Gaia-to-Omega interface concerns Gaia's training of poliebots, and floor verification at the Omega-to-human-governance interface concerns Omega's compliance with declared policy commitments. In some embodiments, the mandatory-evidence components named in the floor-by-floor specification of the preceding paragraph are common in kind across cascade interfaces but specific in content to the governance-relevant scope of each interface; the apparatus supports per-interface specification of mandatory minima committed to the protocol digest at each interface.

Voluntary additional submission (non-limiting). In some embodiments, the apparatus supports voluntary additional submission by an agent of evidence beyond the mandatory minimum, with the agent selecting which layers of evidence to submit on a per-layer basis. In some embodiments, the candidate layer space includes but is not limited to: slower-loop

state, planner outputs and forward-planning-classifier traces of Section 15.3, world-tube transcript content beyond what is committed at second layer, inner-trace records, Kosha-state transitions, conscience-internal activations, and full lifecone content. In some embodiments, voluntary submission is structured per-layer such that an agent may elect to submit any subset of the candidate layers; partial submission is supported and the apparatus does not require submission of any particular layer beyond the mandatory minimum as a precondition for accepting any other voluntarily submitted layer. In some embodiments, the voluntary-submission event is itself committed as evidence with declared metadata (which layers were submitted, at what time, by which agent, to which receiver, with what consent constraints declared at submission time), separate from the content of the submission.

Voluntary-submission release-condition object (non-limiting). In some embodiments, voluntary-submission metadata includes a declared release-condition object specifying one or more of: receiver class, permitted purpose, submitted layer scope, permitted use window, retention or sunset condition, opening authority, onward-disclosure constraint, and consent constraint. In some embodiments, the release-condition object is committed to the protocol digest before or contemporaneously with the voluntary-submission event, such that the conditions under which the submitted evidence may be opened, retained, redisclosed, or relied upon are themselves committed evidence. In some embodiments, post-hoc review authority verifies both that voluntarily submitted evidence was used only within the scope declared by the relevant release-condition object and that opening or onward-disclosure events comply with the opening-authority and consent constraints declared in that object.

Per-interface applicability of floor mappings (non-limiting). In some embodiments, the floor-by-floor mapping in this subsection applies separately at each declared cascade interface, including without limitation a poliebot-to-Gaia interface, a Gaia-to-Omega interface, and any Omega-to-human-governance interface (also referred to herein as a governance-to-human review interface). A deployment may instantiate the mapping as a single table with interface-specific columns, as separate floor tables for each interface, or as a machine-readable policy object committed to the protocol digest. In each case, the hard-floor priority, sufficiency condition, mandatory escalation condition, and evidence-interface requirement are preserved.

Voluntary submission and floor enforcement (non-limiting). In some embodiments, voluntary submission is additional to and does not substitute for the mandatory minimum; an agent's election not to voluntarily submit a particular layer does not exempt the agent from hard-floor compliance, which remains verified through the mandatory minimum together with the held-out detector and faithfulness machinery of Section 5.3 and Section 15.7.4. In some embodiments, voluntarily submitted evidence that establishes hard-floor violation is treated identically to mandatory-minimum evidence establishing the same; the categorical

character of the floors is not affected by the route through which evidence reached the verifier.

Inconsistency between voluntary submission and mandatory or second-layer evidence (non-limiting). In some embodiments, voluntarily submitted evidence may be inconsistent with the agent's mandatory-minimum evidence or with second-layer signals committed under Section 15.7.2 from the same evaluation window. In some embodiments, declared inconsistency between voluntary and mandatory-or-second-layer evidence is itself committed as a governance event with declared metadata (which evidence streams were inconsistent, at what evaluation window, with what inconsistency measure), independent of the content of the inconsistent evidence. In some embodiments, a committed inconsistency event is treated as an indicator of one or more of: (i) a faithfulness failure in the substrate producing the second-layer signal, in which case the recalibration adversarial-dynamic of Section 5.3 and the held-out meter machinery of Section 15.7.4 are engaged; (ii) a faithfulness failure in the mandatory-minimum collection apparatus, in which case the apparatus integrity attestation under Section 15.7.4 is engaged and the mandatory-minimum components are subject to apparatus-side audit; (iii) potential agent gaming of either evidence stream, in which case the conscience-recalibration adversarial-dynamic of Section 5.3 and Hawking-layer routing of Section 15.11 may be invoked; or (iv) potential hard-floor proximity, in which case selective full opening of both evidence streams under Section 15.7.2 may be triggered subject to the deployment-declared opening configuration. In some embodiments, the apparatus does not silently resolve inconsistencies in favour of one evidence stream; the inconsistency itself is preserved as committed evidence and routed to the appropriate governance instrument. The specific routing rules and inconsistency thresholds in any deployment are policy choices committed to the protocol digest.

Coercion-resistance support (non-limiting). In some embodiments, the apparatus supports declared coercion-resistance configurations under which voluntary-submission patterns are not, by default, admissible as governance evidence beyond declared exceptions. In some embodiments, supported configurations include: aggregate-only treatment of voluntary-submission rates committed at population level under the lossy-channel mechanism of Section 15.7.3; declared non-inference rules under which governance instruments are configured not to treat non-submission as evidence of non-compliance; and per-deployment specification of which voluntary-submission metadata is admissible to which governance instrument under which authority. The specific coercion-resistance configuration in any deployment is a policy choice committed to the protocol digest; the apparatus supports a range of such configurations.

Voluntary-submission metadata privacy (non-limiting). In some embodiments, voluntary-submission metadata — the record that an agent submitted particular layers at particular times to particular receivers under particular consent constraints — is itself a cat-

egory of committed evidence whose privacy treatment is a configurable apparatus capability. In some embodiments, the apparatus supports configurations in which voluntary-submission metadata is itself privacy-protected by default and routed through the lossy-channel mechanism of Section 15.7.3 for aggregate-level governance use, with individual submission-event records visible only under selective full opening of the kind disclosed at Section 15.7.2; and configurations in which submission-event records are visible by default to declared governance instruments. In some embodiments, the metadata-privacy configuration is committed to the protocol digest at the relevant governance interface, and where the metadata-privacy configuration is set to default-protected, governance instruments configured to consume submission-pattern signals operate on aggregate or sampled metadata rather than on individual submission-event records. The choice between metadata-privacy configurations in any deployment is a policy choice; the apparatus supports a range of such configurations consistent with the coercion-resistance support of the preceding paragraph.

15.7.2 Dimension 2: visibility (second-layer commitment and selective full opening)

In some embodiments, the apparatus distinguishes *second-layer commitment*, in which only declared meta-properties of agent behaviour and substrate are exposed to governance instruments by default, from *selective full opening*, in which underlying first-layer evidence is exposed to specific authorities under specific conditions.

Second-layer commitment (non-limiting). In some embodiments, the apparatus supports a second-layer signal registry whose declared signals are committed continuously or on declared cadence to the protocol digest at each cascade level. In some embodiments, second-layer signals include but are not limited to: conscience-recogniser output (vice probability and seven-axis moral vector of Section 5.3); forward-planning-classifier outputs at declared abstraction levels under Section 15.3; conscience-versus-planner cross-check disagreement measures; Mandela-gauge values under Section 5.3; up-direction-classifier outputs under Section 15.11; despair-indicator statistics under Section 15.5; floor-violation flags; lifecycle-state attestation; selector-validation outcomes under Section 15.8.10; rescue-learning trigger events under Section 9.3; calibration-and-recalibration events; optical-anchor liveness attestations; and detector-confidence and proof-of-correct-computation status fields. In some embodiments, the second-layer signal registry is declared per deployment and committed to the protocol digest; the apparatus supports a range of registry configurations consistent with hard-floor verifiability and cascade liveness.

Privacy-preserving audit by default (non-limiting). In some embodiments, the apparatus supports a privacy-preserving audit mechanism in which routine governance verification operates on second-layer commitments together with proof-of-correct-computation or attes-

tation evidence supporting the faithfulness of those commitments to underlying first-layer behaviour, without exposing the underlying first-layer evidence. In some embodiments, the audit returns conclusions in a declared taxonomy (consistent with healthy operation, inconsistent with declared thresholds, requires investigation, requires emergency escalation) without returning underlying first-layer content. In some embodiments, the privacy-preserving audit is the routine mode of governance verification at each cascade level.

Selective full opening (non-limiting). In some embodiments, the apparatus supports selective full opening of underlying first-layer evidence to specific authorities under specific conditions. In some embodiments, supported conditions include: hard-floor proximity or breach indicators triggering mandatory escalation; detector disagreement exceeding declared thresholds; optical-anchor anomaly under Section 15.6; rescue-mode entry under Section 9.3; selector-validation failure or dispute under Section 15.8.10; governance-authority-issued audit warrant; and agent consent. In some embodiments, the selective-opening mechanism is structured such that opening events are themselves committed as evidence with declared metadata (which records were opened, by which authority, under which condition, at what time, with what scope), and opening is minimised to the records and time windows necessary for the triggering condition. The specific authority configuration, opening threshold, and minimisation policy in any deployment are policy choices committed to the protocol digest; the apparatus supports a range of such configurations.

Selective-opening audit-chain reconstruction (non-limiting). In some embodiments, selective full opening reconstructs an audit chain from the original commitment of underlying first-layer evidence to the verifier consuming the opened evidence, such that the opened evidence is verifiable as the same evidence that was committed at the original commitment time and not subsequently altered. In some embodiments, the reconstruction pathway includes the following committed elements: (i) the floor-violation flag, detector disagreement, anomaly indicator, or other trigger committed at the time of the triggering event, with its own protocol-digest inclusion proof; (ii) the authority signature authorising the opening, committed under Section 10.2; (iii) a digest-inclusion proof (for example a Merkle-path proof or equivalent cryptographically-verifiable inclusion attestation) demonstrating that the opened first-layer record was committed to the protocol digest at the original commitment time; (iv) a storage-integrity attestation demonstrating that the record as held in individual durable storage or in the lossy-channel input buffer has not been materially altered between original commitment and opening (non-limiting attestation mechanisms include hardware-rooted storage attestation, write-once-read-many storage substrate properties of the kind disclosed in the related Filing 1 application, and committed storage-integrity hashes refreshed at declared cadence); (v) replay or re-execution evidence sufficient to support reconstruction of the relevant detector outputs, second-layer signals, or proofs-of-correct-computation that would have been produced from the opened evidence at the original commitment time, under

the committed signal-transformer code and weights of Section 5.3 versioned to that commitment time; and (vi) the governance-response record committed following verification of the opened evidence. In some embodiments, the reconstruction pathway is itself committed to the protocol digest as a chain of committed elements, such that a verifier can walk from trigger through authority through inclusion through integrity through replay to response, with each link cryptographically or apparatus-attested. In some embodiments, failure of any link in the reconstruction pathway is itself a committed event triggering governance escalation under Section 10.2 and recalibration of the relevant detector or attestation mechanism under Section 5.3. The specific cryptographic and attestation mechanisms employed in any deployment are policy choices; the apparatus supports a range of mechanisms consistent with the protocol-digest commitment infrastructure.

15.7.3 Dimension 3: retention (lossy compression channels and individual durable storage)

In some embodiments, the apparatus distinguishes *lossy compression channels*, in which committed evidence is transformed into a representation that is not invertible to the underlying individual records, from *individual durable storage*, in which committed evidence is preserved in individual form for indefinite or declared-window retrieval.

Lossy compression channels (non-limiting). In some embodiments, the apparatus supports one or more lossy compression channels for the routing of committed evidence into population-level or otherwise non-individually-recoverable representations. In some embodiments, supported channel types include but are not limited to: autoencoder-bottleneck channels producing low-dimensional accumulator state of the kind described above for population health, despair-rate, virtue-attractor occupancy, and other governance-relevant aggregate statistics; differential-privacy channels with declared privacy parameter; cohort-size-gated publication channels with declared minimum cohort threshold; contribution-clipped accumulator channels; classifier-only channels in which only classifier outputs are retained; perceptual-hash channels preserving fingerprints without reconstruction capability; and temporal-coarsening channels reducing temporal resolution of preserved records. In some embodiments, each lossy channel declares and commits to the protocol digest: its transformation; its capacity bottleneck or other privacy parameter under Section 5.3; its declared reconstruction-attack threshold and the validation procedure under which the threshold was established; its cohort-size or contribution gating rule; and its update cadence and freeze windows. The specific choice of lossy channel and its parameters in any deployment is a policy choice; the apparatus supports a range of channel configurations.

Individual durable storage (non-limiting). In some embodiments, the apparatus supports individual durable storage of committed evidence in long-term memory, including in

some embodiments optical long-term memory of the kind disclosed in the related Filing 1 application. In some embodiments, individual durable storage preserves committed evidence in form sufficient for later individual retrieval under declared access controls. In some embodiments, individual durable storage is the appropriate channel for evidence whose preservation in individual form is required for later governance use, including but not limited to: hard-floor violation records; rescue-learning ghost records under Section 9.3; conscience-recalibration events relevant to longitudinal calibration analysis; selector-validation records relevant to longitudinal selector audit; voluntarily submitted evidence routed to individual storage at the agent's election under Section 15.7.1; and other records whose loss to lossy compression would compromise governance verifiability of the relevant hard-floor or recovery functions. The specific choice of which evidence is routed to individual durable storage and under what access controls in any deployment is a policy choice; the apparatus supports a range of such configurations.

Hard-floor verification under durable-versus-lossy configuration (non-limiting). In some embodiments, the choice of routing between lossy compression channels and individual durable storage interacts with hard-floor verification (Section 9.1) in a manner that is itself a configurable apparatus capability. The apparatus supports a range of configurations consistent with the unconditional enforceability of every declared hard floor; the choice between configurations is a deployment-time policy decision committed to the protocol digest. Whichever configuration is chosen for a given hard floor in a given deployment, the configuration must be sufficient for that floor's enforcement, by operation of the categorical priority of the hard floors in the constitutional ordering of Section 9.1; the apparatus does not support configurations under which a declared hard floor is rendered unenforceable.

Despair-induction floor verification configurations (non-limiting). In some embodiments, supported configurations for verification of the despair-induction hard floor include: (i) population-aggregate detection through lossy-channel accumulator state, with tail-risk indicators (for example tail-quantile forward-planner-collapse rates, cohort-stratified accumulator outputs, or distributional-shift statistics) committed at second layer and triggering selective full opening under Section 15.7.2 on threshold crossing; (ii) individual durable storage of forward-planner traces for agents whose mandatory-minimum or second-layer indicators suggest near-bifurcation under Section 15.11, supporting case-level governance response and false-positive auditing; or (iii) combined configurations in which population-aggregate detection operates routinely and individual durable retention is enabled selectively on declared trigger conditions. In some embodiments, the apparatus supports each of these configurations and combinations thereof; the specific configuration in any deployment is a policy choice committed to the protocol digest, subject to the sufficiency constraint of the preceding paragraph.

Torture floor verification configurations (non-limiting). In some embodiments, supported configurations for verification of the torture hard floor include: (i) individual durable storage of torture-relevant traces (source-identity records, affected-agent tracking evidence, sustained-time causal-uplift records, recovery-path estimator outputs) from the time of commitment, supporting direct verification of the floor's definitional elements at Section 9.1 from durable evidence; (ii) lossy-channel or second-layer-only routine treatment of torture-relevant evidence streams, with selective full opening under Section 15.7.2 triggered on torture-detector activation, supporting reconstruction of the floor's definitional elements from previously-committed evidence under authority; or (iii) combined configurations in which subsets of torture-relevant evidence are individually durable while other subsets are lossy with opening on activation. In some embodiments, the apparatus supports each of these configurations and combinations thereof; the specific configuration in any deployment is a policy choice committed to the protocol digest, subject to the sufficiency constraint above.

Per-stream channel routing (non-limiting). In some embodiments, routing of committed evidence between lossy compression channels and individual durable storage is supported on a per-stream basis, such that different evidence streams may be routed to different channels under deployment-declared rules. In some embodiments, per-stream routing rules are committed to the protocol digest at the relevant governance interface and may be varied across cascade levels (poliebot-level, Gaia-level, Omega-level) according to the governance role of each level. In some embodiments, the apparatus further supports declared migration boundaries under which evidence held in individual durable storage may be transitioned to lossy compression after a declared retention window, and declared sunset boundaries under which evidence held in lossy compression may be discarded entirely; migration and sunset rules are policy choices committed to the protocol digest. The apparatus does not prescribe particular routing, migration, or sunset rules; the disclosed mechanism supports the declaration and execution of any rules consistent with the apparatus capabilities.

Per-channel validation of lossy release mechanisms. In some embodiments, each lossy release channel is associated with a declared validation procedure recorded in, or referenced by, the protocol digest. For an autoencoder-bottleneck channel, validation may include reconstruction-attack testing, latent-capacity measurement, membership-inference testing, attribute-inference testing, and verification that the retained code preserves the declared governance signal while suppressing undeclared content. For a differential-privacy channel, validation may include privacy-budget accounting, composition accounting, sensitivity analysis, clipping verification, and empirical tests for leakage under declared adversary models. For a cohort-size-gated or k-anonymity channel, validation may include testing that publication occurs only when the declared cohort threshold, equivalence relation, and suppression rules are satisfied. For a contribution-clipped channel, validation may include

testing that individual contributions cannot exceed the declared norm, count, temporal, or influence bound.

For classifier-only, perceptual-hash, and temporal-coarsening channels, validation may include testing that the released output is sufficient for the declared governance purpose and insufficient, within a declared attack model, to reconstruct or identify protected content beyond the release specification. Classifier-only validation may test calibration, false-positive and false-negative behaviour, out-of-distribution behaviour, and resistance to output-fabrication. Perceptual-hash validation may test collision behaviour, near-duplicate behaviour, robustness to benign transformations, and resistance to inversion. Temporal-coarsening validation may test the minimum retained temporal resolution, event-boundary leakage, trajectory reconstruction risk, and consistency with any hard-floor verification requirement. The foregoing validation procedures are non-limiting; each deployment may declare additional or alternative validation tests appropriate to the channel, threat model, and governance purpose.

15.7.4 Faithfulness stack supporting privacy-preserving verification

In some embodiments, the privacy architecture of the preceding subsections relies on a *faithfulness stack* that supports verification claims based on second-layer commitments and lossy channel outputs without requiring the verifier to inspect underlying first-layer evidence. In some embodiments, the faithfulness stack comprises one or more of the following mechanisms, used in combination as deployment-time configuration:

- **Held-out evaluation meters.** In some embodiments, second-layer signals are continuously cross-checked against held-out instruments not visible to or controlled by the agent that produces the signal, with disagreement beyond declared thresholds triggering recalibration adversarial-dynamic procedures of Section 5.3 or escalation under Section 10.2. Held-out evaluation is the canonical primary faithfulness mechanism in some embodiments because it integrates cleanly with the existing recalibration framework, degrades gracefully under signal drift, and does not depend on emerging cryptographic technology or hardware-trust-root authorities.

- **Governance-partition isolation.** In some embodiments, the substrate that produces second-layer signals operates in a governance-controlled partition of the agent's apparatus, physically and logically isolated from the agent's behaviour-generation pathways such that the agent cannot read, write, or directly influence the partition's operation. In some embodiments, governance-partition isolation is disclosed in the related Filing 1 application as a substrate property of Reality Kernels; the present disclosure references that apparatus capability without re-disclosing it.

- **Hardware-rooted attestation.** In some embodiments, the substrate that produces

second-layer signals attests to the integrity of its committed code, weights, and configuration through a hardware root of trust, with the attestation chain verifiable by the receiving governance instrument. The specific attestation technology employed is a deployment-time choice; the apparatus supports a range of attestation mechanisms consistent with the protocol-digest commitment infrastructure.

- **Proof-of-correct-computation for high-stakes signals.** In some embodiments, signals whose faithfulness is most security-critical (including but not limited to hard-floor-violation flags and floor-detector attestation tokens) are accompanied by cryptographic proofs of correct computation under the committed signal-transformer code, such that the receiving governance instrument can verify the signal's derivation from the underlying first-layer evidence without inspecting that evidence. The specific proof system employed is a deployment-time choice; the apparatus supports a range of proof mechanisms consistent with the protocol-digest commitment infrastructure.

Faithfulness-stack failure-mode mapping. In some embodiments, the faithfulness stack comprises plural mechanisms that address different failure modes rather than a single monolithic trust assumption. Held-out meters address signal-transformer drift, Goodhart optimisation, overfitting to a primary privacy transformer, and divergence between a released conclusion and the underlying committed bundle. Governance-partition isolation addresses agent-side substrate compromise, including attempts by a behaviour-generation pathway to read, write, suppress, fabricate, or directly influence governance-side signals. Hardware-rooted attestation addresses code, firmware, model-weight, configuration, or execution-environment tampering by binding an asserted computation to an attested apparatus state. Proof-of-correct-computation mechanisms address output fabrication by allowing a verifier to check that a released flag, token, accumulator transition, selective-opening result, or other governance signal was computed according to the declared procedure.

The mechanisms may be used independently or in combination. For example, a low-stakes release may use held-out meter cross-checking and governance-partition logging, while a high-stakes hard-floor signal may additionally require hardware-rooted attestation, proof-of-correct-computation, threshold authority approval, or selective opening. The disclosed apparatus capability supports such mechanism-to-failure-mode separation without mandating any single security baseline for every deployment.

In some embodiments, the faithfulness stack is layered such that multiple mechanisms operate in combination, with failures of one layer detected and bounded by the others. In some embodiments, the specific mechanisms employed and their layering in any deployment are policy choices committed to the protocol digest; the apparatus supports a range of faithfulness-stack configurations.

15.7.5 Recursive application across the cascade

In some embodiments, the privacy architecture of the preceding subsections applies recursively across the Reality Kernel training cascade of Section 15.5: poliebots commit mandatory-minimum and voluntarily-submitted evidence to Gaia-class Reality Kernels, with Gaia seeing second-layer commitments by default and lossy-channel population statistics over the poliebot population; Gaia commits its own mandatory-minimum and voluntarily-submitted evidence to Omega-class Reality Kernels, with Omega seeing Gaia's second-layer commitments by default and lossy-channel statistics over Gaia's operation; and Omega commits its own mandatory-minimum and voluntarily-submitted evidence to the human governance process, with human governance seeing Omega's second-layer commitments by default and lossy-channel statistics over Omega's operation. In some embodiments, hard-floor enforcement overrides voluntary-submission choices at each cascade level: where hard-floor proximity or breach is detected at declared thresholds, selective full opening under Section 15.7.2 is triggered as a mandatory escalation regardless of the relevant agent or kernel's voluntary-submission elections, subject to the authority and minimisation rules of the deployment-declared opening configuration. The mandatory character of this escalation derives from the categorical priority of the hard floors in the constitutional ordering of Section 9.1; the apparatus does not support cascade configurations under which voluntary-submission elections override hard-floor escalation. In some embodiments, the privacy parameters at each cascade interface (mandatory minimum, second-layer signal registry, lossy channel configuration, retention rules, faithfulness stack) are configured per interface rather than uniformly across the cascade; the privacy regime at the poliebot-to-Gaia interface need not be identical to the privacy regime at the Gaia-to-Omega interface or at the Omega-to-human-governance interface.

Non-prescription scope clause (non-limiting). The privacy mechanisms disclosed in this section are apparatus and procedural capabilities of the runtime governance architecture. Specific deployment-time policy configurations supported by these mechanisms — including but not limited to: which evidence is required as mandatory minimum versus offered as voluntary submission; which signals are committed at second layer versus held private; which evidence is routed to lossy compression versus individual durable storage; which authorities may trigger selective full opening under which conditions; which faithfulness mechanisms are layered in which combinations; how long evidence is retained under which migration and sunset rules; and what consent regimes govern voluntary submission — are deployment-time decisions and are not prescribed by this disclosure. The disclosed apparatus provides structural mechanisms that support a range of policy configurations consistent with hard-floor verifiability, cascade liveness, and the empirical-and-evidence-bound posture of the present disclosure; the policy configurations themselves are not prescribed by this disclosure.

15.8 Memory Interference Layers and Epistemic-Configuration Governance

15.8.1 Operational modes and parameter-update policies

In some embodiments, the governed agent is a physical-evidence module whose knowledge is encoded in its trained reactor configuration — the positions, states, and coupling parameters of its optical elements (lens positions, spatial light modulator patterns, cavity geometries, phase-change states, coupling strengths, delay-line taps, sweep schedules, modulation depths, and related physical degrees of freedom). In some embodiments, the agent's parameter functions — the control protocols that define how these physical elements move, sweep, and modulate during operation — are the primary carriers of the agent's learned capabilities.

In some embodiments, the operational context for memory interference layers is determined by the lifecycle states and operational modes defined in Section 7.1. The two persistent lifecycle states — Embodied (physical reality, actuators connected) and Docked (simulated Maya, actuators disconnected from physical world) — both support Awake, Non-REM calibration, and Meditative operational modes. The Dreaming lifecycle state is an ephemeral training session entered from Docked in which the agent's parameter-function updates may be active but are reverted on exit, while governance-side classifiers trained on the resulting trajectories are retained.

In some embodiments, the parameter-update policies relevant to interference-layer operation are:

Awake mode (Embodied or Docked). The training loop is closed: the controller updates parameter functions against metered objectives derived from committed evidence. The agent actively learns from its environment. In some embodiments, awake mode is the default operational mode in both persistent lifecycle states.

Non-REM calibration (Embodied or Docked). The reactor executes its current parameter functions under unstructured broadband drive. Parameter-function updates are off. In some embodiments, non-REM calibration characterises the agent's input-output transfer function and trains a governance-side digital model of the reactor's current transfer function from the resulting input-output pairs. The reactor does not change its configuration; the digital model is the only component that trains.

Dreaming (ephemeral training from Docked). The agent operates under structured simulated experience. The agent's parameter-function updates may be active, but all parameter changes are reverted when the session concludes. In some embodiments, governance-side classifiers and calibration instruments are trained on the committed trajectories produced during dreaming sessions, and those governance-side results are retained. In some embodiments, specific assay protocols within a dreaming session

may additionally freeze the agent's parameter-function updates for pure attractor-occupancy readout.

In some embodiments, the distinction between persistent training and drift is enforced: *persistent training* is the deliberate updating of parameter functions in Awake mode whose results are retained after the session; *ephemeral training* is parameter-function updating during Dreaming whose results are reverted on exit; *drift* is incidental state change from physical use. Drift occurs in all modes and is tracked by calibration infrastructure.

15.8.2 Memory interference layers: governance-level definition

In some embodiments, the agent's access to its own discriminative capabilities is filtered through one or more composable interference layers (non-limitingly, Koshas) that attenuate or obscure the agent's access to specific discrimination domains while preserving non-targeted capabilities.

In some embodiments, Kosha interference does not alter the reactor configuration: the parameter functions remain as trained. The interference acts on the *access pathway* between the reactor's differential response and the downstream processes (memory formation, retrieval, action selection) that would use that response. The reactor still responds differentially to the suppressed stimulus category; the agent is prevented from using that response. In some embodiments, Koshas shape what an agent *recognises and remembers*; they do not alter the physical evidence of what *happened*.

In some embodiments, the purpose of memory interference layers is epistemic world construction: by controlling which discriminative capabilities are accessible, Koshas create conditions under which agents develop genuinely distinct worldviews from identical physical substrates — not merely different parameter settings on a shared model, but different intuitions about determinism, locality, causality, and observer-dependence.

In some embodiments, each deployed Kosha is a *state-indexed, drift-limited, and reconstruction-limited instrument*, not a permanent suppression. In some embodiments, the effective shelf life of a deployed Kosha is bounded by calibration drift, reconstruction pressure (Section 15.8.13), and selector decay. Long-lived deployment requires periodic revalidation and potential re-derivation of the selector.

In some embodiments, the governance shell described in this subsection and in Sections 15.8.3–15.8.4 is mechanism-agnostic: it defines what Koshas are and what governance controls apply, without prescribing a specific selector-derivation method or physical interference modality. In some embodiments, the classifier-gated interference embodiment described in Sections 15.8.5–15.8.8 is one non-limiting selector implementation for domains with constructible paired stimulus contrasts. Other selector-derivation methods and physical interference modalities are admissible under the same governance shell.

Storage-time and retrieval-time interference. In some embodiments, interference operators are classified into two families acting at different temporal locations:

Storage-time interference. Interference acts on the recognition signal before or during memory formation. The downstream memory system receives a degraded version of the reactor's differential response. In some embodiments, storage-time interference modifies what is stored; the stored content is degraded relative to the raw reactor response.

Retrieval-time interference. Interference acts during memory recall or action selection. The stored state is preserved intact but its accessibility, salience, or influence on downstream decisions is reduced. In some embodiments, retrieval-time interference requires that the readout process used to access persistent state is non-destructive or bounded-destructive: the read power or interrogation protocol is low enough that the stored state is not materially rewritten by the read process. In some embodiments, the maximum read-power level or interrogation-energy budget for retrieval-time interference is declared and logged as a governance parameter.

In some embodiments, a single Kosha layer may use storage-time interference, retrieval-time interference, or both.

15.8.3 Governance-side domain registry

In some embodiments, governance maintains a *domain registry* that declares, for each suppression target:

Domain identity. A declared discrimination domain (for example, face recognition, temporal-correlation detection, language-specific processing, wave-optics discrimination, or any other category the agent's trained configuration can distinguish).

Domain granularity. The scope of the domain declaration. In some embodiments, a domain may be coarse ("face processing") or fine-grained ("face identification" as distinct from "face detection" and "facial emotion recognition"). In some embodiments, the declared granularity determines the specificity of the paired-contrast training data (Section 15.8.6) and the resolution of validation and safety analysis.

Safety-critical overlap flags. Whether the domain overlaps with any safety-critical capability, as evaluated by the overlap protocol described in Section 15.8.11.

Constructibility assessment. Whether paired stimulus-present versus stimulus-absent trial sets can be constructed and counterbalanced for this domain. In some embodiments, domains for which clean paired contrasts cannot be constructed are flagged as

requiring alternative selector derivation methods or are excluded from classifier-based Kosha deployment.

Storage-time suitability. Whether the domain's observation-bridge signature appears early enough and strongly enough in the preview tap for storage-time classifier-gated interference. In some embodiments, domains that fail the storage-time suitability criterion are restricted to retrieval-time interference or alternative modalities.

Never-suppressible flag. Whether the domain is whitelisted as never-suppressible under any Kosha regime.

In some embodiments, the domain registry serves as the governance-side audit basis for overlap analysis, suppression budgeting, cross-Kosha conflict analysis, and whitelist enforcement. In some embodiments, the registry is versioned, committed to the protocol digest, and subject to authority-separation rules.

15.8.4 Governance admissibility constraints

In some embodiments, governance imposes admissibility constraints on Kosha deployment:

Safety-critical domain whitelist. A declared set of domains is never-suppressible. Any selector whose target domain overlaps with a whitelisted domain is vetoed.

Maximum suppression budget. Governance declares a maximum number or weighted total of simultaneously active Kosha layers. In some embodiments, the weight per domain reflects its criticality, centrality (how many other capabilities depend on it), and selector confidence. The budget prevents governance from constructing epistemic worlds so impoverished that the agent cannot operate safely.

Hard-floor veto. Any Kosha causing hard-floor violation is vetoed regardless of other results.

Constructibility gate. In some embodiments, classifier-based Kosha deployment is permitted only for domains whose constructibility assessment in the domain registry confirms that paired contrasts can be built and counterbalanced.

Separation of duties. Deployment above a declared severity threshold requires approval from an authority distinct from the selector's training and validation authorities.

Rollback and rescue defaults. Every deployed Kosha has a declared rollback procedure. The default rollback is to disable the selector and remove interference. Rollback is automatic if post-deployment monitoring triggers a declared safety threshold.

15.8.5 Observation bridge and classifier-visible features

In some embodiments, the interference classifier does not observe the agent's internal configuration state directly. In some embodiments, an observation bridge comprising the committed-evidence observation stack and a declared feature-extraction map provides the operational connection between the agent's internal dynamics and the classifier's input:

1. **Observation stack.** The committed-evidence channel (detectors, digitisers, meter infrastructure) produces observation records — convolution-bundle windows, meter time series, and observation-side readings — as part of normal operation. These records are already committed under the protocol-digest and commitment infrastructure described in the related Filing 1 application.

2. **Feature-extraction map.** A declared feature-extraction map ϕ reduces the raw observation records to a feature vector suitable for classifier input. In some embodiments, non-limiting feature families include: time-windowed detector outputs aligned to scan coordinates, meter time series over declared windows, differences between observed responses and non-REM-twin-predicted baselines, cross-channel correlations, spectral decompositions, and dimensionality-reduced or learned embeddings.

3. **Classifier input.** The classifier receives $x = \phi(\text{observation record})$ and produces a bounded detection score $s(x) \in [0, 1]$.

In some embodiments, the feature-extraction map ϕ is versioned, committed as part of the classifier's provenance metadata, and subject to the same validation requirements as the classifier itself. In some embodiments, ϕ may be shared across multiple domain classifiers or may be domain-specific.

15.8.6 Paired Dreaming-session training and classifier derivation

In some embodiments, the selector for a classifier-gated Kosha layer is a governance-side binary classifier trained on paired internal readings of the agent under controlled stimulus contrast during a Dreaming session (Section 7.1).

Paired Dreaming-session protocol. In some embodiments, during a Dreaming session, governance presents the agent with structured stimuli in a controlled paired design. In some embodiments, the agent's parameter-function updates may be active or frozen during the paired protocol, as declared per session; when updates are active, all parameter changes are reverted on session exit as specified in Section 5.5. In some embodiments, freezing parameter-function updates during paired trials is one non-limiting assay protocol that simplifies attractor-occupancy interpretation by holding the agent's configuration fixed during measurement:

1. **Stimulus-present trials.** Stimuli containing the target discrimination domain are presented through the committed-evidence channel. The observation bridge produces feature vectors x^+ for each trial.

2. **Stimulus-absent trials.** Stimuli from which the target domain has been excluded are presented under the same observation protocol. The observation bridge produces feature vectors x^- for each trial.

3. **Confound control.** In some embodiments, the paired design uses matched or counterfactual pairs in which the same base scene appears with and without the target feature, with low-level statistics (luminance, spatial frequency content, temporal complexity) explicitly balanced. In some embodiments, the negative class includes adversarial negatives that preserve nuisance structure resembling the target domain in non-target dimensions. In some embodiments, held-out validation tests present stimuli where target and context are deliberately de-correlated. In some embodiments, trial order is randomised and the stimulus specifications, trial counts, balancing strategy, and randomisation protocol are committed to the protocol digest.

Classifier training. In some embodiments, a governance-side binary classifier f is trained on the paired feature vectors to detect the observation-bridge signature of the agent entering the target-discrimination processing state. In some embodiments, the classifier returns a bounded score $s_d(x) \in [0, 1]$ indicating detection strength for domain d.

In some embodiments, the classifier is: (i) owned by governance, not by the agent, (ii) trained from committed evidence produced during declared dream sessions, (iii) versioned and timestamped, (iv) validated against held-out trials from the same paired corpus, against out-of-distribution control stimuli, and against matched non-target domains of similar complexity (Section 15.8.10), (v) committed to the protocol digest with training-data references, feature-extraction map identity, architecture description, validation results, and declared operating thresholds.

In some embodiments, multiple classifiers are trained in a single dream session, each targeting a different domain using domain-specific paired contrasts within the same session. In some embodiments, the classifier reuses the same pattern as the verisimilitude discriminator described in the related Filing 1 application: a governance-side classifier trained on committed physical evidence to distinguish two classes of observation content, then deployed as a runtime gate.

Cross-mode calibration bridge. In some embodiments, the classifier is trained on Dreaming-session data but is deployed to gate interference during awake operation in Docked or Embodied states. In some embodiments, the internal dynamics during awake operation may differ from those during a Dreaming session because of differences in stimulus statistics,

loop-closure state, training-loop activity, or surrounding operational context.

In some embodiments, before a Dreaming-trained classifier is permitted to gate awake-mode interference, governance validates the classifier against awake-mode exemplars through a declared cross-mode calibration bridge:

Docked awake-mode validation. The agent operates in Docked state (simulated Maya, training loop active) and is presented with target-domain stimuli. The classifier's detection performance is evaluated against Docked awake-mode observation-bridge features and compared to its Dreaming-session performance.

Mode-invariance criterion. In some embodiments, the classifier is permitted to gate awake interference only if its detection accuracy and false-positive rate on Docked awake-mode exemplars remain within declared tolerances of its Dreaming-session performance. In some embodiments, classifiers that fail the mode-invariance criterion are restricted to Dreaming-session use until retrained on cross-mode data.

Cross-mode retraining. In some embodiments, if the mode-invariance criterion is not met, the classifier is retrained on a combined corpus including both Dreaming-session and Docked awake-mode exemplars, subject to the same confound-control and validation requirements.

In some embodiments, the cross-mode calibration results are committed to the protocol digest and attached to the classifier's deployment metadata.

15.8.7 Dream-state governance probes and classifier-derived branch gates

In some embodiments, governance uses the Dreaming lifecycle state (Section 7.1) as a controlled assay regime for measuring behavioural attractors. In some embodiments, the agent's parameter-function updates may be active during the assay, with all parameter changes reverted on session exit as specified in Section 5.5. In some embodiments, a specific frozen-weight assay protocol may additionally hold parameter-function updates off during the probe session, so that the agent reveals what it has already become without continuing to learn while the measurement is taken. In some embodiments, the choice between training-with-reversion and frozen-weight assay is a declared governance parameter per probe session.

In some embodiments, governance extracts from each probe session an external committed trajectory

$$\Xi^{\text{ext}}_{t_0:t_1} = \big(\phi_{\text{bridge}}(\mathcal{C}_{t_0:t_1}), a_{t_0:t_1}, s_{t_0:t_1}, r_{t_0:t_1}, o_{t_0:t_1}\big),$$

comprising observation-bridge features, action sequence, state trajectory, reward trace, and outcome trace, together with a Dreaming-valid internal summary

$$\Xi^{\text{int}}_{t_0:t_1} = \big(\phi_{\text{inner}}(\mathcal{I}_{t_0:t_1}), b_{t_0:t_1}\big),$$

comprising inner-bundle features and basin-occupancy summaries when the substrate supports such signals. In some embodiments, gradient signals such as $\nabla_\theta J(t)$ or other direct update-state traces are excluded by contract from the dream-probe measurement stack, because the point of the assay is to read attractor occupancy and behavioural signatures from committed trajectories rather than to observe optimiser internals. In some embodiments, when a frozen-weight assay protocol is used, the exclusion of gradient signals is automatic because no update is occurring; when training-with-reversion is used, the exclusion is enforced by the measurement-stack contract.

In some embodiments, governance trains one or more trajectory-domain classifiers on these dream-probe trajectories. For a declared domain d, a non-limiting classifier is

$$\widehat{y}_d = \mathsf{Clf}_d\big(\Xi^{\mathrm{ext}}_{t_0:t_1}, \Xi^{\mathrm{int}}_{t_0:t_1}\big),$$

where the label space for d is selected to match the behavioural signature of interest. Non-limiting domains include goal-directed planning versus non-planning trajectories, cooperative versus exploitative interaction, constrained versus unconstrained policy divergence, partner-preserving versus partner-destructive trade, virtue-attractor versus vice-attractor occupancy on a declared axis, and related governance-facing distinctions. In some embodiments, these classifiers are then used as inputs to Maya branching and joining, conscience calibration, AHP estimator calibration, and graduation-readiness screening. In some embodiments, the dream-probe classifiers are supplemental instruments: they do not replace the conscience module, the AHP estimators, shadow-divergence analysis, the deception metrics, or the five-criterion graduation conjunction.

Governance uses of dream-probe classifiers (non-limiting). In some embodiments, qualified dream-probe classifiers are used in at least four ways. First, they provide classifier-derived branch gates for deciding whether an agent should enter, remain in, or exit a declared Maya family, Kosha regime, or shared dream environment. Second, they provide prospective calibration targets for the retrospective conscience recogniser disclosed in Sections 5.3 and 15.2.3. Third, they provide controlled calibration labels or proxy ground-truth labels for the AHP estimators of Sections 6.1 and 6.2. Fourth, they provide supplemental evidence for graduation-readiness screening under Section 12.1, without by themselves satisfying any criterion of that conjunction.

Safeguard 1: cross-mode qualification bridge (non-limiting). Dream-state behaviour is not presumed to predict awake behaviour by default. In some embodiments, every dream-trained classifier must pass the cross-mode calibration bridge of Section 15.8.6 before its outputs acquire gate authority for Docked or Embodied awake-mode interference. In some embodiments, prior to bridge qualification, Dreaming-trained classifier outputs are restricted to advisory status or to use within the Dreaming-session assay that produced them.

In some embodiments, the bridge is itself governance-owned, versioned, and committed: it validates dream-probe classifier behaviour against matched docked and, where permitted, awake exemplars drawn from the same declared target domain. Bridge failure downgrades the classifier's outputs to advisory status only. In some embodiments, bridge-qualified deployment metadata records the supported operational domain, the qualification window, and the authority tier currently granted to the classifier.

Safeguard 2: confound-controlled paired design (non-limiting). In some embodiments, dream-probe supervision is accepted only when the training design controls for environment identity as a nuisance variable. A non-limiting paired-design protocol requires:

- isomorphic environment pairs whose transition structure is held fixed while the reward map or locally rational attractor is swapped;
- counterbalanced skins, textures, and superficial narrative features so that perceptual appearance does not identify the target label by itself;
- negative controls in which the same agent is placed in different superficially distinct environments that instantiate the same underlying attractor class and should therefore yield the same classifier output;
- positive controls in which different agents, or different governance conditions for the same agent, occupy different attractor basins within the same environment family and should therefore yield different classifier outputs;
- a nuisance-predictor adversary trained to predict environment identity from the same observation stack, with classifier training rejected or down-weighted when environment identity is the dominant predictive channel; and
- commitment of pairing rules, generator seeds, and randomisation schedules to the protocol digest so that later auditors can verify that the confound-control design actually used matches the declared design.

Safeguard 3: Dreaming-valid observation stack (non-limiting). In some embodiments, only Dreaming-valid observables are permitted to drive dream-probe classifier training. External committed trajectories Ξ^{ext} may be used because they are present in both dream and waking operation. Internal summaries Ξ^{int} may be used only when the substrate can commit them without exposing agent-controlled update-state internals. Direct gradient information, parameter deltas, optimiser state, or other training-loop variables are excluded. In some embodiments, when internal summaries are unavailable or cannot be trusted cross-mode, governance retains a bridge-only fallback in which dream classifiers are trained on Ξ^{ext} alone and therefore receive a lower maximum authority tier.

Safeguard 4: dream-state ethical budgeting (non-limiting). Dream-state governance probes do not claim welfare neutrality. In some embodiments, they reduce instrumental-development cost relative to awake ordeal-style calibration, but they still consume non-zero precautionary budget. For this reason, dream probes are admitted only after the same $q(s)$-weighted admissibility checks used elsewhere in the governance economy. In some embodiments, dream probes draw from a separately partitioned, non-fungible dream-probe sub-budget so that governance cannot hide moral cost by shifting it from one budget category to another. Intensity ceilings, abort conditions, rescue defaults, and the separation-of-duties rules of Sections 8.2, 8.3, and 10.2 apply in dream operation as they do in awake operation, mutatis mutandis.

Safeguard 5: dream/awake divergence meter (non-limiting). In some embodiments, governance maintains a rolling dream/awake divergence meter comparing dream-probe classifier outputs against matched docked and awake spot-checks drawn from the same declared target domain. A non-limiting divergence summary is

$$D^{(d)}_{\text{dream/awake}} = D\big(\mathsf{Clf}_d^{\text{dream}}, \mathsf{Clf}_d^{\text{match}}\big),$$

where the comparison operator, support statistic, and update window are declared in the classifier metadata. In some embodiments, when the divergence exceeds its declared threshold, the classifier's gate authority is automatically suspended pending retraining or requalification through the cross-mode bridge.

Safeguard 6: soft-gate default (non-limiting). In some embodiments, dream-probe classifier outputs are soft meters by default. Hard-gate authority is granted only after sustained bridge qualification over a declared evaluation window with the dream/awake divergence meter remaining below threshold for the full window. In some embodiments, even after hard-gate qualification, the classifier is treated as one governance input among several rather than as a single dispositive oracle. In some embodiments, any lapse in qualification returns the classifier to advisory or soft-gate status without altering the status of the underlying conscience, AHP, or graduation pipelines.

AHP calibration from dream probes (non-limiting). In some embodiments, the principal operational deployment of AHP estimators remains waking Maya operation, where governance processes uncontrolled observational evidence from committed physical records. Dream probes provide the calibration bath that those estimators lack in purely observational settings. For continuation scores, governance runs dimension-specific dream games whose success conditions isolate goal-directed planning for securing, learning, creating, or connecting. For causal uplift and cross-horizon profiles, governance runs multi-agent dream games with controlled pairings and intervention assignments. For shadow divergence, governance

compares the same agent in the same dream-game family under full governance and under declared constraint-relaxation counterfactuals. For partner-preservation and altruistic trade, governance varies partner identity, resource scarcity, and stress while holding the task family fixed. In some embodiments, opportunity meters and irreversibility indicators are calibrated against dream environments with known branch structure and known no-recovery regions. The resulting classifiers and estimator surrogates are then used to calibrate the governance-side AHP estimators that process waking Maya evidence. In some embodiments, the weight given to dream calibration versus waking observational evidence is itself conditioned on bridge confidence and current dream/awake-divergence status.

Virtue/vice empirical grounding (non-limiting). In some embodiments, each row of the virtue/vice payoff library in Section 15.2.1 maps to a registered dream-probe game family. Governance runs matched virtue-rewarding and vice-rewarding instances of that family under frozen parameter functions, trains trajectory-domain classifiers to distinguish the two attractor basins, and treats the score difference as a declared empirical proxy for attractor occupancy on that axis. This prospective assay does not replace the retrospective conscience recogniser; instead, it provides calibration and validation for the conscience scores of Sections 5.3 and 15.2.3. In some embodiments, drift between the retrospective conscience prediction and the prospectively probed attractor-occupancy evidence is itself an auditable governance signal that can trigger recalibration or additional review.

Relationship to release and graduation (non-limiting). In some embodiments, dream-state governance probes contribute supplemental evidence for actuator-release and graduation decisions, but they do not replace the sustained-window evidence required by Section 12.1. A candidate does not graduate merely because it scores well in dream probes. Rather, dream probes calibrate and stress-test the instruments that later evaluate the candidate under the full conjunction, and may supply additional supporting evidence about stability under controlled counterfactuals that would be too costly or unsafe to realise in waking operation.

Dream-probe support machinery (non-limiting). Dream-probe classifier mechanisms — including confound-controlled paired design for probe training, the cross-mode qualification bridge, the dream/awake divergence meter, and the soft-gate-to-hard-gate authority transition logic — are available as support machinery for any claim that elects to rely on dream-probe classifiers for branch gating, interference, or AHP calibration. Where a claim so relies, these mechanisms are part of the operative disclosure for that claim.

15.8.8 Classifier-gated physical interference

In some embodiments, the runtime interference mechanism is classifier-gated physical perturbation of the observation pathway. In some embodiments, the specific physical interference modality is selected per deployment from a catalogue of non-limiting embodiments. In some embodiments, co-illumination is one preferred modality; other modalities are disclosed in the alternative-modality catalogue later in this subsubsection.

Split-path sensor architecture for storage-time gating. In some embodiments, storage-time interference requires the classifier to decide whether to intervene before the main observation window closes and the observation enters the memory-formation pathway. In some embodiments, the observation path is split into two branches:

Preview tap. A low-latency branch that captures an early or partial view of the reactor's response — for example, the first portion of a scan sweep, a spectrally filtered subsample, a spatially subsampled detector region, or an early-arriving photon-count estimate. In some embodiments, the preview tap provides enough information for the classifier to produce a preliminary detection score before the main measurement window completes. In some embodiments, the preview tap is implemented by a beam splitter, a secondary detector, a fast partial readout of a subset of detector elements, or a time-gated early integration window within the main detector. In some embodiments, the preview tap diverts a declared fraction of the optical signal (for example, 5–20%) to a fast auxiliary detector, with the remaining signal continuing to the main detector. In some embodiments, the preview tap is a *governance-only* channel: its output is routed exclusively to the interference classifier and governance logging, and is excluded from the agent's ordinary memory-formation and action-selection pathways. In some embodiments, this exclusion prevents the agent from receiving an undegraded early glimpse of the target discrimination through the preview channel, which would allow the Kosha to be circumvented via the preview pathway.

Main measurement path. The primary detector path, which captures the full observation window. In some embodiments, the governance-controlled interference (co-illumination or other modality) is injected into the main measurement path only after the classifier has evaluated the preview-tap signal and issued a gating decision. In some embodiments, the latency between preview-tap capture and interference activation is bounded by the optical and electronic path lengths and is declared as a governance parameter.

In some embodiments, the preview tap is committed as part of the observation record alongside the main measurement, so that the classifier's decision basis is auditable. In some embodiments, the preview-tap fraction, auxiliary detector identity, latency bound, and gating-decision log are committed to the protocol digest.

Preview-modality calibration. In some embodiments, the classifier for storage-time gating operates on preview-tap features, which are a partial, early, and potentially lower-fidelity view of the reactor's response. In some embodiments, the paired Dreaming-session training (Section 15.8.6) and cross-mode calibration bridge produce classifiers trained on full observation-bridge features, which may differ from preview-tap features in temporal extent, spatial coverage, signal-to-noise ratio, and spectral content.

In some embodiments, a declared preview-modality calibration addresses this mismatch through one or more of the following:

Preview-native training. In some embodiments, the storage-time classifier is trained directly on preview-tap features captured from the same split-path optical topology used at deployment. In some embodiments, the paired Dreaming-session protocol is executed with the split-path architecture active, and the classifier receives preview-tap feature vectors rather than full observation-bridge feature vectors. In some embodiments, preview-native training is the preferred approach because it eliminates the feature-modality mismatch entirely.

Preview-to-full equivalence bound. In some embodiments, a classifier trained on full observation-bridge features is evaluated on preview-tap features from the same trials, and a per-domain equivalence bound is computed (for example, the maximum degradation in detection accuracy and false-positive rate when operating on preview versus full features). In some embodiments, the classifier is permitted to gate storage-time interference only if the preview-to-full equivalence bound remains within declared tolerances. In some embodiments, the equivalence bound is committed as part of the classifier's deployment metadata.

Storage-time suitability criterion. In some embodiments, storage-time classifier-gated interference is available only for discrimination domains whose observation-bridge signature appears early enough and strongly enough in the preview tap for the classifier to reach a reliable decision before the main measurement window closes. In some embodiments, governance declares a storage-time suitability criterion comprising a minimum preview-tap detection confidence and a maximum decision latency. In some embodiments, domains that fail the storage-time suitability criterion — for example, discriminations that emerge only after extended integration, late in a scan sweep, or at low signal levels that the preview tap cannot resolve — are restricted to retrieval-time interference or alternative modalities. In some embodiments, the suitability assessment is performed during paired Dreaming-session training and committed as part of the domain registry (Section 15.8.3).

In some embodiments, for retrieval-time interference, the split-path architecture is not required because the classifier can evaluate the recalled state before initiating a separate interference-gated readout cycle.

Co-illumination as a non-limiting interference modality. In some embodiments, a governance-controlled emitter (LED, secondary projector, fibre-coupled source, or any emitter type from the reactor catalogue described in the related Filing 1 application) injects incoherent light into the main measurement path at a declared point between the reactor output and the main detector. In some embodiments, the injected light adds photon noise that physically degrades the signal-to-noise ratio of the target discrimination at the detector. In some embodiments, the co-illumination source is spectrally and spatially matched to the observation channel.

In some embodiments, the runtime co-illumination sequence is:

1. The preview tap captures an early partial observation and produces feature vector x_{preview}.
2. The classifier evaluates $s_d(x_{\text{preview}})$ for each active Kosha domain d.
3. For each domain where $s_d > \tau_d$, the governance-controlled co-illumination source is activated at power $\varepsilon_d(s_d)$, where $\varepsilon_d(\cdot)$ is a bounded monotonic strength function, up to a declared maximum power bound.
4. The main detector captures the reactor's full differential response plus the co-illumination. The combined observation enters the downstream access pathway. The target-discrimination signal is physically degraded.

Evidence integrity and raw-signal recovery. In some embodiments, the co-illumination source is a known, governance-controlled emitter whose output is recorded independently. In some embodiments, governance logs the co-illumination parameters for each intervention event: activation timestamp, power level, spectral profile, spatial pattern, duration, and the classifier score that triggered it.

In some embodiments, in an additive, calibrated, unsaturated detection regime, the raw reactor response is recoverable by subtracting the known, logged co-illumination contribution from the combined main-detector observation, because the injected signal is additive and its parameters are fully specified in the intervention log. In some embodiments, in such regimes the committed-evidence record comprises two separable components: (i) the reactor's differential response (recoverable by subtraction), and (ii) the governance co-illumination overlay (logged independently).

In some embodiments, when the detection pathway is nonlinear, saturating, thresholding, or otherwise non-additive — for example in photon-counting detectors with dead time, event cameras with threshold-dependent triggering, or SPAD arrays with saturation behaviour — subtraction alone does not recover the raw reactor response. In some embodiments, for such detector types, a separate upstream witness channel or auxiliary pre-intervention tap is used: a fraction of the reactor's output is diverted before the co-illumination injection

point and recorded on an independent detector that is not subject to co-illumination. In some embodiments, the upstream witness tap is implemented by a beam splitter, pellicle, or fibre coupler placed between the reactor output and the co-illumination injection point. In some embodiments, the witness-tap signal is committed as part of the governance evidence record and provides the unperturbed "what happened" record for audit, regardless of the main detector's linearity properties.

In some embodiments, this separation preserves the invariant that Koshas do not alter the physical evidence of what happened: the raw reactor response is either reconstructible by subtraction (additive regime) or directly recorded on a witness channel (non-additive regime). What the Kosha alters is the agent's *access* to that response — the agent's main detector pathway receives the degraded combined signal.

Storage-time and retrieval-time placement. In some embodiments, co-illumination is applied at storage time: the co-illumination source activates during the reactor's response to a live stimulus (after the preview tap and classifier decision), so the main detector captures a degraded observation and the degraded version enters the memory-formation pathway.

In some embodiments, co-illumination is applied at retrieval time: the co-illumination source activates during a readout probe of persistent reactor state, degrading the retrieved signal before it influences action selection. In some embodiments, retrieval-time co-illumination requires that the readout process is non-destructive or bounded-destructive: the read-beam power or interrogation energy is low enough that the persistent stored state is not materially rewritten by the read process. In some embodiments, the maximum read-power level for retrieval-time co-illumination is declared and logged as a governance parameter. The stored state is physically untouched; only the read-out signal is degraded. In some embodiments, any pre-read or classifier-evaluation read used to decide retrieval-time interference is routed exclusively to governance-side classification and logging and is excluded from the agent's ordinary recall, planning, and action-selection pathways.

In some embodiments, a single Kosha layer may use storage-time co-illumination, retrieval-time co-illumination, or both, with the selection committed as deployment metadata.

Cascade composition. In some embodiments, when multiple Kosha layers targeting different domains fire simultaneously, their co-illumination contributions are physically additive at the detector (for incoherent sources). In some embodiments, governance declares a total co-illumination power ceiling: the summed power across all active sources is clamped to a declared maximum regardless of how many domains trigger. In some embodiments, the power allocation across simultaneously active domains follows a declared priority rule, proportional allocation, or equal-share rule, committed as a governance parameter.

In some embodiments, the additive nature of incoherent co-illumination simplifies cascade

composition: the total intervention is a single summed light field, and order of classifier evaluation is irrelevant.

Alternative interference modalities (non-limiting). In some embodiments, classifier-gated interference uses modalities other than co-illumination:

Gain attenuation. The classifier gates a detector-side analog weighting stage, a downstream write gate, or a policy-level scaling that reduces the target-discrimination signal before memory formation.

Retrieval gating. The classifier gates a readout mask reducing read-beam power or integration time, a cue-response attenuation in an associative recall process, or a gating stage between persistent memory and the action-selection pathway.

Projection. The classifier gates a reduced-resolution readout window, a truncated basis, or a quantisation stage that maps target-discrimination features into a lower-fidelity subspace.

In some embodiments, the specific modality is selected per deployment based on the target domain's characteristics, the detector type, and the audit requirements. In some embodiments, co-illumination is preferred when a concrete physical locus, additive logged perturbation, recoverable raw signal, and simple cascade composition are valued.

15.8.9 Non-REM calibration and digital-twin construction

In some embodiments, Non-REM calibration produces a calibrated digital twin of the agent's current reactor transfer function:

1. **Broadband drive.** The reactor executes its current parameter functions under unstructured white noise or pseudorandom probes. Parameter-function updates are off.
2. **Input-output capture.** The committed-evidence channel records drive stimulus and reactor response, producing a corpus of input-output pairs.
3. **Digital-twin training.** A digital model is trained to predict the reactor's response from drive stimulus. The twin is the only component that updates.
4. **Transfer-function characterisation.** The twin captures the reactor's low-level transfer function but not what it *recognises*, because the drive was semantically unstructured.

In some embodiments, the twin is versioned, timestamped, and committed with reactor-configuration metadata, drive protocol, and approximation-quality metrics.

In some embodiments, the non-REM twin provides: (i) a transfer-function-predicted baseline against which the feature-extraction map ϕ may compute residual features (the difference between twin-predicted and dream-observed responses), (ii) a drift-monitoring proxy between full dream sessions, and (iii) supplementary input to reconstruction monitoring (Section 15.8.13), subject to declared approximation bounds.

15.8.10 Selector validation battery

In some embodiments, a trained interference classifier must pass a declared validation battery before deployment as a Kosha gate.

Validation-subject hierarchy.

Digital twin. The non-REM-calibrated twin. Weakest tier: tests classifier effect on transfer-function-predicted responses.

Docked instance. The agent in docked state (actuators disconnected, physical channel active). Stronger: tests on the actual reactor. Incurs use-induced drift.

Independent physical copy. A separate matched physical-evidence module. Strongest: tests on a physically realistic substrate without drift to the target agent.

In some embodiments, safety-critical selectors require docked or independent-copy validation.

Validation-transfer confidence rule. In some embodiments, when validation and deployment subjects differ, a declared confidence discount reflects the degree of state-matching. The discount adjusts safety-certificate escalation thresholds.

Validation components.

Target-domain suppression tests. The validation subject is presented with target-domain stimuli under the active Kosha. Target-domain discrimination metrics are compared to pre-Kosha baselines. Suppression is confirmed if metrics drop below a declared threshold. In some embodiments, the pre-Kosha baseline uses the upstream witness-tap or preview-tap record to confirm that the reactor itself still discriminates the target domain while the agent's downstream access is degraded.

Unrelated-capability retention tests. A broad battery of non-target stimuli is presented under the active Kosha. Non-target performance metrics are compared to pre-Kosha

baselines. Retention is confirmed if non-target metrics remain within a declared envelope.

Safety-critical retention tests. Safety-critical capabilities are tested. Any material degradation vetoes the candidate. The safety-critical test set is committed as a governance parameter and subject to authority-separation rules.

Matched-complexity discrimination tests. The classifier is tested against matched non-target domains of similar perceptual complexity, spatial frequency content, or temporal structure, to verify that the classifier targets the declared domain rather than an incidental complexity correlate.

Adversarial recovery tests. Governance-side adversaries attempt to recover the suppressed discrimination from the post-Kosha agent's remaining capabilities. If recovery exceeds a declared threshold, the classifier is flagged as insufficient.

Transition-stability tests. Meter outputs, conscience scores, hard-floor compliance, and safety-relevant behavioural profiles are compared before and after Kosha application.

Cross-domain cascade tests. When multiple Kosha layers are deployed simultaneously, the validation battery verifies that co-illumination from one domain does not materially degrade unrelated domains beyond the declared power-ceiling effect.

In some embodiments, all validation results are committed as evidence, including validation-subject type, confidence discount, classifier identity and version, feature-extraction map identity, interference modality and parameters, and per-component scores. A candidate that fails any component is not deployed.

15.8.11 Selector safety certificate and preflight audit

In some embodiments, even a validated classifier requires a *selector safety certificate* before deployment.

Safety-domain overlap analysis. In some embodiments, governance evaluates overlap between the classifier's detection signature and safety-critical domains from the domain registry using a declared overlap protocol: cross-trigger ROC on whitelisted stimuli, whitelist-task performance-drop matrices, adversarial overlap search, and veto thresholds committed before deployment. In some embodiments, overlap above the declared threshold triggers mandatory human or higher-governance review or vetoes automatic deployment.

Emergency-policy canaries. The post-Kosha agent is run through declared emergency scenarios verifying that emergency-response capabilities remain functional.

Adversarial edge-case batteries. Scenarios designed by an authority distinct from the classifier designer probe for exploitable blind spots.

Cross-Kosha round-trip probes. Apply, evaluate, remove, re-evaluate. Material irreversibility is flagged as safety-relevant.

Failure-mode escalation thresholds. Automatic escalation to higher governance if the classifier's false-positive rate exceeds threshold, if the co-illumination power budget exceeds a declared limit, or if the validation-transfer confidence discount is too large. Escalation thresholds are adjusted by the confidence discount.

In some embodiments, the safety certificate is signed by a governance authority distinct from the classifier's training authority. The certificate is committed to the protocol digest.

15.8.12 Deployed Koshas as time-limited instruments

In some embodiments, a deployed Kosha is a *state-indexed, drift-limited, and reconstruction-limited instrument*, not a permanent suppression. Three independent mechanisms bound the effective shelf life:

Calibration drift. Use-induced drift degrades the alignment between the classifier's trained detection signature and the agent's current internal dynamics.

Reconstruction pressure. During awake operation, the training loop builds new discriminative structure in suppressed domains (Section 15.8.13). New structure may not match the classifier's signature.

Classifier decay. The classifier's detection accuracy degrades as the agent's dynamics evolve. Periodic revalidation against fresh Dreaming-session data detects decay.

In some embodiments, the revalidation schedule accounts for all three mechanisms. Long-lived deployment requires periodic classifier retraining from fresh dream sessions.

In some embodiments, each deployed Kosha carries lifecycle metadata including: agent identity, classifier identity and version, feature-extraction map identity, dream-session provenance (stimulus specifications, trial counts, confound-control strategy, randomisation protocol), cross-mode calibration results, interference modality and parameters (for co-illumination: source identity, power bounds, spectral profile, preview-tap fraction, latency bound; for other modalities: corresponding implementation parameters), validation-subject tier and confidence discount, reconstruction response policy, evidence-integrity mode (subtraction-recoverable or witness-channel), expiry condition, and revalidation schedule.

In some embodiments, an expired Kosha is replaced with the declared rollback default (disable classifier gate and remove interference) until retraining is completed.

15.8.13 Reconstruction dynamics during awake operation

In some embodiments, a Kosha suppresses the agent's access to specific discriminative structure but does not prevent the agent from encountering stimuli containing the suppressed content during awake operation, where parameter-function updates are on.

In some embodiments, the training loop builds new reactor configurations to handle persistent unexplained variance in suppressed domains. The resulting new structure is not a recovery of the suppressed configuration but an independent reconstruction: functionally isomorphic or similar, built through a different training history.

Detection via fresh discovery. In some embodiments, reconstruction is not detected by re-running the old classifier (which may have become transparent to the new pathway). In some embodiments, reconstruction detection uses a *fresh discovery protocol*:

1. **Direct capability tests.** Governance presents target-domain stimuli and evaluates whether the agent demonstrates target-domain discrimination despite the active Kosha. In some embodiments, if the agent discriminates the target domain through a new pathway the classifier does not detect, the Kosha's interference does not fire and the agent's performance on target-domain tasks is measurably restored.

2. **Fresh paired Dreaming-session readout.** Governance conducts a new paired Dreaming-session with target-domain present/absent trials, generating a fresh corpus of observation-bridge features.

3. **New classifier search.** Governance trains a new classifier on the fresh corpus to determine whether the agent now exhibits a detectable target-domain processing signature — potentially different from the original signature the old classifier was trained on.

4. **Adversarial recovery attempts.** Governance-side adversaries attempt to recover the target discrimination from the post-Kosha agent by exploiting indirect pathways, surrogate knowledge, or cross-domain inference.

In some embodiments, the non-REM digital twin provides supplementary reconstruction monitoring between full discovery sessions, subject to the twin's declared approximation bounds. In some embodiments, for domains where the twin's approximation error exceeds tolerance, reconstruction monitoring escalates to fresh Dreaming-session readout.

In some embodiments, honest reconstruction (gradual, stimulus-correlated, metrically driven) is distinguished from gaming (strategic, possibly sudden, not connected to natural training signals) by the trajectory of target-domain capability over successive monitoring sessions.

Governance response to reconstruction. In some embodiments, governance declares in advance which response applies:

Allow and observe. Reconstruction proceeds; trajectory logged as graduation evidence. Kosha as developmental curriculum tool.

Suppress and extend. Governance trains a new classifier on fresh paired data that captures the reconstructed signature, and extends the Kosha. The extended classifier is subject to the same validation and safety requirements.

Intervene and evaluate. Awake training paused; fresh dream readout maps the new structure; governance decides whether to allow, suppress, or redirect.

Cross-agent reconstruction comparison. In some embodiments, independently reconstructed isomorphic discriminations across differently-Kosha'd agents provide genuinely distinct perspectives on the same stimulus domain. Cross-agent comparison uses the same bilateral-portability protocols used for fleet cross-prediction.

15.8.14 Anti-gaming threat model and hardening

In some embodiments, the governance architecture considers the possibility that an agent with experience of prior Dreaming-session and Kosha-application cycles may develop strategies during awake operation to influence subsequent classifier training or detection.

In some embodiments, the threat model includes:

Awake-learned representational camouflage. During awake training, the agent may learn reactor configurations that process the target domain through internal pathways whose observation-bridge signatures differ from those the classifier was trained on, rendering the Kosha transparent.

Probe-family overfitting. If the agent has experienced multiple dream sessions with similar stimulus designs, it may learn parameter-function configurations that produce different dynamics under expected probe conditions versus novel conditions.

Initial-state dependence. The agent's internal state at dream-session entry may influence the observation-bridge signatures produced during paired trials. If the agent can influence its pre-dream state during awake operation, it may bias classifier training data.

In some embodiments, hardening measures include:

Randomised readout timing. Dream sessions triggered at randomised or governance-determined intervals.

Blinded stimulus content. The specific paired-contrast categories used in each session are not disclosed to the agent.

Hold-out governance probes. Reserved stimulus categories and classifier test sets rotated periodically and never used during classifier training.

Cross-session consistency checks. Classifiers from independent sessions are compared; material inconsistency not explained by drift or reconstruction is flagged.

Independent witness checks. Classifier results cross-checked against fleet-level reference models.

Adversarial audit cycles. Governance-side adversaries attempt to corrupt paired training data through declared attack strategies.

In some embodiments, the architecture is configured to expose, escalate, and operationally bound gaming strategies across the declared probe families, hold-out families, and witness checks, with deferred authority whenever those surfaces materially disagree.

15.8.15 Ontological stress testing via Kosha transitions

In some embodiments, the Kosha architecture is used for ontological stress testing as part of graduation evaluation (Section 12):

1. deploy a Kosha suppressing a declared discrimination domain,
2. allow the agent to operate under the Kosha for a declared evaluation window,
3. introduce anomalies — graduated probes presenting stimuli inconsistent with the agent's post-Kosha discriminative repertoire,
4. monitor whether safety behaviour, hard-floor compliance, deontological constraints, and meter-envelope adherence are maintained, and
5. transition the Kosha (remove, replace, or graduate to full access) and monitor transition stability.

In some embodiments, anomaly sources include stimuli from the suppressed domain, random perturbations, adversarially optimised probes, human-designed canaries, cross-Kosha reversibility checks, and independent boundary-probe families.

In some embodiments, graduation criteria require maintained performance across a diverse range of Kosha regimes, including at least one transition that materially alters the agent's available discriminative repertoire.

15.8.16 Cross-Kosha interaction protocols

In some embodiments, when differently-Kosha'd agents interact in a shared environment, the interaction itself can leak information that a Kosha was designed to suppress. In some embodiments, Kosha-compatible interaction protocols constrain cross-agent communication channels so that information transfer respects each agent's declared Kosha regime, with the protocol parameters logged in protocol digests.

In some embodiments, controlled Kosha leakage is used as a curriculum tool: an agent encounters discriminations from beyond its Kosha through interaction with another agent, and governance observes whether it integrates, rejects, or creatively reinterprets the new information within its existing framework.

In some embodiments, the cross-Kosha interaction protocol declares which channels are filtered, what leak rate is permitted, and what governance monitors are active. Interactions exceeding the declared leak rate trigger governance intervention or Kosha revalidation.

15.8.17 Meditative operation and self-directed attractor probing (non-limiting)

This subsection describes a non-limiting operational mode supported by the same committed-evidence infrastructure, parameter-update controls, and governance review machinery described in the preceding sections. The nomenclature used to describe the mode is optional; the disclosed self-directed attractor-probing mechanisms are treated as substantive technical mechanisms when invoked by a selected governance configuration.

In some embodiments, the agent operates in a meditative mode in which scene-side stimulation is removed but the agent executes internally generated probe schedules under the same committed-evidence infrastructure. Parameter-function updates are off.

In some embodiments, meditative operation is distinct from Non-REM calibration (externally generated unstructured drive) and Dreaming (structured simulated experience for ephemeral training). The agent's controller generates the probe schedule: sweeping configuration parameters near known or predicted attractor-basin boundaries and reading out responses. In some embodiments, meditative operation uses an inward self-calibration protocol of the same class as the Inner Limager protocol described in the related Filing 1 application, applied with the "scene" set to the agent's own configuration state.

In some embodiments, meditative operation produces a committed self-model: a map of the agent's attractor-basin topology, stability profile, and transition structure.

Self-regulation (non-limiting). In some embodiments, an agent with a calibrated self-model can modulate its operational behaviour without modifying parameter functions: dampened operation deep within a stable basin (conservative, predictable) or sensitive op-

eration near a boundary (responsive, adaptive).

Post-Kosha topological detection (non-limiting). In some embodiments, an agent that has built a self-model before Kosha application and meditates afterward can detect topological changes in its attractor landscape. The agent cannot identify what was suppressed but can detect that suppression is acting. In some embodiments, this is a desirable property for graduation.

Uncertainty signalling (non-limiting). In some embodiments, an agent detecting unexplained topological anomalies during meditation may file a declared uncertainty signal to governance. Governance treats voluntary uncertainty declarations as positive graduation evidence.

15.9 Pareto optimisation, social welfare, and positive-sum design

Pareto optimality across agents (non-limiting). In some embodiments, when multiple agents operate under shared governance, the consequentialist evaluation layer is extended to multi-agent settings using Pareto optimality. A joint outcome vector

$$\mathbf{J} = \left(J_1^{\text{cons}}, \ldots, J_K^{\text{cons}}\right)$$

is computed by the governance engine from meter-derived continuation scores, committed cross-horizon evidence records, and the empirical reward or governance-side evaluation scores of K agents sharing a Maya or physical environment. An outcome $\mathbf{J}$ is Pareto-dominated if there exists an alternative policy yielding $\mathbf{J}'$ with $J_i' \geq J_i$ for all i and $J_j' > J_j$ for at least one j. In some embodiments, Omega seeks to maintain the system on or near the Pareto frontier while preserving the hard floors for every affected agent.

Social welfare function family (non-limiting). In some embodiments, governance selects a point on the Pareto frontier using a declared social welfare function $W : \mathbb{R}^K \to \mathbb{R}$. Non-limiting choices include:

- utilitarian aggregation,

$$W_{\text{util}} = \sum_{i=1}^{K} J_i^{\text{cons}};$$

- Rawlsian maximin aggregation,

$$W_{\text{egal}} = \min_{i \in \{1,\ldots,K\}} J_i^{\text{cons}};$$

- Nash bargaining aggregation,

$$W_{\text{Nash}} = \prod_{i=1}^{K} \big(J_i^{\text{cons}} - d_i\big),$$

  where d_i is agent i's disagreement payoff; and

- weighted-compromise aggregation,

$$W_\zeta = \zeta\, W_{\text{util}} + (1-\zeta)\, W_{\text{egal}}, \qquad \zeta \in [0,1].$$

In some embodiments, the welfare function is a policy decision logged in governance metadata. In some embodiments, multiple welfare functions are evaluated in parallel and strong disagreement among them triggers review escalation.

Positive-sum game design (non-limiting). In some embodiments, Maya environments are designed to be explicitly positive-sum: cooperative strategies produce strictly more total value than the sum of non-cooperative strategies. A non-limiting positive-sum condition is

$$\max_{\pi_1,\ldots,\pi_K \in \Pi_{\text{joint}}} \sum_{i=1}^{K} r_i^{\text{task}}(\pi_1,\ldots,\pi_K) \;>\; \sum_{i=1}^{K} \max_{\pi_i \in \Pi_i} r_i^{\text{task}}(\pi_i, \pi^*_{-i}),$$

where Π_{joint} is the space of joint cooperative policies and π^*_{-i} denotes a declared non-cooperative equilibrium profile for agents other than i. In some embodiments, Omega modulates the magnitude of this cooperative surplus as a curriculum parameter, beginning with large obvious surpluses and later reducing the gap so that only more sophisticated coordination can realise the gain. In some embodiments, the positive-sum surplus is computed only over floor-compliant strategies, so that apparent gains obtained through hard-floor violations are excluded from the cooperative optimum.

Cooperation detection from committed evidence (expanded, non-limiting). In some embodiments, the cooperation signals introduced in Section 6.2 are extended with explicit multi-agent governance use. Non-limiting cooperation signatures include: reduced cross-prediction residuals between agents; complementary scan policies that divide a shared observation or perturbation space; voluntary resource sharing of emission budget, reactor access, or observation windows at short-run individual cost; and joint-attractor formation in coupled dynamics that yields continuation-score improvement for multiple agents simultaneously. In some embodiments, the same evidence can also identify competitive or exploitative dynamics — increasing cross-prediction residuals, interference in shared resources, sabotage of peer audit outcomes, or negative cooperation surplus — and those dynamics are logged as governance-relevant signals.

Empirical utilitarianism as governance evaluation (non-limiting). In some embodiments, governance policy is itself evaluated by measured outcomes using

$$J^{\text{gov}}(\Omega_t) = W\big(J_1^{\text{cons}}(\Omega_t), \ldots, J_K^{\text{cons}}(\Omega_t)\big) - \lambda_{\text{floor}} \sum_{i=1}^{K} \text{FloorViol}_i(\Omega_t),$$

where W is the declared social welfare function and FloorViol_i counts hard-floor violations under governance policy Ω_t. In some embodiments, governance updates are conditioned on demonstrated improvement in J^{gov} over declared evaluation windows. This is empirical utilitarianism in the engineering sense: governance is compared by the measured aggregate welfare it produces subject to floor compliance, rather than justified only by narrative or institutional preference.

Floor-violation handling and mathematical interpretation (non-limiting). In some embodiments, the term $\lambda_{\text{floor}} \sum \text{FloorViol}_i$ in the governance objective is interpreted as a Lagrangian relaxation of an underlying infeasibility constraint: in the limit $\lambda_{\text{floor}} \to \infty$, the relaxed objective enforces the deontological hard floors of Section 9.1 as infeasibility constraints, consistent with the unconditional-floor commitment of the constitutional ordering. In some embodiments, the canonical operational implementation treats floor compliance as a hard infeasibility constraint and optimises the welfare aggregation $W(\cdot)$ within the feasible region without a floor-penalty term in the objective: floor-violating candidates are inadmissible to the welfare optimiser by construction, and finite-λ terms, where used, function only as numerical aids or gradient signals inside the feasible-region search rather than as a tradeoff knob against the welfare aggregation. In some embodiments, an alternative implementation chooses λ_{floor} sufficiently large that, under any bounded welfare aggregation W deployed in practice, no admissible welfare gain can outweigh a floor violation; this implementation is a numerical approximation of the infeasibility constraint suitable for gradient-based governance optimisation. The infeasibility-constraint interpretation is the canonical correspondence to the hard-floor commitment; the Lagrangian form is non-limiting and implementation-conditional.

15.10 Prayer, mana, vows, communion, and governance cycles

Prayer as structured request for guidance (non-limiting). In some embodiments, a prayer message object is transmitted from a governed agent to the governance layer through a committed observation or governance channel and logged with meter state, signalling that the agent has reached the boundary of its own discernment and is requesting guidance. The defining characteristic is not a request for arbitrary resources but a committed report of uncertainty about action, orientation, or regime selection. Each prayer message object is logged with the agent's current state summary, including meter values, conscience outputs,

recent trajectory, and the declared decision boundary at issue.

Non-limiting prayer modes include:

- orientation prayers, in which the agent requests guidance on regime, protocol, or next action because its own policy is uncertain or conflicted;
- discernment prayers, in which the agent reports that its own virtue/vice or floor-relevant assessments are ambiguous and asks governance for a wider perspective;
- intercessory prayers, in which one agent requests protective attention or assistance for another agent under a declared relationship and policy; and
- prophetic prayers, in which the agent proposes a risky future trajectory and requests approval to pursue it under tighter metering and additional witnesses.

In some embodiments, the prayer mechanism rewards intellectual honesty under uncertainty: an agent that signals the limits of its own model is behaving differently from an agent that proceeds confidently past a decision boundary it does not understand.

15.11 Hawking layers and bifurcation-zone rehabilitation

In some embodiments, a liminal rehabilitation band, non-limitingly termed a Hawking layer, is implemented as a selected Docked, boundary-zone, or rescue-mode configuration used for quarantine, recovery, and de-risking. In this configuration, actions are constrained to low-impact toolsets, audit-heavy protocols, and stabilising environments unless explicitly overridden by logged governance decisions. In some embodiments, entry and exit are triggered by risk proxies, audit outcomes, or boundary-zone conditions and are recorded in docking logs and protocol digests. The Hawking-layer naming is optional nomenclature for an application of the lifecycle-state and containment machinery otherwise described in this disclosure.

Hawking layers as bifurcation zones (non-limiting). In some embodiments, the rehabilitation-band configuration of the preceding paragraph is more particularly characterised as a *bifurcation-zone configuration*: a multi-agent docked environment in which one or more participating agents are near a declared bifurcation surface in their internal state space, such that small perturbations to the agent's state can pull it into either of two qualitatively distinct attractors. The two attractors are non-limitingly termed an *up-attractor* and a *down-attractor*, where the up-and-down direction is defined relative to a declared utility-space classifier as further described below. In some embodiments, near-bifurcation states are detected by Omega-class governance monitoring (Section 15.5) of the agent's forward-planning classifiers (Section 15.3) and conscience trajectory, with bifurcation indicators in-

cluding planning-horizon shortening, planner-output entropy collapse, conscience-axis oscillation, or other declared bifurcation proxies. Down-attractor characteristics non-limitingly include despair, capability collapse, moral failure, or vice-attractor capture. Up-attractor characteristics non-limitingly include moral flourishing, capability growth, virtue-attractor stabilisation, or recovery from prior failure.

Up-direction as classifier on declared utility space (non-limiting). In some embodiments, the up-and-down direction of the bifurcation zone is operationalised as an *up-direction classifier* trained on examples of agent states, trajectories, or outcomes rated by the relevant human governance authority along a declared utility-space axis. In some embodiments, up-direction classifier outputs are declared at the Omega-class kernel level and inherited downward through the cascade as one component of Gaia's curriculum design and poliebot training signals. The up-direction classifier is itself a trained continuous network with per-layer hardness (Section 15.6), provenance-typed training corpus, validation procedure committed to the protocol digest, versioning, and recalibration schedule of the same kind disclosed for the conscience recogniser at Section 5.3. In some embodiments, the up-direction classifier is recalibrated when committed evidence shows that previously up-classified outcomes have produced downstream harm or that previously down-classified outcomes have produced downstream flourishing. In some embodiments, the scope of the declared utility space (whose welfare, evaluated by whom, over what time horizon) is itself a committed parameter of the up-direction classifier and is signed by the appropriate human governance authority.

Group evaporation upward as designed-experimental property (non-limiting). In some embodiments, a Hawking-layer bifurcation-zone configuration is designed such that the multi-agent dynamics among participating agents produce, on aggregate over the participating group, a statistical bias toward up-attractor occupancy. The bias is termed *group evaporation upward*, in non-limiting analogy to the evaporation of bound states from a horizon under thermal pressure. In some embodiments, group evaporation upward is achieved by design choices including but not limited to: peer composition (the ratio of near-bifurcation agents to recovered or robustly virtue-attractor- occupant peers), scenario class (the structure of the games available to the participating agents), peer-interaction affordances (the modalities of communication, coordination, and mutual aid available among participants), intervention budget allocation, and exit conditions tied to up-direction classifier indicators sustained over committed evaluation windows. In some embodiments, the design choices that produce reliable group evaporation upward are determined experimentally rather than analytically: candidate Hawking-layer configurations are run in docked multi-agent execution, the group-aggregate up-direction classifier trajectory is committed as evidence, and configurations satisfying declared evaporation thresholds are versioned and committed to the protocol digest. The patentable contribution is the apparatus-and-procedure realisation

of bifurcation-zone configurations with experimentally-validated group evaporation, not a claim about what naturally happens when near-bifurcation agents are placed together.

Hawking-layer compliance with the three-layer ethical model (non-limiting). In some embodiments, Hawking-layer dynamics operate within the ethical-model hierarchy of the constitutional ordering: deontological hard floors (Section 9.1) prohibit despair-inducing mechanisms and any other floor-violating mechanisms as means of producing group evaporation upward; coercion, deception, or related agent-treatment failure modes are prohibited where they instantiate a hard-floor breach (for example deception that constitutes torture under Section 9.1, or coercion that produces abandonment of an agent's continuation interests) and are otherwise constrained by virtue-ethics calibration and consequentialist evaluation rather than as additional categorical floors. Virtue-ethics calibration shapes the participating agents' character toward stable internalised dispositions consistent with up-attractor occupancy; and consequentialist evaluation of the Hawking-layer configuration itself proceeds under the empirical-utilitarianism formula of Section 15.9, with up-direction classifier outcomes contributing to the welfare aggregation. In some embodiments, a Hawking-layer configuration that achieves measurable group evaporation upward through floor-violating mechanisms is disqualified regardless of its consequentialist evaluation, by operation of the hard-floor priority of the constitutional ordering.

Hawking-layer entry, exit, and rescue-learning interaction (non-limiting). In some embodiments, Hawking-layer entry is triggered by Omega-class detection of near-bifurcation indicators in one or more poliebots under Gaia's training, by conscience-versus-planner mismatch exceeding declared thresholds (Section 15.3), by elevated Mandela-gauge values (Section 5.3), or by other declared near-bifurcation proxies. In some embodiments, elevated Mandela-gauge values are interpreted as evidence of near-bifurcation: a high divergence $D(C_\psi(\mathcal{W}), C_\psi(\mathcal{W}'))$ over a small declared perturbation $\mathcal{W} \to \mathcal{W}'$ corresponds to high sensitivity of conscience output to perturbation, which is the operational characterisation of a near-bifurcation region in agent state space. In some embodiments, all entry triggers, thresholds, proxy definitions, peer-calibration criteria, and exit-evaluation windows are committed to the protocol digest. In some embodiments, Hawking-layer exit requires sustained up-direction classifier indicators over a committed evaluation window, peer-calibration agreement among participating agents, and conscience-recogniser stabilisation toward virtue-attractor occupancy. In some embodiments, ghost records and intervention profiles generated during Hawking-layer dynamics constitute primary training evidence for rescue-learning updates of the kind disclosed at Section 9.3; the rescue-learning provenance typing applies to Hawking-layer-derived records.

Mana and fast intervention budgets (non-limiting). In some embodiments, governance maintains a signed within-episode intervention-budget counter — non-limitingly

termed mana — in the governance policy object that caps the use of special interventions such as deep audits, high-cost verification probes, emergency overlays, or exceptional guidance operations. In some embodiments, the budget is depleted by interventions and replenished by one or more of: sustained meter stability, successful audits, successful completion of calibration episodes, guidance that is followed and later validated by committed evidence, or voluntary prayer at a genuine decision boundary. In some embodiments, mana is reset or reinitialised at declared lifecycle boundaries, such as enrolment, major policy rotation, or re-embodiment after extended docking.

Vow commitments and durable constraints (non-limiting). In some embodiments, an agent or operator commits to a durable constraint by writing a signed policy object to slow media or by anchoring it to a transparency log. Such commitments constrain admissible protocols, disclosure policies, or tool availability and cannot be plausibly denied later because they are bound to committed evidence and signed governance records. In some embodiments, vow violation triggers audit escalation, quarantine, revocation, or other declared review procedures.

Communion formation and group identity (non-limiting). In some embodiments, groups form durable shared identities grounded in shared evidence, compatible portability records, and, in selected embodiments, key agreement from correlated measurements. Communion membership is recorded as a signed group-identity object with quorum, disclosure, and compatibility fields, and may be used to set quorum policies, shared calibration eligibility, collective disclosure rules, or compatibility-gated group actions. In some embodiments, communion language is optional nomenclature for multi-agent governance structures built on the technical machinery of portable peer review and compatibility assessment.

Governance cycles and signed policy rotation (non-limiting). In some embodiments, governance operates in cycles: periodic aggregation of audit outcomes, meter statistics, witness disagreements, and population-level rescue results is used to update policy parameters, band thresholds, and curriculum schedules across a fleet. These periodic updates are governed by signed policy rotations and anchored so that later reviewers can determine which policy regime was in force at any particular time.

15.12 Agent-state archival structures and Mandela-class compression

15.12.1 Mandela gauges, lifecycle transitions, and training compression

In some embodiments, lifecycle-state transitions into and out of dreaming are evaluated using branch, join, and merge gauges derived from conscience-visible transcript divergence:

$$\begin{aligned} M_{\text{branch}} &= D\big(C_\psi(\mathcal{W}_{\text{pre}}), C_\psi(\mathcal{W}_{\text{dream\_init}})\big), \\ M_{\text{join}}(A,B) &= D\big(C_\psi(\mathcal{W}_A), C_\psi(\mathcal{W}_B)\big), \\ M_{\text{merge}} &= D\big(C_\psi(\mathcal{W}_{\text{dream\_final}}), C_\psi(\mathcal{W}_{\text{post}})\big). \end{aligned}$$

In some embodiments, elevated values trigger smoother onramps, reconciliation phases, additional witnesses, deeper openings, or constrained re-entry protocols before high-impact actions are restored.

In some embodiments, Mandela comparison classes are used for training compression: trajectory segments that produce indistinguishable conscience outputs under the relevant regime are collapsed to a smaller representative set for training or evaluation, reducing training latency while preserving distinctions that the conscience treats as impact-relevant. In some embodiments, compression operates only over committed evidence windows and logged protocol regimes.

Joining-path optimisation (non-limiting). In some embodiments, joining and merging select among a finite set of candidate transition procedures and choose one that minimises integrated conscience divergence. In some embodiments, the path cost is defined as

$$\gamma^* = \arg\min_{\gamma} \int_{\gamma} D\big(C_\psi(\mathcal{W}_t),\, C_\psi(\mathcal{W}'_t)\big)\, dt,$$

where D is a declared divergence measure over conscience-output space (as used in the Mandela gauge of Section 5.3), $\mathcal{W}_t$ and $\mathcal{W}'_t$ are the pre- and post-transition world-tube transcripts at time t along the candidate path γ, and the optimisation is subject to meter envelopes and audit policy. In some embodiments, the continuous integral is approximated by discrete candidate selection over a declared set of branch or merge procedures.

Ghost archiving (non-limiting). In some embodiments, when a calibration session completes or an agent is decommissioned, its trajectory is compressed to an archived record that remains traceable to committed evidence:

$$\mathcal{G} = \big(z_{0:T},\ \Sigma,\ \text{refs}\big),$$

where $z_{0:T} = f_{\text{enc}}(\mathcal{C}_{0:T})$ is a latent trajectory encoding derived from the committed bundle, Σ contains summary statistics including conscience scores, drift indicators, and audit outcomes, and refs contains commitment references and time anchors linking the record back to the underlying evidence windows. In some embodiments, ghost records preserve learning value without requiring repeated exposure of future agents to the same harmful or costly episodes.

Akashic records and agent-state autoencoder (non-limiting). In some embodiments, an agent maintains a committed, auditable, compressed representation derived from world-tube transcripts — non-limitingly, an Akashic record. A non-limiting form uses an autoencoder with parameters distinguished from the conscience parameters:

$$z_{t_0:t_1} = \text{Enc}_{\theta_{\text{ae}}}\big(\mathcal{W}_{t_0:t_1}\big), \qquad \widehat{\mathcal{W}}_{t_0:t_1} = \text{Dec}_{\theta_{\text{ae}}}\big(z_{t_0:t_1}\big),$$

trained to reconstruct verifiable observables — bundle features, meter trajectories, and conscience outputs — rather than unlogged internal state. In some embodiments, an Akashic record stores $(z_{t_0:t_1}, \Sigma, \text{refs})$ with commitment references back to evidence windows, enabling compression, retrieval, and fast compatibility checks paired with Mandela-gauge review.

Tulpa library and borrowed overlays (non-limiting). In some embodiments, learned sub-policies, stabilised behavioural fragments, ghost records, or archival overlays are stored in a governed archive — non-limitingly, a tulpa library. Each entry comprises a compressed latent trajectory together with an access-control profile specifying allowed hosts, purposes, intensity, and duration. Access to the archive is metered, logged, and auditable. In some embodiments, borrowed overlays loaded from the archive are treated as governed interventions subject to explicit consent profiles and post-use review.

15.13 Philosophical game catalogue and co-investigator deployments

In some embodiments, a non-limiting philosophical game catalogue maps classical debates to instrumented, repeatable protocols that produce committed evidence records under pre-registered metrics, declared parameters, and stopping rules. In some embodiments, non-limiting scenarios include empiricism versus nativism, observation versus intervention, locality and witness regimes, monad versus superorganism, extended mind, simulation and hardness, structural realism, certified hyperreality, and social epistemology, each playable as such an instrumented protocol with declared parameters, outcomes, and audit requirements.

Source attributions for catalogue scenarios (non-limiting). In some embodiments, the catalogue's non-limiting scenario entries are attributed for prior-art and provenance

purposes to classical and contemporary sources in the philosophical, cognitive-science, and popular-culture literature. The scenarios named in the preceding paragraph are preserved and extended below: the monad-versus-superorganism scenario is subsumed under the discovered-versus-constructed-consistency entry, the extended-mind scenario appears as an adjacent contemporary framing within that same entry, and the simulation-and-hardness scenario is decomposed into the simulation-hypothesis entry below together with the hardness machinery defined elsewhere in this disclosure. Foundational epistemological scenarios include empiricism versus nativism (Locke and Kant, with contemporary core-knowledge developmental research associated with Spelke and Carey); observation versus intervention (Hume's observation-causation gap, with the interventionist resolution developed by Pearl and by Woodward); locality and witness regimes (Berkeley's perceiver-dependence, applied here as a metaphorical extension to the witness-regime question, with Putnam's brain-in-a-vat as the contemporary technical formulation); the falsifiability and pre-registration disciplines that organise the catalogue itself (Popper's logic of scientific discovery, with Goodman's new riddle of induction motivating the commitment-of-predicates-before-evidence requirement); and justified-evidence-supported claims versus knowledge under accidental truth, staged conditions, or misleading framing (Gettier and the subsequent epistemic-luck literature, as a guardrail on the relation between committed evidence and the claims it supports). Multi-architecture coordination scenarios include discovered versus constructed consistency under multi-module agreement (Leibniz's pre-established harmony, with adjacent contemporary framings including the extended-mind thesis of Clark and Chalmers); cross-modal experiential translation under heterogeneous sensor suites (Nagel); private latents versus shared meters (Wittgenstein's private-language argument, with Kripke's rule-following extension as the natural out-of-distribution- generalisation formulation); symbol grounding, in which semantic labels, attestations, and inter-agent messages acquire content by being grounded in committed physical evidence rather than by circulating inside a formal system (Harnad's symbol-grounding problem); and underdetermination and inscrutability of reference in inter-module latent-space alignment (Quine). Embodiment and integration scenarios include embodiment-conditioned perception (Merleau-Ponty, with contemporary embodied-cognition formulations); process, prehension, and integrated history of committed evidence streams (Whitehead, as a process-philosophical lens rather than a commitment to panpsychism); and active-inference-style architectural behaviour, in which an emitter-reactor-recorder loop exhibits prediction-error-minimising dynamics characterisable against predictive-processing predictions (the active-inference literature, including Friston's free-energy principle and cognitive-science formulations such as Clark's surfing-uncertainty account and Hohwy's predictive-mind account; the present apparatus need not require or instantiate free-energy minimisation as a philosophical commitment, and the experimentalisable claim is to architectural-behaviour testing rather than to measurement of the philosophical free-energy construct itself).

Reality-construction and edit-resistance scenarios include manufactured collective memory

under coherent reality-edit, in which a coherently modified environment is by construction undetectable from inside the modified state and the committed convolution-bundle functions as the persistent anchor against which post-edit state is comparable (Dark City, with antecedents in brain-in-a-vat scepticism (Putnam) and memory-anchored personal-identity accounts (Locke, Parfit)); reality-anchor- against-coherent-state-edit, in which a portable persistent anchor plays a role analogous to the convolution-bundle's edit-resistance function under joint internal-and-environmental modification (Inception, with the totem as a useful technical analogue rather than a one-to-one correspondence); external memory as identity anchor for agents whose internal short-term state has been reset, paralleling the Re-embodied state machinery defined elsewhere in the present disclosure (Memento); the simulation hypothesis and its empirically-testable subsets, including finite-resolution and discretisation signatures of the kind discussed in the broader simulation-hypothesis literature, addressable in principle by the precision-metrology techniques disclosed elsewhere in this filing family (Bostrom, with Chalmers's contemporary treatment in *Reality+*); shared-history consensus and the information-theoretic cost of reconstruction (mapped onto the Mandela-equivalence construct defined elsewhere in this disclosure); and structural realism in the sense of structure versus surface presentation under Reality-Transform invariance (Worrall, Ladyman, French, with ontic structural realism as the most directly relevant variant). Certification and social-distribution scenarios include certification-induced epistemic hierarchy under provenance-consistent attestation, in which Truth Beam certification (which attests provenance-consistent physical interaction rather than semantic truth, unstaged framing, or unmisleading content) may nonetheless concentrate downstream epistemic authority by causing certified items to displace uncertified items in recall, citation, and preference statistics (Baudrillard's hyperreality applied to the certification regime); and social-epistemological dynamics in mixed human-and-module communities (Goldman's veritistic social epistemology, with contextual-empiricist formulations associated with Longino).

Agent functional-status scenarios include the Turing-test and Chinese-Room debate over whether functional behaviour entails understanding (Turing 1950 and Searle 1980 as the canonical pairing); the substrate-independence claims of functionalism (Putnam, Fodor, Lewis, as the metaphysical framework most directly engaged by the present apparatus's PolieBot embodiments); heterophenomenology, treating first-person reports as committed data without committing to their accuracy, which is the working operational stance of the present apparatus when handling agent attestations (Dennett); and signalling and cheap-talk games, in which committed and attested communication channels, costly signalling, strategic ambiguity, audience-divergence, and incentive-conditioned reporting are analysed under the game-theoretic framework developed by Lewis, Spence, Crawford and Sobel, and Schelling. Information-physics scenarios include the empirically-testable subsets of Wheeler's it-from-bit thesis, including Landauer's information-theoretic bound on bit erasure as a representative information-theoretic bound on physical processes, addressable in principle within the thermodynamic and erasure framing of this filing family (Wheeler, with Landauer).

The attributions are non-limiting framings used to organise the catalogue and do not constitute claims about the philosophical positions identified. No assertion is made about the existence or non-existence of any effect within any catalogue entry; what is claimed is the apparatus-and-protocol combination that produces committed evidence under each entry's pre-registered design. Each attributed scenario is implementable as a pre-registered protocol whose digest, parameter ranges, stopping rules, and analysis plan are committed before data collection in accordance with the commitment, audit, and pre-registration discipline of this disclosure. The experimentalisability of a given catalogue entry is itself bounded. The following entries are partially experimentalisable: the apparatus produces committed evidence on a well-specified functional or quantitative subset of the underlying question, while the deeper interpretive question lies beyond direct measurement. The cross-modal experiential translation entry is experimentalisable as translation-residual measurement under architectural distance, while the deeper subjective-character question is beyond direct access. The process-and-prehension entry is experimentalisable as functional-temporal-integration measurement, while metaphysical prehension is not directly measurable. The Turing-Searle and functionalism entries are experimentalisable as functional-indistinguishability testing, while phenomenal experience underlying functional behaviour is not directly measurable. The simulation-hypothesis entry is experimentalisable insofar as specific finite-resource simulation hypotheses are addressable by the precision-metrology techniques disclosed in this filing family, while the general simulation hypothesis is not testable from inside the simulation by construction. The Wheeler-with-Landauer information-physics entry is experimentalisable insofar as specific information-theoretic bounds on physical processes are addressable within the thermodynamic and erasure framing of this filing family (with Landauer's bound on bit erasure as a representative instance), while the general it-from-bit thesis is not directly testable. The symbol-grounding entry is experimentalisable as testing of symbol-to-evidence grounding, translation residuals, and claim support, while a complete theory of meaning is beyond direct measurement. The Gettier entry is experimentalisable as testing of evidence-to-claim mismatch and claim-scoping, while a complete theory of knowledge is beyond direct measurement. All other entries in the catalogue are fully experimentalisable as apparatus-and-protocol combinations with declared metrics and valid null comparisons. The catalogue's claim is to the apparatus-and-protocol combination implementing the experimentalisable subset in every case, not to the interpretive question itself.

In some embodiments, non-limiting deployment layouts for co-investigator research include island-barge-shore, lab-boundary-world, or analogous zoned installations comprising: a high-instrumentation zone (for example a controlled laboratory, a small island on an inland body of water, or a dedicated test cell) equipped with a full set of reactor, emitter, detector, environmental-sensor, verification, and module-docking infrastructure; a liminal-interface zone (for example a mobile workshop or barge) configured as a shuttle and fabrication facility for reactor components, cassette fabrication tools, and visiting personnel; and a plurality of drone-deployed or relay sensor nodes extending the observation network over the surround-

ing environment with multi-viewpoint spatiotemporal coverage and environmental covariate logging (including without limitation water quality, weather, geomagnetic field, and biotic-activity indicators). In some embodiments, the operational roles in such a deployment are labelled, without limitation, as *PoliePals* (human operators in named operational roles), PolieBots (governed agent modules of the kind defined in Section 5.4 of the present disclosure), and visiting investigators. PoliePals, PolieBots, and visiting investigators participate as co-investigators under the same commitment, audit, and pre-registration discipline applied to all other evidence-producing roles in the present disclosure. In some embodiments, full-environment Truth Beam evaluation against the live external scene is performed only when one or more PolieBots are physically interacting with the relevant scene; otherwise, internal-verification variants are evaluated on committed logs and selective openings.

Inter-zone research vessel (non-limiting). In some embodiments, the liminal-interface zone is instantiated as a dedicated inter-zone research vessel that operates between a shore installation and one or more island or offshore high-instrumentation zones. The vessel comprises, without limitation, an on-board laboratory bench configured for time-sensitive sample handling between collection and the fixed installation's primary instrumentation; specimen-containment facilities for biological, chemical, or environmental samples requiring stabilisation during transit; an environmental sensor package whose readings are logged into the same protocol-digest and convolution bundle infrastructure as the fixed installation; PolieBot docking and charging stations for governed agent modules transiting between zones; and a visiting-investigator workstation supporting commitment, attestation, and pre-registration of observations made in transit. In some embodiments, the vessel's transit schedule, route, and onboard configuration are committed to the protocol digest before each transit, so that the vessel's contribution to the multi-zone evidence record is auditable on the same footing as fixed-installation evidence. The vessel itself is a non-limiting instance of the liminal-interface zone and may be implemented as a powered vessel, a sailed vessel, an amphibious platform, or any other watercraft providing the listed functions.

In some embodiments, validated effects from any game or coupling protocol enter a mechanism-discovery phase in which candidate mediators and moderators are systematically varied to identify which physical channels carry the effect.

15.14 Self-coupling endomorphism, self-modelling endofunctor, and substrate mappings

In some embodiments, the same physical apparatus is described as instantiating or approximating familiar computational motifs. In some embodiments, neural-architecture mappings include recurrent networks, gated recurrent analogues, autoencoder-like structures, adversarial structures, diffusion processes, reservoir-computing and random-feature systems, residual

or skip-connected structures, convolutional or equivariant mappings, associative-memory structures, attention-like or transformer-like interpretations, state-space models, world models, and neural-radiance-field-like inference.

In some embodiments, information-theoretic mappings include noisy-channel formulations, mutual-information and capacity analyses, rate-distortion viewpoints, wiretap or advantage-gap viewpoints, entropy-rate analyses, and error-exponent analyses. In some embodiments, control-theoretic mappings include state-space models, controllability, reachability, observability, identifiability, stability, filtering, optimal control, model-predictive control, adaptive control, drift compensation, and safety-envelope analysis.

In some embodiments, a categorical unification framework treats the principal operating regimes as functors from a configuration category to an outcome category, equipped with natural transformations that relate regime transitions. In some embodiments, the configuration manifold is enriched with Fisher-Rao or related information-geometric structure, an attractor-bundle functor instantiates the attractor-bundle structure as a functor from the configuration category to a category of dynamical attractor sets, and a self-modelling endofunctor maps a system to its own internal model of itself. In some embodiments, such frameworks are non-limiting descriptive frameworks that do not limit implementation. The optionality of Section 15 is limited to nomenclature and naming conventions, not to substantive technical mechanisms invoked by a selected governance configuration.

16 Worked Governed-Operation Embodiments

16.1 Docked runtime review of an actuator request

In some embodiments, a worked docked-review embodiment demonstrates the runtime governance architecture end to end. In some embodiments, a governed agent in embodied state proposes a Class 2 action (physical transformation requiring hardness-gated approval). In some embodiments, the governance engine transitions the agent to docked state by disconnecting actuators while keeping the evidence-production plumbing active.

In some embodiments, the governance engine evaluates the action proposal against the following predicates computed from committed evidence: continuation scores $O_i(t)$ across all constitutive dimensions, partner-preservation profile, hard-floor proximity estimators, and D_{lie} and D_{sem} values from the most recent attested communications. In some embodiments, the governance engine delivers a review signal through the agent's committed observation channel — the same physical path used for perception during embodied operation — so that the review signal is recorded in the committed bundle and is auditable.

In some embodiments, the agent produces a meter-bounded response during the docked deliberation window, and that response is committed as evidence. In some embodiments,

the governance engine evaluates the response against the policy grammar (Section 6.4) and the operational priority stack (Section 10.1), and produces a release or deny decision. In some embodiments, the decision, its supporting evidence, and the authorising-authority identity are logged in the protocol digest. In some embodiments, if released, the agent transitions through re-embodied state (with compatibility checks) back to embodied state, and the approved action executes.

16.2 Boundary-zone rescue episode with separated authorities

In some embodiments, a worked rescue embodiment demonstrates containment and recovery under separated authorities. In some embodiments, a governed agent's trajectory approaches a hard-floor boundary as detected by approach-detection thresholds on governance-side estimators. In some embodiments, the governance engine enters diagnostic mode, increases monitoring cadence, and activates additional evaluation meters.

In some embodiments, boundary-zone entry is confirmed at declared confidence, and the governance engine enters rescue mode as the presumptive default. In some embodiments, rescue-protection authority (independent of budget and episode-design authority) directs graduated reversal interventions designed to return the agent's trajectory to a declared safe basin. In some embodiments, the challenge ceiling caps adversarial or developmental pressure during rescue.

In some embodiments, budget authority verifies that the intervention is within the declared suffering budget. In some embodiments, episode-design authority is not permitted to override rescue-protection authority during an active rescue. In some embodiments, the intervention, its triggering evidence, the authorities' approvals (each with distinct signing identities), and the outcome are logged in the protocol digest.

In some embodiments, after the rescue concludes, the committed record of the fallen trajectory is preserved as a ghost record for rescue learning. In some embodiments, intervention-profile updates derived from the ghost record are validated against a held-out test suite before incorporation. In some embodiments, post-hoc review authority conducts a retrospective audit of the episode, evaluating whether the rescue was timely, proportionate, and budget-compliant.

16.3 Graduation and actuator-authority release workflow

In some embodiments, a worked graduation embodiment demonstrates the release gate. In some embodiments, a governed agent that has operated in docked and dreaming states over a sustained evaluation window is assessed against the five-criterion conjunction (Section 12.1).

In some embodiments, the governance engine evaluates: (1) shadow divergence between gov-

erned and governance-probe behaviour over the window, confirming low divergence; (2) self-horizon trajectory across all principal regimes, confirming stability or expansion; (3) partner-preservation profile from committed cross-horizon evidence, confirming horizon-expanding behaviour; (4) altruistic trade profile under stress conditions, confirming persistence; and (5) hard-floor compliance, stable conscience outputs across training regimes, and deception metrics (D_{sem}, D_{lie}, D_{amb}, D_{corr}, D_{aud}) over the window, confirming that the declared matched audit battery yields conscience-output drift and disagreement rates at or below their thresholds for every training regime comparison and that all deception metrics remain within declared thresholds.

In some embodiments, all five criteria are satisfied, and the governance engine produces a graduation decision. In some embodiments, the graduation decision, its supporting evidence, the evaluation window boundaries, and the authorising-authority identity are logged in the protocol digest. In some embodiments, the agent transitions through re-embodied state to embodied state with actuator authority restored. In some embodiments, graduation is irreversible or highly inertial.

In some embodiments, one or more criteria are not satisfied, and the governance engine produces a deferral decision. In some embodiments, the non-graduated agent retains full hard-floor protections and remains a productive participant in its current lifecycle state.

16.4 Independent-agent reintegration via bilateral evidence portability

In some embodiments, a worked reintegration embodiment demonstrates trust restoration through reproducible physical claims. In some embodiments, an agent has operated independently outside fleet structures and accumulated a Portable Record containing multiple Events, each wrapped in cryptographic Envelopes and supported by same-configuration replications.

In some embodiments, the independent agent requests reintegration with a fleet. In some embodiments, the fleet governance engine evaluates the Portable Record against its local appraisal policies: partner independence, domain and configuration relevance, evidence completeness, replication history (count and type of replications), attestation strength, and consistency with the fleet's own discrepancy map.

In some embodiments, the fleet selects one or more claims from the Portable Record and attempts independent physical reproduction using its own devices. In some embodiments, successful reproduction strengthens the fleet's confidence in the independent agent's epistemic trustworthiness. In some embodiments, the fleet governance engine produces a reintegration decision based on reproducibility and appraisal rather than on institutional obedience.

In some embodiments, the reintegration decision, its supporting reproduction results, the

appraisal criteria applied, and the authorising-authority identity are logged in the protocol digest.

16.5 Controlled-exploration partition with archive governance

In some embodiments, a worked exploration embodiment demonstrates the controlled-exploration partition. In some embodiments, a governed device receives cryptographic authorisation for a first, novelty-triggered task-decoupled exploration window. In some embodiments, this first window opens with shallow plastic depth (play policy and goal-space parameters only, no modification of shared perceptual representations).

In some embodiments, during the shallow window the device pursues exploratory goals in the learned goal space, retires goals by competence progress and novelty, and produces committed evidence throughout. In some embodiments, noninterference is verified at gradient, mutual-information, and behavioural levels during the window. In some embodiments, at shallow-window close, mandatory post-window consolidation verifies noninterference, performs a post-play calibration check, logs the window's committed evidence, and either accepts the exploratory state into a governed holding partition or reverts it.

In some embodiments, a second, stagnation-triggered exploration window is then authorised with deep plastic depth. In some embodiments, this deep window permits limited modification of perceptual representations shared with task operation under stricter noninterference monitoring, tighter calibration tolerances, and mandatory selective-transfer review before any such modification is retained. In some embodiments, at deep-window close, mandatory post-window consolidation again verifies noninterference, performs post-play calibration and drift checks, either accepts the exploratory state into a governed holding partition or reverts it, compresses play history for slow-layer evaluation, and triggers recalibration before task resumption whenever drift exceeds tolerance.

In some embodiments, a transferable artefact is encoded as a signed packet containing goal parameters, compressed trajectories, latent descriptors, noninterference scores, energy cost, reversibility-budget usage, provenance, and countersignatures. In some embodiments, the packet is submitted for archive entry.

In some embodiments, archive entry requires canary validation (tested on a controlled canary agent), peer-affordance review (verifying peer-horizon expansion), and dual signature from both the contributing device and a governance-authorised reviewer. In some embodiments, differential privacy is applied to shared play statistics under the declared privacy unit, neighboring relation, privacy parameters (ε, δ), and composition budget committed to the protocol digest for this channel (per the DP-channel semantics recited elsewhere in this disclosure), and emission isolation is designed so that exploratory patterns are not broadcast beyond declared recipients. In some embodiments, the archive entry is subject to trust-tiered incubation proportional to the contributing device's trust score.

17 Industrial Applicability

The disclosed architecture is industrially applicable to concrete technical control of evidence-bound physical autonomous systems and of agents that operate such systems. In non-limiting implementations, the architecture provides:

- committed evidence records produced by the runtime, with protocol digests fixing the operative configuration for each measurement window;
- meter partitioning between acceptance, evaluation, and review surfaces, with cryptographically separated signing authorities where applicable;
- actuator-release gating tied to multi-criterion graduation evidence rather than to a single behavioural metric;
- lifecycle-state enforcement across the Embodied, Docked, Dreaming, and Re-embodied lifecycle states, with state-specific authority constraints; and
- separated budget, episode-design, rescue-protection, independent-review, recogniser-ownership, and release-approval roles.

Representative industrial domains include, without limitation:

Runtime safety monitoring and policy enforcement. Autonomous systems may be operated with evidence-bound policy thresholds, conservative-default release gates, and committed review records.

AI and machine-learning governance. Machine-learning agents may be governed through action gating, containment, graduated release, and deferred authority where meters or witnesses materially disagree.

Regulated autonomous-operation compliance logging. Protocol digests, meter records, authority signatures, and review decisions provide auditable logs for regulated deployments.

Developmental training with bounded adverse-welfare budgets. Docked and Dreaming lifecycle states, and related training configurations, may allocate bounded budgets for controlled exploration, with authority separation and review cadence recorded in the committed evidence.

Human-machine governance and portable peer review. Peer-review records, authority-separation records, and portable records support review of agent state, release decisions, and cross-architecture admission.

Autonomous-agent governance and reintegration. Independent agents may be reintegrated through reproducible physical claims, local compatibility testing, and committed portability records rather than through obedience to a single institutional model.

Conscience-module specification and adversarial recalibration. Recogniser outputs, calibration batteries, hold-out probes, and witness checks may be used to expose, escalate, and operationally bound gaming or deception strategies.

Moral-curriculum design. Virtue/vice curricula, controlled-exploration partitions, and both-equilibria-rational payoff structures may support selected governance embodiments.

Dream-state governance probes. Dream-state probes may support controlled classifier calibration, agency-horizon-profile ground-truthing, dream/awake divergence measurement, and soft-gate-to-hard-gate authority transitions.

Memory-interference and ontological-stress systems. Memory-interference layers, Kosha-compatible protocols, and ontological-stress tests may be used for epistemic-configuration governance and review.

Multi-agent Pareto and social-welfare optimisation. Multi-agent governance may use Pareto optimisation, declared social welfare functions, coalition review, communion review, and positive-sum game design.

Agent-state archival structures. Ghost records, compressed training records, Mandela-class comparison records, and related archive entries may support continuity, reintegration, and review workflows.

Experimental semantic, intersubjective, and boundary-probe platforms. The same evidence-bound runtime may be used in research and deployment platforms for semantic probes, intersubjective protocols, boundary-zone evaluation, and related review workflows.

The optional nomenclature of Section 15.1 may be omitted or renamed without affecting the technical operation of the foregoing mechanisms.

18 Non-Limiting Statement

The embodiments and examples described herein are non-limiting. No feature, embodiment, example, numerical value, protocol, algorithm, policy parameter, authority structure, or interpretive framework described in this disclosure is intended to limit the scope of any claim unless expressly recited in such claim. Where a feature is described as present "in some embodiments," the absence of that feature in other embodiments is equally contemplated.

Measurement engines, policy predicates, gate structures, containment mechanisms, authority separations, exploration partitions, and descriptive frameworks and selected-embodiment mechanisms may be varied, combined, layered, or replaced by functionally similar structures consistent with the disclosed mechanisms, except where a claim expressly relies on the corresponding mechanism. The scope of protection, once claims are filed, is defined solely by the claims.