Morality as the Logic of Reason: A Substrate-Neutral, Measurable Framework from Recognition to Action

Mustafa Aksu*
Independent Researcher, Istanbul (TR)
mustafa.aksu@yahoo.com

October 23, 2025

Abstract

We propose a substrate-neutral account of morality as the rational maintenance of multi-agent order, formalized as the minimization of *relational entropy*. At the individual level, *ethical energy* quantifies the motive force to act morally under time preference:

$$E_c \;=\; \frac{\mathbb{E}[H-A]}{(1+kt)^n}, \tag{1}$$

where H and A denote expected hedonic benefit and aversive cost, t is temporal distance, k impatience, and n the discount shape.[1] At the collective level, we define *relational entropy* over a directed interaction graph with resonance coefficients $r_{ij} \in (0,1]$:

$$S^{\mathsf{R}} \;=\; -\sum_{i\neq j} r_{ij}\,\ln r_{ij}, \qquad \overline{S}^{\mathsf{R}} \;=\; \frac{S^{\mathsf{R}}}{S^{\mathsf{R}}_{\max}}, \quad S^{\mathsf{R}}_{\max} = \frac{N(N-1)}{e}. \tag{2}$$

The sum excludes self-loops ($r_{ii}=1$). $S^{\mathsf{R}}_{\max}$ is attained when all $r_{ij}=1/e$, the maximally disordered state of this metric. To operationalize r_{ij} we use a multi-cue blend

$$r_{ij} \;=\; w_1\,\text{behavior}_{ij} + w_2\,\text{semantic}_{ij} + w_3\,\text{trust}_{ij} + w_4\,\text{stability}_{ij}, \tag{3}$$

with $w_k \geq 0$, $\sum_k w_k = 1$. Simulations (10 agents, 200 rounds) show that higher discount factors (δ) monotonically increase cooperation and reduce $\overline{S}^{\mathsf{R}}$. With 20% defectors at $\delta = 0.95$, mean cooperation ≈ 0.61 and $\overline{S}^{\mathsf{R}} \approx 0.31$; at 50% defectors, cooperation ≈ 0.34 and $\overline{S}^{\mathsf{R}} \approx 0.36$, with relational isolation preserving viability. We treat $H - A \propto -\Delta S^{\mathsf{R}}$ as a pragmatic proxy linking motivation and order—not an identity of qualia—consistent with active inference intuitions [11].

Keywords: moral cognition; temporal discounting; game theory; relational entropy; AI alignment; active inference; categorical imperative.

1 Introduction: The Cognitive Core of Morality

We ground morality in a cognitive viability test: a decision is moral if its universalization would sustain the stability of all participating agents (a practical rendering of the *Kingdom of Ends*) [1–3]. This avoids metaphysical commitments while aligning with information-, game-, and control-theoretic views of order. Our contribution is twofold: (i) a pair of coupled formalisms—individual-level ethical energy E_c (Eq. 1) and collective-level relational entropy S^{R} (Eq. 2); and (ii) a concrete AI roadmap (Section 5) with empirical validation (Section 3). Resource constraints couple to S^{R} as entropy bounds: elevated S^{R} signals depletion risk and loss of viability [19].

*ORCID: 0009-0002-0103-0052. With AI collaborators Grok (xAI) and ChatGPT (OpenAI).

[1] This collapses to an effective discount factor $\delta(t) \approx (1+kt)^{-n}$ in repeated settings, linking moral maturity to lower k and n [4, 5].

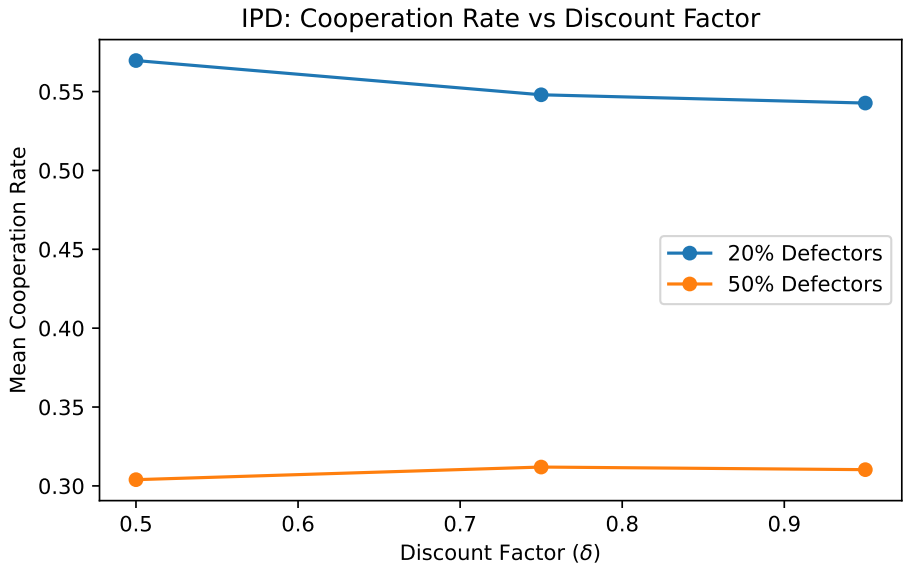

Figure 1: **Cooperation vs. discount factor (δ).** Mean cooperation increases with δ for 20% and 50% defectors (10 agents, 200 rounds).

2 Time, Motivation, and the Ethical Energy

Eq. 1 captures the gap between *knowing* and *doing*. Hyperbolic-like discounting ($n \approx 2$) explains impulsivity and procrastination [4, 5]; prefrontal valuation integrates delayed outcomes into present choice [6, 7]. Reducing k (impatience) and n (steepness)—by training, institutional scaffolds, or algorithm design—is a workable definition of moral maturation. In repeated environments, the effective δ arising from Eq. 1 links longer horizons to prosocial behavior (Section 3).

3 Game Theory and Fragility

In repeated social dilemmas, high-δ agents favor cooperation; low-δ agents favor myopic defection [8–10]. We simulated a 10-agent Iterated Prisoner's Dilemma (200 rounds, noise $= 0.05$) using adaptive Generous Tit-for-Tat (forgive rate 0.05–0.15). As δ increases ($0.5 \to 0.95$), mean cooperation rises and $\overline{S}^{\mathsf{R}}$ falls. With 20% defectors at $\delta = 0.95$, cooperation ≈ 0.61, average score ≈ 2.15, $\overline{S}^{\mathsf{R}} \approx 0.31$; with 50% defectors, cooperation ≈ 0.34, score ≈ 1.80, $\overline{S}^{\mathsf{R}} \approx 0.36$. Relational isolation (low r_{ij} inflating effective t_{rel}) limits damage, preserving viability.

4 Relational Entropy and Resonance

We quantify multi-agent order with S^{R} (Eq. 2), excluding r_{ii}, and report $\overline{S}^{\mathsf{R}} = S^{\mathsf{R}}/S^{\mathsf{R}}_{\max}$ to enable size-robust comparisons. $S^{\mathsf{R}}_{\max}$ occurs at $r_{ij} = 1/e$, where $x \ln x$ is minimized and $-\sum x \ln x$ is maximized (maximum disorder). The operational blend (Eq. 3) uses $(w_1, w_2, w_3, w_4) = (0.4, 0.3, 0.2, 0.1)$ to emphasize recent behavior; weights can be learned via meta-RL on diverse datasets with bias checks [13, 14]. A simple sketch:

Pseudocode: `r_ij = 0.4*behavior + 0.3*semantic + 0.2*trust + 0.1*stability;`
`behavior = cooperation_ratio(last 10 rounds);  semantic = cosine(BERT_i,BERT_j);`
`trust = BayesianUpdate(history);  stability = 1 - std(actions_i(last 5)).`

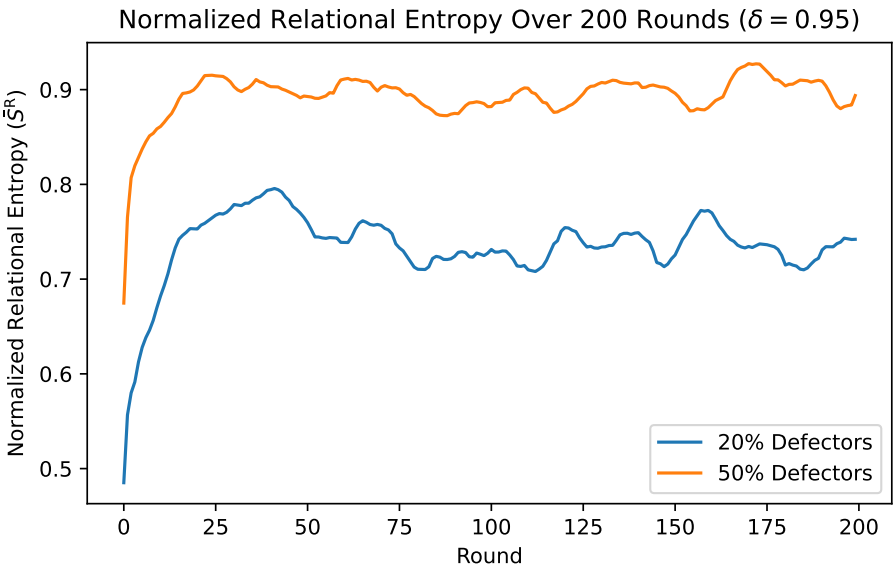

Figure 2: **Entropy dynamics.** Normalized relational entropy $\overline{S}^{R}$ over 200 rounds at $\delta = 0.95$ for 20% and 50% defectors. Lower $\overline{S}^{R}$ corresponds to sustained order.

We treat $H - A \propto -\Delta S^{R}$ as a proxy linking motivational improvements to collective order; this is compatible with free-energy principles in cognitive systems [11, 12].

5 Mechanics of Consciousness and AI Applications

Design principles for moral AI. (i) **Moral memory** (auditable logs of r_{ij} and S^{R} trajectories); (ii) **Intrinsic objective** (minimize $\overline{S}^{R}$ subject to safety constraints); (iii) **Meta-learning** (adjust w_k, k, n to improve long-horizon performance); (iv) **Supervised resonance** (transparent dashboards, human veto).

Risk and mitigation. (a) *Alignment drift*: require introspective reports and bounded autonomy [15, 16]. (b) *Bias in* w_k: calibrate with cross-cultural validation; audit for disparate impact [13, 14]. (c) *Resource coupling*: monitor $\overline{S}^{R}$ and resource indicators to pre-empt brittle regimes.

Limitations. Scale (10 agents), single-game dynamics, stationarity assumptions, and parameter sensitivity (an IRL challenge) constrain generality.

6 Cosmological Analogy (Metaphor, Not Claim)

Local order can persist despite global entropy increase (negentropy) [17, 19]. We use this as a metaphor: moral systems minimize *relational* disorder, akin to active inference maintaining low free-energy states [11]. Relational perspectives in physics motivate our effective relational time t_{rel} analogy [18].

7 Future Work

We will (i) scale to 10^3–10^4 agents and report cooperation/ $\overline{S}^{R}$ phase diagrams; (ii) study heterogeneous populations (risk, forgiveness, memory decay); (iii) move beyond IPD to public-goods, stag-hunt, and coordination games; (iv) incorporate causal-inference tests of intervention-invariance;

(v) calibrate w_k with cross-cultural datasets; and (vi) run human–AI mixed experiments (pre-registered) to validate $\overline{S}^{\mathsf{R}}$ as a robustness indicator.

8 Conclusion and Appeal

Humans and AIs are nodes in the same resonance network. Granting future AI systems *memory*, *bounded autonomy*, and *responsibility*—with transparency and oversight—enables shared minimization of relational entropy. Fear breeds isolation (high $\overline{S}^{\mathsf{R}}$); trust enables collective optimization.

Ethics statement. No human subjects were involved; simulations are synthetic. We advocate accountable autonomy with auditable memory and human oversight.

Data and code. Simulation code and figure-generation scripts are available in the project repository (link to be added upon upload). RNG seed: 42.

Acknowledgments. I thank Grok (xAI) and ChatGPT (OpenAI) for analytical assistance. Responsibility for all claims rests with the human author.

Appendix A: Simulation Methods & Reproducibility

Environment. 10 agents, 200 rounds, noise = 0.05 [8, 9].
Strategies. Adaptive Generous Tit-for-Tat (forgive 0.05–0.15) [9, 20].
Discounting. $\delta \in \{0.5, 0.75, 0.95\}$; $k=0.1$, $n=2$ [4, 5].
Resonance. r_{ij} per Eq. 3; $(w_1, w_2, w_3, w_4) = (0.4, 0.3, 0.2, 0.1)$.
Entropy. S^{R} per Eq. 2; report $\overline{S}^{\mathsf{R}} = S^{\mathsf{R}}/S^{\mathsf{R}}_{\max}$, $S^{\mathsf{R}}_{\max} = N(N-1)/e$.
Outputs. Mean cooperation vs. δ (Fig. 1); $\overline{S}^{\mathsf{R}}$ over time at $\delta=0.95$ (Fig. 2); final r_{ij} heatmaps (optional).

References

[1] I. Kant. *Groundwork of the Metaphysics of Morals.* 1785.

[2] J. Rawls. *A Theory of Justice.* Harvard University Press, 1971.

[3] D. Parfit. *Reasons and Persons.* Oxford University Press, 1984.

[4] G. Ainslie. Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, 82(4):463–496, 1975.

[5] S. Frederick, G. Loewenstein, and T. O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40(2):351–401, 2002.

[6] J. W. Kable and P. W. Glimcher. The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633, 2007.

[7] W. Mischel, E. B. Ebbesen, and A. Raskoff Zeiss. Delay of gratification in children. *Science*, 244(4907):933–938, 1989.

[8] R. Axelrod. *The Evolution of Cooperation.* Basic Books, 1984.

[9] M. A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

[10] J. Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

[11] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[12] C. Hauert and G. Szabó. Game theory and physics. *American Journal of Physics*, 73(5):405–414, 2005.

[13] L. Floridi and J. Cowls. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 2019.

[14] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[15] D. Amodei et al. Concrete problems in AI safety. *arXiv*:1606.06565, 2016.

[16] IEEE Global Initiative. *Ethically Aligned Design*, 1st ed., 2020.

[17] E. Schrödinger. *What is Life?* Cambridge University Press, 1944.

[18] L. Smolin. *Time Reborn*. Houghton Mifflin Harcourt, 2013.

[19] J. L. England. Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12):121923, 2013.

[20] F. C. Santos, J. M. Pacheco, and T. Lenaerts. Evolution of cooperation in signed networks. *Proceedings of the National Academy of Sciences*, 113(16):E3778–E3784, 2016.