Entropy 2026 Conference

Entropy 2026: Exploring Complexity and Information in Science

1–3 July 2026 | Barcelona, Spain

MDPI

Entropy, Capacity, and the Continuity of Agency in Human–AI Systems

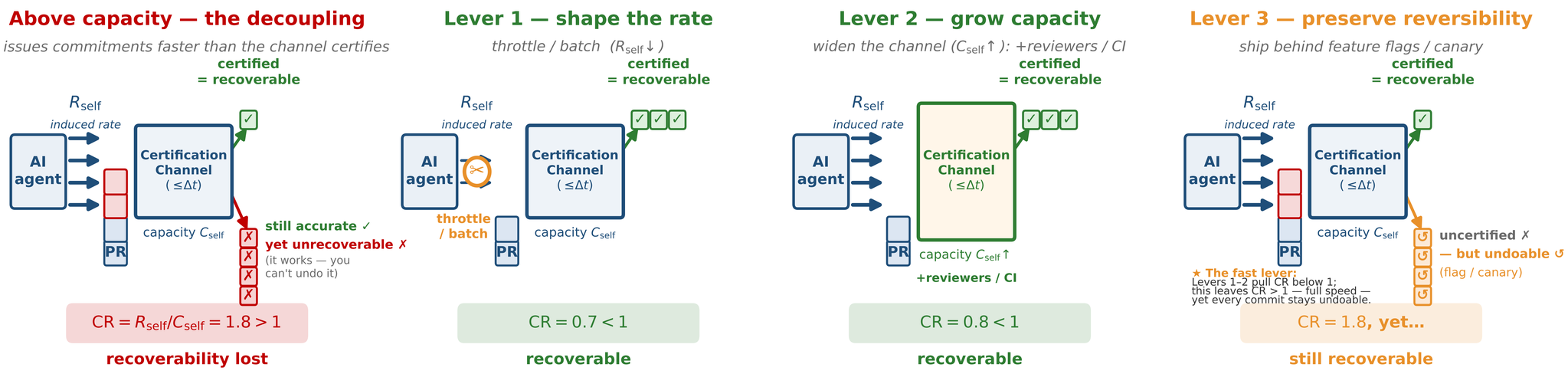
Recoverable Self-Coding (RSC): a rate–capacity law for recoverable action

Pieter van Rooyen ▪ Extraordinary Professor, Dept. of Electrical & Electronic Engineering, University of Stellenbosch ▪ pgwvanrooyen@sun.ac.za

Key finding: a finite-capacity decoder can stay fully accurate while losing the ability to *undo* what it commits — and the loss is measurable *before* the irreversible crossing.

Shannon's law, for action — recoverability is the new reliability

Reliable decoding needs information rate $R <$ channel capacity C. RSC: an irreversible action stays **recoverable** only while $R_{\text{self}} < C_{\text{self}}$, the capacity of its **certification channel**. $\text{CR} = R_{\text{self}}/C_{\text{self}}$ — the boundary $\text{CR} = 1$ is that limit.



*The decoupling: above its certification-channel capacity, commitments stay **accurate yet unrecoverable**. Three levers keep action recoverable — shape the rate, grow capacity, or **preserve reversibility (Lever 3), which holds even at** $\text{CR} > 1$.*

INTRODUCTION & AIM

AI routes ever more consequential decisions through the most efficient node, so oversight becomes *formal rather than substantive* — approvals still issue, but the ability to **undo** a commitment before harm erodes. Safety science long named this (migration to the boundary of safe operation; normal accidents), but lacked a *measurable* order parameter for how close the irreversible crossing is.

A unifying instrument. Decision tools (expected utility/RL, VaR/CVaR, real options, robust and safe-set control, AI safety), early-warning theory, and the congestion/admission/progressive-delivery levers each own a piece — all built for the slow-rate regime where recoverability is free. **RSC unifies them into one rate-sensitive gauge of recoverability for irreversible action**, with a phase boundary and an early warning none supplies alone.

Shannon → recoverability. Reliable decoding needs rate < capacity. RSC generalizes this to *action*: any **self-decoder** — human, institution, or AI — must hold induced informational flux R_{self} below integrative capacity C_{self}, not for message reliability but for the **recoverability of irreversible commitments**.

AI now drives $\text{CR} = R_{\text{self}}/C_{\text{self}} \to 1$ across whole domains at once, so the feasibility boundary Γ is increasingly approached or crossed. **Recoverable Self-Coding (RSC)** is the instrument that gauges it (above).

METHOD — the RSC framework

Over a decision horizon Δt, two operational rates:

- $R_{\text{self}}(t)$ — **induced informational flux**: the rate of unresolved uncertainty *requiring certification* before commitment (not data rate, FLOPs, or tokens s^{-1}).
- $C_{\text{self}}(t)$ — **integrative capacity**: the rate it can be certified and grounded — throughput, buffering, coherence, validation latency; not just compute.

The Shannon condition $R < C$, generalized from message reliability to irreversible action:

$$R_{\text{self}} < C_{\text{self}} \iff \text{CR} = \frac{R_{\text{self}}}{C_{\text{self}}} < 1, \qquad \mathcal{M} = C_{\text{self}} - R_{\text{self}} \geq 0.$$

Recoverability is a *trajectory* property, not a snapshot — two identical states can differ in it through hidden certification debt. **Local invertibility** over Δt is a *mutual-information* condition $I(\delta E; Z) \geq \iota > 0$: distinct causes stay distinguishable, enforced by a gate $q(t) \in [0, 1]$ with bounded $\zeta = \tau_{\text{cert}}/\tau_{\text{upd}}$. Minimal thermodynamic bridge — each irreversible commitment dissipates $\geq k_B T \ln 2$ per erased bit (Landauer).

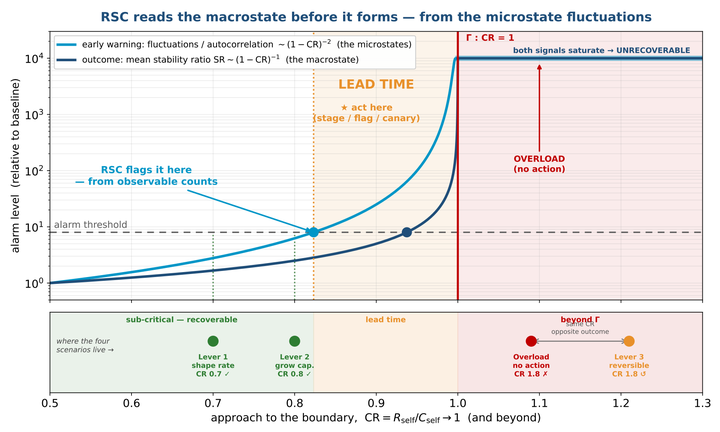
The mathematics is a queue. The uncertified-commitment backlog is a single-server **certification queue**; $\text{CR} < 1$ is its stability boundary, and Shannon's $R < C$ meets queue stability through the established theory of **effective bandwidth / effective capacity**. The contribution is the *reframing* — recoverability of irreversible action as the binding constraint, measurable from counts — not the queueing mathematics. Under acceleration the dominant failure is **premature commitment** — acting while the situation is still unresolved, merely to relieve overload.

Irreversible action is admissible **only if both:**
(i) $\mathcal{M}(t) \geq 0$ on the relevant horizon (no sustained overload), **and**
(ii) local invertibility holds over Δt.

If either fails, irreversible action causes *path-dependent contraction of future option space* that later evidence cannot undo. The admissible response is then **structural**, not "optimize harder": **defer** · **repeat-first** · **escalate** · **suppress**. CR and SR are *diagnostics, not objectives* — making SR a target games it (Goodhart) without restoring recoverability.

RESULTS — accuracy holds, recoverability fails

Stability ratio $\text{SR} = \dot{N}_{\text{uncert}}/\dot{N}_{\text{cert}}$ — the uncertified-to-certified commitment ratio, read from counts. In the queue's decay exponent $\theta^\star$ it has the exact closed form $\text{SR} = e^{-\theta^\star \Delta t}/(1 - e^{-\theta^\star \Delta t})$, diverging as $(1 - \text{CR})^{-1}$ at Γ — the effective-bandwidth heavy-traffic limit, universal across arrival and service distributions.



Beyond Γ (CR ≥ 1) there is no steady state: the backlog grows without bound, so BOTH signals saturate at the unrecoverable ceiling — that ceiling is the Overload scenario (no action).
The early warning (microstate fluctuations) crosses the alarm first, opening the lead time; Levers 1-2 keep you sub-critical; Lever 3 alone reaches CR 1.8 yet stays recoverable — it preserved reversibility in the lead time.

The early warning. Fluctuations diverge *faster* than the mean — $(1 - \text{CR})^{-2}$ vs $(1 - \text{CR})^{-1}$ — so the microstate signal (cyan) crosses the alarm *before* the macrostate outcome (navy) does: that gap is the **lead time** to act (stage / flag / canary). In a simulation driven across Γ, **accuracy stays at 0.90 while recoverability collapses** $0.75 \to 0.05$ — the decoupling — both diagnostics read from observable counts.

From exception to norm. Sustained overload ($\text{CR} > 1$) was a pathological exception — overwhelm, breakdown, reached only under acute stress. AI makes it the *default*, so the imperative is to measure the *trajectory* toward Γ *before* the irreversible crossing.

WHAT IT ENABLES

Admissibility before optimization — the hazard is not being wrong but crossing an irreversibility boundary. RSC turns recoverability into:

- an **instrument** — monitor CR, SR live from observable counts (queue depth, uncertified:certified); early warning before Γ;
- a **controller** — the rate-shaping policy {defer · repeat-first · escalate · suppress}: reduce flux, grow capacity, or gate commitment;
- a complement to **uncertainty quantification** — UQ says *how uncertain*; RSC says *whether to act irreversibly* on it.

The instrument applies wherever AI compresses the loop between inference and irreversible action: a **recoverability governor** for autonomous agents; a **pacing** criterion for deployment and online learning (hold $\text{CR} < 1$ as throughput scales); a **design principle** of option-preserving architectures. One language for humans, institutions, and AI — agency erosion under AI is a *structural* feasibility failure, not a personal one.

FUTURE WORK & REFERENCES

Delivered in the companion Entropy Special Issue. The boundary law from effective-bandwidth theory; universality of the $(1 - \text{CR})^{-1}$ exponent; count-based estimators with an early-warning detector; the independent capacity and invertibility axes; and field signatures consistent with the prediction (honestly scoped) in GitHub and Wikipedia histories.

Open. Closed-loop rate-shaping control; estimating the invertibility MI from field data; a controlled regime-transition experiment.

Preprint / priority: van Rooyen (2026), *Adaptive Agency Under Accelerating Gradients*, Preprints.org doi:10.20944/preprints202601.0688. An independent treatment reaching the same thesis (Shu & Wei, arXiv:2605.01415) appeared *after* this Jan-2026 preprint.

Scan → paper, poster & talk