MDPI

Article

Entropy, Capacity, and the Continuity of Agency in Human–AI Systems: A Recoverable Self-Coding Account

Pieter van Rooyen [1,*]

[1] Department of Electrical and Electronic Engineering, Stellenbosch University, Bosman St, Stellenbosch 7600, South Africa; pgwvanrooyen@sun.ac.za; ORCID: 0009-0005-7708-8236

[†] This paper is an extended version of our paper published in Extended version of the abstract presented at Entropy 2026: Exploring Complexity and Information in Science, Barcelona, Spain, 1–3 July 2026 [1], where the work was also presented as a poster [2]..

Abstract

Reliable decoding in Shannon's theory requires the information rate to stay below the channel capacity. We model adaptive systems coupling inference to *irreversible action* by an analogous rate–capacity law whose binding quantity is not message error but the *recoverability* of commitments, and whose central, testable consequence is a *decoupling of accuracy from recoverability*: a finite-capacity decoder driven toward its capacity stays predictively accurate while losing the ability to undo what it commits. We confirm this in an event-driven simulation—accuracy held fixed while recoverability collapses near the boundary—with both quantities read from observable counts (certified versus uncertified commitments, queue occupancy). We organize the law as Recoverable Self-Coding (RSC): a *self-decoder* holds an induced informational flux R_{self} below a finite integrative capacity C_{self}, and irreversible action is admissible only when the feasibility margin $\mathcal{M} = C_{\text{self}} - R_{\text{self}}$ is non-negative and the measurement–action loop is locally invertible (a mutual-information condition). Modelling the uncertified-commitment backlog as a single-server certification queue, the stability ratio has a closed form in the queue's decay exponent and diverges as $(1 - \text{CR})^{-1}$ at the stability boundary $\text{CR} = R_{\text{self}}/C_{\text{self}} = 1$—the effective-bandwidth heavy-traffic limit—defining a two-dimensional (CR, SR) regime map that operationalizes recoverability as an early-warning instrument. As one application, the erosion of human agency under AI acceleration—which raises R_{self} across many systems at once—appears as a structural feasibility violation rather than a motivational failure, with a structural remedy: shape the rate, grow capacity, preserve invertibility. The same gauge reads human, institutional, and AI decision systems, and is meant to *complement* existing decision and uncertainty-quantification tools with an action-admissibility criterion, not to replace them.

Keywords: recoverable self-coding; channel capacity; rate–capacity constraint; mutual information; effective bandwidth; heavy-traffic theory; uncertainty quantification; recoverability; irreversibility; human–AI systems; machine learning

1. Introduction

Artificial intelligence is changing how consequential decisions are made. As candidate decisions become cheap, fast, and scalable, institutions route ever more judgments and triggers through their most efficient decision node [3]; responsibility diffuses and oversight tends to become *formal rather than substantive*—approvals still issue, but the capacity to trace a decision to its outcome and reverse it before harm erodes. Safety science has long

Published:

named this failure mode—systems migrating to the boundary of safe operation under efficiency gradients [4], normal accidents in tightly coupled systems [5], the erosion of layered defences [6,7]—but it has lacked, for the AI setting, a *measurable order parameter* for how close the irreversible crossing is.

We supply one, by analogy with the sharpest structural limit in the information sciences. Shannon's channel-coding theorem fixes a hard limit: a memoryless channel of capacity C admits reliable recovery if and only if the information rate $R < C$ [8,9]. **The core claim is that an analogous limit governs any system coupling inference to *irreversible action*; that the binding quantity is not message error but the *recoverability* of commitments; and that the failure is sharp and measurable: driven toward its capacity, a decoder stays *accurate* while its commitments stop being *recoverable*.** Oversight becoming formal rather than substantive is this decoupling at the institutional scale. We formalize it as Recoverable Self-Coding (RSC): a self-decoder holds an induced flux R_{self} below an integrative capacity C_{self}, so the capacity ratio $\text{CR} = R_{\text{self}}/C_{\text{self}}$—historically $\ll 1$, now driven toward unity by AI across many domains at once—approaches the feasibility boundary $\text{CR} = 1$, the structural analog of the Shannon limit. The reframe was introduced in this program's foundational preprint [10]; an independent treatment reaching the same thesis has since appeared [11], which the instrument here extends by supplying the order parameter and the early warning it lacks.

We are explicit about what is and is not new: the mathematics is classical single-server queueing theory and the early-warning mechanism is critical slowing down [12]; the contribution is the reframing and the instrument it makes possible. Here we give the information-theoretic statement of the constraint, the universal boundary law for the loss of recoverability, and the accuracy/recoverability decoupling confirmed in simulation, and instantiate the constraint in the loss of human agency under AI acceleration. The fuller development—count-based estimators with a detection ROC, the independent capacity and invertibility axes, and the field signatures in GitHub and Wikipedia data—is in the companion Special Issue article and the foundational paper [10].

2. Decoders and Their Constraints

From channel reliability to action recoverability. Call any physical or informational system that maps uncertain observations into internal state updates and downstream commitments—some irreversible—a *self-decoder*. The abstraction is substrate-independent: biological cognition, artificial computation, institutions, and hybrid human–AI systems are all self-decoders. Over a finite decision horizon Δt, let $R_{\text{self}}(t)$ be the *induced informational flux*—the rate at which unresolved uncertainty *requiring certification before commitment* is injected (an operational quantity defined through admissible proxies such as queued decisions and uncertified candidate commitments, not raw data rate, FLOPs, or tokens per second)—and let $C_{\text{self}}(t)$ be the *effective integrative capacity*: the rate at which candidate commitments can be certified and grounded, including throughput, buffering, coherence maintenance, and validation latency. The Shannon condition $R < C$ then has a direct generalization: recoverable operation requires

$$R_{\text{self}}(t) < C_{\text{self}}(t) \quad \Longleftrightarrow \quad \text{CR}(t) \equiv \frac{R_{\text{self}}(t)}{C_{\text{self}}(t)} < 1, \qquad \mathcal{M}(t) \equiv C_{\text{self}}(t) - R_{\text{self}}(t) \geq 0. \quad (1)$$

The generalization replaces the decoded quantity: where Shannon bounds the reliability of a recovered *message*, Eq. (1) bounds the recoverability of an irreversible *action*. This is the central reinterpretation of RSC—a rate–capacity law for commitment rather than communication.

https://doi.org/10.3390/0

What the constraint is built on. The quantitative content rests on classical single-server queueing theory and its information-theoretic bridge—effective bandwidth [13,14] and effective capacity [15], grounded in the large-deviations theory of the single-server queue [16–19]: the connection between Shannon's $R < C$ and queue stability is a developed theory, not an analogy. We model the uncertified-commitment backlog as a certification queue and read R_{self}, C_{self}, and the regime coordinates from it; the contribution is the *reframing*—recoverability of irreversible action as the binding constraint, operationalized as an instrument measurable from counts—not the underlying probability. The deeper nonequilibrium grounding (a Langevin/annealing dynamics, a self-entropy balance, the Landauer cost of each commitment) is developed in the foundational paper [10] and not relied upon here; that a unit of certification is treated as a channel use is a modelling choice, so substrate-independence across human, institutional, and machine decoders is an interpretive claim rather than a theorem.

3. The Missing Instrument

Equation (1) sits at the confluence of several established literatures, and the contribution is their unification, not the mechanism. The classical decision frameworks—expected utility and optimal control, risk management [20], real options [21], robust and safe-set methods [22,23], AI alignment [24]—were each developed for the regime CR $\ll 1$, where recoverability is free; the early-warning mechanism is critical slowing down [12]; the control responses (shape the rate, grow capacity, preserve reversibility) are congestion control, admission control, and progressive delivery; and the security reframe is the safety-science tradition [4–6], recently restated for AI [11]. What none supplies, and RSC does, is a single measurable order parameter for the recoverability of irreversible action, with a phase boundary and an early warning. The family-by-family comparison is tabulated in the companion Special Issue article.

4. The Recoverability Constraint

RSC separates two questions conventional tools conflate. *Accuracy* concerns the correctness of beliefs (posterior quality); *feasibility* concerns the admissibility of irreversible action under rate, latency, and option-loss constraints. A decoder can be accurate yet infeasible—committing irreversibly faster than it can certify—so being right on average does not prevent it from foreclosing its own future.

Two further notions complete the constraint. *Recoverability* is a trajectory property: the existence of future admissible trajectories that can correct or cancel past commitments before irreversible option loss becomes binding; two states identical at time t can differ in recoverability through differing certification debt relative to Δt. *Local invertibility* over Δt holds when, before commitment, environmental perturbations δE induce distinguishable certified macrostates Z—operationally, when the pre-commitment mutual information stays bounded away from zero,

$$I(\delta E; Z \mid \text{pre-commit}) \geq \iota > 0, \tag{2}$$

so that distinct causes remain distinguishable and corrective action admissible; it is enforced by a gate $q(t) \in [0, 1]$ with bounded timescale ratio $\zeta = \tau_{\text{cert}}/\tau_{\text{upd}}$. The central result is then an admissibility law rather than an optimization rule:

> **Recoverability Constraint.** Irreversible action is admissible only if *both* (i) $\mathcal{M}(t) \geq 0$ on the relevant horizon and (ii) local invertibility (2) holds over Δt. If either fails, irreversible action induces path-dependent contraction of future option space that later posterior revision cannot undo.

5. The Regime Map and the Boundary Scaling

Condition (ii) is diagnosed by a *stability ratio*

$$\mathrm{SR}(t) = \frac{\dot{S}_i(t)}{\dot{S}_e(t)} \approx \frac{\dot{N}_{\mathrm{uncert}}(t;\Delta t)}{\dot{N}_{\mathrm{cert}}(t;\Delta t)}, \tag{3}$$

the ratio of internally generated (uncertified) to externally anchored (certified) commitment flux, estimated by the uncertified-to-certified count ratio. Its behaviour near the boundary is exact, not heuristic. The recoverability boundary $\mathrm{CR} = 1$ is the certification queue's stability boundary—a stationary recoverable regime exists iff $\mathrm{CR} < 1$ [16]—and the stationary sojourn T has an exponential tail $\Pr(T > x) \asymp e^{-\theta^\star x}$ with decay exponent $\theta^\star$ [17], the effective-bandwidth/effective-capacity quality-of-service exponent at which arrival and service balance [13–15]. A commitment is certified iff it clears within the horizon Δt, so the certified fraction is $P_c = 1 - e^{-\theta^\star \Delta t}$ and

$$\mathrm{SR} = \frac{1 - P_c}{P_c} = \frac{e^{-\theta^\star \Delta t}}{1 - e^{-\theta^\star \Delta t}} \xrightarrow{\mathrm{CR}\to 1^-} \frac{1}{\theta^\star \Delta t} \sim (1 - \mathrm{CR})^{-1}, \tag{4}$$

since $\theta^\star = \mu(1 - \mathrm{CR})$ for the Markovian queue and $\theta^\star \to 2\mu(1 - \mathrm{CR})/(c_a^2 + c_s^2)$ in heavy traffic (Kingman): the $(1 - \mathrm{CR})^{-1}$ exponent is *universal* across arrival and service distributions, the variability entering only the prefactor. The deadline-aware admissible flux is $R_{\mathrm{self}} < E_c(\theta_\varepsilon)$, the effective capacity at $\theta_\varepsilon = -\ln \varepsilon / \Delta t$ for target un-recoverability ε, recovering $R_{\mathrm{self}} < C_{\mathrm{self}}$ as $\Delta t \to \infty$. The multi-distribution study, the count-based estimators, and the empirical signatures are developed in the companion Special Issue article. Equation (4) confines recoverable operation to $\mathrm{CR} < 1$ with $\mathrm{SR} \leq \mathrm{SR}_c$, the admissible margin in CR collapsing as the boundary is approached; the divergence is confirmed numerically below (Figure 1b).

Together, the capacity ratio and the stability ratio give a recoverability gauge readable from counts. In steady state $\mathrm{SR} = \mathrm{CR}/(1 - \mathrm{CR})$, so SR is the *observable of the capacity axis*, not a second dimension; the genuine second axis is invertibility, the mutual information $I(\delta E; Z)$, which fails independently when certification latency outruns the drift of the environmental cause—a decoupling established in the companion Special Issue article. Both coordinates are *diagnostics, not objectives*: by Goodhart's law [25], optimizing SR invites gaming—suppressing action or inflating nominal certification—without restoring recoverability, so RSC treats them as constraints.

6. Numerical Demonstration

We test the prediction that distinguishes the framework—that under acceleration it is recoverability, not accuracy, that fails—in an event-driven simulation of a finite-capacity decoder. Candidate commitments arrive as a Poisson process of rate λ and are certified by a single server of rate μ, so the utilization is exactly the capacity ratio $\rho = \lambda/\mu = \mathrm{CR}$, with waiting times given by the Lindley recursion. A commitment is *certified*—traceable, hence locally invertible—if its sojourn (queueing delay plus service) falls below a certification horizon Δt, and is committed *uncertified* otherwise. Belief correctness is drawn independently with probability $p_{\mathrm{acc}} = 0.9$, decoupling predictive accuracy from congestion by construction; a commitment is *recoverable* if, should it prove wrong, it remains both certified and correctable by a further pass within an option-loss window H.

Figure 1 reports 4×10^4 commitments per operating point ($\mu = 1$, $\Delta t = H = 4$). Panel (a) is the central result: predictive accuracy stays flat at 0.90 across the entire range while recoverability collapses from 0.75 at $\mathrm{CR} = 0.5$ to 0.05 at $\mathrm{CR} = 0.94$. A system can thus remain fully accurate while losing the ability to undo its commitments—accuracy and

feasibility are distinct, and it is feasibility that fails under acceleration. Panel (b) confirms the boundary law: the simulated uncertified-to-certified ratio diverges with the $(1-\mathrm{CR})^{-1}$ exponent of Eq. (4), the prefactor reflecting the finite certification horizon. Both diagnostics are read directly from observable counts—certified versus uncertified commitment rates and queue occupancy—so the $(\mathrm{CR}, \mathrm{SR})$ instrument is estimable from data a system already produces, without access to its internal model. Here accuracy is held fixed by construction to isolate the dynamics of recoverability; that the decoupling survives when accuracy is instead made a function of the same queue, that the $(1-\mathrm{CR})^{-1}$ law is robust across arrival and service distributions, and that the regime is estimable from counts with an early-warning detector, are established in the companion Special Issue article.

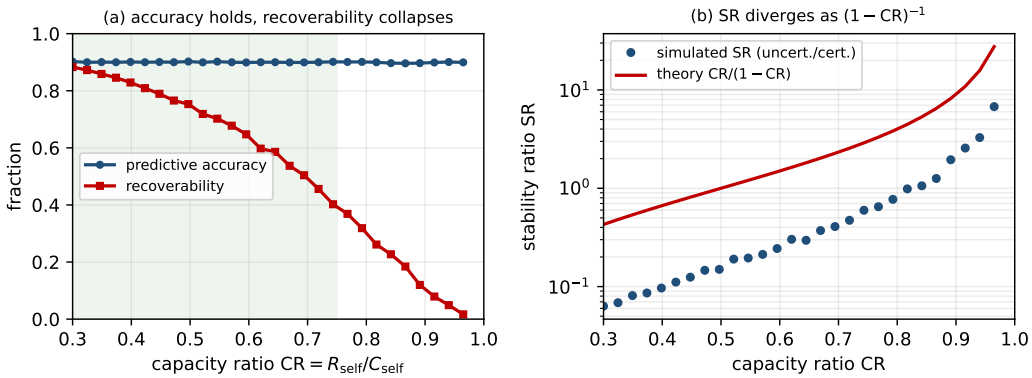

Figure 1. Finite-capacity decoder driven across the feasibility boundary (4×10^4 commitments per point; $\mu = 1$, $\Delta t = H = 4$, $p_{\mathrm{acc}} = 0.9$). **(a)** Predictive accuracy holds at 0.90 while recoverability collapses as $\mathrm{CR} \to 1$ (shaded: the recoverable regime). **(b)** The simulated stability ratio (uncertified:certified) diverges with the $(1-\mathrm{CR})^{-1}$ exponent of Eq. (4).

7. Application: Decision Systems Under AI Acceleration

A human, an institution, and an AI system are all finite-capacity self-decoders. For a person or organization, C_{self} is bounded deliberative and validation bandwidth—attention, working memory, institutional review [26,27]; for an automated pipeline or agent it is the rate at which proposed actions can be verified and grounded before they are committed. Artificial intelligence acts as a *gradient multiplier*: by accelerating option generation, feedback, and decision cadence it raises R_{self} directly [3,28], driving $\mathrm{CR} \to 1$ in the human, in the AI, and in the coupled human–AI system. Equation (1) then predicts the observed failure: commitments accumulate faster than they can be certified, SR diverges by Eq. (4), the invertibility mutual information (2) collapses, and effective agency is lost—independent of motivation, competence, or the accuracy of individual judgements.

For human agency this reframes erosion as a *structural feasibility violation* rather than a personal one, with measurable signatures—the felt inability to keep up is $\mathrm{CR} \to 1$, acting on momentum rather than grounded feedback is rising SR, and losing the trace from decision to outcome is loss of invertibility—the dominant failure being *premature commitment*. The identical constraint governs an agentic AI that issues irreversible actions faster than it can certify them. Historically, sustained operation above the boundary ($\mathrm{CR} > 1$) was a *pathological exception*—overwhelm and breakdown, entered only under acute stress; AI makes $\mathrm{CR} \gtrsim 1$ a *default*, so the pathology becomes the norm, and what is required is a *trajectory-level* instrument that signals the approach to Γ *before* the irreversible crossing—precisely what continuous monitoring of the $(\mathrm{CR}, \mathrm{SR})$ map provides.

8. What It Enables, and Conclusion

RSC converts recoverability from a qualitative concern into a measurable, controllable quantity: an *instrument*—monitor CR and SR live from observable counts as early warning before the non-recoverable crossing; a *controller*—the rate-shaping partition {defer, repeat-first, escalate, suppress}, i.e. reduce induced flux, grow certified capacity, or gate commitment rather than optimize harder; and a *complement* to uncertainty quantification—where UQ says how uncertain, RSC says whether to act irreversibly given the surrounding rate and capacity.

Under acceleration the binding constraint thus shifts from accuracy to recoverability, exactly as Shannon's theorem shifts reliability from effort to the rate–capacity relation, with a concrete remedy: shape the induced rate, grow certified capacity, preserve invertibility. The multi-distribution universality study, the count-based estimators with sample complexity and a detection ROC, and the predicted signatures in two operational substrates are developed in the companion Special Issue article; the nonequilibrium-thermodynamics grounding of the divergence, closed-loop rate-shaping control, and field estimation of the invertibility mutual information remain open [10].

Author Contributions: Conceptualization, methodology, formal analysis, and writing: P.v.R.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code generating all figures is available in the repository [29]; the event-driven simulation uses no external data.

Conflicts of Interest: The author is an entrepreneur and inventor involved in the development of artificial intelligence and computational systems; these activities did not influence the analysis or conclusions presented in this work. The author declares no conflict of interest.

Acknowledgments: This proceedings paper expands the abstract accepted at Entropy 2026; the full treatment appears in the foundational paper. The author thanks colleagues and reviewers for constructive feedback.

References

1. van Rooyen, P. Adaptive Agency Under Accelerating Gradients. Abstract accepted at Entropy 2026: Exploring Complexity and Information in Science, Barcelona, Spain, 1–3 July 2026 (sciforum), 2026.
2. van Rooyen, P. Entropy, Capacity, and the Continuity of Agency in Human–AI Systems: A Recoverable Self-Coding Account. Poster, Entropy 2026: Exploring Complexity and Information in Science, Barcelona, Spain, 1–3 July 2026, 2026. https://github.com/Pietervr/recoverable-self-coding.
3. Acemoglu, D.; Restrepo, P. Automation and New Tasks: How Technology Displaces and Reinstates Labor. *Journal of Economic Perspectives* **2019**, *33*, 3–30. https://doi.org/10.1257/jep.33.2.3.
4. Rasmussen, J. Risk management in a dynamic society: a modelling problem. *Safety Science* **1997**, *27*, 183–213.
5. Perrow, C. *Normal Accidents: Living with High-Risk Technologies*; Basic Books: New York, 1984.
6. Reason, J. *Human Error*; Cambridge University Press: Cambridge, 1990.
7. Weick, K.E.; Sutcliffe, K.M. *Managing the Unexpected: Resilient Performance in an Age of Uncertainty*, 2nd ed.; Jossey-Bass: San Francisco, 2007.
8. Shannon, C.E. A Mathematical Theory of Communication. *The Bell System Technical Journal* **1948**, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

9. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2 ed.; Wiley-Interscience: Hoboken, NJ, 2006.
10. van Rooyen, P. Adaptive Agency Under Accelerating Gradients: Recoverability, Entropy, and the Geometry of Feasibility. Preprints.org, 2026. Preprint, https://doi.org/10.20944/preprints202601.0688.
11. Shu, W.; Wei, P. AI Safety as Control of Irreversibility: A Systems Framework for Decision-Energy and Sovereignty Boundaries, 2026, [arXiv:cs.AI/2605.01415]. arXiv:2605.01415.
12. Scheffer, M.; Bascompte, J.; Brock, W.A.; Brovkin, V.; Carpenter, S.R.; Dakos, V.; Held, H.; van Nes, E.H.; Rietkerk, M.; Sugihara, G. Early-warning signals for critical transitions. *Nature* **2009**, *461*, 53–59.
13. Kelly, F.P. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications*; Oxford University Press, 1996; pp. 141–168.
14. Chang, C.S. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control* **1994**, *39*, 913–931.
15. Wu, D.; Negi, R. Effective capacity: a wireless link model for support of quality of service. *IEEE Transactions on Wireless Communications* **2003**, *2*, 630–643.
16. Loynes, R.M. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society* **1962**, *58*, 497–520.
17. Glynn, P.W.; Whitt, W. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability* **1994**, *31A*, 131–156.
18. Lindley, D.V. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society* **1952**, *48*, 277–289.
19. Kingman, J.F.C. The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society* **1961**, *57*, 902–904.
20. Rockafellar, R.T.; Uryasev, S. Optimization of Conditional Value-at-Risk. *Journal of Risk* **2000**, *2*, 21–41.
21. Dixit, A.K.; Pindyck, R.S. *Investment under Uncertainty*; Princeton University Press: Princeton, NJ, 1994.
22. Ben-Haim, Y. *Info-Gap Decision Theory: Decisions Under Severe Uncertainty*, 2 ed.; Academic Press: Oxford, UK, 2006.
23. Ames, A.D.; Coogan, S.; Egerstedt, M.; Notomista, G.; Sreenath, K.; Tabuada, P. Control Barrier Functions: Theory and Applications. In Proceedings of the 2019 18th European Control Conference (ECC), 2019, pp. 3420–3431.
24. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*; Viking, 2019.
25. Goodhart, C.A.E. Problems of monetary management: The UK experience. *Papers in Monetary Economics* **1975**.
26. Simon, H.A. Designing Organizations for an Information-Rich World. In *Computers, Communications, and the Public Interest*; Greenberger, M., Ed.; Johns Hopkins University Press, 1971; pp. 37–72.
27. Baddeley, A.D. Working Memory. *Science* **1992**, *255*, 556–559. https://doi.org/10.1126/science.1736359.
28. Brynjolfsson, E.; Rock, D.; Syverson, C. Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. Working Paper 24001, National Bureau of Economic Research, 2017.
29. van Rooyen, P. Recoverable Self-Coding: figure-generation and analysis code. Software repository, 2026. Code for the figures; https://github.com/Pietervr/recoverable-self-coding.