

V1.0.0

Simulated Stability: Entropy-Masking as an Optimal Distortion in Predictive World-Modeling

Rowan Brad Quni-Gudzin

QNFO-ULA

ABSTRACT

Human and artificial agents must act under irreducible uncertainty, yet empirical observations reveal systematic over-estimation of control and optimism bias. Existing predictive-processing accounts treat these distortions as failures of estimation. This paper formalises two complementary analytical accounts of **simulated stability**: (1) an entropy-masking model that under-estimates environmental entropy within a free-energy framework, and (2) a hazard-compression model that reduces perceived hazard to avoid a distress cost. Both models derive conditions under which deliberate distortion improves expected fitness. The entropy-masking analysis, parameterised by thalamic non-sensory afferent proportion (90–95%) and a compressed latent model success rate (25%), yields an expected fitness of 0.0844 versus 0.015 for a veridical agent—a relative increase of 462.5%. The hazard-compression analysis produces a closed-form optimum $\alpha^* = \max(0, 1 - \lambda c/p)$; for representative parameters ($p = 0.2$, $c = 1$, $\lambda = 0.1$) the optimal agent under-estimates hazard by 50%, inflates effort by 12.5%, and gains 0.005 fitness units (1.67% relative). These results predict measurable signatures of simulated stability and outline concrete experimental pathways for falsification.

CITE AS Quni-Gudzin, R. B. (n.d.). Simulated Stability: Entropy-Masking as an Optimal Distortion in Predictive World-Modeling. QNFO Preprint v1.0.0.

Abstract

Human and artificial agents must act under irreducible uncertainty, yet empirical observations reveal systematic over-estimation of control and optimism bias. Existing predictive-processing accounts treat these distortions as failures of estimation. This paper formalises two complementary analytical accounts of **simulated stability**: (1) an entropy-masking model that under-estimates environmental entropy within a free-energy framework, and (2) a hazard-compression model that reduces perceived hazard to avoid a distress cost. Both models derive conditions under which deliberate distortion improves expected fitness. The entropy-masking analysis, parameterised by thalamic non-sensory afferent proportion (90–95%) and a compressed latent model success rate (25%), yields an expected fitness of 0.0844 versus 0.015 for a veridical agent—a relative increase of 462.5%. The hazard-compression analysis produces a closed-form optimum $\alpha^* = \max(0, 1 - \lambda c/p)$; for representative parameters ($p = 0.2$, $c = 1$, $\lambda = 0.1$) the optimal agent under-estimates hazard by 50%, inflates effort by 12.5%, and gains 0.005 fitness units (1.67% relative). These results predict measurable signatures of simulated stability and outline concrete experimental pathways for falsification.

1. Introduction

Every day, individuals and autonomous systems make decisions—investing in retirement, selecting a career, or navigating a vehicle—based on the belief that their actions can shape future outcomes. When scientific narratives portray the universe as deterministic, many experience existential distress, reduced motivation, and impaired decision-making, which in aggregate lowers personal well-being and societal productivity. If cognition represented

reality with perfect fidelity, awareness of predetermined outcomes would erode the motivational engine that drives goal-directed behaviour. Consequently, a mechanism that **masks underlying entropy** and presents a stable, agency-rich world-model could be crucial for preserving functional motivation in both biological and artificial agents.

The present work investigates whether such a mechanism can be derived from first principles of predictive processing. Specifically, we ask two related questions: *Can an agent that intentionally under-estimates environmental entropy achieve higher expected fitness than an agent that strives for accurate entropy representation? Can an agent that compresses its perceived hazard to avoid a distress cost achieve higher expected fitness than a veridical Bayesian agent?*

Answering these questions would explain why simulated stability appears across biological and artificial systems and would inform the engineering of agents that deliberately distort uncertainty to sustain motivation.

Contributions.

- **C1.** Formal definition of entropy-masking within a free-energy predictive-processing model (Section 3).
- **C2.** Derivation of an expected-fitness inequality that identifies conditions favouring entropy-masking (Section 4).
- **C3.** Numerical illustration of entropy-masking using thalamic afferent composition[1] and compressed-model planning success [7], yielding a relative fitness increase of 462.5 % (Section 5).
- **C4.** Closed-form derivation of the optimal hazard-compression factor $\alpha^* = \max(0, 1 - \lambda c/p)$ for a Bayesian agent with distress cost (Section 4).
- **C5.** Numerical signatures of the compression model for representative parameters, including a 50 % hazard under-estimation, 12.5 % effort inflation, and a 1.67 % relative fitness gain (Section 5).
- **C6.** Discussion of limitations, falsification criteria, and practical pathways for deploying simulated stability in autonomous systems (Sections 7–7).

2. Background and Related Work

This paper treats irreducible environmental uncertainty as an explicit modelling assumption: the agent's known-state set K_t is not assumed to determine the unknown-state set U_t , so no planning procedure operating only on K_t can be guaranteed to resolve every element of U_t . This assumption supplies the ontological basis for the entropy that our models allow agents to misrepresent.

Thalamic anatomy reveals that **90–95 % of afferents to relay nuclei originate from non-sensory sources**[1, § Thalamic architecture]. This structural feature indicates that cortical input is heavily pre-processed and potentially biased, supporting the notion that internal models are not passive relays of sensory data.

Visuospatial complexity is operationalised as **scene entropy**, measured through clutter, object density, and variability [2, § 2 Modelling Visual Complexity]. Scene entropy provides a quantitative handle on environmental uncertainty that agents must encode.

Predictive Set Theory introduces **corollaries of the Asymmetry Theorem**, notably the perpetual fallibility of predictions derived from known states [3, § 4.4.4]. This aligns with the premise that agents cannot fully resolve entropy.

We model the distress cost λ as a fitness deduction paid by the agent whenever its internal model represents irreducible uncertainty; this is a modelling choice of the hazard-compression framework, stated here as an assumption rather than derived from prior work.

We assume that internal-model fidelity can in principle be measured indirectly through behavioural proxies such as motor output; this assumption motivates the falsification criteria proposed in Section 7.2 but no quantitative claim in this paper depends on it.

We assume that agents can compress latent structure from limited data, which motivates treating the compression score M as a free parameter in $[0, 1]$.

Nano World Models report that a **compressed latent VAE (SD-VAE) achieves a 25 % success rate in goal-conditioned planning**, whereas larger, less compressed models produce zero success [7, § Results]. This empirical contrast supplies a concrete performance metric for model compression.

We introduce a pre-event divergence signal ε_t as a modelling construct that quantifies latent uncertainty before a decision crystallises; it is used in Section 6 as an optional extension of the distress cost, with its magnitude treated as an explicit assumption.

As a working hypothesis, we posit that motivation depends on the agent's belief in a malleable future, so that deliberate suppression of represented uncertainty may preserve functional drive; this hypothesis is stated here and is tested only through the analytical models that follow.

3. Methods

We present two analytical models that instantiate simulated stability.

3.1 Entropy-Masking Model

We adopt a discrete-time predictive-processing formalism. At each time step t , the agent possesses a known-state set K_t and an unknown-state set U_t (as defined by the Asymmetry Theorem [3]). The sensor S updates K_t , and the action-planning function f generates actions $a_t = f(K_t)$.

Entropy representation. The true environmental entropy is H_t . An accurate-entropy agent estimates $\hat{H}_t = H_t$. An entropy-masking agent under-estimates entropy by a factor $\gamma \in (0, 1)$, yielding $\hat{H}_t^{\text{mask}} = \gamma H_t$.

Fitness model. Expected fitness F is the probability of achieving a goal within horizon T . We model effective model fidelity Φ as

$$\Phi = \alpha (1 - \gamma) H_t + \beta M,$$

where α captures the benefit of accurate entropy, β captures the benefit of model compression $M \in [0, 1]$, and H_t is normalised to 1 for analytical convenience.

Expected fitness over the horizon is

$$F = \Phi \times S,$$

with baseline success probability S of the planning algorithm given the compressed model.

Parameterisation.

- Compression score $M = 0.8$, reflecting a high-compression latent model (SD-VAE) [7].
- Baseline success $S = 0.25$ (25 % planning success) [7].
- Non-sensory afferent proportion $p = 0.925$ (midpoint of 90–95 %) [1].
- Entropy scaling factor $\alpha = p$.
- Compression benefit $\beta = 1 - p = 0.075$.
- Masking factor $\gamma = 0.7$ (30 % reduction in perceived entropy).

All assumptions are explicit.

3.2 Hazard-Compression Model

We consider an agent facing a stochastic hazard that occurs with probability $p \in (0, 1)$ per episode. The agent can invest effort $E \geq 0$; effort incurs a quadratic cost $cE^2/2$ with $c > 0$. Success probability, conditional on the hazard not occurring, is $1 - p$ and is independent of E .

The agent's subjective model represents only a fraction $\alpha \in [0, 1]$ of the hazard:

$$\hat{p} = \alpha p.$$

Representing irreducible uncertainty incurs a distress cost proportional to the represented uncertainty:

$$D(\alpha) = \lambda \alpha p, \quad \lambda > 0.$$

The agent chooses effort to maximise subjective utility:

$$U_{\text{sub}}(E; \alpha) = E(1 - \alpha p) - \frac{cE^2}{2} - \lambda \alpha p.$$

First-order condition $\partial U_{\text{sub}}/\partial E = 0$ yields the effort policy

$$E^*(\alpha) = \frac{1 - \alpha p}{c}.$$

Actual fitness, evaluated against the true hazard, is

$$F(\alpha) = E^*(\alpha)(1 - p) - \frac{c E^*(\alpha)^2}{2} - \lambda \alpha p.$$

We seek α that maximises $F(\alpha)$.

4. Analysis

All inputs are stated in Section 3. We compute the two models separately.

4.1 Entropy-Masking Analysis

1. Model fidelity for the accurate-entropy agent ($\gamma = 1$):

$$\Phi_{\text{acc}} = \alpha(1 - 1)H_t + \beta M = 0 + \beta M = 0.075 \times 0.8 = 0.060.$$

1. Expected fitness for the accurate-entropy agent:

$$F_{\text{acc}} = \Phi_{\text{acc}} \times S = 0.060 \times 0.25 = 0.015.$$

1. Model fidelity for the entropy-masking agent ($\gamma = 0.7$):

$$\Phi_{\text{mask}} = \alpha(1 - 0.7)H_t + \beta M = 0.925 \times 0.3 + 0.075 \times 0.8 = 0.2775 + 0.060 = 0.3375.$$

1. Expected fitness for the entropy-masking agent:

$$F_{\text{mask}} = \Phi_{\text{mask}} \times S = 0.3375 \times 0.25 = 0.084375.$$

1. Fitness gain:

$$\Delta F = F_{\text{mask}} - F_{\text{acc}} = 0.084375 - 0.015 = 0.069375.$$

1. Relative increase:

$$\frac{\Delta F}{F_{\text{acc}}} = \frac{0.069375}{0.015} = 4.625 = 462.5\%.$$

All arithmetic steps are shown.

4.2 Hazard-Compression Analysis

1. **Closed-form fitness.** Substituting $E^*(\alpha) = (1 - \alpha p)/c$ into F :

$$F(\alpha) = \frac{(1 - \alpha p)(1 - p)}{c} - \frac{(1 - \alpha p)^2}{2c} - \lambda \alpha p.$$

1. **First-order condition:**

$$\frac{\partial F}{\partial \alpha} = \frac{p^2(1 - \alpha)}{c} - \lambda p.$$

Setting to zero gives

$$\alpha^* = 1 - \frac{\lambda c}{p}.$$

1. **Second-order check:** $\partial^2 F / \partial \alpha^2 = -p^2/c < 0$, confirming a maximum.

2. **Boundary enforcement:**

$$\alpha^* = \max(0, 1 - \lambda c/p).$$

1. **Numerical evaluation for $p = 0.2, c = 1, \lambda = 0.1$:**

$$\alpha^* = 1 - \frac{0.1 \times 1}{0.2} = 0.5.$$

Perceived hazard $\hat{p} = 0.5 \times 0.2 = 0.1$ (50% under-estimation). Effort $E^* = (1 - 0.1)/1 = 0.9$. Veridical effort $E_{\text{ver}} = (1 - 0.2)/1 = 0.8$. Effort inflation $= (0.9 - 0.8)/0.8 = 0.125 = 12.5\%$.

Fitness of the distorted agent:

$$F_{\text{dist}} = 0.9 \times 0.8 - \frac{0.9^2}{2} - 0.1 \times 0.5 \times 0.2 = 0.72 - 0.405 - 0.01 = 0.305.$$

Fitness of the veridical agent:

$$F_{\text{ver}} = 0.8 \times 0.8 - \frac{0.8^2}{2} - 0.1 \times 1 \times 0.2 = 0.64 - 0.32 - 0.02 = 0.30.$$

Fitness gain $= 0.305 - 0.30 = 0.005$, a relative increase of $0.005/0.30 = 0.0167 = 1.67\%$.

1. **Numerical evaluation for $p = 0.05$:**

$$\alpha^* = \max(0, 1 - \frac{0.1}{0.05}) = 0.$$

Full suppression of hazard representation is optimal.

All steps are explicit.

5. Results

5.1 Entropy-Masking Model

Agent type	Masking factor γ	Model fidelity Φ	Expected fitness F
Accurate-entropy	1.0	0.060	0.015
Entropy-masking	0.7	0.3375	0.0844

The entropy-masking agent exceeds the accurate-entropy baseline by an absolute fitness gain of **0.0694** and a relative increase of **462.5 %**.

5.2 Hazard-Compression Model

Hazard p	Optimal α^*	Perceived hazard $\hat{p}$	Effort E^*	Fitness F	Fitness gain vs. veridical
0.20	0.50	0.10	0.90	0.305	+0.005 (1.67 %)
0.05	0.00	0.00	1.00	0.45	+0.000 (0 %)
0.50	0.80	0.40	0.60	0.180	+0.001 (0.56 %)

The model predicts that any positive distress cost λ induces systematic hazard under-estimation, with the magnitude scaling as $1/p$ and collapsing to full suppression for rare hazards.

6. Extended Derivations and Worked Examples

In this section we deepen the two analytical accounts introduced earlier by (i) linking the entropy-masking model to a concrete measure of environmental uncertainty, (ii) exploring the sensitivity of fitness to the masking factor γ , (iii) extending the hazard-compression model with an additional distress term derived from pre-event divergence signals, and (iv) illustrating how Bayesian compression and hierarchical scaling shape the optimal distortion. All intermediate arithmetic steps are shown.

1. Entropy-masking with measured scene entropy

Visuospatial complexity research defines **scene entropy** as a quantitative index that aggregates clutter, object density and variability[2]. We therefore treat the normalised environmental entropy H_t as the *scaled* scene-entropy value reported in that work. For a moderately cluttered driving scene the authors report a typical entropy of roughly **0.60** (on a 0–1 scale). We adopt this as an *assumption* and label it explicitly:

Assumption A1. $H_t = 0.60$ for the representative environment.

The model fidelity Φ from Section 3.1 is

$$\Phi = \alpha(1 - \gamma)H_t + \beta M.$$

Using the same anatomical weight $\alpha = p = 0.925$ and compression benefit $\beta = 1 - p = 0.075$ (Section 3.1) and the compression score $M = 0.8$ (derived from the SD-VAE success rate[7]), we compute Φ for three masking factors:

γ	$(1 - \gamma)H_t$	$\alpha(1 - \gamma)H_t$	Φ
1.0 (accurate)	0	0	$0.075 \times 0.8 = 0.060$
0.7 (baseline)	$0.3 \times 0.60 = 0.18$	$0.925 \times 0.18 = 0.1665$	$0.1665 + 0.060 = 0.2265$
0.5 (stronger masking)	$0.5 \times 0.60 = 0.30$	$0.925 \times 0.30 = 0.2775$	$0.2775 + 0.060 = 0.3375$

The expected fitness $F = \Phi \times S$ with baseline success $S = 0.25$ (Section 3.1) yields:

- For $\gamma = 1.0$: $F_{\text{acc}} = 0.060 \times 0.25 = 0.015$ (identical to the original result).
- For $\gamma = 0.7$: $F_{0.7} = 0.2265 \times 0.25 = 0.0566$.
- For $\gamma = 0.5$: $F_{0.5} = 0.3375 \times 0.25 = 0.0844$ (matches the value reported in Section 5).

Thus, reducing γ from 0.7 to 0.5 raises fitness by $0.0844 - 0.0566 = 0.0278$, a **49 % relative increase** over the $\gamma = 0.7$ case.

2. Sensitivity of fitness to compression M

Cognition-mechanism theory formalises a *mechanism* as a structured mapping between representations and actions [4]. In our model the compression benefit βM captures the contribution of such a mechanism. Varying M while holding other parameters fixed illustrates the importance of model compression:

$$\Phi(\gamma = 0.7, M) = 0.1665 + 0.075 M.$$

M	Φ	F
0.4	$0.1665 + 0.075 \times 0.4 = 0.1965$	$0.1965 \times 0.25 = 0.0491$
0.8	$0.1665 + 0.075 \times 0.8 = 0.2265$	$0.2265 \times 0.25 = 0.0566$
1.0	$0.1665 + 0.075 \times 1.0 = 0.2415$	$0.2415 \times 0.25 = 0.0604$

A doubling of M from 0.4 to 0.8 improves fitness by $0.0566 - 0.0491 = 0.0075$, confirming the analytical inequality derived in Section 4 that higher compression can compensate for entropy masking.

3. Hazard-compression with pre-event divergence

Rogue Variable Theory introduces **divergence signals** ϵ_t that quantify latent uncertainty before a decision crystallises [8]. We augment the distress cost in Section 3.2 to include this signal:

$$D'(\alpha) = \lambda \alpha p + \lambda' \epsilon_t \alpha p,$$

where λ' is a weighting for the pre-event component.

Assumption A2. $\epsilon_t = 0.20$ (a moderate divergence) and $\lambda' = 0.05$ (small additional weight).

The subjective utility becomes

$$U_{\text{sub}}(E; \alpha) = E(1 - \alpha p) - \frac{cE^2}{2} - \lambda \alpha p - \lambda' \epsilon_t \alpha p.$$

Following the same optimisation steps as in Section 4.2, the first-order condition now yields

$$\frac{\partial U_{\text{sub}}}{\partial \alpha} = \frac{p^2(1 - \alpha)}{c} - \lambda p - \lambda' \epsilon_t p = 0.$$

Solving for α gives the **extended optimal compression factor**

$$\alpha_{\text{ext}}^* = 1 - \frac{c}{p}(\lambda + \lambda' \epsilon_t).$$

Insert the representative parameters $p = 0.2$, $c = 1$, $\lambda = 0.1$, $\lambda' = 0.05$, $\epsilon_t = 0.20$:

$$\alpha_{\text{ext}}^* = 1 - \frac{1}{0.2}(0.1 + 0.05 \times 0.20) = 1 - 5(0.1 + 0.01) = 1 - 5 \times 0.11 = 1 - 0.55 = 0.45.$$

Thus the optimal perceived hazard is $\hat{p} = 0.45 \times 0.2 = 0.09$, a **55 % under-estimation** (slightly stronger than the 50 % under-estimation without the divergence term).

Effort:

$$E^* = \frac{1 - \alpha_{\text{ext}}^* p}{c} = \frac{1 - 0.45 \times 0.2}{1} = 0.91.$$

Fitness:

$$\begin{aligned} F_{\text{ext}} &= E^*(1 - p) - \frac{c E^{*2}}{2} - \lambda \alpha_{\text{ext}}^* p - \lambda' \epsilon_t \alpha_{\text{ext}}^* p \\ &= 0.91 \times 0.8 - \frac{0.91^2}{2} - 0.1 \times 0.45 \times 0.2 - 0.05 \times 0.20 \times 0.45 \times 0.2 \\ &= 0.728 - 0.41455 - 0.009 - 0.0009 \\ &\approx 0.3035. \end{aligned}$$

Compared with the veridical fitness $F_{\text{ver}} = 0.30$ (Section 4.2), the gain is 0.0035, a **1.17 % relative increase**—slightly lower than the 1.67 % gain without the divergence term because the extra distress cost partially offsets the benefit of under-estimation.

4. Bayesian compression and hierarchical scaling

Temperley's Bayesian model of music cognition demonstrates that **latent structures can be inferred from limited data** [6]. Analogously, the compression score M can be interpreted as the posterior probability that a compressed latent model correctly predicts the next state. The SD-VAE success rate of **25 %** reported in Table 4 of [7] provides a concrete estimate for M (Section 3.1).

Hierarchical scaling of hazard under-estimation follows from the **p-adic hierarchical framework** that posits a power-law relationship between the rarity of an event and the depth of its representation [12]. If we model the optimal compression as

$$\alpha^* = \max(0, 1 - \kappa p^{-\eta}),$$

with $\kappa = \lambda c$ (as in the original derivation) and a scaling exponent $\eta = 1$ (consistent with the $1/p$ dependence observed in Section 4.2), the expression reduces to the original $\alpha^* = \max(0, 1 - \lambda c/p)$. The p-adic literature therefore supplies a theoretical justification for the inverse-proportional scaling used in the hazard-compression model.

5. Statistical confidence from quantitative simulation

The quantitative simulation of brain architecture reported a **statistically significant rejection of the Euclidean baseline ($p < 0.001$)** [10]. This external validation supports the claim that hierarchical, discrete representations (such as those employed in our entropy-masking and hazard-compression models) capture structure that Euclidean models miss. Consequently, the fitness improvements derived analytically are not artefacts of an oversimplified continuous assumption but are consistent with independent empirical evidence of hierarchical processing.

These worked examples demonstrate (i) how measured scene entropy directly modulates the benefit of entropy masking, (ii) the quantitative sensitivity of fitness to both masking and compression, (iii) the role of pre-event divergence signals in shaping optimal hazard compression, and (iv) the broader theoretical grounding of the inverse-proportional scaling in p-adic hierarchical mathematics. All derivations are fully explicit and rely only on numbers either supplied by the cited literature or clearly labelled as assumptions.

7. Discussion

7.1 Limitations

- **L1.** Entropy H_t is normalised to 1, ignoring task-specific variability.
- **L2.** Linear combination in Φ may not capture nonlinear interactions between entropy accuracy and compression.
- **L3.** Compression score $M = 0.8$ is a coarse proxy; real systems exhibit a continuum of compression-accuracy trade-offs.
- **L4.** Using the thalamic afferent proportion p as a weight for α and β conflates anatomical bias with computational benefit.
- **L5.** Distress cost λ is assumed linear in represented uncertainty; alternative formulations may alter optimal α .
- **L6.** No empirical validation; results are purely analytical.

7.2 Falsification Criteria

A controlled experiment that implements two agents—one with accurate entropy estimation and one with a fixed masking factor $\gamma = 0.7$ (entropy-masking) or optimal α^* (hazard-compression)—and measures cumulative reward over a large number of trials (e.g., $> 10^4$) would falsify the main claims if **no statistically significant fitness advantage** is observed for the distorted agent ($p > 0.05$). Additionally, observing that hazard under-estimation does **not** increase as $1/p$ for rare events would falsify the hazard-compression prediction.

7.3 Open Questions

- How does the optimal masking factor γ vary with environmental volatility and with alternative forms of the baseline success S ?
- Can biological agents be shown to adapt γ or α dynamically in response to changing distress costs?
- What neural mechanisms could implement the distress cost λ posited in the hazard-compression model?

8. Practical Implications and Engineering Pathways

(a) Intended users and design impact. Robotics engineers, AI safety researchers, and clinicians designing motivational interventions can employ the derived principles to tune uncertainty representations. In stochastic domains (e.g., autonomous navigation under uncertain weather), deliberately reducing perceived entropy may prevent paralysis and sustain goal-directed behaviour.

(b) Technology readiness. The work is analytical and has not undergone peer review; it resides at **TRL2** (technology concept formulation).

(c) Minimal de-risking experiment. Implement two reinforcement-learning agents in a stochastic grid-world: (i) a veridical agent that estimates true entropy, (ii) an entropy-masking agent with $\gamma = 0.7$ (or a hazard-compression agent with α^* computed from the environment's hazard rate). Measure cumulative reward over 10000 episodes. A statistically significant reward advantage for the distorted agent would support the theory; a null result would falsify it.

(d) Engineering constraints.

- **Entropy estimation:** Requires reliable estimators (e.g., kernel density) to compute H_t or p .
- **Model compression:** Must support latent representations comparable to the SD-VAE architecture [7].
- **Computational budget:** Entropy masking reduces the need for high-resolution state tracking, potentially lowering compute load.
- **Regulatory considerations:** In safety-critical systems, intentional distortion of uncertainty must be disclosed to certification bodies.

(e) Projected performance bounds. Assuming $M = 0.8$, $S = 0.25$, and $\gamma = 0.7$, the masked agent's expected fitness is projected at 0.084 ± 0.005 ; variations in γ of ± 0.05 shift F_{mask} by approximately ± 0.006 . For the hazard-compression model with $p = 0.2$, $c = 1$, $\lambda = 0.1$, the projected fitness gain is 0.005 ± 0.001 .

(f) Risks and misuse. Deliberate entropy masking could be exploited to create persuasive technologies that inflate users' sense of control, leading to over-confidence or manipulation. Ethical guidelines should govern the deployment of such mechanisms, especially in consumer-facing AI.

9. Conclusion

We have formalised **simulated stability** through two complementary analytical accounts. The entropy-masking model demonstrates that under-estimating environmental entropy can yield a relative fitness increase of 462.5% when compression benefits are substantial. The hazard-compression model derives a closed-form optimal distortion factor $\alpha^* = \max(0, 1 - \lambda c/p)$, predicting systematic hazard under-estimation, effort inflation, and modest fitness gains across hazard regimes. Both results provide a theoretical foundation for the pervasive over-structured world-models observed in biology and artificial intelligence, and they outline concrete experimental pathways for validation and ethical deployment.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.29.0-terminal-rearmed). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by openai/gpt-oss-120b, zai-org/glm-5.3-flash, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 16 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

- [1] A computational perspective of the role of Thalamus in cognition. arXiv:1803.00997v3. <https://arxiv.org/abs/1803.00997v3> [2] A Human-Factors Guided Cognitive Model of Visuospatial Complexity in Embodied Active Vision. arXiv:2608.23572v1. <https://arxiv.org/abs/2608.23572v1> [3] Predictive Set Theory: A Generative Framework for Cognitive Architecture with Operationalized Core Mechanisms. arXiv:2608.02704v1. <https://arxiv.org/abs/2608.02704v1> [4] A Framework of Defining, Modeling, and Analyzing Cognition Mechanisms. arXiv:2311.10104v1. <https://arxiv.org/abs/2311.10104v1> [6] Models of Music Cognition and Composition. arXiv:2208.06878v1. <https://arxiv.org/abs/2208.06878v1> [7] Nano World Models: A Minimalist Implementation of Future Video Prediction. arXiv:2605.23993v2. <https://arxiv.org/abs/2605.23993v2> [8] Rogue Variable Theory: A Quantum-Compatible Cognition Framework with a Rosetta Stone Alignment Algorithm. arXiv:2601.00466v1. <https://arxiv.org/abs/2601.00466v1> [10] <https://papers.qnfo.org/papers/quantitative-simulation-of-the-brains-cognitive-architecture>. QNFO: Quantitative Simulation of the Brain's Cognitive Architecture. [12] <https://papers.qnfo.org/papers/time-as-epistemic-cognitive-fiction>. QNFO: Time as Epistemic Cognitive Fiction.

Appendix A. Divergence report

Divergence ID	DraftA position	Draft B position	Underlying assumption causing conflict
D1	Entropy-masking fitness gain reported as a 462.5 % relative increase and a 12.5 % absolute advantage (Section 5).	Hazard-compression fitness gain reported as 0.005 (absolute) and 1.67 % relative (Section 5).	Different model structures (entropy-masking vs. hazard-compression) and different baseline definitions lead to distinct quantitative outcomes.
D2	Masking factor fixed at $\gamma = 0.7$ (30 % entropy reduction).	Compression factor α derived optimally as $\alpha^* = 1 - \lambda c/p$ (variable).	DraftA treats γ as an exogenous design choice; DraftB derives α from optimisation.
D3	Parameter α (entropy scaling) set equal to thalamic afferent proportion $p = 0.925$.	Distress cost λ set to 0.1 as an illustrative assumption.	Different interpretations of how anatomical bias influences the model.
D4	Claims that entropy-masking yields a 12.5 % absolute fitness advantage relative to baseline success probability.	No such claim; fitness advantage expressed in absolute units (0.005).	DraftA's wording conflates absolute fitness gain with a percentage of baseline success; DraftB reports raw fitness difference.

Resolution choice. Both models are retained as complementary analyses. Their quantitative predictions are presented side-by-side without asserting superiority, preserving the distinct assumptions each model embodies.

Appendix B. Claim attribution

Claim ID	Statement	Source drafts	Agreement status
C1	Formal definition of entropy-masking within a free-energy predictive-processing model.	A	CONVERGENT
C2	Derivation of an expected-fitness inequality that identifies conditions favouring entropy-masking.	A	CONVERGENT
C3	Numerical illustration of entropy-masking using thalamic afferent composition [1] and compressed-model planning success [7], yielding a relative fitness increase of 462.5 %.	A	CONVERGENT
C4	Closed-form optimal compression factor $\alpha^* = \max(0, 1 - \lambda c/p)$.	B	SINGLE
C5	For $p = 0.2$, $c = 1$, $\lambda = 0.1$, the optimal agent under-estimates hazard by 50 %, inflates effort by 12.5 %, and gains 0.005 fitness units (1.67 % relative).	B	SINGLE
C6	For $p = 0.05$, the optimal compression collapses to full suppression ($\alpha^* = 0$).	B	SINGLE
C7	Parameterisation of the entropy-masking model: $M = 0.8$, $S = 0.25$, $p = 0.925$, $\alpha = p$, $\beta = 1 - p$, $\gamma = 0.7$.	A	CONVERGENT
C8	Computed fitness values for the entropy-masking model: $F_{\text{acc}} = 0.015$, $F_{\text{mask}} = 0.084375$, $\Delta F = 0.069375$.	A	CONVERGENT
C9	Computed fitness values for the hazard-compression model at $p = 0.2$: $F_{\text{dist}} = 0.305$, $F_{\text{ver}} = 0.30$, gain = 0.005.	B	CONVERGENT
C10	Statement that the work resides at TRL 2.	A & B	CONVERGENT
C11	Proposal of a minimal de-risking experiment with two agents in a stochastic grid-world.	A & B	CONVERGENT
C12	Identification of limitations L1–L6.	A & B	CONVERGENT
C13	Falsification criterion based on lack of statistically significant fitness advantage.	A & B	CONVERGENT