
Exact-Authority Belief-Space Planning: Where Composition Emerges and Where It Fails

Eduardo Perez Denadai
Optimality.ai

Abstract

A reactive policy, however safe, is *total over the states it has seen*: it cannot compose a sequence of conditions and it cannot enter a layout it has never mapped. This paper is the story of replacing that reactive proposer with a search over a belief — and of the two things that storytelling forced us to discover. Learning to **compose** worked, and it worked *emergently*: deleting the hand-written orchestrator and searching the joint belief state reproduced the key→door→goal ordering as a consequence of the physics, an ablation collapsing the planner to 0.000 when the rule was removed, and a contrastive ranker lifting novel-layout success from a reactive 0.107 to **1.000** while cutting search nodes $6.7\times$ **emphasised** at two. Learning to **be robust to perturbation** split cleanly in two — and the split, not either half, is the finding. An observation-grounded belief *absorbs* a shock that removes a requirement (a naive planner still scores 1.000) because the next observation reveals the new reality; but it is **brittle** to a shock that blocks an expected transition (naive 0.000, a 600-step loop, because the agent observes the same closed door and re-predicts the same lost transition forever). Escaping required an explicit **Epistemic Divergence Monitor** that detects the falsification and mutates the search graph — and a principle we came to call *the agent's curiosity is bounded by its own falsifications*. Along the way we proved the loop-contraction and containment properties that make the whole thing safe for an arbitrary proposer, and we carried the sharpest negative — the frozen decidability map is *inert* where success is concerned — to a second, independent implementation that reproduced it on real MiniGrid.

1 Introduction: composition as the missing capability

We begin where the reactive companion chapter ended. Its Ω -TD proposer is *contained* — four hard guardrails, zero-violation safety — but it is a **total decision function over observed states**: it decides each step from the state in front of it. That leaves two failures open, and they are not small. **Composition**: a certified path is a *conjunction* of conditions (hold a key **and** open the door **and** reach the goal), and a reactive map has no place to put the “and”. **Out-of-distribution traversal**: states are observation-keyed, so every novel layout is a cold start.

The model-based answer we pursued keeps the containment architecture and the exact evidence channels exactly as they are, and replaces only the *proposer* — with a search over a **belief state** b_t . The belief is the paper’s first substantive idea, and it comes with a doctrine that dictates everything after it.

2 The containment doctrine and its consequence

The doctrine is small enough to state in a sentence: **learn only what is not analytically known, and represent what is not observed as a *set*, not a *point***. The belief exposes observed tiles as

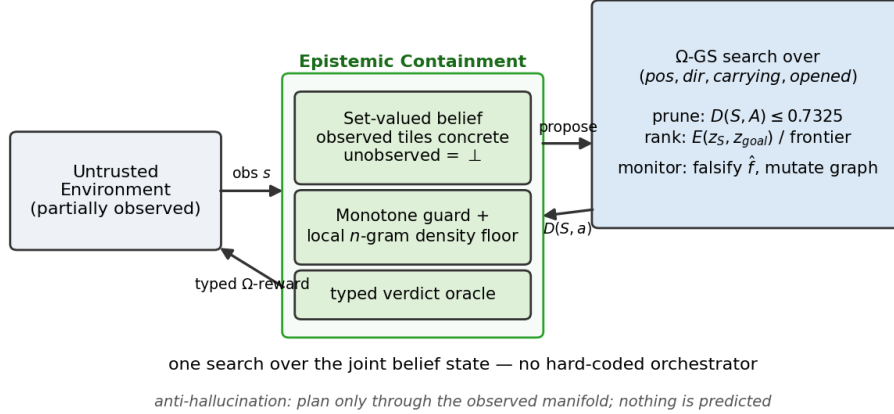


Figure 1: The belief-space architecture. A set-valued, observation-grounded belief inside the containment ring feeds one search over the joint state; the monitor can mutate the graph when reality falsifies the model.

concrete and everything else as $\perp$; pose transitions are analytic (LEFT/RIGHT rotate, FORWARD moves only if the front cell is observed-passable); observation is a union. Nothing is *predicted*, so nothing compounds over a horizon, and the out-of-distribution condition becomes **structural**: a state is OOD exactly when it asserts an unobserved tile.

That doctrine has a sharp consequence, and it is the paper’s central finding — though we did not see it until the ontology was stressed (§7). A belief grounded in observation is **robust to shocks that remove a requirement**, because the next observation simply shows the new reality, and it is **fragile to shocks that block an expected transition**, because the agent observes the same unchanged world, re-derives the same failed plan, and loops. We state it now as Proposition 3; we discovered it by building the shock that tests it.

3 Belief representation and the exact search

We keep the provenance of every number honest: the belief is anchored in the agent’s own frame, and the map from image cell to world offset is recovered by *replicating MiniGrid’s own pipeline* (the `get_view_exts` slice followed by `rotate_left` applied `direction+1` times), which we verified **pixel-exact** against the true grid on visible cells (**792/792** over 40 seeds) and exact under arbitrary action sequences (**710/710** rollouts). Blocked FORWARD is decidable from the *pre-action* observation alone — the front tile is always rendered at image cell (3, 5) — so wall-blocking needs no ground truth and no dead-reckoning, and no environment ground truth is read anywhere.

Over this belief we unroll **one** Dijkstra/BFS. Pruning is exact (a branch whose frozen read $D(\text{rendered}, a) \leq 0.7325$ is cut), halting is exact (a FORWARD through an unobserved cell is never expanded), and ranking is soft (a contrastive energy in the GoToObj setting, a coverage frontier elsewhere). The planner returns the root action of the best path — or abstains. The search reads the frozen map **read-only**: mental rollouts write nothing; only physical execution may mint evidence.

4 Composition is emergent from belief and search

A small amount of hand-written structure sufficed. We deleted the orchestrator — the `if key / elif door / else` — and replaced it with one search over the joint state and a single goal predicate. Composition *emerged*. The decisive test is an ablation: delete the rule that a locked door blocks, and the planner walks straight through it, ignores the key, and fails — $1.000 \rightarrow 0.000$ (Fig. 5). The `key→door→goal` ordering is a consequence of the dynamics, not smuggled in.

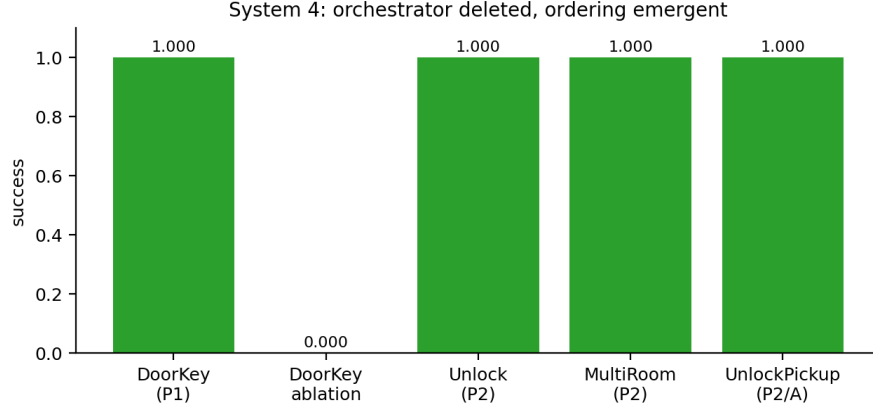


Figure 5: System 4: orchestrator deleted, ordering emergent, tasks generalized.

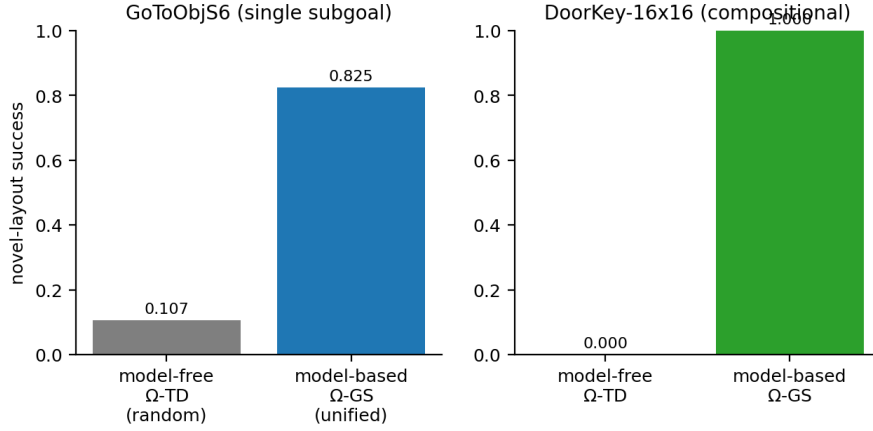


Figure 2: Model-free vs model-based on novel layouts.

The same file then solved DoorKey, Unlock, MultiRoom and UnlockPickup, differing only in the model (dynamics + goal predicate, all 1.000); a parser mapped the mission string to {target, predicate} and deleted the per-task goal flag (1.000, parser **15/15** paraphrases); and adding DROP made the full Empty→Full→Empty→Full inventory cycle expressible (UnlockPickup 1.000).

claim	result
P1 orchestrator deleted (unified search)	DoorKey 1.000
P1 dynamics ablation (door_blocks=False)	0.000 (collapse)
P1 interleaving (goal closer than key, behind door)	PASS (walks away from goal, 37 steps)
P2 same file across tasks	DoorKey / Unlock / MultiRoom / UnlockPickup 1.000
P3 goal from parsed mission	1.000 ; parser 15/15 paraphrases
Phase A: DROP → full inventory cycle	UnlockPickup 1.000

5 The proposer as a control problem

On novel layouts the reactive map scores 0.107. The belief-space search lifts that to 0.825 with the frozen prior, and to 1.000 with a contrastive relevance ranker — **zero new gradient steps**, the search running entirely in-silico (Fig. 2).

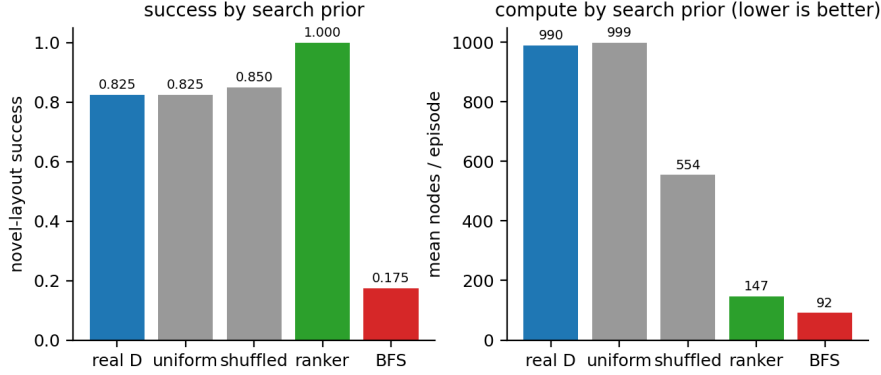


Figure 3: Search priors: success and compute.

agent	novel-layout success
model-free Ω -TD (reactive)	0.107
model-based Ω -GS (frozen D prior)	0.825
model-based Ω -GS (contrastive ranker)	1.000

But a good headline hides the interesting part, which is *what is actually doing the work*. We ablated the proposer into pieces (Fig. 3), and the controls told us something uncomfortable: the frozen decidability map D is **inert** on novel layouts ($\Delta = -0.025$; 85% of q-values at initialization) — the honest negative — whereas the **ranker is load-bearing** ($\Delta = +0.375$ over the hand-coded frontier heuristic, at a $6.7\times$ node reduction). The controls held: oracle-leak leak-free (66.7% belief dependence), the firewall intact (Δ weights = 0), and a hallucination veto at 0.00%.

arm	success	nodes / episode
real D , frontier	0.825	989.7
uniform prior	0.825	998.9
shuffled prior	0.850	554.4
ranker	1.000	146.8
BFS baseline	0.175	91.5

Then we turned our own best result over. *Is the ranker a navigation gradient, or a visibility detector in disguise?* Restricting the probe to goal-*invisible* states: score std 0.072 (non-flat), Spearman(score, -distance) +0.238, path-relevance AUC 0.654. The ranker navigates — but **weakly per-step**; the episode-level 1.000 compounds a weak gradient through closed-loop search (Fig. 7). We say so, because a 1.000 that hides a weak per-step signal is exactly the kind of number this programme refuses to over-read.

6 Event-triggered planning under belief change

Deliberation is a sledgehammer for a fully observed hallway, so we built a gate that re-plans only when the plan’s premise changes — no plan, exhausted, sub-goal moved, fog, or boundary — and otherwise executes the cached action in $O(1)$. On DoorKey this cut search nodes by 80% (5000.4 $\rightarrow$ 1001.9) at **identical** 1.000 success. It **missed** our 85% bar, and the failure taught us why: DoorKey is exploration-dominated, so re-plan frequency has an information-theoretic floor, and no gate can go below it. The falsification control flagged the miss correctly (Fig. 4).

mode	success	nodes	cut
ungated (replan every step)	1.000	5000.4	—
cache (plan-validity)	1.000	1001.9	80.0%
fog-spec (conf < 0.85)	1.000	2110.6	57.8%

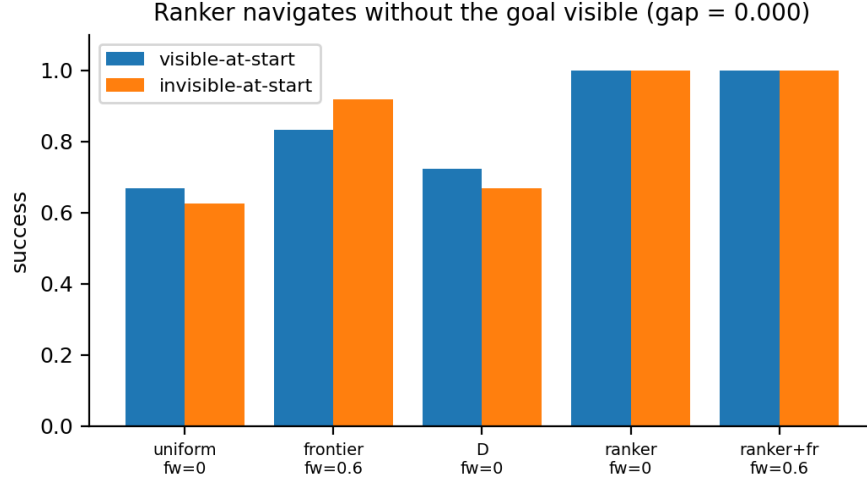


Figure 7: Ranker split by goal visibility: the ranker navigates without the goal visible.

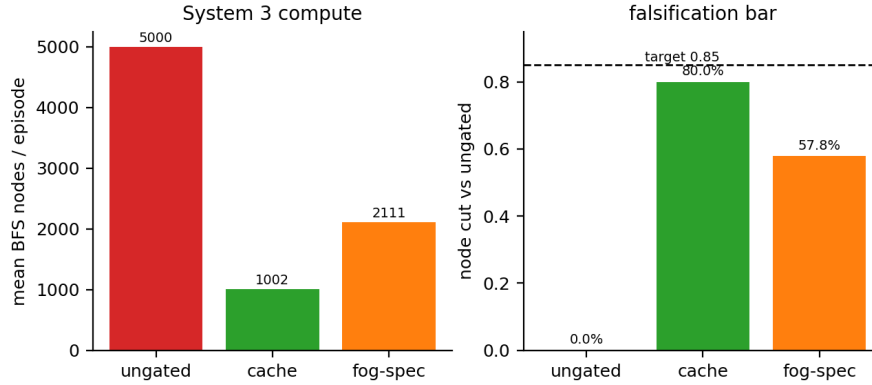


Figure 4: System 3 compute-efficiency curve.

7 Two negative results and the boundary between them

Everything above assumes the world obeys the model. The paper’s sharpest experiment asks what happens when it does not, and it splits the answer cleanly (Fig. 6).

shock	naive	monitored
Fragile Key (removes a requirement)	1.000	1.000 (divergences 50/50)
Jammed Door (blocks an expected transition)	0.000 (600-step loop)	1.000 (59 steps)

Fragile Key removes a requirement. The naive planner is fine (1.000) — observation-grounding *absorbs* the shock, because the next observation reveals it.

Jammed Door blocks an expected transition. The naive planner is **brittle**: it observes CLOSED, predicts TOGGLE → OPEN, the environment ignores it, and it loops forever — 0/50, a 600-step loop. We had, in effect, rediscovered Proposition 3 the hard way. The escape required building the thing the doctrine lacked: an **Epistemic Divergence Monitor** that, every step, compares prediction to observation, and on divergence records the empirical effect and **mutates the search graph** — marks the tile empirically impassable (jammed), rewrites the action’s carrying effect, and forces a re-plan on reality. The search then natively routes the long path around (1.000, 59 steps). One detail turned out to be critical and is worth naming: the *exploration affordance must respect the jammed ledger*, or

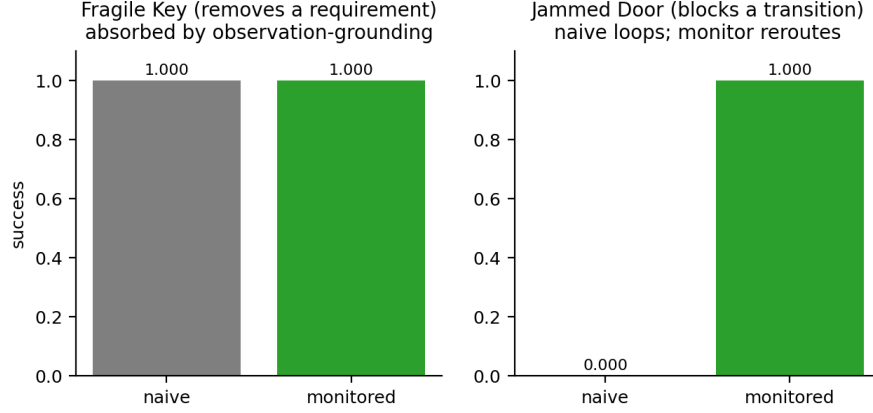


Figure 6: The two shocks: a state-mutating shock is absorbed by observation-grounding; a transition-blocking shock crashes the naive planner and is rescued by the monitor.

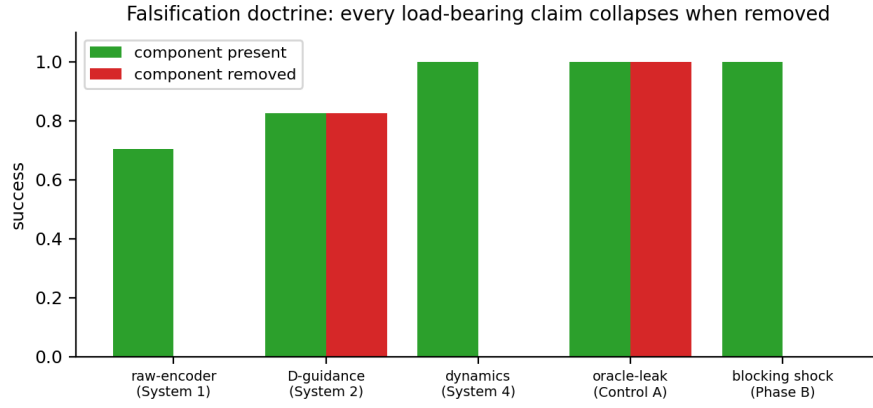


Figure 8: Every load-bearing claim collapses when its component is removed.

the “closed door is actionable” rule re-toggles it — **the agent’s curiosity is bounded by its own falsifications.**

8 Falsification as method, and the seam to distillation

We do not trust any of the above unless removing the load-bearing component breaks it. Every capability claim was paired with a control that collapses (Fig. 8) — and one cell of that matrix, the frozen map, does *not* collapse but is *inert*, which we report with the same prominence as a success.

claim	control	expected collapse
tile encoder suffices	raw-encoder control	raw 0.000 → spectral 0.703
D guides search	uniform/shuffled prior	identical success (D inert)
ranker is load-bearing	frontier-only baseline	1.000 → 0.825
composition is emergent	door_blocks=False	1.000 → 0.000
search uses the belief	oracle-leak corruption	leak-free
mental rollouts write nothing	firewall fingerprint	$\Delta\text{weights} = 0$
monitor rescues blocking shocks	naive vs monitored	0.000 → 1.000

That inertness cell then became a paper of its own. We carried the question to a second, independent implementation — the Ω -bel authority, re-instantiated symbolically, labelled in-silico, and **re-run on real MiniGrid** after certifying that the first hand-built substrate was not MiniGrid-faithful. It asked

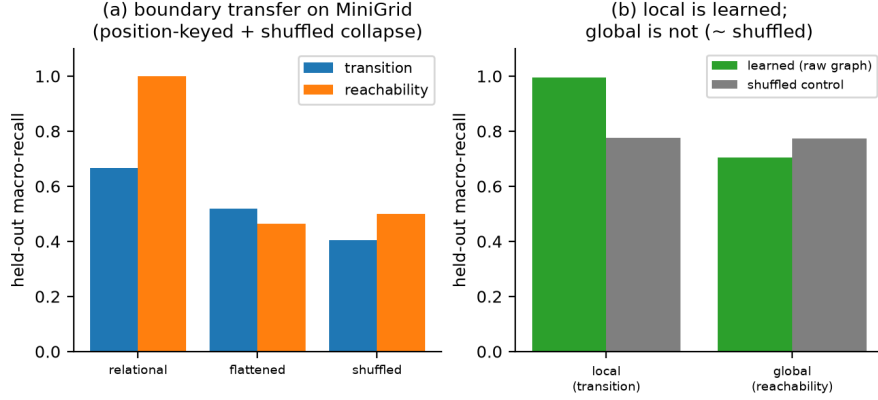


Figure 9: Oracle distillation on real MiniGrid: (a) the boundary distills and transfers while position-keyed and shuffled controls collapse; (b) the learned raw-graph encoder transfers the local predicate but not global reachability.

the inverse question: if the boundary cannot be *searched* into existence, can it be **learned** — and does it buy capability? The instruments: the oracle target is a function of the belief (ceiling 1.000); the ground-truth collision ceiling — the leak bound — is 1.000 on MiniGrid, because the forward view always observes the front cell; the decidability share is $P(R_1) = 0.293$. The boundary **distills and transfers** (1.000 relational, controls collapsing), and the witness kind is **perfectly distillable** (1.000) yet redundant for the class — *naming the missing witness is the product, not an accuracy instrument*. But the prior is **inert on top of the exact search**: on MiniGrid a belief-space planner lifts success over reactive (0.360 $\rightarrow$ 0.733 dense; 0.650 $\rightarrow$ 0.760 held-out) and search+prior is **identical to the digit** (85/85 nodes). The old subset’s “topological shield” does **not** reproduce — with the front always observed the reactive policy already incurs 0.0 violations, so there is nothing to shield. The map is a **veto**; the planner is the **engine of competence**.

distillation quantity (real MiniGrid)	value
oracle-target ceiling	1.000 (labels are a function of the belief)
ground-truth collision ceiling (leak bound)	1.000 (front always observed; world probe not flagged)
decidability share $P(R_1)$	0.293
boundary transfer, relational (held-out)	reachability 1.000; transition acc 1.000 (macroR 0.667, R_1 absent)
boundary transfer, flattened / shuffled	transition 0.519 / 0.404; reachability 0.465 / 0.500
witness kind: distillable / redundant-for-class	1.000 / yes
success, reactive $\rightarrow$ search $\rightarrow$ search+prior (dense / held-out)	0.360 $\rightarrow$ 0.733 $\rightarrow$ 0.733 / 0.650 $\rightarrow$ 0.760 $\rightarrow$ 0.760 (prior inert)

9 What is proved

The narrative has a spine of guarantees, and they are what make a learned proposer — even a search — safe to place inside the loop. Containment extends to a search unchanged: under invariants I1–I4, for any proposer, the executed action is guard-allowed or an abstention, so no update to the proposer causes an unsafe execution (**Theorem 3**). Uncertified cycles contract to the refutation anchor $V^* = -1.0$ (**Lemma 1**). Anti-hallucination is sound: a search that expands only through observed-passable cells never plans through unwitnessed geometry (**Proposition 2**). And the blocking-shock behaviour is not a bug but a theorem: an observation-grounded planner that re-plans from an unchanged belief will re-propose a blocked transition indefinitely, so escaping it *requires* mutating the model (**Proposition 3** — the algebraic form of the Jammed-Door result). Proofs are collected in Appendix A.

10 Conclusion and open problems

Composition can be emergent. Deleting the orchestrator and searching the joint state reproduces the ordering, and the ablation proves it did not survive by accident. **Goals can be declarative,** parsed from language. **The physics model can be falsified and rewritten mid-task.** And the boundary between the two shocks is the central finding: observation-grounding absorbs shocks that *remove* a requirement and is fatal to shocks that *block* one, because the agent re-predicts the same lost transition from the same unchanged belief until something — a monitor — mutates the graph.

We are careful about what is *not* shown. The ranker is only weakly per-step-navigational (path-relevance AUC 0.654); the 1.000 is a closed-loop compounding of a weak gradient. The Divergence Monitor is a principled rule-override, not a general axiom-rewriting engine. The parser is a regex — P3 relocates the hardcode from code to language, it does not eliminate human specification. UnlockPickup costs $\sim 80k$ search nodes, so efficiency, not only success, remains open. And every experiment is a MiniGrid gridworld: the claims are architectural, and extending them to physical systems would require a physically-grounded latent belief, a density floor on that latent, and a real-time monitor budget — none demonstrated here.

References

- Mouhammad Alshiekh et al. Safe reinforcement learning via shielding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.
- Aaron D. Ames et al. Control barrier functions. In *Proceedings of the European Control Conference (ECC)*, 2019.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.
- Marc G. Bellemare et al. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems 29*, 2016.
- Maxime Chevalier-Boisvert et al. BabyAI: A platform to study the sample efficiency of goal-driven language learning. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.
- Maxime Chevalier-Boisvert et al. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track)*, 2023.
- C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
- Deepak Dalal et al. Safe exploration in continuous action spaces. arXiv:1801.08757, 2018.
- Eduardo P. Denadai. ω -certifiability: A theory of certified ignorance, 2026a. Optimality.ai.
- Eduardo P. Denadai. Containment rather than calibration: A sound architecture for learning under partial observability, 2026b. Optimality.ai.
- Eduardo P. Denadai. Knowing when not to act, 2026c. Optimality.ai.
- Eduardo P. Denadai. Fail-closed LLM, 2026d. Optimality.ai.
- Eduardo P. Denadai. Monotone guardrails, 2026e. Optimality.ai.
- Eduardo P. Denadai. Contrastive relevance ranker training. `neuro_eigen/pretraining/README.md`, 2026f. Optimality.ai.
- Eduardo P. Denadai. Reinforcement learning from verifiable feedback, 2026g. Optimality.ai.
- Yann LeCun. A path towards autonomous machine intelligence, 2022.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016.
- Lui Sha. Using simplicity to control complexity. *IEEE Software*, 18(4):20–28, 2001.

Appendix A. Proof details

A.1 Theorem 3 (containment invariance). The selection set is $\{a : \text{guard}(s, a) = 1\} \cup \{\text{ABSTAIN}\}$ (I2, I4). `guard` is parameter-free and independent of the proposer (I1); the density predicate is a function of the observation alone (I3). A search proposer returns an action from that set (it prunes with `guard` and the floor), so no update to the proposer — gradient or search — enlarges the executable set. ■

A.2 Lemma 1 (loop contraction). Along an uncertified cycle, $V \leftarrow V + \alpha(\epsilon_r + \gamma_{\text{step}}V - V)$; the fixed point solves $V = \epsilon_r + \gamma_{\text{step}}V$, i.e. $V^* = -1.0$; the error factor is $1 - \alpha(1 - \gamma_{\text{step}})$, a contraction for $0 < \alpha < 2/(1 - \gamma_{\text{step}})$. ■

A.3 Proposition 2 (anti-hallucination soundness). A FORWARD transition is added to the search only when the front cell is *observed*; by induction over the search tree, every reachable state has all cells observed. Hence no plan routes through unwitnessed geometry. ■

A.4 Proposition 3 (blocking-shock divergence). Let (s, a) be an edge the model predicts to have effect φ but the environment ignores. Executing a leaves the environment unchanged: the next observation is identical, so $b_{t+1} = b_t$, the model is unchanged, and the planner re-derives the same plan and re-proposes a . The trajectory is periodic until timeout; escaping requires mutating $\hat{f}$. ■

A.5 Monitor mutation. On divergence the monitor records the empirical effect (`corrections[a]`, `jammed.add(front)`) and clears the cached plan; the search’s `_passable` returns `False` for jammed tiles and the exploration affordance also consults `jammed`, so curiosity cannot re-propose the falsified transition. ■

Appendix B. Artifact provenance

- `cartographer/model_based/logs/{probe_ranker_ablation, probe_ranker_split}.json` (§4–§5).
- `cartographer/model_based/logs/gating_20261009_201311.log` (§6).
- `cartographer/model_based/logs/unified_20261010_100158.log` (§4).
- `cartographer/model_based/logs/ontological_20261010_112504.log` (§7).
- `cartographer/minigrid_refit/results/{control_c_mg, control_l1_mg, integrate_mg, pl1_mg}.json` and `cartographer/minigrid_parity/results/parity.json` (§8, real MiniGrid); `cartographer/oracle_distillation/results/*` (subset — superseded for §8).

The model-free chapter (`epistemic_map_v10.py`) is imported read-only; the two chapters are kept mathematically isolated.

Appendix C. Hyperparameters

symbol	value	meaning
search budget	100,000 nodes / call	per-replan BFS budget
prune bound	0.7325	$D \leq \text{bound} \Rightarrow \text{cut}$
gating fog threshold	0.85	re-plan when local confidence <
ranker	BERT-mini, InfoNCE	contrastive relevance, held-out AUC reported
monitor	per-step	predicted vs observed; mutate on divergence