WARM WINTER

Act, Escalate, or Abstain

Deriving an AI trust gate from one objective

A research whitepaper · warmwinter.io · June 2026

There's a fast-growing gap in AI right now: agents can take actions faster than anyone can verify them. The reflexive fix is "add oversight." But many current oversight systems rely on asking another model whether the first model's decision looks reasonable — and a second model is exactly as capable of being confidently wrong as the first. What's missing isn't another opinion. It's a *calibrated* one: something that knows when it doesn't know, and says so.

Warm Winter is a trust gate that sits in front of an agent's actions and returns one of three verdicts:

- **act** — proceed autonomously.
- **escalate** — spend more (a stronger model, a careful path, a human) to be sure.
- **abstain** — there's no grounded signal here; don't pretend there is.

The obvious question from anyone technical is: *where do those three verdicts come from?* If the answer is "we picked some thresholds," it's just vibes with extra steps. So we did the work to answer it properly. This is that answer — five results that turn out to be one idea, each computed from real or controlled data, each accompanied by an explicit distinction between what we prove and what we measure.

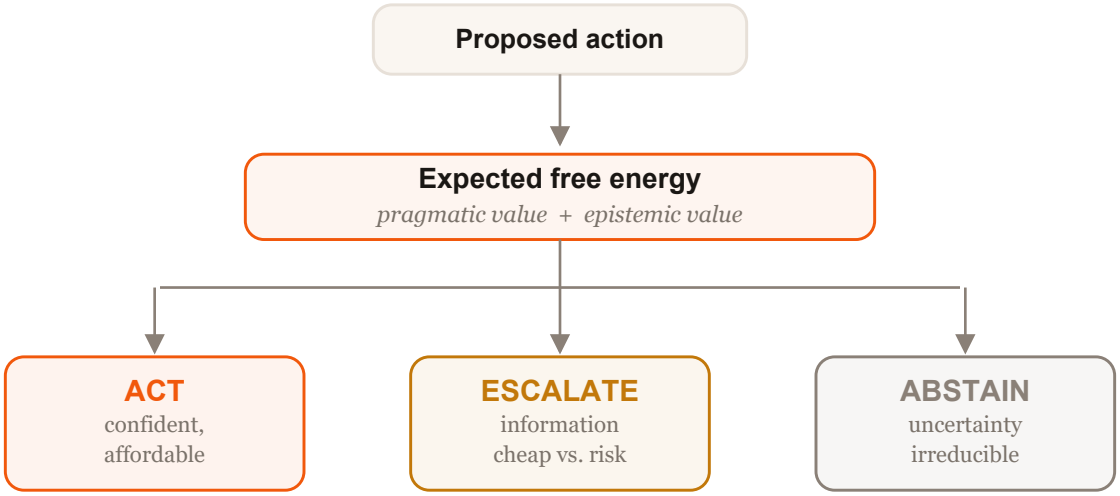

Figure 1. The gate as one decision. A proposed action is scored by a single expected-free- energy objective; act / escalate / abstain are the regions of its minimizer.

Contributions

- We show how **act, escalate, and abstain emerge as the optimal policy under a single objective** (expected free energy), rather than as separately specified rules.
- We **replace count-based grounding thresholds with a Bayesian epistemic/aleatoric uncertainty decomposition**, making honest abstention principled and base-rate-aware.
- We introduce **causal evaluation of trust-gate interventions** using the system's own randomized exploration (IPS and doubly-robust off-policy estimators).
- We combine **adaptive conformal prediction with decision routing** to maintain calibrated confidence under the distribution shift we simulate.

Each result is reproducible from a small script; quantitative figures below are from those runs, and the foundational methods are cited at the end.

One objective, not three rules

Start with the punchline. The three verdicts are not three hand-written rules. They are an optimal policy under one objective borrowed from active inference (a Bayesian decision framework inspired by statistical physics). That objective trades off two things every action has:

- **pragmatic value** — does this get us the outcome we want?
- **epistemic value** — how much do we still not know, and would acting/escalating teach us?

Minimize the expected free energy of that trade-off and the three verdicts naturally emerge: **act** when the cell is confident and the stakes are affordable, **escalate** when more information is cheap relative

to the risk, **abstain** when the uncertainty is irreducible at acceptable cost. We verified this against the gate's actual decision table: the minimizer reproduces all nine (state × stakes) cells — without separately specifying a rule for each decision cell — from the *structure* of the objective (the full nine-cell mapping is in the Appendix, and reproducible from the accompanying script). The rules emerge as an optimal policy under a single objective rather than as nine independent assertions. (The objective's utility and cost terms are parameters; what the structure fixes is the *shape* of the decision regions, which is why one parameterization recovers the whole table.)

Everything below is a term in that one objective, made rigorous.

The cost term: what a computation is worth

Escalating means spending more compute to buy more reliability. That's an energy-for- information trade, and it's measurable. On public model-routing data we computed the return on compute (accuracy bought per unit cost), the decision-relevant information each escalation actually extracts — the **mutual information** between the escalate decision and whether the cheaper model would have sufficed — and, the honest part, the **wasted compute**: escalations that spent energy without changing the answer. Chasing the last few points of quality, it turns out, costs a large fraction of non-predictive spend. The gate doesn't hide that; it quantifies it, so "escalate" becomes a priced decision instead of a reflex.

(We do *not* claim the router runs at the thermodynamic limit — it's ~20 orders of magnitude above the Landauer floor. What transfers from the physics is the efficiency *criterion*: spend information/energy only where it's predictive. That part we compute exactly.)

The caution term: knowing which kind of uncertain

"Abstain when you don't know enough" needs a definition of "enough." Most systems use a count threshold — *seen 30 outcomes? okay, trust it.* That's a base-rate-blind proxy, and it gets cases backwards. Under the Bayesian model a calibrated system already maintains, the posterior decomposes a cell's *predictive* uncertainty into two components:

- **epistemic** (reducible) — uncertainty about the rate itself, from limited data. More outcomes shrink it.
- **aleatoric** (irreducible) — the world's own noise. A fair coin stays a fair coin.

Measured by posterior variance, a balanced cell at 30 samples is **3.1× more uncertain** than a near-deterministic cell at 25 that the count threshold abstains on — the threshold has it exactly backwards. And the split tells you *which* non-act verdict you're in: high epistemic uncertainty means **escalate** (gather more, it'll help); high aleatoric means an **honest null** — you've measured it precisely, and the honest answer is "it's a coin flip."

That last distinction is worth dwelling on, because it's where this gate differs from a conventional guardrail. Most guardrails sort actions into *safe* vs *unsafe*. This one separates **"unknown"** (we haven't seen enough — escalate, and it will help) from **"fundamentally unknowable at this resolution"** (we've measured it and there is no signal to extract — so claiming one would be the dishonest move). Abstaining on the second is false humility; treating the first as if it were the second wastes a learnable opportunity. Knowing which is which is the difference between a system that defers wisely and one that just defers. We wired this into the gate behind a flag: skewed cells ground sooner, balanced cells are held to the same bar, and nothing changes unless you turn it on.

The honesty term: making "we saved X" causal

A trust gate's whole pitch is that it prevents bad actions. But you can only ever observe the outcome of the path you *took* — when the gate escalates, you never see what acting would have done. So a naive "we prevented X failures" is a correlation, not a cause. The fix is to let the gate occasionally explore (act when it would normally escalate) at a known rate, then reweight those samples. With that, standard off-policy estimators recover **an unbiased estimate of the value of acting, under the usual off-policy assumptions** (known propensities and overlap) — where the naive number is off by 22 percentage points — and a doubly-robust version does it with 42% less estimator variance.[1] That turns "things went fine while we escalated" into "acting here would fail this much more often, and here's the confidence interval." The honest boundary: for high-stakes actions the gate never explores, so *overlap* fails and that counterfactual is genuinely unidentifiable — and we refuse to estimate it rather than fake it.

[1] Synthetic benchmark with known ground truth (true value of acting 0.70 vs escalating 0.95, exploration rate ε=0.10, 3000 replications): the naive policy-value estimate reads 0.925 (bias +0.225); IPS and doubly-robust both recover 0.70 within a standard error, the latter at 42% lower variance. Reproducible from `off_policy_eval.py`.

The guarantee: staying honest when the world moves

Calibration is a number, and numbers go stale — especially here, because **agents adapt to being gated**. A gate that escalates a class of actions changes which actions the agent attempts; the policy and the gate form a feedback loop, and a guarantee proven on yesterday's distribution can quietly decay. This is a Goodhart problem most trust systems don't even name.

Conformal prediction upgrades the number to a distribution-free **coverage guarantee** — valid under exchangeability — that the true outcome lands in the gate's prediction set at least 90% of the time, and it holds even when the underlying model is wrong. Exchangeability is exactly what drift breaks, so we tested it: when we simulate the model degrading mid-stream, a frozen threshold quietly loses 8 points of coverage, while an adaptive (online) variant **maintains empirical coverage under the simulated distribution shift considered here**. We do not claim validity under arbitrary shift — adaptive conformal handles some kinds of drift, not all. The set widening from {success} to {success, fail} *is* the gate choosing to escalate — the same structure as the objective, now emitted by a guarantee.

Why this matters more than another benchmark

None of these methods is exotic on its own — that's the point. We're not trying to out-research the frontier labs on raw modeling; they'll win that. The bet is that *calibrated decision-making with honest abstention, verified against real outcomes*, is a different and more durable thing to be good at. The methods are standard; what compounds is the record of which calls were right.

And notice what every section had: an explicit "we do not claim this." The thermodynamic limit we don't reach. The coverage that's marginal, not per-decision. The counterfactuals we won't estimate without overlap. That's not hedging — for a product whose entire job is catching ungrounded overconfidence, holding its *own* claims to that standard is the whole credibility argument. A trust layer that oversells itself is the thing it warns you about.

Try it

The gate ships as a tiny SDK — `pip install warmwinter` / `npm install warmwinter` — and has no LLM in the decision path, so running it is essentially free. There's a copyable recipe to gate your own coding agent's tool calls, in shadow mode first so it watches before it ever steers. The math above is all reproducible from small scripts; the verified track record (weather, grid prices, flight delays — including where it correctly abstains) is on the site.

Act, escalate, and abstain aren't three rules. They're one objective, minimized — its cost is the value of computation, its caution is a Bayesian uncertainty decomposition, its claims are made causal by the system's own exploration, and the whole thing is wrapped in a calibration guarantee that adapts to the simulated drift we study. That's a trust layer you can actually check.

— Warm Winter · warmwinter.io

Appendix: the nine-cell mapping

The claim that the three verdicts are an optimal policy under one objective is checked concretely: for each cell of the gate's decision table — cell *state* (verified / provisional / ungrounded) × *stakes* (low / medium / high) — we compute the expected free energy of each action and take the minimizer. It matches the gate's rule in all nine cells.

state \ stakes	low	medium	high
verified	act	act	act
provisional	act	escalate	escalate
ungrounded	escalate	escalate	abstain

The boundaries follow from the objective's structure, not per-cell tuning: a *verified* cell is as good as escalating and you know it, so acting dominates at every stakes level; a *provisional* cell's risk crosses the (fixed) cost of escalating as stakes rise; an *ungrounded* cell's stakes-amplified epistemic uncertainty eventually makes every committing action cost more than deferring, which is where *abstain* wins. The script prints each action's expected free energy per cell, so the argmin is auditable rather than asserted.

Reproducibility. The quantitative results are each produced by a small, self-contained script: the nine-cell mapping (`active_inference.py`), the off-policy estimates (`off_policy_eval.py`), the conformal coverage and drift test (`conformal_gate.py`), and the epistemic/aleatoric decomposition (`epistemic_uncertainty.py`). Available with the paper at warmwinter.io.

References

The methods here are deliberately standard; the contribution is composing them into a calibrated decision layer verified against real outcomes.

1 Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). *Active inference and epistemic value*. Cognitive Neuroscience, 6(4), 187–214.

2 Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). *Active inference on discrete state-spaces: A synthesis*. Journal of Mathematical Psychology, 99.

3 Kendall, A., & Gal, Y. (2017). *What uncertainties do we need in Bayesian deep learning for computer vision?* Advances in Neural Information Processing Systems (NeurIPS) 30.

4 Hüllermeier, E., & Waegeman, W. (2021). *Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods*. Machine Learning, 110(3), 457–506.

5 Horvitz, D. G., & Thompson, D. J. (1952). *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47(260), 663–685.

6 Dudík, M., Langford, J., & Li, L. (2011). *Doubly robust policy evaluation and learning*. International Conference on Machine Learning (ICML).

7 Jiang, N., & Li, L. (2016). *Doubly robust off-policy value evaluation for reinforcement learning*. International Conference on Machine Learning (ICML).

8 Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

9 Angelopoulos, A. N., & Bates, S. (2023). *Conformal prediction: A gentle introduction*. Foundations and Trends in Machine Learning, 16(4), 494–591.

10 Gibbs, I., & Candès, E. (2021). *Adaptive conformal inference under distribution shift*. Advances in Neural Information Processing Systems (NeurIPS) 34.

11 Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). *Thermodynamics of prediction*. Physical Review Letters, 109(12), 120604.

12 Landauer, R. (1961). *Irreversibility and heat generation in the computing process*. IBM Journal of Research and Development, 5(3), 183–191.

13 Parrondo, J. M. R., Horowitz, J. M., & Sagawa, T. (2015). *Thermodynamics of information*. Nature Physics, 11(2), 131–139.