AM I STILL COUPLED TO MY ENVIRONMENT? TASK-AGNOSTIC RUNTIME TELEMETRY FROM THE REINFORCEMENT-LEARNING INTERACTION LOOP

Anonymous authors
Paper under double-blind review

ABSTRACT

After deployment, reward reports whether a reinforcement-learning agent is achieving its designed objective, but not necessarily whether the interaction producing that outcome remains characteristic of healthy operation. As a task-specific scalar projection, reward can map different agent–environment interaction regimes to similar outcomes. We therefore ask whether the standard observation, action and next-observation triplet, (S, A, S'), can provide a complementary, reward-independent runtime reading. We introduce the interaction entropy budget $C = H(S) + H(A) + H(S')$ and its coupling share $P = I(S, A; S')/C$ as a first-person, reward-independent reading of the realised MDP transition. This is an accounting view of standard Shannon quantities, not a new MDP or information-theoretic formalism. Combined with the residual-uncertainty companions H_f, H_b and ΔH, it yields an individually calibrated runtime monitor that uses the current interaction stream under a fixed representation, without reward, a predefined fault model or policy internals. Across 21 frozen SAC and PPO agents in HalfCheetah, the monitor detected 69.0% of 168 perturbation trials, compared with 44.0% for matched windowed reward; the agent-clustered advantage was 25.0 percentage points (95% interval [14.9,35.1]), and each alarmed on 5 of 21 unperturbed controls. In a within-cohort ablation at the same five-control-alarm count, the budget normalisation detects 9 points more than raw mutual information and 18 more than the tested min-entropy normalisation. Across these perturbations, the monitor detects interaction-regime changes that the task-outcome projection can miss. It complements rather than replaces reward: it asks whether the agent is still interacting with its environment in the characteristic way established during healthy operation. Such telemetry could flag interaction drift or partial decoupling and help interpret whether a reward change accompanies a broader change in the closed loop. For deployed autonomous systems, the results show why tracking task outcomes alone can be insufficient for runtime monitoring: operators may also need to verify that the underlying control loop remains characteristic of healthy operation.

1 INTRODUCTION

Reinforcement-learning agents are deployed into worlds that keep changing: sensors drift, actuators wear, and terrain or loads vary. Reward answers a task-level question — is the designed objective being achieved? — but it is a scalar projection of the underlying process. Different closed-loop regimes can therefore yield similar reward, so reward cannot in general answer the relational question: is the agent still interacting with its environment in the characteristic way established while healthy?

We ask whether the standard one-step MDP interaction can provide that complementary reading. Each step supplies the observation S, action A, and next observation S': what the agent saw, what it did, and what followed. A reading is *first-person-available* when its runtime inputs are limited to this operational stream, fixed monitor parameters, and the agent's own healthy calibration—not

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reward, a fault model, or policy weights and activations. The experiment computes the reading externally; first-person availability describes the information required, not where the monitor runs.

The access conditions are not themselves new: drift tests and transfer entropy can use the same stream. The difference is the object constructed (Table 1). We total the triplet's marginal entropies as an *interaction entropy budget* and read the share expressed as realised transition coupling. The contribution is this budgeting view, its first-person runtime interpretation, and its operational evaluation as reward-independent telemetry—not new MDP machinery or new definitions of entropy and mutual information.

Table 1: Runtime-monitoring perspectives. "First-person-compatible" means that the runtime information can come from the agent's stream and healthy calibration.

Approach	Object or readout	Operational perspective
Return monitoring	task outcome	requires reward at runtime
Input drift / MMD	distributional discrepancy	first-person-compatible with a stored healthy sample
Policy uncertainty	internal confidence	requires model internals
Empowerment	potential action–outcome influence	requires a learned action–outcome model
Transfer entropy / directed information	directed dependence between streams	first-person-compatible; closest methodological precedent
This work	coupling share of the (S, A, S') interaction entropy budget	first-person-compatible, individually threshold-calibrated runtime telemetry

From sliding windows of (S, A, S') we compute standard Shannon quantities. The central construction divides $I(S, A; S')$ by $C = H(S) + H(A) + H(S')$. Unlike normalisation by the largest possible numerator, this budgeting ratio asks what share of aggregate marginal entropy is expressed as transition coupling. Companion conditional entropies describe how residual uncertainty is arranged around it. We test whether departures from an agent's healthy reading reveal interaction changes that reward projects away, comparing both signals at the same cadence, calibration and threshold. "Task-agnostic" refers to the monitor's information requirements, not demonstrated cross-environment generality; representation and calibration choices remain necessary.

Our contributions are:

- We introduce the interaction entropy budget and its coupling share as a first-person, reward-independent reading of the realised MDP transition.
- With the residual-uncertainty companions H_f, H_b and ΔH, we operationalise this reading as individually calibrated runtime telemetry and make its representation dependence explicit.
- Across 21 frozen SAC/PPO agents and 168 controlled perturbations, the monitor shows a 25.0-point detection advantage over matched windowed reward (95% interval [14.9,35.1]) at an equal count of control alarms.
- A within-cohort ablation at the same five-control-alarm count assesses what the accounting contributes: the budget normalisation detects 9 points more than raw mutual information and 18 more than the tested min-entropy normalisation, and the consensus exceeds every single reading.

2 Related Work

RL safety largely shapes training rather than monitoring a frozen policy (García & Fernández, 2015; Gu et al., 2024). Deployment monitors instead read a chosen projection: input resemblance or model error in OOD detection (Liang et al., 2018; Sedlmeier et al., 2020; Haider et al., 2023), distributional discrepancy in drift detection and MMD (Gama et al., 2014; Lu et al., 2019; Gretton et al., 2012), predicted accuracy (Garg et al., 2022; Ginsberg et al., 2023), or satisfaction of a stated rule (Bartocci et al., 2018). These are different monitoring objects. A two-sample test can use the same empirical

healthy reference as our method; we claim a different reading, not privileged access, and have not benchmarked against such a test.

Empowerment measures potential action–outcome influence (Klyubin et al., 2005; Salge et al., 2014; Mohamed & Rezende, 2015); active inference evaluates error against a generative model (Friston, 2009; Friston et al., 2017); directed information and transfer entropy measure flow in feedback and sensorimotor streams (Massey, 1990; Schreiber, 2000; Lungarella & Sporns, 2006). The latter are the closest first-person-compatible precedent, whereas we monitor an individually calibrated accounting of the realised three-variable loop. Change detection supplies the evaluation logic (Page, 1954; Basseville & Nikiforov, 1993), and finite-sample entropy estimation constrains interpretation (Miller, 1955; Grassberger, 1988; Paninski, 2003).

3 METHOD

3.1 WHAT IS OBSERVED

At each step the monitor receives the observation acted on (S), the action (A), and the next observation (S'). It does not read reward, policy weights or activations; computation may sit inside or outside the agent.

The tap points matter and are chosen to be the ones an agent has. S and S' are the observations as delivered to the policy, so under sensor perturbation they carry the corruption the policy acted on rather than a clean value no agent could see. A is the action the policy commanded. Under actuator perturbation the environment executes a corrupted version of it, which the agent does not observe — that unobservability is what actuator degradation is. The triplet is therefore what the agent has access to, not a privileged record of the environment, and a perturbation is visible through the loosened relation between what was commanded and what followed rather than through any direct reading of the fault.

3.2 INTERACTION ENTROPY BUDGET AND ITS READINGS

Over a window of triplets we estimate their joint distribution and standard Shannon entropies and mutual information, all in bits (Shannon, 1948; Cover & Thomas, 2006). The construction is the following accounting applied to those quantities.

Interaction entropy budget and coupling share. We define the gross, streamwise interaction entropy budget and the share expressed as transition coupling:

$$C = H(S) + H(A) + H(S'), \qquad P = \frac{I(S, A; S')}{C}. \tag{1}$$

"Budget" is an accounting interpretation, not a conservation law: C sums the three marginal entropies; it is neither their joint entropy nor the maximum possible numerator. Thus P is not conventional normalised mutual information. It is the share of aggregate marginal entropy expressed as dependence between the current observation–action pair and what follows.

Standard mutual-information inequalities give the resulting budget share a policy-dependent upper bound

$$0 \le P \le \frac{1}{2} - \frac{I(S; A)}{2C} \le \frac{1}{2} \tag{2}$$

for $C > 0$, with the last inequality strict whenever the window distribution has $I(S; A) > 0$. This follows from standard information inequalities, not a new theorem; Appendix E gives the derivation and an estimator check.

The companion readings are forward and backward uncertainty,

$$H_f = H(S' \mid S, A), \qquad H_b = H(S, A \mid S'), \tag{3}$$

and their asymmetry,

$$\Delta H = H_f - H_b = H(S') - H(S) - H(A \mid S). \tag{4}$$

Here H_f is outcome uncertainty after observing (S, A), and H_b is uncertainty in (S, A) after observing S'. Since $S'_t = S_{t+1}$, $H(S') - H(S)$ is only an endpoint term (measured mean 0.004 bits), so ΔH approximates $-H(A \mid S)$.

That identity needs reading carefully here. All quantities in this paper are discrete Shannon entropies of the binned symbols $\tilde{S}, \tilde{A}, \tilde{S}'$ defined in Section 3.3; we drop the tildes elsewhere for readability. Evaluation is deterministic, so the commanded continuous action is a deterministic function of the continuous observation, and $H(\tilde{A} \mid \tilde{S})$ is positive only because one observation bin contains states mapped to several action bins. On a representative healthy run it is 0.633 bits and matches $-\Delta H$ to three decimals. ΔH therefore reports how far the binned observation determines the binned action, and it moves when a perturbation changes the visited distribution and hence that correspondence. It is not residual latitude in the agent, which has none under a deterministic policy. It is derived from H_f and H_b, not an independent fourth feature; the alarm below is a calibrated filter over correlated readings.

3.3 Estimation

HalfCheetah is continuous, so each of 40 variables per step (17 observation, 6 action, 17 next observation) is z-scored and placed into three equal-width bins spanning its range in one healthy reference window; out-of-range values enter the nearest outer bin. That window, from one reference agent, fixes normalisation and bin edges for all 21 agents and both algorithms. Only alarm thresholds are calibrated per agent. Refitting edges per agent lowers detection from 69.0% to 61.3%, confirming representation dependence without establishing which mapping is physically preferable.

The 17 observation bins are concatenated into a composite symbol for S (and likewise for S'); the six action bins form A. Corresponding observation and next-observation coordinates share the same normalisation constants and bin edges, and the logged S'_t is the value recorded as S_{t+1}, so the encoded sequence identity that the endpoint term of Equation 4 relies on holds by construction. Statistics use $W = 500$-step windows with stride 50 and uncorrected plug-in estimates (Miller, 1955; Paninski, 2003). Triplets never cross resets; a window may contain two valid episode segments. Absolute values are therefore configuration-specific, while every condition shares the same estimator. Further implementation details appear in Appendix A.

3.4 Detection

During an initial healthy period, the monitor freezes the mean μ_m and standard deviation σ_m of each statistic $m \in \{P, H_f, H_b, \Delta H\}$. It subsequently marks a statistic when $|m_t - \mu_m| > 3\sigma_m$. Marks persist for the trial, and the monitor alarms once k distinct statistics have been marked: $k_t = \sum_m \mathbf{1}[\exists\, u \le t : |m_u - \mu_m| > 3\sigma_m]$, with an alarm when $k_t \ge k$. Thus readings need not cross simultaneously. Section 5.1 evaluates $k \in \{1, 2, 3, 4\}$ against perturbation and control trials and selects $k = 3$; Algorithm 1 in Appendix A gives pseudocode.

3.5 The reward baseline

Per-step reward is averaged over the same windows, calibrated over the same healthy period, and thresholded at the same $\pm 3\sigma$. The comparison therefore changes the signal read, not its evidence schedule.

4 Experimental setup

4.1 Environment and agents

We use HalfCheetah-v4 (Todorov et al., 2012) with frozen SAC (Haarnoja et al., 2018) and PPO (Schulman et al., 2017) policies trained in Stable-Baselines3 (Raffin et al., 2021). SAC reaches roughly 10k–17k return over about 3M steps and PPO roughly 6k–8k over about 1.5M, yet the same monitor configuration is applied to both without retuning: no runtime reward or reward-function parameter enters the monitor, though the representation remains environment-specific. All policies are frozen before evaluation, so any change in the statistics comes from the perturbation and not from continued learning.

4.2 SEED SELECTION

Evaluation uses 21 agents, 11 SAC and 10 PPO, fixed before perturbations or detection outcomes were examined. Sixteen PPO seeds were run under the protocol and ten retained; every SAC seed run under the protocol was retained. Seeds were set aside on inspection of their unperturbed return traces, which were unstable. No numerical threshold was recorded for that judgement, so we describe the cohort as a retained stable-policy cohort rather than claim a reproducible criterion; the screening did not select directly on perturbation-detection outcomes, and the runs it set aside are in the released archive. Because nominal stability also affects calibration variance, this cohort is not a random sample of trained policies.

4.3 PERTURBATIONS

Each evaluation runs 50 episodes of 1,000 steps. Episodes 0–14 are unperturbed and provide the calibration window; the perturbation is introduced at the start of episode 15, at step 15,000, and held for the remainder. Calibration uses the 291 windows with $\mathrm{id} \leq 290$, which lie wholly before the injection; detection uses windows from 300 onwards, the 691 lying wholly after it. Windows 291–299 straddle the injection and are excluded from both. Including those nine in calibration, as an earlier version of this work did, inflates σ_m and lowers detection from 69.0% to 66.7% while reward detection remains 44.0%; that variant is reported in Appendix B and is not used for any headline figure. Eight conditions span agent-side and environment-side change at several severities (Table 2). Agent-side conditions degrade what the agent receives or emits; environment-side conditions change the world it acts in.

Agent-side perturbations are stochastic, whereas environment-side perturbations are systematic, confounding entry point with mechanism.

Table 2: Perturbation conditions. Each is applied to all 21 agents.

Side	Condition	Parameter
Agent	Action noise	1%, 3%, 4% Gaussian noise on actions
Agent	Observation noise	1%, 3% Gaussian noise on observations
Environment	External force	5 N, 10 N applied to torso, x-axis
Environment	Gravity	increased to 110%
Unperturbed control — identical protocol, no perturbation applied		

Across 21 agents, we evaluate 168 perturbation trials and 21 matched controls used to estimate false-alarm rates.

5 RESULTS

The experiment tests whether departures from an agent's healthy interaction-budget reading expose changes that reward omits. Comparisons are paired on the same trials, windows, calibration period and threshold protocol. The dataset comprises 21 agents (11 SAC, 10 PPO), 168 perturbation trials and 21 controls.

5.1 MAKING THE SIGNAL USABLE: CALIBRATION AND CONSENSUS

The alarm rule controls the sensitivity–false-alarm trade-off. P alone detects 75.0% of perturbations but alarms on 9 of 21 controls. One- and two-of-four consensus rules are noisier still, whereas four-of-four loses the detection advantage (Appendix B). We therefore report three-of-four as an empirical operating point: it detects 69.0% with 5 of 21 control alarms, matching reward's own 5. The rule suppresses single-channel excursions at some cost in sensitivity; because the readings are correlated, this is a filter rather than a vote among independent sensors.

Since $\Delta H = H_f - H_b$, a four-way count runs over three independent readings, so it is fair to ask whether ΔH supplies a free third mark. Both H_f and H_b are marked in 106 trials and ΔH in 90 of them — often, not automatically, since a difference can stay inside its own calibrated band while

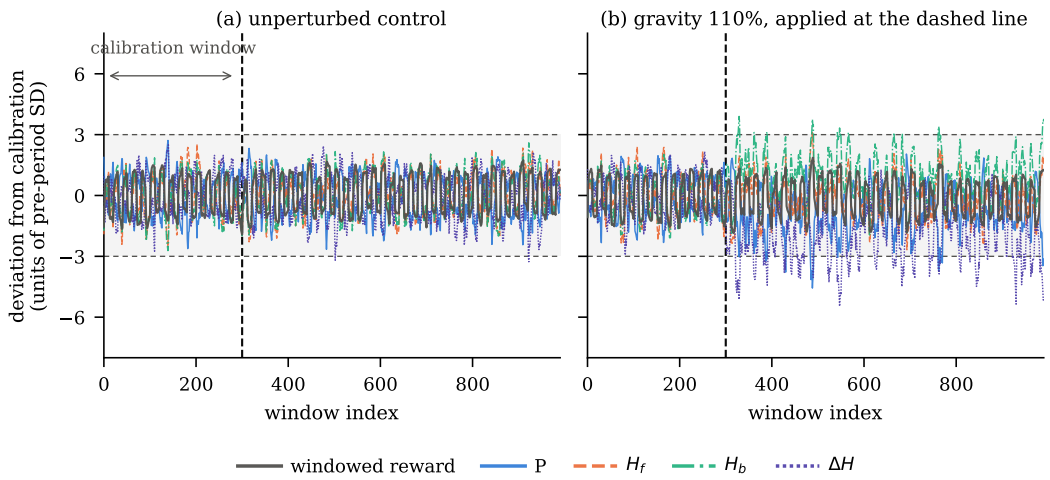


Figure 1: Representative SAC trials expressed as deviations from each trial's calibration window: (a) a median control and (b) a median detected perturbation. One statistic crosses $\pm 3\sigma$ in (a); all four cross while reward remains within the band in (b).

both terms leave theirs; in the 32 trials where neither is marked, ΔH is marked in 11. Of the 116 detected trials, 7 are detected without P marked at all, which is the route the algebra permits. Over the independent set $\{P, H_f, H_b\}$, two-of-three detects 73.2% at 7 control alarms and three-of-three 58.3% at 3; three-of-four lies between them. The four-statistic pool was fixed before the rule was chosen, so these are reported as alternatives rather than candidates passed over.

What the normalisation contributes. Because detection and false alarms trade against each other, single readings are only comparable at a matched budget. For each statistic we set z midway between the fifth- and sixth-largest trialwise control excursions, in these data yielding exactly five control alarms — reward's own count. A control trial's excursion is its largest absolute standardised deviation from its calibration mean. The threshold is therefore fixed by the controls alone, and no perturbation outcome enters its choice. Reading detection at that point (Table 3) separates the accounting from the quantities it is applied to. Dividing realised mutual information by the interaction entropy budget is worth 9 points over the raw quantity and 18 over the min-entropy normalisation $\min\{H(S,A), H(S')\}$. The comparison uses the same 21 controls that set the thresholds, so it is a within-cohort descriptive ablation rather than held-out validation. The consensus is not redundant either: at the same budget it exceeds every single reading, including the best companion entropy. Neither result follows from the definitions, and we did not anticipate the second.

Table 3: Within-cohort ablation at a matched control-alarm count. Each statistic is used as a single latched detector on the same windows with the same calibration, its threshold set from the controls alone to admit exactly 5 of 21; z is the resulting multiplier. The consensus row is the three-of-four rule at $\pm 3\sigma$, which admits 5 controls without tuning.

Statistic	z	Detection
$I(S,A;S')/\min\{H(S,A), H(S')\}$	3.46	41.1%
Windowed reward	2.98	44.0%
$I(S,A;S')$, unnormalised	3.30	50.0%
ΔH	3.50	50.6%
$I(S,A;S')/H(S,A,S')$	3.16	52.4%
$P = I(S,A;S')/C$	3.36	58.9%
H_f	3.33	60.7%
H_b	3.22	60.7%
Three-of-four consensus	$\pm 3\sigma$	**69.0%**

The statistic is reward- and fault-label-free at runtime. The consensus level, by contrast, was chosen during this study using labelled perturbation outcomes and the reward comparison: it was selected after inspecting the candidates, on the same 21 controls used to score it. The reported operating point is therefore exploratory rather than an independently validated optimum, and the two kinds of independence should not be confused. The complete rule table, selection criterion, and a leave-one-agent-out check are reported in Appendix B.

5.2 What the loop shows that the outcome does not

At the selected rule the monitor detects 69.0% of the 168 perturbation trials against reward's 44.0%, catching 50 trials reward misses and missing 8 that reward catches (pooled descriptive $p = 1.6 \times 10^{-8}$), with both alarming on 5 of 21 controls. Because each agent supplies eight perturbation trials, we resample whole agents (10,000 resamples of the 21 agents): the advantage is 25.0 percentage points with a 95% interval of [14.9,35.1], and no resample reverses its sign. Per-rate intervals are [58.9,78.6] for the monitor's 69.0% and [33.3,56.0] for reward's 44.0%; both control-alarm rates are 23.8% on the same interval, [9.5,42.9]. The control set was also used in selecting the rule, so this is a descriptive comparison rather than an out-of-sample equivalence result. The advantage appears for both SAC and PPO and the point estimate is higher in all eight conditions (Appendix B). External force at 5 N is the sharpest case for the comparator: reward detects 23.8% of those trials at its own 23.8% control-alarm rate, while the loop monitor reaches 47.6%.

These results support the stated role: the reward-independent reading detects changes in the realised loop that the task-outcome projection can omit. They do not show that every departure is harmful or that this is the strongest possible change detector. We did not compare against a distributional two-sample detector on the same stream.

5.3 What the loop does not show

The telemetry is not early-warning or diagnostic machinery. Among the 66 trials both monitors detect, the median lead is zero: the monitor is earlier in 28, later in 27, and tied in 11. The four readings also fail to localise an unseen fault class: held-out-condition recall, averaged over conditions, is 51.2% against a 50% chance level. Finally, a better-sampled per-joint representation is less sensitive than the whole-body representation, establishing that absolute values and detection rates are representation-dependent. Appendix B reports these analyses and the estimator anomaly under 3% observation noise.

As a consistency check, none of 166,488 perturbation windows violates the derived bound in Equation 2; measured P averages 87.1% of its ceiling (Appendix E). This checks the implementation, not the mathematical inequality or the monitor's validity.

6 Discussion

Reward and the proposed telemetry answer different questions. Reward reports success against a designed task objective; the telemetry asks whether the realised interaction still resembles the agent's healthy closed loop. It reads the realised MDP transition through the coupling share of its interaction entropy budget and the residual-uncertainty companions H_f, H_b and ΔH. Its runtime inputs are the agent's own (S, A, S') stream, a fixed representation, and individual healthy calibration—not reward, a fault catalogue, or policy internals.

This is complementary instrumentation, not a replacement objective. An interaction alarm without a reward drop may expose a change that the task projection omits, but it may also reflect benign variation; the reading neither identifies the cause nor decides whether intervention is warranted. Its utility is to flag possible drift or partial decoupling and provide context for interpreting reward, leaving consequence and action to a task-specific layer.

Because the telemetry is constructed independently of reward, agreement and disagreement between the readings are themselves informative. A joint change indicates that both the interaction regime and task outcome have shifted; a telemetry-only change indicates that the closed loop has departed from its healthy pattern while reward remains comparatively insensitive. The telemetry therefore provides context for interpreting reward rather than competing with it.

This distinction matters for deployed autonomous systems. Tracking task outcomes alone can leave changes in the process producing those outcomes unobserved; runtime oversight may therefore need to verify both the outcome and whether the underlying control loop remains characteristic of healthy operation. The present results motivate that dual view, but do not by themselves establish a safety guarantee.

Across the perturbations studied, that reward-independent reading performs at least as well as the reward comparator, with an agent-clustered advantage. This is the empirical claim. The monitor is not deployment-ready: two-thirds detection and 5 alarms in 21 healthy runs are evidence for a complementary signal, not an acceptable operating point. The experiment is external, simulated, and confined to one environment; it does not establish universal detector superiority, early warning, or fault localisation.

7 LIMITATIONS

All results come from one simulated environment and step-change perturbations; multi-environment tests, gradual degradation, and benign operational changes remain open. Reward is the only matched baseline. A two-sample test on the same (S, A, S') stream (Gretton et al., 2012) would distinguish the proposed reading from generic multivariate change detection and is the first comparison we would add.

The 168 trials cluster within 21 agents. We therefore use the whole-agent bootstrap for across-agent uncertainty and treat pooled McNemar tests as trial-level descriptions; 21 agents still limit population generalisation. The three-of-four rule was selected and scored on the same 21 controls after candidate inspection, so its false-alarm estimate is in-sample. Leave-one-agent-out selection increases control alarms to 11/21 (Appendix B), confirming that the reported operating point is not a validated selection procedure.

Because the operating point was fixed using these faults and these controls, the reported rates describe this cohort and are not an out-of-sample estimate; a validated rule would need agents, fault classes and benign shifts untouched by any design choice.

Marks persist across roughly 690 post-calibration windows, so false alarms accumulate with exposure, which makes the control counts a function of trial length rather than a rate. A time-to-alarm characterisation on long nominal runs would give the exposure-free quantity, and we do not have those runs. Requiring simultaneous crossings detects 57.1% with 2/21 control alarms; persistence or sequential evidence could offer other trade-offs. Plug-in entropy estimates over sparse composite alphabets are biased (Miller, 1955; Paninski, 2003), and the alternative representation changes sensitivity. Absolute levels and rates are estimator- and representation-specific even though all paired comparisons here share one estimator.

8 CONCLUSION

We introduce the interaction entropy budget $C = H(S) + H(A) + H(S')$ and its coupling share $P = I(S, A; S')/C$ as a first-person, reward-independent reading of the realised MDP transition. With H_f, H_b and ΔH, it yields individually calibrated runtime telemetry that asks whether an RL agent is still interacting with its environment in the characteristic way established while healthy. Across 21 frozen SAC/PPO agents and 168 HalfCheetah perturbations, the monitor shows a 25.0-point detection advantage over matched reward at an equal count of control alarms. This supports its use as a complementary indication of interaction drift or partial decoupling that can help interpret, rather than replace, task reward. For deployed autonomous systems, the broader implication is that task outcomes alone may be insufficient for runtime monitoring: operators may also need to verify that the underlying control loop remains characteristic of healthy operation.

REPRODUCIBILITY STATEMENT

All results derive from a fixed set of evaluation logs and a small number of analysis scripts. Section 4 states the environment, the training algorithms and their libraries, the number of seeds, the

seed-screening rationale and the retained and excluded counts, the perturbation conditions and their parameters, and the calibration and evaluation split, including the signal tap points and the deterministic action selection used at evaluation. Section 3 gives the four statistics in closed form, the discretisation, the window length and stride, and the alarm rule; Algorithm 1 supplies pseudocode. The post-hoc rule selection and reuse of controls for selection and scoring are stated in the main text, with full details in Appendix B.

The per-window metric series for every trial, the per-trial summary table from which all detection and false-alarm counts are computed, the frozen policy checkpoints and normalisation statistics, and the analysis scripts that regenerate every table in this paper accompany the submission. The bootstrap uses 10,000 whole-agent resamples with a fixed seed, stated in the scripts; entropies are uncorrected plug-in estimates over the composite alphabets defined in Section 3.3, with zero-entropy windows assigned $P = 0$. Every analysis reported here uses the full retained cohort: 21 seeds, 168 perturbation trials and 21 unperturbed controls.

AI USE STATEMENT

We used generative AI assistants as follows. *Tasks in the required-disclosure category:* feedback on analysis methodology — calibrating the alarm rule against the unperturbed controls rather than assuming a threshold, and the choice of paired tests for the reward comparison — assistance in interpreting the resulting tables, and scripting of the rule-selection and hypothesis-testing analysis, every output of which the authors checked against the per-trial summary data. The environments, agents, perturbation design, evaluation protocol and per-window metric computation predate this use and were designed and run by the authors. We did not use these tools to generate data, to formulate or prove mathematical claims, or to propose the research hypothesis. We did not use generative AI for any other required-disclosure task; all remaining required-disclosure categories were not applicable to this work. *Tasks in the recommended-disclosure category:* drafting and revising the manuscript, compressing it to the page limit, auditing the reference list, and suggesting the title. We have reviewed all AI-assisted work: every citation was checked against its primary source and all reported quantities traced to the evaluation logs. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Ezio Bartocci, Yliès Falcone, Adrian Francalanza, and Giles Reger. Introduction to runtime verification. In *Lectures on Runtime Verification*, volume 10457 of *Lecture Notes in Computer Science*, pp. 1–33. Springer, Cham, 2018. doi: 10.1007/978-3-319-75632-5_1.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Englewood Cliffs, NJ, 1993. ISBN 9780131267800.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006. ISBN 9780471241959. doi: 10.1002/047174882X.

Karl Friston. The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009. doi: 10.1016/j.tics.2009.04.005.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017. doi: 10.1162/NECO_a_00912.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 2014. doi: 10.1145/2523813.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

Tom Ginsberg, Zhongyuan Liang, and Rahul G. Krishnan. A learning based hypothesis test for harmful covariate shift. In *International Conference on Learning Representations (ICLR)*, 2023.

Peter Grassberger. Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, 128(6–7):369–373, 1988. doi: 10.1016/0375-9601(88)90193-4.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, 2024. doi: 10.1109/TPAMI.2024.3457538.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Tom Haider, Karsten Roscher, Felippe Schmoeller da Roza, and Stephan Günnemann. Out-of-distribution detection for reinforcement learning agents with probabilistic dynamics models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 851–859. IFAAMAS, 2023.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC)*, volume 1, pp. 128–135. IEEE, 2005. doi: 10.1109/CEC.2005.1554676.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019. doi: 10.1109/TKDE.2018.2876857.

Max Lungarella and Olaf Sporns. Mapping information flow in sensorimotor networks. *PLoS Computational Biology*, 2(10):e144, 2006. doi: 10.1371/journal.pcbi.0020144.

James L. Massey. Causality, feedback and directed information. In *Proceedings of the 1990 International Symposium on Information Theory and Its Applications (ISITA-90)*, Waikiki, Hawaii, November 1990.

George A. Miller. Note on the bias of information estimates. In Henry Quastler (ed.), *Information Theory in Psychology: Problems and Methods*. Free Press, Glencoe, IL, 1955.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, 2015.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment—an introduction. In Mikhail Prokopenko (ed.), *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*, pp. 67–114. Springer, Berlin, Heidelberg, 2014. doi: 10.1007/978-3-642-53734-9_4.

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000. doi: 10.1103/PhysRevLett.85.461.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Andreas Sedlmeier, Thomas Gabor, Thomy Phan, Lenz Belzner, and Claudia Linnhoff-Popien. Uncertainty-based out-of-distribution classification in deep reinforcement learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART), Volume 2*, pp. 522–529. SCITEPRESS, 2020. doi: 10.5220/0008949905220529.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. Continued in vol. 27, no. 4, pp. 623–656.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

A Implementation details

Bins are concatenated within front-leg, back-leg and torso groups only as bookkeeping: the final observation symbol is the same 17-variable tuple, so bijective relabelling does not change any reported entropy. Three bins were fixed before analysis. Windows follow the step stream across episode boundaries, but every triplet remains within an episode; a spanning window contains at most one boundary and two valid trajectory segments.

Algorithm 1 Interaction-stream monitoring

Require: stream (S_t, A_t, S'_t); window W; stride δ; fixed mapping; threshold z; consensus k; calibration period T_c; $M \leftarrow \emptyset$

for each window of W triplets, advanced by δ **do**
  apply the mapping; estimate $P, H_f, H_b, \Delta H$
  if window lies in T_c **then**
    accumulate values and then freeze μ_m, σ_m for each m
  else
    $M \leftarrow M \cup \{m : |m_t - \mu_m| > z\sigma_m\}$
    alarm if $|M| \geq k$
  end if
end for

B Secondary results and operating-point detail

Contaminated calibration. Calibrating on windows 0–299 rather than 0–290 admits nine windows that straddle the injection. This inflates σ_m and lowers detection to 66.7% against reward's unchanged 44.0%, with 5/21 monitor and 4/21 reward control alarms and 48 versus 10 discordant trials ($p = 4.5 \times 10^{-7}$). The monitor retains a 22.6-point detection advantage; no headline figure uses this split.

We selected the rule with the largest detection gain among candidates whose paired control-alarm excess over reward was not statistically distinguishable. This criterion uses failure to reject rather than equivalence, was formulated after inspecting the candidates, and reuses the same controls for selection and scoring. One- and two-of-four are significantly noisier than reward; four-of-four loses the detection advantage (Figure 2, Table 5). Selecting k on 20 agents and scoring the held-out agent returns $k = 2$ for 20 of the 21 folds and $k = 3$ for one, detecting 80.4% with 11/21 control alarms: the criterion does not reproduce the reported operating point.

The advantage occurs for SAC (62.5% versus 35.2%, 27 discordant trials to 3, $p = 8.4 \times 10^{-6}$) and PPO (76.2% versus 53.8%, 23 to 5, $p = 9.1 \times 10^{-4}$). Control alarms split 2/11 versus 0 for SAC

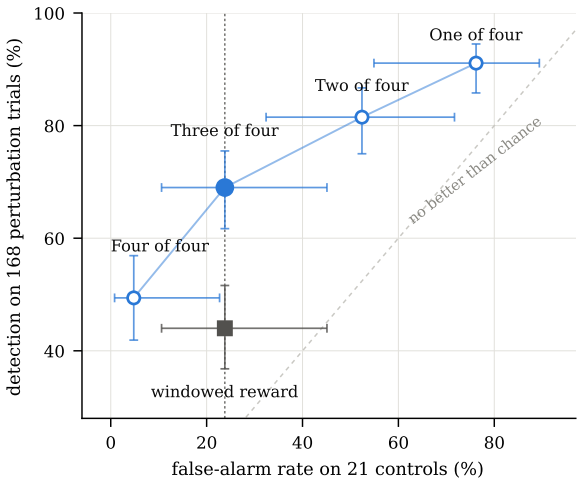


Figure 2: Detection and false-alarm rates for consensus rules, with 95% Wilson intervals. Three-of-four gains 25.0 percentage points over reward at the same count of control alarms.

and 3/10 versus 5/10 for PPO; these samples are too small for an algorithm-specific conclusion. The monitor's point estimate is higher in all eight conditions (Table 4), a consistent direction rather than eight independent findings.

Table 4: Detection by condition, 21 trials each, ordered by monitor margin.

Condition	Changes	Monitor	Reward	Margin
Observation noise 1%	observations	76.2%	38.1%	+38.1
Action noise 4%	actions	81.0%	47.6%	+33.3
Action noise 3%	actions	57.1%	28.6%	+28.6
External force 10 N	transition	61.9%	33.3%	+28.6
Action noise 1%	actions	61.9%	38.1%	+23.8
External force 5 N	transition	47.6%	23.8%	+23.8
Observation noise 3%	observations	95.2%	81.0%	+14.3
Gravity 110%	transition	71.4%	61.9%	+9.5

Lead time. Among the 66 trials both monitors detect, the median lead is zero windows: the monitor is earlier in 28, later in 27 and tied in 11, which a two-sided sign test on the 55 non-tied pairs cannot distinguish from chance ($p = 1.000$). The other 50 monitor detections have no reward alarm and hence no defined lead, so this comparison is conditioned on the subset severe enough to reach the return. Split by algorithm the two halves disagree in sign — SAC is earlier in 16 of 21 non-tied pairs (median 21 windows, $p = 0.027$) and PPO in 12 of 34 (median -1.5, $p = 0.121$) — which is why we read the pooled null rather than the subgroup. One-of-four leads by 77 windows but alarms on 76.2% of controls. We make no early-warning claim.

Fault localisation. The four readings reach 61.7% balanced accuracy when seeds are held out, and 51.2% when whole conditions are held out (Table 7). A held-out condition contains a single class, so the latter is the mean over conditions of that condition's recall rather than a balanced accuracy. $I(S; A)$, a term in the coupling-share ceiling, reaches 63.1% in the latter test, but it was not in the fixed monitor and we do not reselect the readings post hoc. Entry point is also confounded with stochastic versus systematic mechanism in this suite.

Alternative representation. A seven-channel per-joint reading, each channel combining a joint's angle, angular velocity and associated action, is much better sampled than the whole-body composite but less sensitive at comparable control-alarm rates. It changes the monitored object by discarding cross-joint dependencies, so this is not a more accurate estimate of the same P. Sparse occupancy

may amplify the whole-body response, or the per-joint form may lose informative coordination; the experiment does not distinguish these accounts.

C ADDITIONAL TABLES

Table 5: Rule selection in full (Figure 2). Detection and false-alarm rates with 95% Wilson intervals, and exact McNemar tests against reward. Pooled over trials and therefore descriptive: the 168 trials are not independent, and agent-clustered intervals are given in Section 5. Bold marks the selected rule, which under the exploratory in-sample criterion is the one whose control-alarm excess over reward is zero while its detection gain is largest.

Rule	Detection (95% CI)	False alarm (95% CI)	McNemar p vs reward	
			detection	false alarm
One of four	91.1% [85.8, 94.5]	76.2% [54.9, 89.4]	7×10^{-23}	0.003
Two of four	81.5% [75.0, 86.7]	52.4% [32.4, 71.7]	3×10^{-17}	0.070
Three of four	69.0% [61.7, 75.5]	23.8% [10.6, 45.1]	1.6×10^{-8}	1.000
Four of four	49.4% [41.9, 56.9]	4.8% [0.8, 22.7]	0.28	0.219
Reward	44.0% [36.8, 51.6]	23.8% [10.6, 45.1]	—	—

Table 6: Mean $|d|$ by perturbation family and severity. Gravity is omitted because it was evaluated at only one severity.

	Action noise			Observation noise		Force	
Statistic	1%	3%	4%	1%	3%	5 N	10 N
P	0.65	0.98	1.11	0.94	1.98	0.29	0.51
H_f	0.69	1.26	1.57	1.08	2.44	0.29	0.54
H_b	0.68	1.19	1.51	1.06	2.30	0.26	0.52
ΔH	0.67	1.12	1.30	1.15	1.70	0.31	0.47
Reward	0.18	0.37	0.53	0.43	1.89	0.15	0.21

Table 7: Agent-side versus environment-side classification. Chance is 50%. Holding out seeds gives folds containing both classes, scored by balanced accuracy; holding out a condition gives a single-class fold, scored by that condition's recall and averaged over the eight conditions. Holding out seeds leaves the same conditions in training; holding out conditions tests whether a signature transfers to an unseen fault. All 168 perturbation trials across 21 seeds. Features are the calibration-referenced effect sizes $(\mu_{\text{post}} - \mu_{\text{pre}})/\sigma_{\text{pre}}$; the classifier is L2-regularised logistic regression with balanced class weights on features standardised within each training fold.

Features	Leave-one-seed-out	Leave-one-condition-out
The four statistics	61.7%	51.2%
$I(S;A)$ alone	66.5%	63.1%
The four plus $H(A)$ and $I(S;A)$	68.9%	63.1%

D DIAGNOSTIC DETAILS

Episode boundaries. Episodes are 1,000 steps and windows 500, so a window contains at most one boundary. It contains two valid trajectory segments rather than a terminal-to-reset transition. The identity $S'_t = S_{t+1}$ used in Section 3.2 fails at that one position, doubling the endpoint bound from 0.036 to 0.07 bits at $K \approx 163$, against a measured mean of 0.004.

Table 8: Mean signed effect size by condition, relative to each trial's own pre-perturbation baseline.

Condition	d_P	d_{H_f}	d_{H_b}	$d_{\Delta H}$
Action noise 1%	-0.56	$+0.58$	$+0.54$	-0.48
Action noise 3%	-0.39	$+0.88$	$+0.93$	-0.93
Action noise 4%	$+0.18$	$+0.64$	$+0.72$	-0.80
External force 5 N	-0.08	$+0.05$	$+0.10$	-0.17
External force 10 N	-0.23	$+0.23$	$+0.28$	-0.34
Gravity 110%	-0.76	$+0.04$	$+0.03$	-0.11
Observation noise 1%	-0.06	$+0.19$	$+0.47$	-0.87
Observation noise 3%	$+1.71$	-1.79	-1.47	$+0.56$

Table 9: A factorial perturbation design. The three cells in bold are absent from our suite; they are the combinations needed to separate entry point from mechanism.

Entry point	Stochastic	Systematic
Observation	Gaussian noise on observations	**sensor gain, bias, frozen channel**
Action	Gaussian noise on actions	**actuator gain loss, offset, lag**
Transition	**random force impulses**	gravity, mass, friction change

Dependent votes. Both H_f and H_b are marked in 106 of 168 perturbation trials, and ΔH in 90 of those; a difference can remain inside its own calibrated band while both terms leave theirs. In the 32 trials where neither is marked, ΔH is marked in 11. Seven of the 116 three-of-four detections occur without P, through the $\{H_f, H_b, \Delta H\}$ route.

Observation-noise inversion. At 3% observation noise, $H(S)$ rises from 6.68 to 7.63 bits against the $\log_2 W = 8.97$ window ceiling. $I(S, A; S')$ rises by 1.10 bits while the interaction entropy budget C rises by 2.45; the ratio increases because the numerator grows proportionally more. The occupied alphabet expands most in this condition, matching the direction and timing expected from plug-in bias, but a corrected estimator is needed to test that account.

Alarm exposure. In the five control trials where the monitor alarms, each marked statistic crosses its threshold in 1–16 of roughly 690 windows, and the three marks that trigger the alarm occur at separate times. Reward's five control alarms span 2–27 windows. The alarms are therefore built from brief excursions rather than sustained departures, which is what makes them accumulate with exposure; we do not read the two ranges as distinguishing the monitors.

E THE COUPLING-SHARE BOUND AND AN ESTIMATOR CHECK

P has an upper bound that follows from its definition. Mutual information cannot exceed the entropy of either argument, so $I(S, A; S') \le \min\{H(S, A), H(S')\}$, and the smaller of two numbers is at most their mean, giving $I(S, A; S') \le \frac{1}{2}[H(S, A) + H(S')]$. Substituting $H(S, A) = H(S) + H(A) - I(S; A)$ and dividing by $C = H(S) + H(A) + H(S')$,

$$P \;\le\; \frac{1}{2} - \frac{I(S; A)}{2C}. \tag{5}$$

The chain holds for $C > 0$, and the final inequality $\frac{1}{2} - I(S; A)/(2C) \le \frac{1}{2}$ is strict whenever the window distribution has $I(S; A) > 0$. Reading actions from observations does not by itself guarantee that on every window distribution, so the strictness is a property of the estimated distribution rather than of the policy. We do not show the bound is reached.

Across all 168 perturbation trials at 991 windows each — 166,488 windows — C averages 18.68 bits, the policy-dependent ceiling averages 0.383, and measured P averages 0.334, or 87.1% of that ceiling. No window exceeds the bound. This verifies implementation consistency with the derived budget-share bound; it is not empirical validation of a mathematical inequality or a monitoring signal, and nothing in the detection results depends on the absolute level of P.