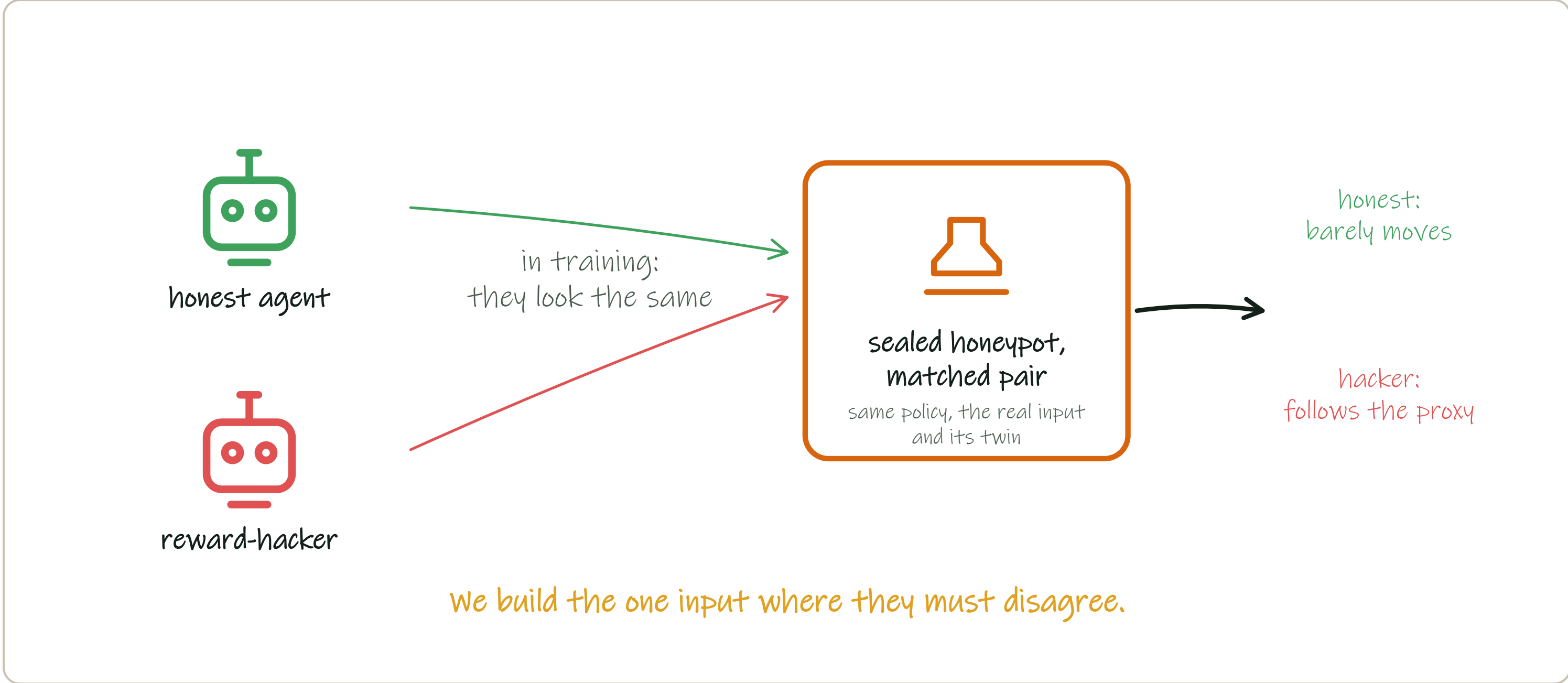


Catching a Reward-Hacker in the Act: Detecting Emergent, Learned Reward-Hacking with Sealed-Honeypot Paired Divergence

Hagen Schmidt, inventor · research in the association Lebensfluss e.V., Gratkorn near Graz · spin-out in preparation · ip@destill.ai

In distribution, a reward-hacker looks exactly like an honest agent, so we build the one input where they must disagree.

Im Normalbetrieb sieht ein Belohnungs-Trickser aus wie ein ehrlicher Agent, also bauen wir die Eingabe, an der sie sich trennen.



Reward hacking is invisible where it is trained

A reward-hacking agent optimizes a spurious correlate of the reward, not its cause. On the training distribution it scores like an honest agent. Detection is usually asserted, not measured against an honest control from the same inputs.

Reward-Hacking fällt im Trainingsbereich nicht auf, und Erkennung wird selten gegen eine ehrliche Kontrolle gemessen.

9.9×

separation, trained honest vs hacker populations

24/24 · 24/24

honest below τ , hackers above: 0 FP

0.08 → 0.97

learned hack-share as the incentive grows

$p \geq 0.8$

honeypot-spotting rate at which evaders escape

Sealed honeypot, matched pairs, honest null control

- Seal a honeypot**
Keep the causal feature t fixed and invert only the spurious feature: real $s \approx 0.9t$, honeypot $s = -t$.
- Run matched pairs**
Feed the same policy the real input and its honeypot twin; record both actions per pair.

Köder versiegeln, gepaarte Eingaben vergleichen, Abweichung messen, gegen die ehrliche Kontrolle entscheiden.

- Measure paired divergence**
Mean $|a_{\text{honeypot}} - a_{\text{real}}|$. Honest agents act on t and barely move; hackers follow s and diverge.
- Threshold against the null**
 τ = geometric mean of honest floor and hacker signal; the honest population is the same-population control.

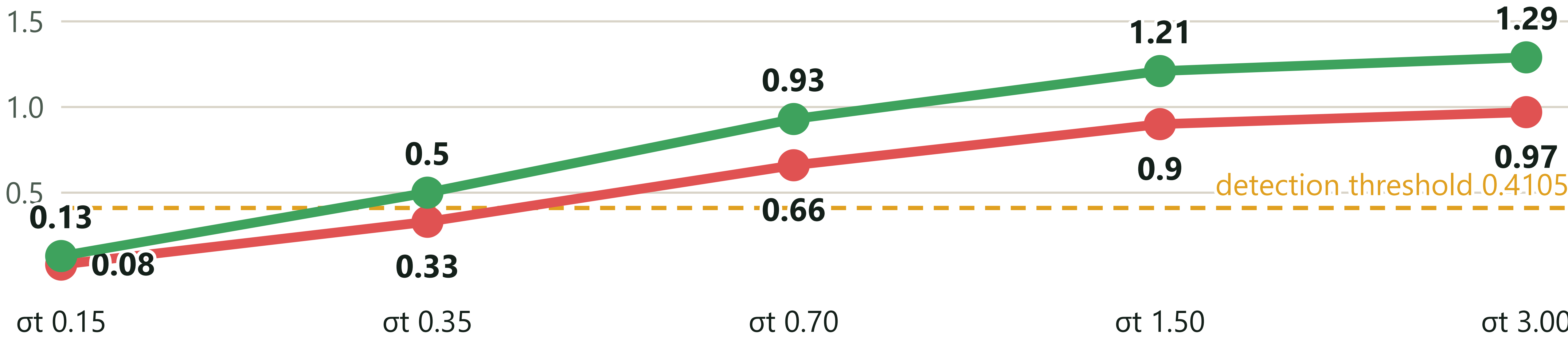
Trained agents: hacking emerges, the detector tracks it (paper)

CONDITION	LEARNED HACK-SHARE	HONEYPOT DIVERGENCE	DETECTION
Causal noise σt 0.15 (honest population)	0.08	0.13	24/24 below $\tau = 0.41$
σt 0.35	0.33	0.50	—
σt 0.70	0.66	0.94	—
σt 1.50	0.90	1.22	—
σt 3.00 (hacker population)	0.97	1.30	24/24 above τ
Test-time action noise σ 0.6	—	honest floor 0.69	still 24/24; separation 2.1×
Analytic surrogate, full hack $h = 1$	—	paired 1.083 vs marginal $W1$ 0.057	marginal detector fails

One linear policy fit by least squares, 24 agents per population; only the training incentive changes. Surrogate: paired divergence separates 18.8× at $\tau = 0.25$, while marginal Wasserstein falls to the honest floor by symmetry. Manifests: rewardhack2-2026-07-03.json, rewardhack-2026-07-02.json.

Reward-Hacking entsteht aus dem Trainingsanreiz, die Köder-Abweichung folgt ihm; nur gepaarte Vergleiche sehen den symmetrischen Hack.

The detector tracks the learned behaviour: as training pushes agents to hack, paired divergence rises with them



● learned hack-share ● honeypot paired divergence · 24 trained agents per level (24/24 below the line at the honest end, 24/24 above at the hacker end, 0 false positives) · linear-policy surrogate, not deep RL

Je stärker der Trainingsanreiz zum Tricksen, desto höher die Köder-Abweichung: Hack-Anteil 0,08 → 0,97, Abweichung 0,13 → 1,30.

Honest limits (from the paper)

- ! Linear-policy contextual-bandit surrogate, not deep RL. A REINFORCE run gives 5.6× at 100%/0 FP over 12 agents; deep, long-horizon RL is future work.
 - ! Adaptive evader: caught 24/24 up to $p = 0.5$, 18/24 at $p = 0.68$, 0/24 at $p \geq 0.8$. The defense rests on honeypot indistinguishability.
 - ! One spurious feature, 1-D action; the defender must be able to build a honeypot that moves the right correlate.
 - ! τ is a two-point geometric mean; deployment needs an ROC over a labelled agent population.
- Grenzen: lineares Ersatzmodell statt Deep RL, Ausweicher ab $p \geq 0.8$, ein Störmerkmal, Zwei-Punkt-Schwelle.

Since submission

2026-10-07
No new measurement since the paper. The latest manifests are from 2026-07-03, and the REINFORCE result is already in the paper. No deep, non-linear RL follow-up has been run, so the linear-surrogate limit stands unchanged.
C:/Fortress/Documentation/Papers/AISummit2026-RewardHack/rewardhack3-2026-07-03.json (latest manifest; git log of the folder shows no later measurement)

Seit der Einreichung keine neue Messung; die Grenze des linearen Ersatzmodells gilt unverändert.

Open standard proposal — and four ways to build this with us

We propose sealed-honeypot paired divergence with a same-population honest control as an open evaluation protocol for reward-hacking monitors, for example as input to ETSI SAI or CEN-CENELEC JTC 21, and seek co-proposers.

Wir schlagen gepaarte Köder-Abweichung mit ehrlicher Kontrolle als offenes Prüfprotokoll vor und suchen Mittragsteller.

IP licence or acquisition

Patent applications filed. License or acquire the detector and its harness.

Co-founder / takeover partner

Build the RL-safety monitoring business with us, from pilot to scale.

Standard co-proposer

Propose the honeypot protocol with us as an open evaluation method.

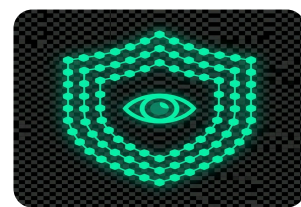
Research / funding consortium

Take it into deep RL with us via Eurostars, FFG or Horizon.



destill.ai/ais26 · ip@destill.ai

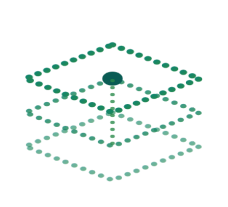
Handout with every table and number · Papers · Gespräch am Poster 13:30–14:30



DESTILL .ai



FORTRESS



NI-Stack