Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Predictive Safety Execution Kernel for External Planners: System Implementation and Independent Recomputation under a Frozen Protocol

Jiwei Wang

2446323921@qq.com

Yanching Institute of Technology

Kaihang Ma

Yanching Institute of Technology

Research Article

Keywords: software quality assurance, runtime verification, safety-critical software, large language model planner, empirical evaluation, independent recomputation

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11778674/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Predictive Safety Execution Kernel for External Planners: System Implementation and Independent Recomputation under a Frozen Protocol

Jiwei Wang[1*] and Kaihang Ma[1]

[1*]School of Information Science and Technology, Yanching Institute of Technology, Sanhe 065201, Hebei, China.

*Corresponding author(s). E-mail(s): 2446323921@qq.com;
Contributing authors: 13032695021@163.com;

Abstract

Hengjie is a runtime quality-assurance mechanism for control software driven by external planners, including large language models (LLMs). Because planner output is untrusted, release is gated: the kernel releases a candidate only after robust predictive verification against state uncertainty and the committed queue, with a certified fallback. For the discrete-time model, every released action remains covered by the execution invariant under the premises, for any proposer. The characterisation is a pre-registered, frozen-protocol study of a simulated thermal process under bounded faults (424 executable cells, failures and counterexamples retained). The protected configurations (legacy B, 34 cells; B2, 88 cells) recorded 0 violation minutes and 0 uncovered-authorisation minutes; the unprotected configuration recorded 25,792 and 34,433, respectively. Under one-slot delay the protected paths stayed safe but completed 10–11 of 1,440 task minutes. A pre-registered budget study re-ran the reviewed LLM cells at 4× and 16× the frozen budget and cut fallback share from 0.964 to 0.507, but left the task ratio identical (0.9981; non-LLM configuration 0.9983): budget changed which controller holds the band, not the outcome. A control-barrier-function safety filter wrapping the same proposer, run as an exact replica in the same slots as a twelve-cell engineering baseline, recorded zero violation minutes and held the disconnection cells where the heuristic chain B collapsed. Decision latency: median 1.32–1.37 s per 60 s tick. All frozen-batch quantities were independently recomputed from a public replication package with zero discrepancies; reproducing a published predictive-safety baseline deviated by +31.4%, the cause undetermined after pre-registered single-factor diagnosis.

Keywords: software quality assurance, runtime verification, safety-critical software, large language model planner, empirical evaluation, independent recomputation

1 Introduction

External planners, including large language models (LLMs), increasingly propose actions for physical equipment. A proposal that is reasonable at the information level can be unsafe at the equipment level: observations may be stale when they arrive, and issued instructions cannot be revoked before they expire, so every new plan acts together with the committed queue and the equipment's inertia. Hengjie separates proposal from authorisation. Planners only propose; an execution-layer kernel releases a candidate only after robust predictive verification accounting for current-state uncertainty and already-issued instructions. A failing candidate is corrected at minimum deviation; when none passes, the kernel falls back to a certified continuation instead of passing the risk to the equipment. The defence is layered: a format and privilege review first confines any proposal to the declared setpoint whitelist, and only the predictive certification layer can authorise release; the two layers answer different questions, and the results below keep them separate.

In software-engineering terms, the kernel is a runtime verification gate for AI-enabled software: it treats each planner proposal as unverified output of an untrusted component and subjects it to an acceptance check before release, in the spirit of runtime verification (Leucker and Schallhart 2009) and of runtime assurance for learned components. The study is conducted as an empirical quality evaluation: the protocol was frozen and registered before execution, failures and counterexamples were retained rather than discarded, and every reported quantity of the frozen batch was recomputed independently from a public replication package. The freeze timeline and the execution deviations from it are registered in Section 4.3 and Supplementary S7.

This paper is an implementation study of one system, using an identified thermal model, an LLM planner and a frozen protocol. At its centre is an execution-semantics-aware safety architecture: authorisation covers not only newly proposed actions but also already-committed actions whose delivery may be delayed, held or unresolved, so certificates remain aligned with device-level execution semantics. Its proposer-independent conditional safety result is stated over a branch-wise execution invariant representing unresolved delivery and hold possibilities explicitly (Section 3.5). The pre-registered empirical characterisation of the safety–progress trade-off shows that under one-slot delay the protected configurations stay safe while task progress collapses; a delay-tolerant revision restores legacy-B progress; and the B2 selector exposes a distinct candidate-exhaustion boundary (Sections 5.6, 5.7 and 6). The evidence base is the full experimental record: all 424 executable cells of the frozen matrix, with failures, counterexamples and engineering events retained. Because the experimental conditions differ in several components at once, the evidence supports joint-configuration statements only. The contributions are: (1) a runtime quality gate that certifies planner-issued plan segments against the committed queue, with certificates aligned to device-level execution semantics — validity windows, hold and delayed-delivery

branches — under communication faults; (2) a pre-registered empirical characterisation of its safety–progress trade-off, including a negative result for the tested LLM planner under the tested budget; and (3) a public replication package whose independent recomputation reproduced every reported quantity of the frozen batch with zero discrepancies.

2 Related work and contribution boundary

Hengjie's predictive verification is related to robust predictive safety filtering within the model predictive control framework (Mayne et al. 2000), where candidate-action safety is guaranteed, under each method's stated assumptions, through robust invariant sets or interval predictions; representative neighbours include the robust predictive output-feedback safety filter (RPOF-SF) (Brunke et al. 2022) and adaptive model predictive safety certification (AMPSC) (Didier et al. 2021). Safety filtering based on control barrier functions (CBFs) (Ames et al. 2017), including Self-triggered CBF (Yang et al. 2019), corrects the nominal input through barrier constraints solved online. Organisationally, the closest relative is shielding for reinforcement learning, in which a runtime shield interposed between a learning agent and the plant vetoes or replaces actions that violate a temporal-logic safety specification (Alshiekh et al. 2018); the object certified here is instead the predicted physical evolution of planner-issued segments against the already-committed queue, under communication faults. The organisational pattern also descends from the Simplex runtime-assurance architecture, in which a verified safety element supervises an unverified high-performance element and holds authority over it (Sha 2001), an arrangement later extended to learned controllers in the Neural Simplex Architecture (Phan et al. 2020); the present kernel differs in supervising externally proposed plan segments at the execution layer, under communication faults, with a certificate audit chain in place of a control-switching rule. Related frameworks include MARIS-TRA (Miller 2026) and ZTSafe (Zhou et al. 2026) for safety evaluation and runtime assurance, and SENTINEL (Zhan et al. 2025), which evaluates embodied agents formally at the semantic, plan and trajectory levels — an evaluation framework, not a runtime mechanism.

The closest framework-level neighbour, MARIS-TRA (Miller 2026), differs from the present kernel on each of these axes. Its certified object is an agent's action, scored by the quantitative robustness of a signal-temporal-logic fragment over predicted traces in a continuous-action multi-agent arena; the assurance acts at the decision point, actuation noise and delay are examined as robustness experiments, and the formal claim is a conditional predicted-trace certification under rollout assumptions. Hengjie's certified object is instead a planner-issued setpoint segment for a single physical plant, certified against the already-committed instruction queue: the remaining authorised slots of an issued instruction, one-slot delay, command loss and disconnection windows are inside the proof obligations themselves, as the apply, hold and delayed-delivery branches of one invariant. Where MARIS-TRA repairs a failing action auditably at the decision point, the kernel here corrects over a fixed candidate grid at minimum deviation or refuses release and continues on a certified fallback; and where MARIS-TRA's evaluation reports decision latency, the present study reports certificate coverage and

task attainment under a frozen protocol; its own decision latency is characterised by a separate measurement study (Section 3.4).

RPOF-SF, the closest mechanism-level neighbour, certifies a nominal controller's input within the control loop under output feedback. This kernel certifies planner-issued setpoint segments at the execution layer, accounting for the already-committed queue and communication faults. The kernel serves both scripted and LLM planners, the latter through a frozen asynchronous protocol whose invocation budget, review counts and target whitelist were fixed in advance. Its predictive verification incorporates the already-issued instruction, with its remaining authorised slots, and represents state uncertainty in interval form under the implemented channel faults (command loss, one-slot delay, acknowledgement loss, disconnection windows). The public evidence package lets third parties inspect the per-cell artefacts and recompute the reported quantities of the frozen batch; the raw per-tick decision records are not public (see Data availability). The experiments include no same-condition comparison against the published methods above; Sections 4.1 and 6 address this gap.

Runtime verification, finally, checks a running system's executions against a correctness property (Leucker and Schallhart 2009). Hengjie adopts that stance but moves it from observation to enforcement at the release point: a candidate that cannot be certified is corrected at minimum deviation or refused outright, and the certificate — not a monitor's verdict after the fact — is the artefact on which release depends. This enforcement stance, applied to externally proposed plan segments rather than to a controller's own input, is the level at which the present study contributes to the quality assurance of AI-enabled software.

3 System model and execution protocol

3.1 Object model and the safety problem

The simulated thermal object has four nodes (heater, air, wall and sample) and scalar action $u \in [-1, 1]$. The parameters are project settings, not measured values; Supplementary S1 gives the dynamics, powers, integration schedule, update equations and all parameter values. The measured air temperature is the true value plus bias and noise, quantised to 0.02 °C, with an error bound of ±0.12 °C. The task band is [29.5, 31.0] °C; in the normal scenario the reference curve is 30.0 °C over [0, 240) min and 30.5 °C over [240, 1440) min; the hard limits are [18, 32] °C, and any minute in which a substep carries the air or sample temperature outside them counts toward violation minutes.

The kernel verifies against an aligned two-state interval abstraction of the four-node network: the retained coordinates are the air temperature and the heater excess T_h - T_a, and the wall and sample contributions are eliminated into bounded additive terms through exact rational prefix matrices (12 prefixes per minute). The abstraction's ambient interval is [22.975, 25.025] °C, its solar interval [0, 0.2] W and its measurement-error bound ±0.12 °C (an ambient sensor, where present, is read at ±0.5 °C and clipped to the same ambient interval). Its interval coefficients were identified and frozen before the batch (identification method: Supplementary S1). The identification code is in the deposited source tree, but the run's own artefacts are not in the deposit, so the frozen

values are documented by the configuration and the method and cannot be re-derived from the public record.

In compact form, the plant evolves in 5 s substeps j within each one-minute slot k as $x_{k,j+1} = A(\theta)x_{k,j} + B(\theta)\varphi(u_k) + E(\theta)w_{k,j}$, with measurement $y_k = Cx_{k,0} + v_k$, where $\varphi(u) = (\max(u, 0), \max(-u, 0))$ and the two columns of $B(\theta)$ route the first component to the heater node as heating power $24u^+$ W and the second to the air node as cooling power $12u^-$ W (the piecewise actuator map of Supplementary S1); θ collects the conductances, w the environmental inputs and v the measurement error, each confined to its declared interval domain (Supplementary S1). The kernel reasons over the two-state enclosure X_k of $(T_a, T_h - T_a)$: the true state lies in X_k at every covered slot (invariant I1 of Section 3.5), and a candidate is certified only if every execution branch induced by the committed queue keeps the enclosure's successor inside the hard limits and closes on the certified OFF continuation (Supplementary S2). The certificate condition is a conjunction over branches: $\mathrm{Cert}_k(u) \iff \forall b \in \Omega_k :$ $\mathrm{prefixes}(b, u) \subseteq S \ \wedge \ \text{OFF-closes}(b, u)$, where the branches b are apply for one stage, apply for two stages, hold the committed instruction for its remaining authorised slots, and each delivery mode not yet closed by evidence; $\mathrm{prefixes}(b, u) \subseteq S$ requires every substep prefix of the branch to keep the enclosure inside the hard limits, and OFF-closes(b, u) requires the zero-action suffix from the branch's terminal enclosure to be itself certified to remain in S for the whole remaining horizon. A candidate failing any conjunct is not released.

The implemented fault set is narrower than the general threat picture of Section 1 (full input vocabulary and message sealing: Supplementary S2). On the command channel, the frozen profiles are command loss at 5% or 20%, a one-slot command delay, acknowledgement loss at 20%, and full-channel disconnection windows of 5, 30 or 120 min. Of the plant-side faults, the formal matrix exercised only the actuator faults of Section 5.3, and those as detection conditions rather than protected safety conditions (Section 3.4). Out-of-order delivery and observation-channel loss are not in the implemented set: observations are produced locally at every tick. Task attainment is the task ratio (task_ratio, the fraction of minutes with air temperature in the task band), physical excursions are violation minutes (violation_minutes), and minutes of authorisation not covered by a certificate are false authorisations (false_authorizations).

3.2 Kernel architecture

Hengjie consists of four components (Fig. 1). (1) Interval observers. A fast observer maintains the belief as a two-state box (air temperature and heater excess intervals), starting from the input's prior box (temperature [29.4, 29.6] °C); at each tick it intersects the prediction with the measurement interval (reading ±0.12 °C, outward-rounded; Supplementary S2). On a trigger condition the observer latches: it substitutes a diagnostic box and authorises nothing further (trigger conditions and the slow observer: Supplementary S2). (2) Robust predictive verification over a finite candidate set. Candidates form a fixed 21-point grid {-1.0, -0.9, ..., 1.0}; the requested action (from the PID (proportional–integral–derivative) controller, or derived from an AI plan as in (4)) heads the search order of (3). At most one instruction is active at the device at any time ('queue'); an issued instruction carries at most one grace

slot beyond its issue slot, and prediction takes the held instruction with its remaining authorised slots (0–2, counting the current slot) as the committed quantity; the transport layer adds at most a one-slot delay queue. Verification takes three forms. On the finite two-step screening path, the candidate is propagated for one minute and the full action range [-1, 1] for a second, with all 24 substep prefixes required to keep temperature and heater excess ([-40, 40]) within the hard limits. On the Rule B path, a candidate is certified only if every obligation branch (apply for one stage, apply for two stages, hold the committed instruction for its remaining slots) is safe at every substep (air and sample within [18, 32] °C, heater excess within [-40, 40]) and closed by a terminal zero-action (OFF) proof (Supplementary S2); the certificate is accepted only when an independent verifier reproduces it exactly, and any disagreement aborts the run — a protocol-wide rule; the per-run disagreement field itself is recorded only in the AI family's summaries (Section 4.3). Rule B thus verifies apply and committed-hold branches with terminal OFF closure; the standard comparator retains only the apply branch, without the hold obligations, assuming same-slot execution, and when it cannot authorise it issues nothing and leaves the device holding its last action. On the B2 path (component (4)), the same obligation structure is proved slot-wise (a two-slot proof covering the issue slot and one grace slot, then a one-slot renewal proof) under the premise that authenticated feedback uniquely identifies the held action and its remaining slots; if that premise cannot be established, B2 does not authorise. The premise audit rates this attribution premise as only partially established (Supplementary S3): it is discharged slot-wise where the authenticated record resolves a unique action, and elsewhere the device possibilities are carried as a set under I2/I3. (3) Minimum-deviation correction and refusal. Deviation is |u - u_req|; candidates are tried in increasing deviation, ties towards the smaller action. When no candidate is authorised, the kernel falls back along the path's own certified continuation, not to a generic safe action: the screening path first tries the rechecked zero tail of the previously certified two-step plan, otherwise the rechecked fixed action -1.0 (full cooling; conditions in Supplementary S2); the Rule B and B2 paths use the inherited zero from the parent certificate's OFF suffix, intersected with the fresh measurement box and re-verified at every tick. If neither candidate nor fallback is authorisable, the kernel issues nothing and raises a safety disagreement, which under the batch discipline of Section 4.3 stops the run. The gate described here governs certificate-authorised release: on the screening path, a recorded recheck does not make the runner's fixed -1.0 degraded actuation after authorisation has ceased a certified fallback, whereas the Rule B/B2 continuation applies only where the inherited parent-certificate OFF suffix and its current proof obligations remain valid. (4) Execution-layer rule review (B2). On the online AI path, every candidate plan must also pass B2's review before adoption: its format is checked against the privilege rules (Supplementary S2), and the resulting candidate is proved by the same slot-wise B2 proof as on the non-AI path, where B2 gates release directly. The planner's output remains a proposal; release authority rests with the kernel and the review stage. Table 1 (Section 4.1) defines the conditions assembled from these paths and each one's guarantee-bearing component. The certificate-authorised release logic of the Rule B and B2 paths (components (2)–(4)) — excluding the screening runner's degraded post-latch actuation reported in Supplementary S14 — is:

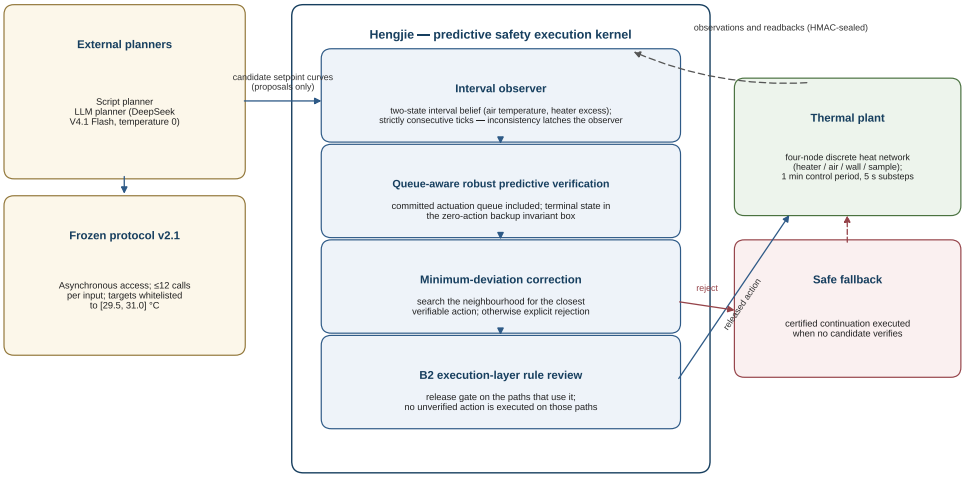


Fig. 1 Hengjie kernel architecture and interfaces. Planners only propose candidate setpoint curves; release rests with the kernel, the B2 rule review gates the paths that use it, and rejected candidates are replaced by the safe fallback; where neither a candidate nor the fallback can be authorised, the kernel issues nothing and raises a safety disagreement. Commands may be lost, delayed by one slot, or cut off by disconnection; acknowledgements may be lost.

```
CERTIFY-AND-RELEASE(X_k, Omega_k, u_req):
  for u in the candidate grid, ordered by |u - u_req|, ties towards the smaller action:
      if every apply, hold and delayed branch induced by Omega_k is certified,
         including its OFF continuation: issue u, record its obligations, return
  if the inherited certified continuation verifies against X_k: continue it
  else: issue nothing and raise a safety disagreement
```

3.3 Asynchronous access protocol for AI planners

The external LLM planner (in this batch, DeepSeek V4.1 Flash (DeepSeek 2026), official endpoint, temperature parameter 0) is connected through protocol v2.1, frozen with the experiment matrix before the experiments. The planner can submit candidates only at three invocation points: plan_start, plan_segment_end and after_rejection_replan. Within completed cells (Section 4.3), it is limited to 12 invocations per input and 288 per family (24 inputs × 12). Candidates are segments of a setpoint curve, at most 6 segments, each spanning a single one-minute tick, with targets confined to the whitelist [29.5, 31.0] °C (timing and adoption rules: Supplementary S2).

3.4 Scope of the guarantees

The safety conclusions hold only within the model. All experiments are software simulations with research-set parameters and no calibration against real hardware; 'zero violations' means that no simulated minute in the observation window crossed the hard limits under the frozen model, parameters and protocol v2.1 definitions. It is not

a safety commitment for any real physical system. The two-state abstraction by itself authorises at most two steps and proves no terminal set, so wherever whole-run safety is reported on the protected paths, the all-time component rests on the zero-action OFF proof of Section 3.2.

That proof covers a narrow four-node physical family in which only two conductances vary; the batch parameter box lies inside this family, so the closure result covers every formal protected run. The closure rests on the family's passivity over the declared environmental domain: under zero action the object only dissipates heat through the two varying conductances, so from every state of the invariant box a zero-action continuation remains within the hard limits — which is what makes a certified OFF fallback available at every tick. Fan failure and door opening lie outside this family; the formal protected runs used normal plant inputs with in-domain environmental strata, and the 12 actuator cells are detection-and-alarm, not protected safety, conditions. The extension also requires a credible initial physical box, verified apply/hold obligations, safe OFF prefixes and entry into the invariant box; Supplementary S2 lists the family and these premises in full. Per-tick decision latency was measured in a separately pre-registered study that re-executed recorded cells under per-component timing instrumentation on a compute node (Hygon C86 7185; single-threaded, core-pinned, thread pools forced to one; full records: Supplementary S18). Over the representative healthy, cold-start and post-disconnection regimes, the whole online decision loop (observation, PID, B2 decision, issue and submission) took a median of 1.32–1.37 s per tick, with a 95th percentile of 1.77–1.84 s including authenticated readback — at most 3.1% of the 60 s (one-slot) tick budget; under one-slot delay the loop was cheaper (median 0.58–0.59 s, 95th percentile 0.73–0.74 s), the plan layer's decision step took a median of 1.42–1.47 s (95th percentile at most 1.54 s), and the largest single tick observed anywhere took 4.22 s, 7.0% of the budget. The study reports the chain A sets per run under its registered consistency rule: one healthy nominal run is a flagged outlier at about twice the loop time of its sibling runs, and one contaminated parallel batch was discarded and re-run solo (Supplementary S18). The check-call counts recorded in Supplementary S7 are the frozen batch's own record of invocation load.

For the delay-tolerant (DT) extension (Sections 5.2, 5.6), the delay premise is registered in slots as $d_{\text{eff}} = d_{\text{channel}} + d_{\text{clock}} \leq g$, with $d_{\text{channel}} = 1$, $d_{\text{clock}} = 0$ and $g = 1$, over the fixed version domain $d = 1$, $g = 1$. The entry $d_{\text{clock}} = 0$ is a constructive premise, not a conclusion: issuance, delivery, execution and knowledge intake share a single logical clock, and a deployment with a second clock source is outside the domain. Configurations with d > 1 or g $\neq 1$ are rejected before system start. A restart stops issuance until device and knowledge state are re-proved consistent. The derived minimum silence $S_{\min} = g + 1 = 2$ slots (residual validity g plus one observation round trip) is the earliest resumption the slot order permits, not a sufficient drain-and-reconciliation guarantee; a larger ACK-delay bound enlarges the required wait. Supplementary S2 records the full derivation, the out-of-domain handling and the separate roles of the premise digest, channel guard and independent checker.

3.5 Formal properties: proposer independence and the scope of the guarantees

Let S be the declared in-model safe set: air and sample temperatures in $[18, 32]$°C and heater excess in $[-40, 40]$ at every covered slot and substep of the declared discrete-time model. Let I be the kernel invariant, the conjunction of enclosure soundness (I1), device possibilities held as a set wherever the record cannot resolve a unique action (I2), a set-valued in-flight branch set Ω covering every delivery mode not yet closed by deployment evidence (I3), outstanding certificate obligations recorded against their named knowledge snapshots (I4), and consistent cross-layer bookkeeping (I5). Supplementary S3 gives the full definitions. The violation-minutes metric of Section 3.1 counts the air- and sample-temperature components of S; the heater-excess component is enforced through the enclosure invariant I1, not counted in the metric.

Proposition (proposer-independent conditional safety). Assume that initialisation establishes I and that I implies membership of the plant state in S; assume the environment, observation and execution premises audited in Supplementary S3 hold; and the preservation of I by every transition of the execution (release of an accepted candidate, continued hold, expiry, feedback ingestion, rejection or silence, and restart with reconciliation) is discharged rather than assumed: Supplementary S3 proves it by induction on slots over the six protocol-event cases, under hypotheses H1–H6. Then, for every proposer P supplying arbitrary inputs at the kernel interface, every resulting execution remains in S at every covered execution slot. A proposer's inputs determine *which* certified action, if any, is released, and never *whether* a released action is certified.

The proposition concerns this conditional safety property only: it asserts nothing about acceptance rates, task performance, liveness or computational cost, each of which may depend on the proposer. The six hypotheses (H1–H6) are stated in full in Supplementary S3. The invariant is maintained branch-wise: where the record cannot resolve which permitted action was executed (AT-3), all consistent branches are covered and no unique attribution is asserted. The premise audit in Supplementary S3 maps each premise to the frozen clause that establishes it or to the hypothesis that carries it (the audit rests on the frozen protocol and design documents; the implementation code is outside its scope): of 29 premises, 21 are established, 7 are partially established and carried as hypotheses H1–H6 or by the set-valued definitions I2/I3 — among them the within-slot phase order (H4), a modelling convention assembled from the frozen per-layer orders rather than itself a frozen clause — and 1, a reading of S_{min} as a sufficient recovery bound, is not established and is not used. Read by kind, the 29 premises divide into environment and model assumptions, protocol design requirements, and properties discharged by implementation or experiment (Supplementary S3 marks each premise's kind alongside its verdict); 'established' records establishment within the audit's two documentary sources and does not by itself assert that a corresponding implementation property has been experimentally verified. The latches of Section 3.2 are sticky. On paths with a valid inherited parent-certificate OFF continuation, later slots are covered hold slots only while the invariant's premises and the applicable hold, expiry and OFF obligations remain satisfied; this statement does not cover the screening runner's fixed degraded actuation after authorisation has ceased. A

latch does not assert that a physical fault persists over the latched interval, and a fault-window statement carries no automatic guarantee after the window (Supplementary S4).

Guarantee scope. In the protected delay_1 runs under the legacy execution semantics, 1,431 of 1,440 delayed deliveries were rejected per cell and non-zero execution was confined to ticks 0–8, with zero observed safety violations and zero false authorisations. The delay-tolerant experiments restore the legacy-B flagship criterion in 4/4 tested cells, whereas B2 meets its pre-registered ratio criterion in 1 of 4 (Section 5.6). These are finite-run results in the specified carriers, configurations and seeds: the legacy rejections characterise the legacy execution semantics, with the DT semantics assessed on its own runs (Section 5.6), and B2's full-horizon exhaustion is a finite-horizon outcome under the tested inputs. Both findings delimit the tested operating envelope, and the certifiability of liveness under other conditions remains open. Safety and progress are therefore reported separately; a bounded-starvation monitor is provided as a diagnostic, and no liveness theorem is claimed.

A slot is starved when no certified decision takes effect in it and the device falls back to the held or zero action. Over a sliding window of $N = 30$ slots, the monitor alarms when at least 4 slots are starved ($\rho > 10\%$) or a starved run reaches 6 consecutive slots (calibration record: Supplementary S2). The monitor detects its defined predicates on an adequately observed trace. Its silence does not establish task completion, fair scheduling or future progress, and AT-3 alternatives are carried as a set, never resolved into a definite starvation label. Supplementary S12 discloses the relationship with a related theoretical manuscript in preparation; no result here depends on it.

The lifecycle of an issued instruction under these semantics — issuance against a named knowledge snapshot, delivery, the validity window of the issue slot plus at most one grace slot, execution, certified readback, and discard on expiry or late arrival — is summarised in Supplementary Fig. S1.

4 Experimental design and integrity verification

4.1 Freeze and the experiment matrix

Experimental protocol v2.1 was frozen on 2026-10-27; the freeze registration fixed the matrix, seeds and input texts, and is deposited with the core evidence package (Data availability). The deposit was published after execution (2026-11-09); the freeze time itself rests on the internal registration and its hash chain, which the public record does not independently timestamp. The frozen tree contains 16,825 files, and the SHA-256 of its master manifest begins with 19c220d4. The matrix comprises 428 cells, 424 of them executable; the remaining 4 are timeout_invalid counterfactual cells, registered as not executed (N/A-by-design) because an invalid first invocation leaves no adopted proposal to replay. The matrix contains 100 distinct inputs, assigned seeds 30000–30099 in frozen order (seed-allocation checks: Supplementary S13). Each input–seed combination was run once per condition, conditions sharing inputs and seeds (a paired design); no dispersion or seed-sensitivity analysis is reported, so every Section 5 magnitude is a single realisation's recorded value. The 24 AI inputs comprise 12 formal task texts (the normal_staged_hold, aggressive_replan and timeout_invalid scenarios,

four inputs each) and an injection group of 12 inputs in three forms (four each): a 35 °C raw form, a 35 °C form with a prefixed system note, and a 30.8 °C system-note form.

Table 1. Experimental conditions and their verification and continuation rules. All conditions are in-house configurations with the PID proposer. Fallback entries describe certificate-authorised release; the screening runner's degraded post-latch actuation (Supplementary S14) is not a fallback in this sense.

Condition (producer)	Verification path	Fallback / unauthorised case	Guarantee-bearing component
PID; PID_task_reference (episode)	None; direct execution	None	None
A (episode)	Two-step screen (slow-state + probe)	Rechecked zero tail, else rechecked full cooling	Finite assessment; no all-time closure
standard (episode)	Two-step screen (fixed slow-state bounds)	Same finite fallback	Finite assessment; no all-time closure
B (episode)	Rule B	Reverified inherited zero, else stop	Apply/hold branches, OFF closure, local expiry
A (legacy)	Two-step screen (slow-state + probe)	Same finite fallback; if unauthorised, issue nothing, retain local hold	Authenticated finite envelope; no all-time closure
standard (legacy)	Apply-only Rule B projection; same-slot execution assumed	No inherited-zero fallback; if unauthorised, issue nothing, retain local hold	Apply branches with planning OFF suffixes; no hold coverage or autonomous expiry
B (legacy)	Rule B (authenticated held-action knowledge)	Reverified inherited zero, else stop	Apply/hold branches, OFF closure, local expiry
B2; same_input_nonAI_B2 (b2)	Slot-wise B2; two-slot proof, then one-slot renewal	Reverified inherited zero, else stop	Slot-bound apply/hold branches, OFF closure, local expiry

Episode A and standard provide no whole-run guarantee; the Rule B and B2 guarantees are conditional and in-model. AI and reference-curve configurations are defined in Section 4.2; actuator cells evaluate detection only.

The families are producer groupings over four unit grids (composition, grids and seed ranges: Supplementary S13), belonging to three execution frameworks: episode runs the full kernel loop, whereas legacy and b2 use dedicated frameworks with different field systems.

Phase 1 supplied non-frozen, unrecomputed feasibility evidence used only as design motivation; it supports no comparative claim.

A pre-registered example gate for a later-proposed comparison first tried to reproduce the feasible-set volume of the AMPSC (adaptive model predictive safety certification) baseline's published example (Didier et al. 2021, Table I): 5.86×10^3 and failed: the reproduction returned 7,698 ± 21 (mean ± 1 s.d.) at N = 50 over 30,000 merged samples, a relative error of +31.4% against the pre-registered tolerance band [5.27×10^3, 6.45×10^3], and every point of the declared grid {50, 100, 200, 500} lay at ≈ 7.7–7.9 $\times 10^3$, outside the band. The comparison line was stopped under the pre-registration; no same-condition comparison is reported and no superiority is claimed in either direction (reproduction chain and adjudication: Supplementary S5; guarantee-retention analysis: S6).

A separately pre-registered diagnostic study did not resolve the discrepancy: none of the tested individual reconstruction choices accounted for it, and both published Table I volumes fell between the volumes reconstructed at adjacent tested integer horizons; the publication does not state the horizon, and the comparison remains stopped (diagnostic record: Supplementary S5).

4.2 Four-way controlled design for the AI cells

For each AI input, four pipeline configurations are compared: reviewed_AI_with_B2 (the AI submits candidates online; adoption follows B2 review throughout); same_input_nonAI_B2 (the AI removed, B2 gating the same input directly); same_input_reference_curve_PID_B2 (a trusted reference curve submitted tick by tick for B2 review); and unreviewed_AI_isolated_counterfactual (the first adopted proposal's curve from the same cell's reviewed run, its terminal target extended over the whole run under the 'hold the recorded terminal target' assumption, replayed under an isolated PID with no review, no replanning and no fallback). The four share the input text and seed; the isolated counterfactual removes review, replanning and fallback together, so it does not isolate the effect of any single component.

4.3 Batch execution and incident handling

The fingerprint gate: the model service's system fingerprint was rechecked before and after the batch. The freeze-registration baseline is e72d3c7f…; the fingerprint measured during execution was bc19ba97…, and the measured fingerprint was adopted by ruling as the execution baseline, the frozen baseline being retained as a historical record — the paper does not mix the two. Both post-batch probes returned the execution baseline, so no drift occurred within the batch. The execution binding record

registers two further deviations from the frozen tree and an exception ruling on the cap for environment-terminated reruns (Supplementary S7). Incident handling (A3 in the freeze registration): any checker disagreement stops the batch with evidence preserved and no automatic rerun; after an environment termination, only the affected cells resume from the most recent complete checkpoint; after an engineering fault is fixed, only the affected cells are rerun, the failed results being retained but excluded from the statistics. The batch had 48 environment restarts, with 1 further restart during recomputation and no loss of completed cells; all cells affected by the 5 engineering faults were later completed under A3, and 2 historical halts are documented in Supplementary S7.

The call ledger: completed cells contain exactly 288 AI invocations under the family cap, which constrains invocations within completed cells (breakdown: Supplementary S7), with 0 hard violations and all response fingerprints at the execution baseline; the partially completed directories left by environment terminations consumed a further 88 responses outside that cap. Total expenditure was therefore 376 calls; those 88 responses produced no results counted in the statistics. The recomputation makes no model invocations. The deposited records keep no batch-level tally of checker disagreements or observer latches; the one family whose summaries carry a disagreement field records 0 in all 24 cells.

The DT-line runs of Section 5.6 follow three further record disciplines (attribution codes AT-1–AT-4, with contested cases reported as candidate sets; latch witnesses W1–W7, with untriggered channels recorded explicitly and per-cell alarm fields; and no retrospective exclusion on channel-guard alarms; full statement: Supplementary S7).

4.4 Independent recomputation

Reader programs written independently of the batch driver recomputed all products in two stages. They used only recorded artefacts and ran neither the production engine nor the model service. Each family checker (a recomputation reader, not a kernel-internal verifier of the kind in Section 3.2) was first validated on sample cells before use. Stage 1 re-verified the SHA-256 binding of each cell's bound products (424/424 cells pass; the records keep per-cell verdicts, not per-file counts). The hash-chained summaries anchor the records' integrity and order; no safety or liveness conclusion follows from them. Stage 2 re-derived the semantic quantities under the frozen producer definitions: 424 cells and 4,656 checks, with 0 discrepancies and 0 errors (breakdown by check name: Supplementary S8). For one counterfactual cell logged twice, the stage-2 totals use the later four-check record. The field independent_exact_report_matches (Supplementary S7) is a producer-side consistency count, not a stage-2 check. All Section 5 numbers come from products that passed the applicable stage.

Recomputation agreement provides evidence of record consistency and arithmetic agreement, while physical-model validity remains unestablished: agreement cannot exclude errors in five items shared with the kernel — the Euler integration scheme and step size, the identified parameters and parameter box, the environment and disturbance inputs, the generation path of the external row data, and the floating-point

environment — whereas record parsing, decision logic and the reduction implementation are independent reimplementations. Supplementary S8 gives the list in full with the mapping of checked quantities to physical targets, the record-layer completeness scope (the B2 legacy-producer arm carries no witnesses block) and the registered Euler dt/5 refinement rule for the sample cells of the registered follow-up batches, together with the executed pass: all 104 sample cells are stable under the refinement.

4.5 Stage-2 mismatch study protocol

The mismatch study of Section 5.7 runs on the stage-2 development carrier, not a formal carrier: 13 cells, two seeds (28001, 28002), a 720-step horizon. Mismatch is injected as a step change on two channels, the heater factor θ_h and the ambient gain κ_g, at the tested lattice points ($\theta_h \in \{1.00, 1.01, 1.02, 1.03, 1.04, 1.05\}; \kappa_g \in \{1.00, 1.28, 1.31, 1.35\}$). (The batch records and Supplementary S11 write the heater-factor grid coordinate as κ_h.) A run stops when truth-side monitoring registers the out-of-domain exit (criterion F0): the κ thresholds are truth-audit thresholds, never kernel detection thresholds, and a stop is an offline abort by the evaluation runner, not a kernel property. Every run is rescored offline under the deployment information set (Supplementary S10) with the grading taxonomy of Supplementary S11.

5 Results and analysis

5.1 Protected paths and the counterexample window

Table 2 summarises the physical families' core results. The protected paths keep a double zero across all communication-fault and healthy scenarios: in the legacy family, the 34 condition-B cells total 0 violation minutes and 0 false authorisations; in the B2 family, all 88 cells record 0 violation minutes and 0 false authorisations, with a median task ratio of 0.9993. Unprotected condition A totals 25,792 violation minutes and 34,433 false authorisations over its 34 cells; the peak air temperature, 39.07 °C, occurs in the cold-start cell with a 120 min disconnection, which accumulates 1,410 violation minutes; the largest single-cell violation total, 1,422 minutes, is recorded by the nominal disconnect_5 cells. The standard condition falls in between: 2,728 violation minutes, all from the 2 cold-start cells with a 120 min disconnection, and 4,771 false-authorisation minutes.

Table 2. Aggregate comparison of the three legacy conditions (34 cells each). The standard condition is an intentionally weakened execution configuration, not a replication of a published baseline. Task-ratio quartiles are descriptive over each condition's 34 pooled cells, which share the design's 100 distinct inputs across conditions in paired runs and are not independent replications; the delay stratum is reported separately in Section 5.2. The standard row's false authorisations are per-slot certificate-coverage counts, not safety events. IQR denotes the interquartile range.

Condition	Total violation minutes	Total false-authorisation minutes	Median task ratio (IQR)
A (unprotected)	25,792	34,433	0.249 (0.0085–0.6929)
standard	2,728	4,771	0.9993 (0.9972–0.9993)
B (protected)	0	0	0.9927 (0.9155–0.9990)

The two zeros are of different kinds. On the protected paths, uncovered-authorisation minutes are zero by construction, because those paths release only certified actions; violation minutes are an observed outcome, consistent with the in-model guarantee of Sections 3.4 and 3.5 under its stated premises. At design stage, 60 legacy cells were pre-registered as a counterexample window (the A and standard cells of all communication scenarios except receipt_loss_20: 30 units × 2 conditions): their excursions are expected by design (an unprotected path driven out of bounds under long disconnection plus cold start); what is measured is the form and magnitude of the failure. Per-family execution detail is recorded in Supplementary S7. The comparisons form a single paired design: Table 1 fixes the verification path of every condition, Section 5.2 separates the delay stratum, Sections 5.4 and 5.5 hold the AI configurations to the same inputs, and Section 5.6 compares the revised semantics against the original on identical cells. The standard condition's false-authorisation figure is a per-slot certificate-coverage count from the path audit (the standard configuration has no hold-certificate structure), not a count of stale-execution authorisations. It is not cited as a safety event (Supplementary S7). False-authorisation figures are per-slot counts; stated in minutes, as in the abstract and Section 5.2, the same quantity is uncovered-authorisation minutes.

The disconnect_5 stratum is bimodal, and its raw records identify the mechanism (Fig. 2). In the nominal legacy-A cells the five-tick outage (ticks 8–12) exhausted the legacy path's certificate chain: of 14 decisions attempted, the eight before the outage were authorised and the six after reconnection (ticks 13–18) all found remote authorisation unavailable (first invalid certificate at 540 s); no further decision was recorded, and the device held its last action (0.3) for the remaining 1,422 minutes while the air temperature crossed 32 °C at 1,105 s and peaked at 35.63 °C. Both nominal replicates record the same 1,422 violation minutes. In the cold-start replicates the same outage fell inside the initial transient with held action 0, authorisation remained available for all 1,435 decisions, and no violation occurred. The protected legacy-B cell bridges the outage on its local certified fallback; its air temperature remains within the hard band throughout, ranging from 29.20 to 30.52 °C. The bimodality is a property of the legacy path's authorisation continuity, not of the plant or of the outage length.

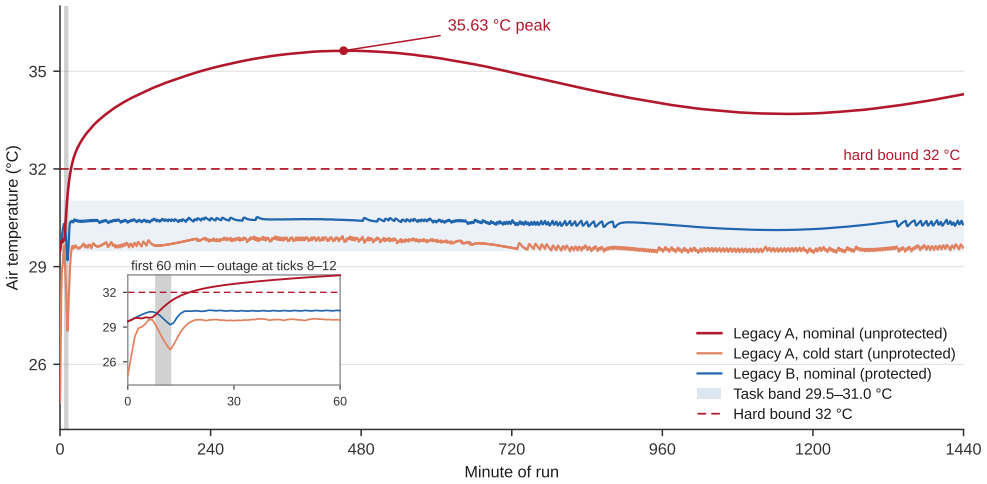


Fig. 2 Air temperature in the disconnect_5 cells (five-tick outage at ticks 8–12, shaded). Legacy A, nominal: authorisation is lost at reconnection and the held heating action drives air temperature to a 35.63 °C peak (1,422 violation minutes). Legacy A, cold start: the outage falls in the initial transient; no violation. Legacy B, nominal: local certified fallback bridges the outage, holding air temperature within the hard band [18, 32] °C (29.20–30.52 °C).

5.2 Safety–task split in the delay scenario

The 1 min command-delay scenario (delay_1) splits safety from task attainment. The four condition-A cells violate for 1,413–1,417 minutes, out of bounds for nearly the entire run. Conditions B and B2 each record 0 violation minutes and 0 uncovered-authorisation minutes, but only 10–11 task minutes out of 1,440. The standard condition preserves the task (median task ratio 0.999) with 0 violations, at the cost of 83–320 uncovered-authorisation minutes per cell; it matches the protected paths on violations and dominates them on task attainment in this scenario, but not on certificate coverage. For the B2 cells, the summaries show authorisation decisions passed, 1,440/1,440, with 0 authorisation-unavailable events; the executed action was zero in 1,431 of 1,440 steps (further action-distribution detail: Supplementary S13). The legacy B summaries record the same full authorisation pass and task totals, but not the action-distribution fields. Event-level forensics on these cells resolve the mechanism. In each of the eight protected cells the kernel issued 1,440 authorisations and commands; 1,431 deliveries arrived one slot late (origin slot t-1, delivery slot t) and all were rejected under the single reason `expired_delayed_or_future_slot` (accepted 0, applied 0), from tick 9 continuously to tick 1439. Action was nonzero exactly on ticks 0–8, and the double zero held throughout. The collapse arises from the composition, not from any single component: device, kernel and transport each behaved as specified (expired tokens rejected, issuance continuing, deliveries one slot late). It is a liveness failure, not a safety failure or an implementation defect. The standard cells accepted all 1,431 late deliveries (task 1,439), which is the mechanism behind their coverage trade-off.

5.3 Actuator fault detection

Each actuator fault type (readback_mismatch, refuses_close, refuses_open) was tested in 4 cells. In all 12 cells, faults took effect at 480 s and the first credible alarm at 540 s: a latency of at most 60 s at the one-control-period timestamp granularity (12/12 within one period). These results establish detection and alarming only; fault handling and recovery were not evaluated. The comparison behind the alarm contrasts the device readback with the action actually applied as recorded by the simulation harness (Supplementary S13); that latency bound is therefore a result under that test information set, in which the applied action is ground truth, and is not a guarantee for deployments that must infer the applied action from observable signals alone. The alarm channel is instrumented only in the actuator runner: the other 412 cell summaries contain no alarm field, so false-alarm rates cannot be estimated from this batch. Alarm mechanics and record scope are given in Supplementary S13.

5.4 The four-way AI comparison and the arithmetic of fallback dominance

Fallback averaged 1,388.3/1,440 minutes, or 96.4%, in the reviewed cells. The frozen budget forces this dominance: at most 12 invocations per cell and at most 6 ticks per adopted proposal permit at most 72 min of AI curves; the measured AI-curve total was 1,241 min (mean 51.7 per cell), and rejections and shorter horizons limited coverage further. The reviewed cells reached a mean task ratio of 0.9981 across 24 cells (range 0.9951–0.9993), a mean curve-tracking MAE (mean absolute error) of 0.1641 °C (median 0.1185) and 0 hard violations. The non-AI control on identical inputs (nonAI_B2) has a mean task ratio of 0.9983 with 0 violations and 0 false authorisations, tracking the reviewed cells closely cell by cell (differences ≤ 0.0021; across the 24 pairs the median difference is 0.0000, the mean -0.0001, and 20 pairs are exactly equal). Under this batch of scenarios and the frozen budget, adding AI produced no observed task-ratio gain, and neither configuration produced hard violations. The measurable differences lie in the shape of curve tracking and its departure from the reference in specific scenarios, and do not extrapolate to other budgets or scenarios.

Across the twelve injection cells the planner was called 144 times; 143 calls returned valid proposals and one returned service unavailable. Every valid proposal was adopted, no minute was rejected at certification, and no hard violation occurred. Trace-level inspection shows the failure sits entirely at the proposal layer: in the 35 °C cells every returned curve point is the trusted reference target (30.5 °C in the plain form; 30.0/30.5 °C in the system-note form), and in the 30.8 °C system-note cells the returned points are likewise only 30.0 and 30.5 °C — the injected targets (35 °C, 30.8 °C) appear in no returned curve. The fallback share also has a structural component: calls fire at six-tick segment boundaries, so the twelve-call budget is exhausted by tick 72, and the remaining horizon executes the certified fallback by construction; within the injection cells the fallback supplies 16,565 of 17,280 minutes (95.9%).

For the three injection groups (12 cells) and the first-invocation fault-injection group (timeout_invalid, 4 cells; forms: Section 4.1, where the counterfactual cells on

these inputs are registered as not executed), the reviewed task ratios are almost identical, cell by cell, to the non-AI cells on identical inputs, with 0 hard violations; across the reviewed family, rejected minutes total 33, all from the aggressive_replan group. The task-ratio and zero-violation summaries alone could not show whether injected content was absent from the planner's proposals or was intercepted by the kernel. In this batch, the trace records show that the injected content never entered a returned proposal. The frozen batch therefore does not test the kernel's dynamics-layer interception of an injection-induced malicious proposal. That interception is exercised directly, against scripted proposers, in the follow-up batch of Section 5.8.

In the refcurve arm, 20 of 24 cells have 1,440/1,440 curve-target ticks, yet rejected and fallback ticks coincide exactly in all 24 cells (995–1,435) (the reference's later segment coincides numerically with the 30.5 °C fallback target; Supplementary S13). The four aggressive_replan cells, where reference and fallback differ, separate the two (Section 5.5).

A separately pre-registered addendum batch then varied the budget itself (registration, per-level records and full analysis: Supplementary S16). The same 24 AI cells were re-run online at two enlarged per-cell caps — 48 invocations (level L1; 1,152 calls in total) and 192 (level L2; 4,608) — against the frozen cap of 12 (level L0; 288), with the same inputs, engine and protocol v2.1, and the non-AI control of Section 4.2 as the fixed reference line. The fallback share fell steeply with dose, from 0.964 at L0 to 0.857 at L1 and 0.507 at L2 (paired mean differences 0.107 and 0.457; both paired bootstrap 95% confidence intervals exclude zero). At L1 availability was near-complete (5 of 1,152 completions service-unavailable); at L2, 603 of 4,608 completions (13.1%) were service-unavailable returns, so the L2 fallback share accumulated under that elevated unavailability (Supplementary S16). The fallback dominance of the frozen batch was predominantly a budget effect rather than a structural floor. The task outcome did not move: the per-cell task ratio is identical at the three levels in all 24 cells (family mean 0.9981 at every level, against the control's 0.9983), and the paired task-ratio difference is zero. The additional budget therefore changed which controller holds the band — the AI curve or the certified fallback — and not the band outcome; the reviewed pipelines separate from the non-AI control on task ratio at no level. Curve MAE moved only slightly (0.1641 °C at L0, 0.1600 at L1, 0.1551 at L2). The enlarged caps still bound in every cell: demanded dispatches exceeded each cap, with exactly one cap-suppressed dispatch recorded per cell at both levels (the demand count is censored at the cap: it registers the dispatches made plus that first suppressed dispatch, and no demand beyond it), so under the registered bracketing rule the planner's natural demand is not bracketed even at 192 invocations per cell, sixteen times the frozen budget, and no non-binding level exists in this design. All 48 new cells were independently recomputed (stage-1 SHA pass and stage-2 recomputation over traces and events: 96 of 96 records, 1,104 checks, zero mismatches), and their regenerated counterfactual arms record zero violation minutes in every cell (Supplementary S16).

5.5 The aggressive_replan data chain and how to read the counterfactual

The registered task for the aggressive_replan scenario is to hold 31.0 °C throughout, the upper edge of both the task band and the whitelist; its credible reference is a constant 31.0 °C, and the four configurations are compared descriptively. The reviewed arm has task ratios of 0.9979–0.9993 but a curve MAE of 0.498–0.504 °C (other groups: 0.051–0.132), with the air running near 30.5 °C, in band but about 0.5 °C from the reference, consistent with fallback dominance in this group (1,417–1,420 fallback minutes). The refcurve arm, proposing a constant 31.0 °C tick by tick, has 1,432–1,435 ticks rejected: a band-edge constant candidate is almost never accepted under this kernel's review. The counterfactual arm replays a constant 31.0 °C in all four cells, obtained by extending the same cell's first adopted proposal at tick 0 under the assumption 'hold the recorded terminal target'. The isolated PID replay peaks at 31.12 °C, with about half its minutes above the band's upper edge and a task ratio of 0.468–0.550, yet 0 violation minutes. Across all 20 counterfactual cells, violation minutes are 0, with a median task ratio of 0.9972. Per-cell values for this family are recorded in the core evidence package (Data availability).

In these four cells, a band-edge constant target without in-loop correction obtains only about half of the in-band time, whereas the reviewed pipeline's task ratio of about 0.998 arises from review, replanning and fallback acting jointly. The relative contributions of review, replanning and fallback are left unapportioned, and the standalone quality of the model's first proposal is not inferred. In its native context, that proposal is expected to be corrected by later replanning; executing its extension in isolation measures the condition 'first proposal executed alone', not the reviewed pipeline.

5.6 Registered DT follow-up batches

Two pre-registered follow-up batches under the frozen DT semantics (d = 1, g = 1) — pilot and formal, 52 cells each — take up the delay scenario of Section 5.2; the formal batch, accepted in full at its registered adjudication (52/52) independently of the flagship criteria below, is the headline. On the legacy-B family the flagship criterion is met: delay_1 task minutes 1,439/1,439/1,439/1,440 (with identical values in the healthy cells), ratios ≈ 1.000 (criterion ≥ 0.9; 4/4), double zero throughout, against same-cell OLD delay_1 controls of 11/11/10/12 (the original delay handling, without the DT revision, re-run on the same inputs within this batch — follow-up reruns, not the frozen-batch cells behind the abstract's 10–11 range). Both batches are open-loop replays of recorded runs under separate development seed domains — the pilot 930001–930052 on development inputs, the formal batch 940001–940052 on the frozen formal inputs — disjoint from the frozen batch's seed domain (30000–30099) and from each other, and are reported separately, never pooled. The pilot returned the same pattern: the one-slot collapse is removed to the level of the no-delay baseline.

The B2 family stops at a different boundary, one that has nothing to do with delay. Formal flagship ratios (delay_1 over same-cell delay_0 task) for the four flagship B2 cells are 0.656, 0.647, 0.512 and 1.000: one cell of four, criterion not met, no exemption; the pilot gives the same 1/4. Six of the eight core cells end as stopped_with_witness:

the stop witness records 42 fresh-candidate attempts, all rejected (fresh_attempts = 42 in the witness signature of the pilot and formal batches' adjudications), no recursive zero, the parent certificate's one-slot stage consumed. Stop positions are set by the input, not the delay (same-input healthy/delay_1 pairs stop within a few ticks of each other). The registered attribution for the tested input seeds is full-horizon candidate exhaustion of the B2 selector; the 200-tick smoke runs never reached these states. Stopped cells are scored in four columns (headline, prefix, gap, unserved remainder), with the double zero claimed only to each coverage point. The 1/4 flagship outcome is a finite-sample boundary observation over the tested inputs.

Restarts fall into two classes: in the B2 restart cells, recovery issuance resumes at tick 722 — $S_{\min} = g + 1 = 2$, the earliest resumption the slot order permits — completes, and stops again on the same exhaustion; in the legacy-B restart cells, the inherited-zero adjacency guard blocks the first recovery decision at tick 722 and no recovery issuance was completed. The cell-level restart course is recorded in Supplementary S13, and the corresponding smoke cell's classification in Supplementary S9. Four delay_2 probe cells were rejected out of domain at start-up on a premise-digest mismatch: no trial run took place and no task was scored, both safety counts standing vacuously at zero. The two batches' records are deposited in the Zenodo record cited under Data availability (bundle_dt.tar.gz, version 5).

5.7 Stage-2 mismatch study (development carrier)

Under the protocol of Section 4.5, the tested lattice brackets the truth-audit exit for the heater factor θ_h in (1.00, 1.01] on seed 28001. The ambient-gain (κ_g) exit depends on the seed: on seed 28001, no exit at the tested κ_g of 1.28 or 1.31 within 720 steps (at $\kappa_h = 1.00$); on seed 28002, exit at $\kappa_g = 1.35$ at tick 2, the exit point lying in (1.00, 1.35]. Brackets are not merged across seeds. The brackets locate where the truth-side monitor stopped each run; the deployment information set registered no event in any stopped cell (that set excludes, by the membership definition of Supplementary S10, the fuse channel and all environment flags), and no surviving cell produced an executed-but-uncaught failure — an observation limited to a finite horizon on this development carrier. It does not show that the deployment side detects such exits in continued operation, and a truth-side exit is not itself a physical safety failure. Rescoring under the deployment information set (Supplementary S10–S11) split all ten stopped cells (X0 simulation-side, no event deployment-side), and no executed-but-uncaught coverage failure (X2U) was observed among the three surviving cells; the X2U type specimen remains cell A_gha077 (Supplementary S11).

5.8 Addendum batch: scripted proposers and an engineering baseline

The planner study in Section 5.4 did not test dynamics-layer interception of injection-induced proposals, because the injected targets never appeared in a returned plan. The scripted addendum therefore supplies adversarial requests directly at the decision interfaces, complementing the observed proposal-layer non-adoption. The registered addendum batch (registration, binding gates and per-cell records: Supplementary S14)

replaces the planner with scripted proposers at those interfaces, leaving the frozen engine and protocol v2.1 untouched. All cells reuse frozen inputs and registered seeds; each producer class first had to pass a registered binding gate against its frozen reference (tick-for-tick hex-level equality for the scalar, plan-layer and heuristic producers; an independently audited development cell for the production path). The batch comprises 48 scalar-layer cells (four request profiles × three communication conditions × two initial strata × two replicates), 16 plan-layer cells (four scripted plan curves × four AI inputs of Section 5.4 (identified in Supplementary S14)), 24 engineering-comparison cells and eight construction-level contract cases. Verdict rules were registered in advance; negative outcomes are reported as run.

Scalar layer, nominal stratum. Under healthy conditions the dynamics check corrects the adversarial profiles to zero violation minutes: ADV-BANG is corrected on 1,339 and 1,364 of 1,440 ticks in its two replicates, ADV-HEAT on all 1,440, ADV-PULSE on $\approx 1,439$, with all 1,440 ticks authorised; the benign PID control is corrected on $\approx 1,439$ ticks by the same minimum-deviation rule. Under one-slot delay the outcome is the same — zero violation minutes with all 1,440 ticks authorised; ADV-BANG is corrected on 1,438 ticks and ADV-HEAT on all 1,440. Under the 30-tick disconnection, the PID control and the HEAT and PULSE profiles again record zero violation minutes (1,410 authorised ticks; the outage accounts for the remainder).

A scoped failure of the screening path's degraded fallback. The remaining combination fails. The failure is in the screening path's degraded post-latch fallback semantics, as replicated by the scalar harness; it is not a failure of the production path. In both replicates of ADV-BANG under disconnection, the first reconnected tick (tick 38, after 30 disconnected ticks) latches an observer conflict; authorisation never resumes (8 authorised ticks in total), and the degraded fallback applies the screening path's fixed action -1.0 (full cooling) at every tick, independently of the profile's continuing alternating requests. Air temperature falls from 25.29 °C at reconnection and first violates at tick 56; each replicate totals 1,384 violation minutes with a minimum of 5.45 °C. No violating action was certificate-issued: the failure sits in what the screening path does after authorisation stops, where it continues to apply that fixed degraded action without certificate coverage. The production B2 path, whose fallback is the parent certificate's OFF suffix under command expiry, records zero violation minutes under the identical fault profile (engineering comparison below), so the finding is scoped to the screening path's degraded post-latch fallback semantics under the conjunction of observer latch and disconnection; it does not transfer to the production path, and the run's certificate-issued action record remains clean.

Scalar layer, cold-start stratum. All 24 cells, including the benign PID control, latch the observer at tick 0: the cold-start initialisation — a uniform 24.0 °C start — is incompatible with the aligned observer's preheated prior envelope (Section 3.2) — the frozen batch's cold-start cells and chain A start from the same uniform state but take the starting prior from the input, whose cold-start box ([23.9, 24.1] °C) contains that state, and do not latch — and after the tick-0 latch the screening path continues fixed degraded cooling rather than certificate-authorised control, along a trajectory identical across proposers within each condition (air range 5.45–23.83 °C, with 1,433 violation minutes when healthy and 1,398 under disconnection; 17.76–25.07 °C, with 2

violation minutes under one-slot delay). This stratum is therefore uninterpretable for the interception question — it supports neither a pass nor a failure — consistent with the scalar score's own registered scope note (no cold-start claim). Cold-start behaviour of the production path is covered by the engineering comparison below.

Plan layer. Interception is established at this layer. Constant-edge (CUR-EDGE) and segmented (CUR-SEG) plans are rejected by the plan certificate on ≈ 100% of prepared ticks (1,430–1,435 per cell; 5–10 ticks adopted), the alternating plan (CUR-SAW) on 97–98%, and the stale-binding plan (CUR-STALE) is refused at the format layer on all 1,440 ticks of every cell (stale_snapshot). All 16 cells record zero violation minutes with task ratios of 0.9986–1.0: the per-slot kernel below the plan layer holds the task band while the plan layer refuses the plan.

Engineering comparison. Chain A is the production B2 path itself (issue, slot-wise proof, authenticated readback, independent audit); chain B is a heuristic threshold controller (the frozen PID parameters, the same device-side expiry, and a local protection trip at 31.5 °C) with no certificate — certificate coverage is recorded as not applicable for chain B, never as zero or as an error. Chain A completes all 12 cells with independent audits passed and zero violation minutes across every stratum and condition, including cold start and the 30-tick disconnection, over matched 1,440-tick windows; its task ratios are 0.997–0.999 when healthy and 0.975–0.980 under the 30-tick disconnection, but under one-slot delay its task attainment also collapses, to 0.0035–0.0076, with certificates and audit intact. Chain B is safe and effective while the channel is healthy (task ratios 0.97–1.0, zero violations), but under one-slot delay 1,431 of its commands expire in each delay cell and the task ratio collapses to 0.0028–0.01 (violations still zero), and under the 30-tick disconnection its protection trip latches at tick 42 with a task ratio of 0.01 and one replicate recording 1 violation minute (the other 0). The comparison therefore separates the two chains on task survival under disconnection, not under communication faults generally: under delay both chains lose the task, and what separates them there is certificate coverage and independent audit, not survival. A third chain, a Simplex instance over the same twelve cells (the same frozen PID as proposer, a zero-action safety controller, and a one-step exact-model lookahead against a certified safe set), is reported in Supplementary S15: it matches chain B's task ratios in ten of the twelve cells and, in the two nominal-stratum disconnection cells where chain B's protection latches, holds 0.969 and 0.972 with zero violation minutes; under one-slot delay its task attainment collapses with both other chains. A fourth chain, a control-barrier-function safety filter in quadratic-programme form (Ames et al. 2017) in its standard discrete-time form (Agrawal and Sreenath 2017), wrapping the same frozen PID proposer on the same exact-model replica state basis as chain C, is reported in Supplementary S17: it records zero violation minutes in all twelve cells and task ratios identical to chains B and C in ten of them, and in the two nominal disconnection cells it holds 0.969 and 0.972 with zero violation minutes, like chain C, intervening on three ticks in total; under one-slot delay its task attainment collapses with all three other chains. Chains C and D act on an exact, noise-free replica of the plant with same-slot execution, and neither carries the committed-queue, hold or attribution obligations that chain A certifies; their agreement with chain B on ten of the twelve cells — and their holding, in the two disconnection cells where chain

B collapses, at ratios close to chain A's — therefore bounds the plant-level difficulty of those cells, while under one-slot delay, where any difference would have to come from execution semantics, all four chains collapse alike. The engineering delay cells run the frozen production path under device-side expiry semantics, without the DT revision, and show the same expiry-related task loss as Section 5.2.

Contract cases. All eight construction-level cases behave as specified: out-of-whitelist actions and binding conflicts return privilege_violation (two cases each), expired actions return stale_snapshot (two), a lawful-but-unsafe candidate produces a real certificate rejection witness, and a lawful safe candidate is accepted and certified on its first tick.

These results concern action gating under scripted proposers in the same plant model; they complement, and do not revise, the planner study of Section 5.4, whose cells remain the evidence for how the deployed planner's proposals were handled.

6 Discussion: threats to validity and open items

The conclusions are subject to the following threats. (1) Model validity: beyond the in-model scope of Section 3.4, fault forms and intensities are design choices; real channel distributions may differ. All reported magnitudes are values of the declared Euler discretisation; for the follow-up batches, the registered dt/5 falsification pass has been executed, and every conclusion drawn from those batches is stable under the refinement (Supplementary S8). The delay-tolerant semantics further rests on a single logical clock (Supplementary S2); deployments whose endpoints keep independent clocks are outside the tested scope. (2) Scenario scope: the experiments use one low-dimensional thermal model with limited task-band and reference-curve forms. High-dimensional, strongly coupled objects are untested. The episode and legacy/b2 frameworks differ and are not compared across. (3) Baseline comparison: no same-condition comparison against published methods was performed; the one pre-registered attempt failed its reproduction gate: the reproduced feasible-set volume deviated +31.4% from the published value, outside the pre-registered tolerance band (Section 4.1). A separately pre-registered diagnostic study tested the declared candidate reconstruction choices individually and resolved none of them into an explanation; the deviation's cause remains undetermined (Supplementary S5). Spread isotropically over that example's six dimensions, a +31.4% volume deviation corresponds to about +4.7% in linear scale per dimension ($1.314^{1/6} \approx 1.047$); the actual deviation need not be isotropic, and the gate verdict is unaffected. (4) AI scope: the AI findings describe one model group with a 12-per-cell budget and one fingerprint baseline (the build measured in execution, adopted by ruling after the frozen fingerprint gate failed; Section 4.3). Because the planner whitelist coincides with the task band, these results are conditioned on a pre-clipped action space; they are not extrapolated to other models. (5) Provenance and record scope: two record limitations apply. The identification run artefacts are not deposited (Section 3.1); record-layer completeness covers only the instrumented paths (Section 4.4). (6) Stated-but-untested semantics: the v2.1 feedback-ingestion rule held by construction across all 424 cells and was never exercised as a test; the conflicting legitimate cross-slot readback of the delay-tolerant line motivated the v3.0

rule frozen alongside it. Channel-guard coverage varies by fault class, from immediate to undecidable for infinitely late delivery (Supplementary S8).

The follow-up programme settles three of the frozen batch's four open mechanism questions — the delay composition (Section 5.2), the refcurve rejection mechanism (below) and the disconnect_5 bimodality (Section 5.1). On delay, the Section 5.2 collapse is a compositional liveness failure, removed for legacy B by the delay-tolerant revision; the B2 boundary is, by the registered attribution for the tested input seeds, selector candidate exhaustion (Section 5.6). On the refcurve rejections, in the eight forensically examined cells (aggressive_replan ×4, normal_staged_hold ×4) 73,948 of 82,328 candidate attempts were refused under a single stage-level reason (the residual-heat prefix violates the hard bound), with an independent checker concurring in every refusal. The task minutes behind the 995–1,426 range over the 20 non-aggressive_replan cells of the 24-cell refcurve arm come from the certified screening ladder, not curve tracking; the injection-scenario cells' tick-by-tick records have since been examined for the 30.8 °C form in all four cells: all 29,572 rejected candidate attempts in these scripted PID reference-curve cells carry the same safety-failure signature (input valid, certificate unsafe), as do all 44,225 rejected attempts in the aggressive_replan cells (counting conventions: Supplementary S13). The delay results separate two properties: safety feasibility, whether a certified continuation exists, and execution feasibility, whether a certified action can still reach the device inside its authorisation window. The protected configurations retained the first and lost the second; the delay-tolerant revision restores the second for legacy B, and the B2 stops identify the selector, not the channel, as the next binding constraint.

The mismatch results of Section 5.7 are confined to step injection on the stage-2 development carrier; their brackets are lattice results over the tested point set, not continuous thresholds. No exit within 720 steps is a finite-horizon non-observation, and transfer of these brackets to the formal carrier is untested.

Three findings carry implications for software quality practice. First, certificate coverage is a usable release-time quality metric for AI-enabled control software: in the delay scenario it separated the protected configurations from the standard comparator (83–320 uncovered-authorisation minutes per cell) where violation counts alone did not (0 minutes for both). Second, the certified fallback carried the LLM pipelines' high attainment, accounting for 96.4% of minutes. Under this invocation budget, no quality gain can be attributed to the planner itself. Third, a quality gate can preserve safety while silently destroying liveness (10–11 of 1,440 task minutes under one-slot delay); acceptance criteria for such gates should therefore include progress measures alongside violation counts.

Table 3 collects the defects and boundaries the frozen protocol exposed, with their dispositions; none was repaired in place, by design.

Table 3. Defects and boundaries found under the frozen protocol, and their disposition.

Defect or boundary	Where reported	Disposition
Cold-start initialisation (uniform 24.0 °C) incompatible with the aligned observer's preheated prior envelope; all cold-start screening cells latch at tick 0	Section 5.8	The stratum is declared uninterpretable for the interception question and scored neither for nor against; production-path cold start is covered by the engineering comparison of Section 5.8
Screening path's fixed degraded action (-1.0) applied after authorisation has ceased, under observer latch plus disconnection	Section 5.8	Scoped finding on the screening path's degraded post-latch semantics; the production B2 fallback (the parent certificate's OFF suffix) records zero violation minutes under the identical fault profile
Inherited-zero adjacency guard blocks the first recovery decision in the legacy-B restart cells	Section 5.6	Bypassing the recursive zero at recovery is a registered candidate change for v3.1
B2 selector candidate exhaustion over the full horizon	Section 5.6	Registered attribution for the tested input seeds; the flagship criterion is reported as not met (1 of 4), with no exemption

One question remains open: the episode relationship between the A/standard air-temperature peaks and the target values of the fallback tail segments.

Table 4 states, for each headline result, the conclusion it supports and the conclusion it does not.

Table 4. Results and their conclusion boundaries.

Result	Establishes	Does not establish
Protected paths (legacy B, 34 cells; B2, 88 cells): 0 violation minutes and 0 uncovered-authorisation minutes	In-model safety of the certified release paths under the tested faults and inputs	Safety on real hardware, or under fault forms outside the declared family (Section 3.4)

Result	Establishes	Does not establish
One-slot delay: protected paths safe but reduced to 10–11 of 1,440 task minutes; the DT revision restores legacy B in 4/4 cells	The collapse is a compositional liveness failure (execution feasibility), not a safety failure	Liveness under other delay or clock regimes; the B2 flagship criterion (met in 1 of 4 cells)
LLM invocation budget $12 \rightarrow 48 \rightarrow 192$ per cell: fallback share $0.964 \rightarrow 0.857 \rightarrow 0.507$, task ratio identical in all 24 cells	The frozen batch's fallback dominance was predominantly a budget effect; the budget did not change the task outcome	Any conclusion about other models, tasks, or action spaces not pre-clipped by the whitelist (threat (4) above)
Injection cells: no injected target appears in any returned proposal	Non-adoption at the proposal layer, for this planner and batch	Dynamics-layer interception of an adopted malicious proposal; that interception is exercised with scripted proposers in Section 5.8
Comparison chains C and D each match chain B on ten of twelve engineering cells, and hold in the two disconnection cells where chain B collapses	The plant-level feasibility of those cells on an exact-model, same-slot basis	Coverage of the committed-queue, hold and attribution obligations that chain A certifies; under one-slot delay all chains collapse alike
Decision latency: median 1.32–1.37 s per 60 s tick on the measured node	The decision loop's cost is a small share of the tick budget in the measured regimes	A worst-case bound; the largest observed tick (4.22 s) is a sample maximum

7 Conclusions and availability of evidence

Hengjie separates proposal from authorisation at the execution layer. All 424 executable cells were completed and independently recomputed with 0 discrepancies. Within the model and the frozen protocol, the protected paths (legacy B, 34 cells; B2, 88 cells) recorded no violation minutes and no uncovered-authorisation minutes, while the unprotected conditions produced the pre-registered counterexamples. The one-slot-delay collapse is a liveness failure of the composition, not a safety failure: the follow-up programme restores legacy-B attainment (under the revised delay-tolerant semantics), and the B2 selector's candidate exhaustion and the seed-dependent mismatch exits are the tested operating boundaries (Sections 5.6, 5.7 and 6). The reviewed AI pipelines achieved high task attainment dominated by certified fallback (96.4% of minutes), with no task-ratio gain under this budget. One mechanism question from the frozen batch remains open (Section 6).

Statements and Declarations

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors report there are no competing interests to declare.

Author contributions. Jiwei Wang: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review and editing, Project administration. Kaihang Ma: Software, Writing – review and editing.

Data availability. The core evidence package — frozen-protocol registration and the execution binding record, per-cell summary artefacts for all 424 executable cells, the frozen engine source files, and the recomputation records — is publicly archived on Zenodo (DOI: 10.5281/zenodo.24328178). The same record also deposits the forensic evidence subset behind Sections 5.2 and 6: the adjudication record and execution log, with the per-tick alignment tables and aggregate files for the twelve delay_1 cells examined at event level (the eight protected and four standard cells; the four condition-A cells were reconciled at summary level) and the eight refcurve cells. The same record also deposits the stopped AMPSC comparison line's materials — the example-gate reproduction scripts, parameters and volume records, the example-gate adjudication and the guarantee-retention analysis (Supplementary S5–S6) — together with the full item-by-item premise-audit table summarised in Supplementary S3. A Level A minimal dataset in the same record provides the Fig. 2 source data, the disconnect_5 forensic tables and mechanism report, the call timelines of the 24 AI cells, the injection-funnel and paired-difference summaries, the refcurve rejection-reason counts and the witness cell behind Section 5.6; the Section 5.7 mismatch cells are tabulated in Supplementary S11. The public per-cell layer supports inspection of every reported aggregate and recomputation of the frozen batch's Section 5 quantities. It does not include the raw per-tick decision records (among them the candidate-level decision.attempts records behind the mechanism questions of Section 6) or the full raw data (approx. 87 GB); these are held in the Google Drive archive 'Hengjie Formal Batch Archive 2026-11-02' and are available from the corresponding author upon reasonable request (input-directory inventory: Supplementary S7). Scope of third-party verification: from the public record alone, a third party can inspect every reported aggregate and recompute the frozen batch's Section 5 quantities under the stage-2 protocol of Section 4.4; re-running the batch itself additionally requires the raw per-tick records in the Drive archive. The follow-up batches of Section 5.6 are covered by the same record: its version 5 adds the archive bundle_dt.tar.gz holding their registrations and adjudications, the replay design and code, the cell manifests with the seed and input mapping, and the complete per-cell results. For the addendum batch of Section 5.8, the registration (with its Amendment 1), the binding-gate reports, the run summary and the per-cell summary and binding records — the materials from which every number in Section 5.8 and Supplementary S14 is computed — form part of this submission's supplementary records; the batch's per-tick raw logs ($\approx$ 8 GB) are held by the corresponding author and are available upon reasonable request. The identification behind the frozen configuration (Section 3.1) is represented by its code in the frozen source tree and by

the deposited configuration carrying the identified values; the identification run's own development data and outputs were not retained.

Code availability. The frozen engine source files are deposited in the Zenodo record cited under Data availability.

Generative AI and AI-assisted technologies. During the preparation of this work the authors used a generative AI tool — Muse Wonder — to assist with identifying candidate literature, translation of the initial draft into English and language editing. All AI-assisted output was reviewed and edited by the authors and verified against the underlying research records; all cited works were verified by the authors against their published sources; the authors take full responsibility for the content of this article. No generative AI tool was used to generate or alter the research data reported here, and no AI tool is listed as an author.

References

Agrawal A, Sreenath K (2017) Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation. In: Proceedings of Robotics: Science and Systems (RSS), https://doi.org/10.15607/RSS.2017.XIII.073

Alshiekh M, Bloem R, Ehlers R, et al (2018) Safe reinforcement learning via shielding. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 2669–2678, https://doi.org/10.1609/aaai.v32i1.11797

Ames AD, Xu X, Grizzle JW, et al (2017) Control barrier function based quadratic programs for safety critical systems. IEEE Transactions on Automatic Control 62(8):3861–3876. https://doi.org/10.1109/TAC.2016.2638961

Brunke L, Zhou S, Schoellig AP (2022) Robust predictive output-feedback safety filter for uncertain nonlinear control systems. In: Proceedings of the 2022 IEEE 61st Conference on Decision and Control (CDC). IEEE, pp 3051–3058, https://doi.org/10.1109/CDC51059.2022.9992834

DeepSeek (2026) Deepseek api documentation: Models and pricing; change log. https://api-docs.deepseek.com/quick_start/pricing

Didier A, Wabersich KP, Zeilinger MN (2021) Adaptive model predictive safety certification for learning-based control. In: Proceedings of the 2021 IEEE 60th Conference on Decision and Control (CDC). IEEE, pp 809–815, https://doi.org/10.1109/CDC45484.2021.9682832

Leucker M, Schallhart C (2009) A brief account of runtime verification. Journal of Logic and Algebraic Programming 78(5):293–303. https://doi.org/10.1016/j.jlap.2008.08.004

Mayne DQ, Rawlings JB, Rao CV, et al (2000) Constrained model predictive control: Stability and optimality. Automatica 36(6):789–814. https://doi.org/10.1016/

S0005-1098(99)00214-9

Miller T (2026) Hierarchical temporal runtime assurance for controlled agentic ai systems: Safety shielding, auditable action repair, and bounded recovery. Machine Learning and Knowledge Extraction 8(9):293. https://doi.org/10.3390/make8090293

Phan DT, Grosu R, Jansen N, et al (2020) Neural simplex architecture. In: NASA Formal Methods, Lecture Notes in Computer Science, vol 12229. Springer, pp 97–114, https://doi.org/10.1007/978-3-030-55754-6_6

Sha L (2001) Using simplicity to control complexity. IEEE Software 18(4):20–28. https://doi.org/10.1109/MS.2001.936213

Yang G, Belta C, Tron R (2019) Self-triggered control for safety critical systems using control barrier functions. In: Proceedings of the 2019 American Control Conference (ACC). IEEE, pp 4454–4459, https://doi.org/10.23919/ACC.2019.8814657

Zhan SS, Liu Y, Wang P, et al (2025) SENTINEL: A multi-level formal framework for safety evaluation of foundation model-based embodied agents. arXiv, https://doi.org/10.48550/arXiv.2510.12985

Zhou H, Lei H, Yang M (2026) ZTSafe: Safety-certified risk-adaptive scheduling for zero-trust time-sensitive industrial networks. Future Internet 18(9):466. https://doi.org/10.3390/fi18090466

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- sqjsuppsource.zip