Institute for Physical AI @ John Bailey Institute · The Charlot Lab

SECURITY · AUTONOMY

Technical Report TR-2026-38
Research / Position · Preprint v1
13 August 2026

SECURITY · AUTONOMY

A Certificate That Bounds Danger Does Not Bound Waste

A safety certificate constrains where a machine may go, and leaves free what it may spend getting there. That freedom is a measurable channel, and the same measurement that would finally make robot energy comparable is the one that closes it.

The Charlot Lab · Institute for Physical AI @ John Bailey Institute
Third in the Security cluster, after TR-2026-36 and TR-2026-37.

computed from measured inputs · cross-checked on 2 action dimensions · a limit of our own detector · the mitigation is a gap we already named

Certificate-gated control is a strong answer to a compromised policy: a backdoor grants an adversary an arbitrary policy, but not an arbitrary trajectory, because every action must still pass a check made on the device. This report asks what that leaves open, and answers with a number. A certificate distinguishes safe from unsafe actions; it does not distinguish cheap ones from expensive ones. On a measured autonomous mobile robot, angular speeds of 1.2 and 0.4 rad/s draw 92.1 W and 19.4 W, a 72.7 W span in which *every* point is certified safe, since slowing down is the safe direction. Treated as a Gaussian channel against the 2.9 W whole-system prediction residual at the platform's own 3 Hz decision rate, that carries 13.9 bits per second. The result that matters is at the other end: a policy modulating *at* the three-sigma threshold where an energy monitor would flag it still carries 5.0 bits per second, roughly 18 kbit per hour. **The detection floor does not close the channel**, and that is a statement about the detector this Institute described in TR-2026-36. A cross-check on a second and genuinely certificate-constrained dimension, detection rate at 6 versus 3 Hz, gives 5.4 bits per second, the headline varies by 2.6x across dimensions and the conclusion does not. The mitigation is not a stricter certificate, because strictness pushes an adversary further into the direction the certificate already blesses. It is the **task denominator**: joules per unit of work completed, which PAI-280 identifies as absent from every

published measurement standard, and which is precisely the quantity that distinguishes a machine doing its job from a machine dawdling.

1. What does certificate-gating actually buy?

An adversary who poisons a policy obtains something powerful and something limited. Powerful: the ability to induce a chosen action on a chosen trigger while the policy remains near-optimal on its nominal task. Limited: if every action must clear an on-device certificate before dispatch, the reachable set is bounded by the certificate rather than by the policy. A backdoor grants an arbitrary policy; it does not grant an arbitrary trajectory.

That is a genuinely strong posture, and it does not require detecting the backdoor. It is the posture the lab's own certificate work assumes [3]. This report asks the complementary question: **what remains reachable when the certificate holds?**

> **What this report does not claim.** Backdoor attacks against reinforcement-learning policies are an active and crowded field, and the question of whether a policy can be backdoored is answered there, repeatedly, by others [4]. This report claims no attack, no trigger design, and no detector. It computes a bound on a channel, which is the standard defensive form: you cannot mitigate a channel whose capacity you have not priced.

2. What does a certificate leave free?

A safety certificate answers one question about each candidate action: is it safe. It has no opinion on how much energy the action costs, and among actions that are all safe there is usually a wide spread of cost. So a compromised policy that never once violates its certificate can still choose *which* safe action to take, and thereby modulate the machine's power draw.

This is far from a subtle effect: it is the difference between a robot doing its job briskly and the same robot doing the same job slowly, and no stability or collision certificate distinguishes them.

3. How wide is that freedom, measured?

A profiled autonomous mobile robot [2] draws **92.1 W to 19.4 W** of system power as its angular speed is reduced from 1.2 to 0.4 rad/s, and TR-2026-36 [1] is where this Institute takes up what that freedom costs. Both are ordinary operating speeds. Since a slower robot is a safer robot, **a safety certificate permits the entire range.** VERIFIED

The observer's side is also measured: whole-system power prediction on that platform is reported at better than 96 % accuracy, which against a ~74 W operating point gives a residual of about **2.9 W**. The platform's own perception tasks run at **3 to 6 Hz**, which sets the symbol rate available to a policy that re-decides each cycle. VERIFIED

4. How much information fits inside it?

Treating the modulation as a Gaussian channel, capacity per use is $\tfrac{1}{2}\cdot\log_2(1 + (A/\sigma)^2)$ for amplitude A against residual σ.

Modulation amplitude	S/N	bits/use	bits/s @ 3 Hz	bits/s @ 6 Hz
72.7 W, the full safe span	628	4.65	**13.9**	27.9
18.2 W, a quarter of it	39	2.67	8.0	16.0
8.7 W, at the 3σ detection floor	9	1.66	**5.0**	10.0
3.6 W, 5 % of the span	1.6	0.68	2.0	4.1

Table 1. Every row is a sequence of individually certified-safe actions. COMPUTED FROM THE §3 INPUTS

The finding, and it is about our own detector. The three-sigma threshold at which the energy monitor of TR-2026-36 would raise an anomaly sits at 8.7 W, which is 12 % of the safe span. A policy that modulates right at that threshold, never tripping it, still carries **5.0 bits per second**, about 18 kbit in an hour. That is ample for a key, a credential, or a model fingerprint. **The detection floor does not close the channel.** Stating that plainly about our own architecture is the point of the report.

5. Does the result survive a second dimension?

One action dimension is one data point, and angular speed is the most favourable one available, precisely because slowing is the safe direction. The check is whether the result survives on a dimension the certificate genuinely constrains.

Dimension	Span	bits/s @ 3 Hz	Certificate-gated?
Angular speed, 1.2 → 0.4 rad/s	72.7 W	13.9	**no**, slower is safer, the whole range is permitted
Detection rate, 6 → 3 Hz	9.7 W	5.4	**partly**, a lower rate cuts the safety margin, so a time-to-collision certificate does bound it

Table 2. Both land in the same order. The headline varies by 2.6x across dimensions; the conclusion does not. COMPUTED

The asymmetry is itself the result. The adversary's best channel is the direction the certificate is structurally unable to object to. Excess caution is always permitted by a safety envelope, so tightening the envelope does not close this, it pushes an adversary further into the region the envelope already blesses. **A certificate that bounds danger does not bound waste.**

6. What closes it?

One quantity closes it, and it is a definition rather than a technology: **joules per unit of work completed**. Three layers are in play today and each bounds something different. The certificate bounds the *trajectory*. The energy monitor of TR-2026-36 bounds *instantaneous power*. The task denominator bounds the ratio between them, which is the quantity a dawdling machine gives itself away on.

A policy exploiting this channel is, by construction, spending more energy than its task requires. A monitor normalised by completed work sees that immediately; a monitor watching watts cannot, because every individual reading is legitimate.

> **And that quantity is exactly the one PAI-280 identifies as missing.** That course's reporting-record lesson finds that no published standard defines robot energy per task: ISO 20140 defines efficiency as a ratio and leaves both terms to the user; the newly published ISO/TS 25213 specifies how to measure a six-axis industrial arm's consumption and does not supply a task denominator either. The gap was identified as a *comparability* problem. This report finds it is also a **security** problem: the missing denominator is the quantity that would close a channel a safety certificate cannot. **That is the constructive result, and it should be read as the report's main claim.**

7. What would sharpen this?

Four things would extend this result, and each is a defined experiment rather than an open question.

Build the decoder. The Gaussian-capacity form gives the ceiling a cooperative receiver could reach. Constructing an actual decoder would give the achievable rate underneath it, and the distance between the two is a useful number for anyone sizing a defence.

Profile a second platform. Both dimensions here come from one robot in one published profiling study, which makes this a worked example. A second power envelope turns one number into a span across machines, and that span is what a standard would be written against.

Add a task-performance monitor and re-measure. The analysis assumes a policy free to choose among safe actions at the decision rate. Any performance monitor already in place narrows the effective span, so measuring how much it narrows is the §6 mitigation quantified rather than asserted.

Read the located prior art in full. The §1 survey was a single sweep and the works it found were read at abstract depth. Reading them fully is what would let this report position itself firmly against that literature, and it is a day of work rather than a research programme.

8. What follows from this

Certificate-gated control converts an unbounded software compromise into a bounded physical one, which is worth having. The bound it

provides is on danger, and the quantity it leaves free is cost. On measured inputs from one autonomous mobile robot, the resulting channel carries 13.9 bits per second at full amplitude and **5.0 bits per second while remaining below the threshold at which an energy monitor would raise an alarm**. A cross-check on a second, genuinely constrained dimension gives 5.4 bits per second, so the conclusion is robust to the choice of dimension even though the headline is not.

The mitigation is a measurement the field does not yet standardise rather than a stricter safety envelope: joules per unit of work completed. This Institute identified that absence as a comparability problem in PAI-280; it is also the thing that closes this channel.

9. The forcing function

A channel that survives a detector is not a reason to stop; it is a specification for the instrument that would close it. The value of pricing this channel is that the closing move becomes obvious and small.

What is bounded	The physics that sets it	The change that moves it	What becomes possible
A safety certificate cannot bound spend	The certified-safe action set has an energy span, and excess caution is always inside it	**A task denominator: joules per unit of work completed**, a measurement definition, not a technology	An autonomous machine whose expenditure is as accountable as its safety, which is what makes autonomy auditable rather than merely permitted
A power monitor cannot see a compliant adversary	Every individual reading is legitimate; only the ratio to work done is not	Normalising the monitor by completed work	A monitor that flags dawdling, which is also the monitor that flags inefficiency, the same instrument serves safety and economics

Table 3. The mitigation is a definition the field has not agreed on, which is the cheapest class of obstacle: it needs consensus, not invention.

The trajectory reads plainly. Bounding an autonomous machine's *behaviour* was **impossible** before certificate-gated control; it is *routine*

now. Bounding its *expenditure* is where behaviour-bounding was a decade ago, desirable, not yet standardised, and blocked by a missing denominator rather than by any physical limit. That is the next thing to define, and defining it costs nothing but agreement.

References

1. Institute for Physical AI @ JBI, TR-2026-36, *Energy Observability in Embodied Systems*. Supplies the residual floor and the plane dependence. The angular-speed power figures come from reference 2, not from this report. INSTITUTE PUBLICATION
2. L. Liu, W. Shi and K. G. Shin, "Power-Efficient Autonomous Mobile Robots," arXiv:2511.20467v1, 25 November 2025. Source of the 92.1/19.4 W angular-speed figures and the 36.5/26.8 W detection-rate figures. READ IN FULL
3. Institute for Physical AI @ JBI, TR-2026-07, *Provable by construction*, and PAI-280, *Measuring Energy in Physical AI*. INSTITUTE PUBLICATIONS
4. Backdoor attacks against reinforcement learning: an active field including TooBadRL (arXiv:2506.09562), SleeperNets (arXiv:2405.20539), BACKDOORL (arXiv:2105.00579), and the test-time defence Plan2Cleanse (arXiv:2605.09638). LOCATED BY ONE SWEEP; NOT READ IN FULL

▶ The companions

Institute for Physical AI @ John Bailey Institute · The Charlot Lab

TR-2026-38 · Preprint v1 · 13 August 2026