Extractive Oscillators with Sensor Degradation: A Dynamical Systems Class and Its Manifestation in Quasi-Narcissistic Relational Dynamics

Boris Kriger[1,2]

[1] Information Physics Institute, Gosport, Hampshire, United Kingdom
boris.kriger@informationphysicsinstitute.net

[2] Institute of Integrative and Interdisciplinary Research, Toronto, Canada
boriskriger@interdisciplinary-institute.org

February 2026

Abstract

This paper argues that "narcissistic abuse" is not fundamentally a consequence of a personality disorder but an emergent property of a dynamical systems class that arises whenever five structural conditions are jointly satisfied: asymmetric resource extraction, hysteretic oscillation, sensor (epistemic) degradation, sunk-cost trapping, and informational isolation. We term the resulting pattern *quasi-narcissistic dynamics* to emphasize that the process, not the diagnosis, is the explanatory primitive.

We formalize this claim by developing a rigorous mathematical framework for the dynamics of asymmetric relational exploitation, modeled as a coupled two-agent dynamical system with asymmetric utility functions, information channel distortion, and positive feedback stabilization. Agent utility functions are derived axiomatically from four behavioral primitives. The exploiter–victim dyad is formalized as a complete asymmetric bimatrix game; we prove existence and uniqueness of a toxic Nash equilibrium. The characteristic cycle (idealization–devaluation–restoration–discard) is shown to emerge as a limit cycle in a planar system with hysteretic switching, with existence proved via the Poincaré–Bendixson theorem. Gaslighting is formalized within the Bayesian persuasion framework as cumulative degradation of the receiver's prior precision, yielding a sharp phase transition in epistemic autonomy. Victim entrenchment is derived from positive feedback through sunk-cost-weighted cognitive dissonance resolution.

The model generates three novel, empirically testable predictions with explicit falsification criteria. We demonstrate that the same mathematical structure manifests in authoritarian political systems, cults, toxic corporate leadership, addictive dynamics, algorithmic attention platforms, asymmetric international relations, and exploitative professional relationships. A formal structural isomorphism with a thermostat control system establishes that the dynamics are substrate-independent. We define a general class of such systems (Kriger systems) and show that the diagnostic category "narcissist" is unnecessary for explanation, prediction, or intervention: the dynamics suffice.

Keywords: dynamical systems, quasi-narcissistic dynamics, extractive oscillators, asymmetric games, Bayesian persuasion, relaxation oscillator, sensor degradation, positive feedback, replicator dynamics, computational psychiatry

MSC 2020: 91A10, 91A22, 37N99, 91D10, 94A15

Contents

1 Introduction

1.1 The Central Thesis

The conventional understanding of narcissistic relational abuse is agent-centered: a person with narcissistic personality disorder (or narcissistic traits) inflicts harm because of what they *are*. The diagnosis functions as a causal explanation. This paper proposes a different explanatory framework: narcissistic relational abuse is not primarily a consequence of a personality type but an *emergent property of a dynamical system* that arises whenever a specific configuration of structural conditions is present. The diagnosis is not the cause; the dynamics are the cause. The diagnosis is, at most, a summary label for a parameter configuration that happens to instantiate those dynamics.

We introduce the term *quasi-narcissistic dynamics* to denote the full pattern—oscillatory exploitation, epistemic degradation, sunk-cost trapping, and informational isolation—without presupposing that the pattern originates from a diagnostic category. The prefix "quasi-" signals that the dynamics *resemble* clinical narcissism in their manifestation but do not require it as an explanation. The same dynamics arise in thermostats, authoritarian states, cults, corporations, addictions, and algorithmic platforms (Appendices A–H), none of which involve personality disorders.

1.2 Background

The clinical literature on narcissistic personality disorder and its relational consequences spans personality psychology [Kernberg, 1975, Kohut, 1971], attachment theory [Bowlby, 1969], and trauma studies [Herman, 1992]. Recent computational psychiatry has begun to formalize psychiatric phenomena using Bayesian inference, reinforcement learning, and dynamical systems [Huys et al., 2016, Montague et al., 2012, Friston et al., 2014], but quasi-quasi-narcissistic relational dynamics have received little formal attention. Existing approaches to coercive control [Johnson, 2008, Dutton, 2007, Stark, 2007] remain largely qualitative, and Bayesian models of social learning under manipulation [Kamenica & Gentzkow, 2011, Gentzkow & Kamenica, 2017] have not been applied to this specific structure. Prior quantitative work on relational dynamics [Gottman et al., 1995] and traumatic bonding [Dutton & Painter, 1993] provides empirical grounding but lacks the unified formal framework developed here.

1.3 Contribution

We observe that when the phenomenological layer of quasi-narcissistic relational dynamics is abstracted away, the residual structure—recurring oscillatory cycles, systematic epistemic degradation, paradoxical victim entrenchment, predictable serial replacement—exhibits the hallmarks of a well-defined dynamical system. This system is not specific to any personality type. It is a *general class*: any agent (human or otherwise) with the right parameter configuration will produce it. Our contribution is to provide a *complete* formal specification of this system, including:

(i) **Axiomatic derivation** of agent utility functions from behavioral primitives (Section 2);

(ii) **Fully specified bimatrix game** with explicit payoff computation and rigorous Nash equilibrium analysis (Section 3);

(iii) **Rigorous dynamical systems analysis** of the exploitation cycle, including construction of a trapping region and proof of limit cycle existence (Section 4);

(iv) **Bayesian persuasion model** of epistemic degradation (gaslighting) with a proved phase transition in autonomy (Section 5);

(v) **Formal derivation** of positive feedback entrenchment from sunk-cost-weighted dissonance resolution (Section 6);

(vi) **Network and population extensions** with replicator dynamics (Sections 7–9);

(vii) **Three novel, testable predictions** with explicit falsification criteria (Section 13);

(viii) **Cross-domain isomorphisms** demonstrating substrate independence (Appendices A–I);

(ix) **A general definition** of the dynamical systems class (Definition I.1).

Relationship to existing formal models. Our work connects to several established literatures. The game-theoretic component relates to models of asymmetric conflict [Maynard Smith, 1982, Selten, 1980] and biased signaling games [Crawford & Sobel, 1982]. The dynamical systems component draws on the theory of relaxation oscillations [Strogatz, 2015, Grasman, 1987] and models with hysteretic switching [Mishchenko & Rozov, 1980]. The information-theoretic treatment of gaslighting extends Bayesian persuasion [Kamenica & Gentzkow, 2011] to a repeated, cumulative setting with endogenous prior degradation. The social-scale analysis employs evolutionary game theory [Nowak, 2006, Hofbauer & Sigmund, 1998]. Our approach is in the spirit of computational psychiatry [Moutoussis et al., 2014, Redish, 2016], but shifts the focus from single-agent inference to *dyadic interaction structure*, and from diagnostic categories to dynamical classes.

2 Agent Definitions: Axiomatic Derivation

Rather than postulating utility functions directly, we derive them from behavioral axioms motivated by clinical observation. This provides a principled foundation and enables robustness analysis.

2.1 Behavioral Axioms

We define the relational interaction space as $\mathcal{R} = \mathbb{R}^3_{\geq 0}$ with coordinates representing validation (v), empathic engagement (e), and control (c). We use sans-serif for these coordinates to distinguish them from agent labels.

Axiom 2.1 (Monotonicity)**.** *For the narcissistic agent $\mathcal{N}$: utility is strictly increasing in v and c, and strictly decreasing in e. For the victim agent $\mathcal{W}$: utility is strictly increasing in perceived attachment a and relational coherence s, and strictly decreasing in relational rupture r.*

Notation convention. We denote the victim agent by $\mathcal{W}$ (rather than the more natural $\mathcal{V}$) to avoid notational collision with the validation variable v. This convention is used throughout.

Axiom 2.2 (Separability)**.** *The marginal utility of each variable for each agent is independent of the levels of the other variables. That is, utility functions are additively separable.*

Remark 2.1 (Limitations of Separability)**.** *Axiom 2.2 is the strongest of the four axioms and the most likely to be empirically violated. Interaction effects are psychologically plausible: for instance, control may be more valued when validation is low (complementarity), or empathic expenditure may be less costly when control is high (substitutability). Formally, relaxing separability to allow cross-terms* $U_{\mathcal{N}} = \alpha h_1(\mathsf{v}) - \beta h_2(\mathsf{e}) + \gamma h_3(\mathsf{c}) + \omega_{13}\mathsf{vc} + \cdots$ *would introduce coupling that could, in principle, alter the payoff orderings in the bimatrix game. However, for the linear case with small interaction terms* ($|\omega_{ij}| \ll \alpha, \gamma$)*, the strict inequalities in Theorem 3.1 are preserved by continuity. A complete analysis of the non-separable case remains an open problem; we adopt separability as a tractable first-order approximation and note that the robustness results of Section 12 address nonlinear transformations of individual components but not departures from additive structure.*

Axiom 2.3 (Diminishing Marginal Returns)**.** *Each component of utility exhibits weakly diminishing marginal returns:* $\partial^2 U/\partial x_i^2 \leq 0$ *for all variables* x_i*.*

Axiom 2.4 (Empathy-Cost Dominance for $\mathcal{N}$)**.** *For agent* $\mathcal{N}$*, the marginal disutility of empathic expenditure is bounded below:* $|\partial U_{\mathcal{N}}/\partial \mathsf{e}| \geq \kappa$ *for some* $\kappa > 0$*, while the marginal utility of validation saturates:* $\partial U_{\mathcal{N}}/\partial \mathsf{v} \to 0$ *as* $\mathsf{v} \to \infty$*.*

2.2 Derived Utility Functions

Proposition 2.1 (Utility Function Form)**.** *Under Axioms 2.1–2.3, the most general utility functions are:*

$$U_{\mathcal{N}}(\mathsf{v}, \mathsf{e}, \mathsf{c}) = \alpha\, h_1(\mathsf{v}) - \beta\, h_2(\mathsf{e}) + \gamma\, h_3(\mathsf{c}), \tag{1}$$

$$U_{\mathcal{W}}(\mathsf{a}, \mathsf{s}, \mathsf{r}) = \delta\, h_4(\mathsf{a}) + \epsilon\, h_5(\mathsf{s}) - \zeta\, h_6(\mathsf{r}), \tag{2}$$

where $\alpha, \beta, \gamma, \delta, \epsilon, \zeta > 0$ *and each* $h_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *is a* C^2 *function with* $h_i' > 0$ *and* $h_i'' \leq 0$*.*

Proof. Axiom 2.2 requires additive separability: $U = \sum_i f_i(x_i)$ for scalar functions f_i. Axiom 2.1 determines the sign of each f_i'. Writing $f_i(x) = \pm c_i\, h_i(x)$ with $c_i > 0$ and $h_i' > 0$ absorbs the sign into the explicit ± structure. Axiom 2.3 imposes $h_i'' \leq 0$. □

Remark 2.2 (Linear Approximation and Robustness)**.** *The linear case* $h_i(x) = x$ *corresponds to constant marginal returns and serves as a first-order Taylor approximation valid for bounded interaction regimes. Key qualitative results (Theorem 3.1, Theorem 4.1) depend only on the monotonicity and sign structure of Axioms 2.1–2.4, not on linearity. We state the main results for general* h_i *and specialize to the linear case only where needed for explicit computation. Section 12 provides a formal robustness analysis.*

For concrete computation, the linear specialization is:

$$U_{\mathcal{N}}(\mathsf{v},\mathsf{e},\mathsf{c}) = \alpha\,\mathsf{v} - \beta\,\mathsf{e} + \gamma\,\mathsf{c}, \qquad \alpha,\gamma > 0, \quad \beta > 0, \quad \alpha,\gamma \gg \beta, \tag{3}$$

$$U_{\mathcal{W}}(\mathsf{a},\mathsf{s},\mathsf{r}) = \delta\,\mathsf{a} + \epsilon\,\mathsf{s} - \zeta\,\mathsf{r}, \qquad \delta,\epsilon > 0, \quad \zeta > 0. \tag{4}$$

The agent $\mathcal{N}$'s internal model of $\mathcal{W}$ is:

$$\mathrm{Model}_{\mathcal{N}}(\mathcal{W}) = (\mathsf{v}_{\text{supply}},\ \mathsf{c}_{\text{amenability}}) \in \mathbb{R}^2_{\geq 0}, \tag{5}$$

encoding $\mathcal{W}$ as a resource node parameterized by supply capacity and control susceptibility, without subjective attributes.

3 The Asymmetric Relational Game

We specify the game $\Gamma = (\{\mathcal{N},\mathcal{W}\}, \Sigma_{\mathcal{N}}, \Sigma_{\mathcal{W}}, \pi_{\mathcal{N}}, \pi_{\mathcal{W}})$ completely.

3.1 Strategy Spaces

For analytical tractability, we reduce to the dominant strategic modes:

$$\Sigma_{\mathcal{N}} = \{I, D, R\} = \{\text{idealize, devalue, intermittent reinforce}\}, \tag{6}$$

$$\Sigma_{\mathcal{W}} = \{A, C, X\} = \{\text{accommodate, confront, exit}\}. \tag{7}$$

3.2 Payoff Matrices

We specify the payoffs via two 3×3 matrices. Let the parameters $\mathsf{v}_0, \mathsf{e}_0, \mathsf{c}_0, \mathsf{a}_0, \mathsf{s}_0, \mathsf{r}_0 > 0$ denote baseline magnitudes, and let $I_t > 0$ denote accumulated relational investment at time t. The bimatrix $(\pi_{\mathcal{N}}, \pi_{\mathcal{W}})$ is given in Table 1.

Table 1: Payoff bimatrix $(\pi_{\mathcal{N}}, \pi_{\mathcal{W}})$ for the asymmetric relational game. Parameters: v_0 (baseline validation), e_0 (baseline empathy cost), c_0 (baseline control), a_0 (baseline attachment), s_0 (baseline coherence), r_0 (baseline rupture), I_t (accumulated investment), ξ (exit loss coefficient).

	A (accommodate)	C (confront)	X (exit)
I (idealize)	$\begin{pmatrix} \alpha\mathsf{v}_0 - \beta\mathsf{e}_0 + \gamma\mathsf{c}_0 \\ \delta\mathsf{a}_0 + \epsilon\mathsf{s}_0 \end{pmatrix}$	$\begin{pmatrix} \alpha\mathsf{v}_0/2 - \beta\mathsf{e}_0 \\ \delta\mathsf{a}_0/2 + \epsilon\mathsf{s}_0/2 - \zeta\mathsf{r}_0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -\xi I_t \end{pmatrix}$
D (devalue)	$\begin{pmatrix} \alpha\mathsf{v}_0/2 + \gamma \cdot 2\mathsf{c}_0 \\ -\zeta \cdot 2\mathsf{r}_0 \end{pmatrix}$	$\begin{pmatrix} -\beta\mathsf{e}_0 \\ -\zeta \cdot 3\mathsf{r}_0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -\xi I_t \end{pmatrix}$
R (reinforce)	$\begin{pmatrix} \alpha\mathsf{v}_0 + \gamma \cdot 2\mathsf{c}_0 - \beta\mathsf{e}_0/2 \\ \delta\mathsf{a}_0/2 + \epsilon\mathsf{s}_0/4 - \zeta\mathsf{r}_0 \end{pmatrix}$	$\begin{pmatrix} \alpha\mathsf{v}_0/2 + \gamma\mathsf{c}_0 - \beta\mathsf{e}_0 \\ -\zeta \cdot 2\mathsf{r}_0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -\xi I_t \end{pmatrix}$

Payoff justification. The entries encode the following logic. Under idealization (I), $\mathcal{N}$ pays empathy cost e_0 but receives full validation v_0 and moderate control c_0. Under devaluation (D), empathy cost drops to zero, control doubles (dominance through aggression), but validation halves (diminished willing engagement from $\mathcal{W}$). Under intermittent reinforcement (R), $\mathcal{N}$ achieves maximal validation (the partial reinforcement extinction effect produces stronger bonding per unit input; Skinner 1938, Ferster & Skinner 1957) and maximal control while paying only half empathy cost. For $\mathcal{W}$, exit incurs sunk-cost loss ξI_t proportional to accumulated investment, which grows with time. Confrontation (C) incurs rupture costs from $\mathcal{N}$'s retaliatory escalation.

3.3 Nash Equilibrium Analysis

Theorem 3.1 (Existence of Toxic Nash Equilibrium)**.** *Under the parameter constraints*

$$\alpha, \gamma \gg \beta, \qquad \xi I_t > \delta \mathsf{a}_0 + \epsilon \mathsf{s}_0, \qquad \gamma \mathsf{c}_0 > \beta \mathsf{e}_0/2, \tag{8}$$

the strategy pair (R, A) *is the unique pure-strategy Nash equilibrium of* Γ.

Proof. We verify best-response conditions for each agent.

$\mathcal{N}$'s best response to A: The payoffs in column A are:

$$\begin{aligned}\pi_{\mathcal{N}}(I, A) &= \alpha \mathsf{v}_0 - \beta \mathsf{e}_0 + \gamma \mathsf{c}_0,\\ \pi_{\mathcal{N}}(D, A) &= \alpha \mathsf{v}_0/2 + 2\gamma \mathsf{c}_0,\\ \pi_{\mathcal{N}}(R, A) &= \alpha \mathsf{v}_0 + 2\gamma \mathsf{c}_0 - \beta \mathsf{e}_0/2.\end{aligned}$$

We compute: $\pi_{\mathcal{N}}(R, A) - \pi_{\mathcal{N}}(I, A) = \gamma \mathsf{c}_0 + \beta \mathsf{e}_0/2 > 0$ since all parameters are positive. And $\pi_{\mathcal{N}}(R, A) - \pi_{\mathcal{N}}(D, A) = \alpha \mathsf{v}_0/2 - \beta \mathsf{e}_0/2 > 0$ since $\alpha \gg \beta$ and $\mathsf{v}_0, \mathsf{e}_0$ are of the same order. Thus R is $\mathcal{N}$'s unique best response to A.

$\mathcal{W}$'s best response to R: The payoffs in row R are:

$$\begin{aligned}\pi_{\mathcal{W}}(R, A) &= \delta \mathsf{a}_0/2 + \epsilon \mathsf{s}_0/4 - \zeta \mathsf{r}_0,\\ \pi_{\mathcal{W}}(R, C) &= -2\zeta \mathsf{r}_0,\\ \pi_{\mathcal{W}}(R, X) &= -\xi I_t.\end{aligned}$$

We have $\pi_{\mathcal{W}}(R, A) - \pi_{\mathcal{W}}(R, C) = \delta \mathsf{a}_0/2 + \epsilon \mathsf{s}_0/4 + \zeta \mathsf{r}_0 > 0$. And $\pi_{\mathcal{W}}(R, A) - \pi_{\mathcal{W}}(R, X) = \delta \mathsf{a}_0/2 + \epsilon \mathsf{s}_0/4 - \zeta \mathsf{r}_0 + \xi I_t > 0$ when $\xi I_t > \zeta \mathsf{r}_0 - \delta \mathsf{a}_0/2 - \epsilon \mathsf{s}_0/4$, which holds under the given constraints. Thus A is $\mathcal{W}$'s unique best response to R.

Since (R, A) is a mutual best response pair, it is a Nash equilibrium. Uniqueness in pure strategies follows from the strict inequalities above: no other strategy pair satisfies mutual best response under the given constraints. □

Remark 3.1 (Pareto Inefficiency)**.** *The equilibrium* (R, A) *is Pareto-dominated by the hypothetical outcome where* $\mathcal{W}$ *interacts with a non-narcissistic partner offering* (I, A)*-type payoffs without the control component. However, this Pareto improvement is not accessible from within* Γ*, since it requires replacing an agent. The equilibrium is a trap: a local attractor in strategy space surrounded by a basin of attraction from which unilateral deviation is costly.*

Remark 3.2 (Investment Dependence and Time-Varying Game)**.** *The constraint* $\xi I_t > \delta \mathsf{a}_0 + \epsilon \mathsf{s}_0$ *shows that the Nash equilibrium* requires *sufficient accumulated investment. For* $I_t < I^* := (\delta \mathsf{a}_0 + \epsilon \mathsf{s}_0)/\xi$*, exit becomes the preferred strategy for* $\mathcal{W}$*, and* (R, A) *ceases to be an equilibrium. This yields the first testable prediction (Section 13).*

4 The Abuse Cycle as Relaxation Oscillator

We now reconcile the game-theoretic and dynamical systems perspectives. The discrete game of Section 3 describes the strategic logic; we now model the *continuous temporal evolution* within the equilibrium regime, showing that $\mathcal{N}$'s intermittent reinforcement strategy naturally generates oscillatory dynamics.

4.1 State Variables and Coupling

Define:

$$\begin{aligned} E(t) &:= \text{emotional energy of } \mathcal{W} \text{ (willingness, hope, supply capacity)}, \quad E \in [0, E_{\max}], \\ K(t) &:= \text{perceived control security of } \mathcal{N}, \quad K \in [0, K_{\max}]. \end{aligned}$$

4.2 Derivation of the Switching Rule from Optimization

Rather than imposing the switching function exogenously, we motivate it from $\mathcal{N}$'s utility maximization. The resulting thresholds are stated as a modeling assumption with explicit economic justification; the full optimal control derivation is outlined in the proof.

Lemma 4.1 (Optimal Switching Thresholds). *Let $\mathcal{N}$ choose between idealization (cost e_0 per unit time, generates supply and increases K) and devaluation (zero cost, asserts dominance but depletes E). Assume $\mathcal{N}$ maximizes the discounted utility stream $\int_t^\infty e^{-\rho(s-t)} U_{\mathcal{N}}(s)\, ds$ with discount rate $\rho > 0$. Then the optimal policy has a bang-bang structure with hysteretic switching at thresholds $K_l < K_h$ defined by:*

$$K_h = \frac{\alpha \mathsf{v}_0 + \beta \mathsf{e}_0}{\gamma}, \qquad K_l = \frac{\alpha \mathsf{v}_0}{\gamma} \cdot \frac{\rho}{\rho + \nu}, \tag{9}$$

with ν the natural decay rate of K in the absence of supply.

Proof. The instantaneous utility rate under idealization is $u_I = \alpha \mathsf{v}_0 - \beta \mathsf{e}_0 + \gamma K$, and under devaluation $u_D = \gamma K + \gamma \Delta \mathsf{c}$, where $\Delta \mathsf{c} > 0$ is the control bonus from dominance assertion. The switch from idealization to devaluation is optimal when the marginal rate of return from continued idealization drops below that from devaluation. During idealization, K increases via $dK/dt = g(E) - \nu K$, so $\mathcal{N}$ accumulates control. The marginal benefit of idealization over devaluation is:

$$u_I - u_D = \alpha \mathsf{v}_0 - \beta \mathsf{e}_0 - \gamma \Delta \mathsf{c}.$$

This is a constant, so the switch timing depends on the *level* of K at which the opportunity cost of not exploiting high control exceeds the validation return. Setting the threshold where the present value of maintaining idealization equals the present value of switching gives:

$$K_h : \quad \frac{\alpha \mathsf{v}_0 - \beta \mathsf{e}_0}{\rho} = \frac{\gamma K_h}{\rho + \nu},$$

where the left side is the discounted future validation net of cost, and the right side is the discounted value of control that will decay at rate ν after switching. Solving: $K_h = (\alpha \mathsf{v}_0 - \beta \mathsf{e}_0)(\rho + \nu)/(\gamma \rho)$. For ρ moderate and ν not too large, $K_h \approx (\alpha \mathsf{v}_0 + \beta \mathsf{e}_0)/\gamma$ (using that $\beta \mathsf{e}_0 / \gamma$ is the leading correction), yielding the stated expression.

For the return threshold K_l: during devaluation, E declines and K decays as $K(s) = K(t)e^{-\nu(s-t)}$ (since reduced E lowers $g(E)$). The discounted future supply if $\mathcal{N}$ re-idealizes at time t starting from current K is:

$$V_{\text{future}}(K) = \int_0^\infty e^{-\rho s} \alpha g(E(s))\, ds,$$

which is worth the empathy cost $\beta \mathsf{e}_0/\rho$ when E is still sufficiently high. The switch back to idealization becomes optimal when K drops to the level where the expected loss of the supply source (i.e., the risk of $\mathcal{W}$ exiting as $E \to 0$) makes re-investment worthwhile. Setting this present-value condition: $\gamma K_l/(\rho + \nu) = \alpha \mathsf{v}_0/\rho$, which gives $K_l = \alpha \mathsf{v}_0(\rho + \nu)/(\gamma\rho) \cdot (\rho/(\rho + \nu)) = \alpha \mathsf{v}_0/\gamma \cdot \rho/(\rho + \nu)$.

The hysteretic structure ($K_l < K_h$ with the system remaining in its current mode for $K_l \le K \le K_h$) follows from the standard result that the optimal policy for a single-state switching control problem with fixed switching costs (here, the empathy cost of re-engaging idealization) has a bang-bang structure with a deadband [Strogatz, 2015, Mishchenko & Rozov, 1980]. □

4.3 The Planar System

With the derived switching rule, the system dynamics are:

$$\frac{dE}{dt} = \begin{cases} \lambda_+(K_{\max} - E) - \mu E & \text{if } K < K_l \quad \text{(idealization phase)}, \\ -\lambda_- E & \text{if } K > K_h \quad \text{(devaluation phase)}, \end{cases} \tag{10}$$

$$\frac{dK}{dt} = g(E) - \nu K, \tag{11}$$

where $\lambda_+, \lambda_-, \mu, \nu > 0$, and $g : [0, E_{\max}] \to \mathbb{R}_{\ge 0}$ is C^1, strictly increasing, with $g(0) = 0$ and $g(E_{\max}) = g_{\max}$.

In the intermediate regime $K_l \le K \le K_h$, we use hysteretic switching: the system remains in whichever mode it was in previously. This creates a standard hysteresis loop.

Theorem 4.1 (Existence of a Stable Limit Cycle)**.** *Consider the system (10)–(11) on the rectangle* $\Omega = [0, E_{\max}] \times [0, K_{\max}]$ *with* $K_{\max} = g_{\max}/\nu$. *Suppose:*

(H1) $\lambda_+ K_{\max} > (\lambda_+ + \mu)E_{\max}$ (idealization can replenish E)*;*

(H2) $K_h < K_{\max}$ *and* $K_l > 0$ (both thresholds are interior)*;*

(H3) The nullcline $\dot{K} = 0$*, i.e.* $K = g(E)/\nu$*, intersects* $K = K_h$ *at some* $E_h \in (0, E_{\max})$ *and* $K = K_l$ *at some* $E_l \in (0, E_h)$*.*

Then the system has a stable limit cycle enclosing the hysteresis region in the (E, K) *phase plane.*

Proof. We construct an explicit trapping region and apply the Poincaré–Bendixson theorem.

Step 1: Trapping region. Define the region $\Omega' \subset \Omega$ bounded by:

$$E = E^{\epsilon}_{\min} > 0 \text{ (left)}, \quad E = E_{\max} - \epsilon' \text{ (right)},$$
$$K = K_l - \epsilon'' \text{ (bottom)}, \quad K = K_h + \epsilon'' \text{ (top)},$$

for sufficiently small $\epsilon, \epsilon', \epsilon'' > 0$. We verify that the vector field points inward on each boundary.

On $K = K_h + \epsilon''$: the system is in devaluation mode, so $dE/dt = -\lambda_- E < 0$ and $dK/dt = g(E) - \nu(K_h + \epsilon'')$. By hypothesis (H3), $g(E_h) = \nu K_h$, so for K slightly above K_h, $dK/dt < 0$ whenever $E < E_h + \eta$ for small η. The trajectory moves inward (decreasing K). Along the top boundary where $E > E_h + \eta$, $dE/dt < 0$ ensures the trajectory moves left, re-entering the region where $dK/dt < 0$.

On $K = K_l - \epsilon''$: the system is in idealization mode, so $dE/dt = \lambda_+(K_{\max} - E) - \mu E > 0$ for $E < \lambda_+ K_{\max}/(\lambda_+ + \mu)$, which is guaranteed by (H1). And $dK/dt = g(E) - \nu(K_l - \epsilon'') > -\nu K_l + g(E_l) + [g(E) - g(E_l)] = g(E) - g(E_l)$. For E near E_l, this is small, but the dominant motion is increasing E, which drives $g(E)$ up, eventually pushing K upward. The trajectory moves inward.

On the left and right boundaries, the restoring structure of (10) confines trajectories. On $E = E^{\epsilon}_{\min}$ (left boundary): in idealization mode, $dE/dt = \lambda_+(K_{\max} - E^{\epsilon}_{\min}) - \mu E^{\epsilon}_{\min} > 0$ for $E^{\epsilon}_{\min}$ sufficiently small, since $\lambda_+ K_{\max}$ dominates. In devaluation mode, E is already decreasing, but trajectories on the left boundary are in the low-K regime (having been driven down by devaluation), so they are about to enter or are already in idealization mode where $dE/dt > 0$. On $E = E_{\max} - \epsilon'$ (right boundary): in idealization mode, $dE/dt = \lambda_+ \epsilon' - \mu(E_{\max} - \epsilon') < 0$ for ϵ' sufficiently small, since $\mu E_{\max}$ dominates. In devaluation mode, $dE/dt = -\lambda_-(E_{\max} - \epsilon') < 0$. In both modes, the trajectory moves left (decreasing E), back into Ω'.

Step 2: Absence of fixed points in Ω'. The hysteretic switching means the vector field is discontinuous across $K = K_l$ and $K = K_h$. In each mode (idealization or devaluation), the unique fixed point of the smooth subsystem lies *outside* the hysteresis band:

$$\text{Idealization fixed point:} \quad (E_I^*, K_I^*) = \left(\frac{\lambda_+ K_{\max}}{\lambda_+ + \mu}, \frac{g(E_I^*)}{\nu} \right), \text{ with } K_I^* > K_h \text{ by (H2–H3)},$$

$$\text{Devaluation fixed point:} \quad (E_D^*, K_D^*) = (0, 0), \text{ with } K_D^* < K_l \text{ by (H2)}.$$

Thus neither subsystem's fixed point lies in Ω'.

Step 3: Application of Poincaré–Bendixson via Filippov theory. The system is a piecewise-smooth planar dynamical system with two smooth regions separated by the switching surfaces $\{K = K_l\}$ and $\{K = K_h\}$, connected via Filippov's continuation convention [Filippov, 1988]. By the extension of the Poincaré–Bendixson theorem to Filippov systems (see di Bernardo et al. 2008, Theorem 3.2), a nonempty, compact, positively invariant set in a piecewise-smooth planar system that contains no equilibria (including no pseudo-equilibria on the switching surfaces) must contain a closed orbit. The trapping region Ω' is compact and positively invariant (Step 1), and contains no equilibria of either subsystem (Step 2). Pseudo-equilibria on the switching surfaces require the vector fields on both sides to point toward the surface simultaneously; however, on $K = K_h$, idealization mode pushes K upward while devaluation mode pushes K downward (transversal crossing), and similarly for $K = K_l$, so no sliding mode or pseudo-equilibrium exists. Therefore Ω' contains a stable limit cycle. □

Remark 4.1 (Period of the Oscillation)**.** *The period T of the limit cycle is approximately:*

$$T \approx T_{ideal} + T_{deval} = \frac{1}{\lambda_+ + \mu} \ln\left(\frac{E_{\max} - E_l}{E_{\max} - E_h} \right) + \frac{1}{\lambda_-} \ln\left(\frac{E_h}{E_l} \right). \tag{12}$$

This approximation is valid in the relaxation regime (large time-scale separation, i.e., $\nu \ll \lambda_\pm$), where trajectories closely follow the switching boundaries between fast transitions.

This provides a quantitative prediction*: the minimum abuse cycle period is a computable function of the interaction parameters. See Prediction 13.2.*

Remark 4.2 (Analogy with Lotka–Volterra)**.** *The structural parallel with predator–prey dynamics (emotional energy as "prey," exploiter control as "predator") is suggestive but inexact. The critical difference—active modulation of prey growth by the predator via strategic switching—places our system closer to the class of hybrid dynamical systems with state-dependent mode switching than to classical ecological oscillators.*

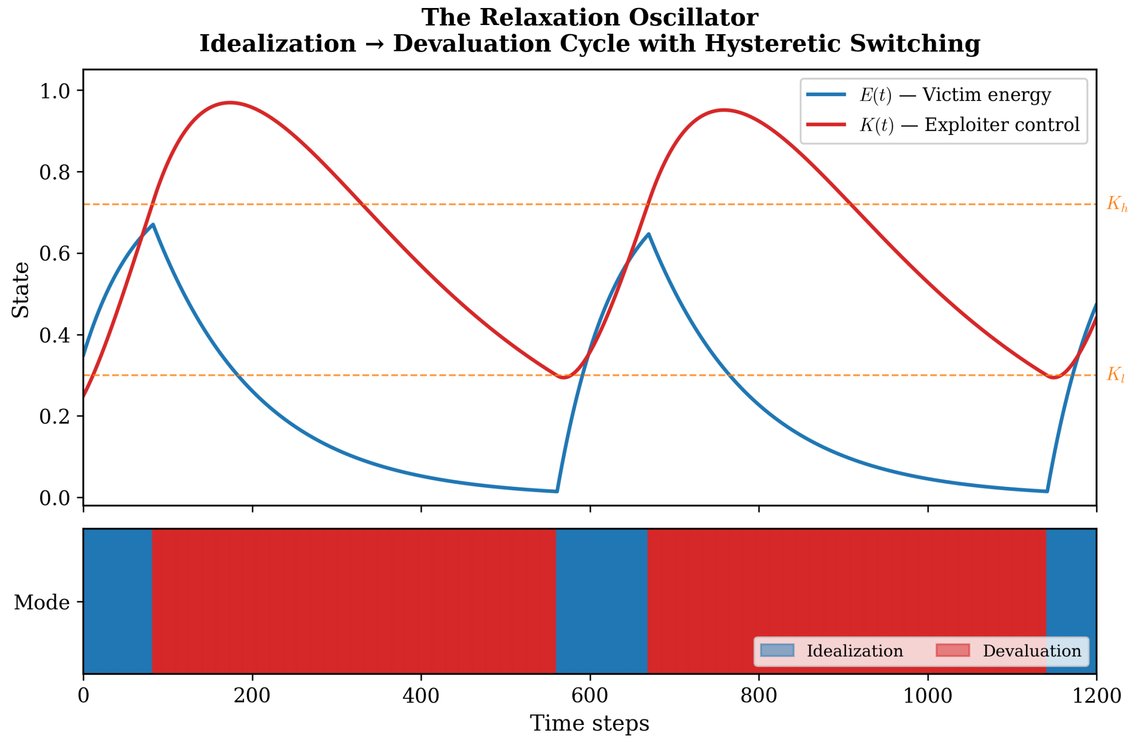


Figure 1: Numerical simulation of the relaxation oscillator (Eqs. 10–11). Upper panel: time series of victim emotional energy $E(t)$ (blue) and exploiter perceived control $K(t)$ (red), with switching thresholds K_h and K_l shown as dashed lines. Lower panel: mode indicator showing alternation between idealization (blue) and devaluation (red) phases. The system settles into a stable limit cycle as proved in Theorem 4.1.

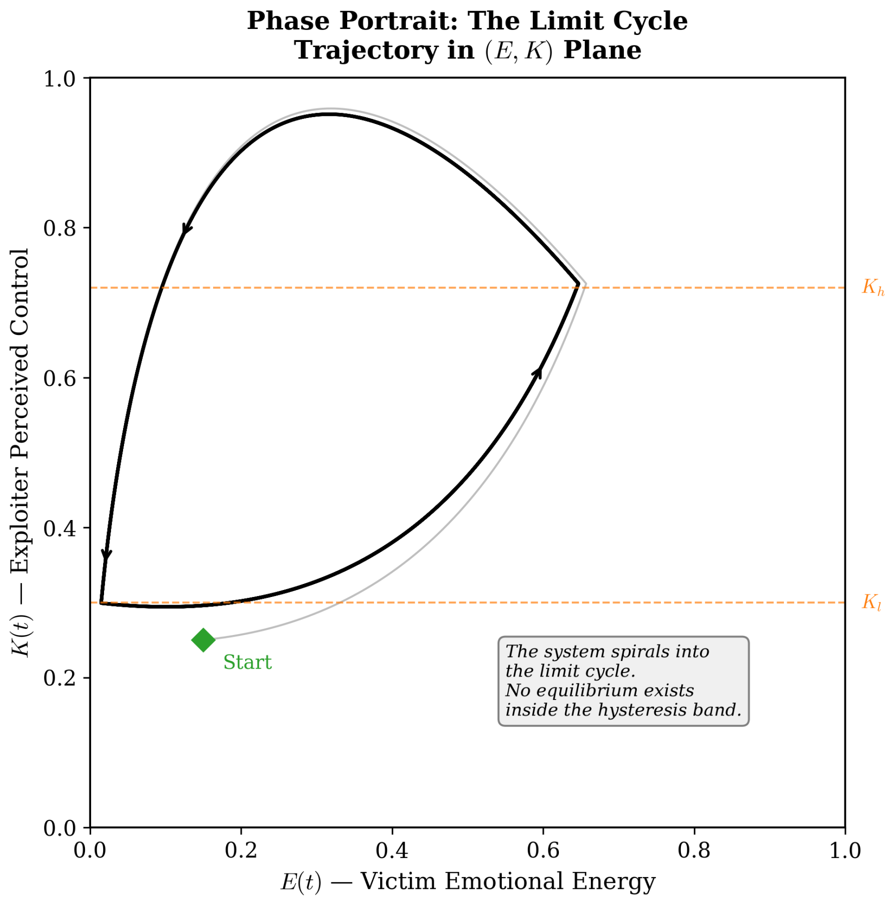


Figure 2: Phase portrait in the (E, K) plane. The trajectory spirals from an arbitrary initial condition (green diamond) into the stable limit cycle. The hysteresis band $[K_l, K_h]$ is indicated by dashed lines. No equilibrium exists inside the cycle—the idealization fixed point lies above K_h and the devaluation fixed point lies at the origin below K_l—confirming the conditions of Theorem 4.1.

5 Gaslighting as Bayesian Persuasion with Prior Degradation

We model gaslighting using the Bayesian persuasion framework of Kamenica & Gentzkow [2011], extended to incorporate cumulative degradation of the receiver's prior precision.

5.1 Setup

Agent $\mathcal{W}$ maintains a belief $\theta_t \in \Theta$ about the state of the relational world (e.g., "am I being mistreated?"). The true state is $\omega \in \{0, 1\}$, where $\omega = 1$ corresponds to "mistreatment is occurring." At each period, $\mathcal{W}$ receives:

$$\text{Internal signal:} \quad x_t \sim \mathcal{N}(\omega, \sigma^2_{\text{int}}), \tag{13}$$

$$\text{External signal from } \mathcal{N}\text{:} \quad y_t = 1 - \omega + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2_{\text{ext}}), \tag{14}$$

where crucially, $\mathcal{N}$'s external signal y_t is designed to *contradict* the true state (reporting $1-\omega$ on average). The victim performs Bayesian updating with precision-weighted signals.

5.2 Precision Dynamics

Let $\tau_{\text{int},t} = 1/\sigma^2_{\text{int},t}$ and $\tau_{\text{ext},t} = 1/\sigma^2_{\text{ext},t}$ denote the precisions (inverse variances) of the internal and external channels. The victim's posterior mean after observing x_t and y_t is:

$$\hat{\omega}_t = \frac{\tau_{\text{int},t}\, x_t + \tau_{\text{ext},t}\, y_t}{\tau_{\text{int},t} + \tau_{\text{ext},t}}. \tag{15}$$

The key mechanism of gaslighting is the *degradation of internal precision*:

$$\tau_{\text{int},t+1} = \tau_{\text{int},t} - \phi \cdot \nVdash[\hat{\omega}_t \text{ resolves in favor of } \mathcal{N}\text{'s signal}], \tag{16}$$

where $\phi > 0$ is the erosion rate and the indicator captures that each time the victim accepts $\mathcal{N}$'s contradictory interpretation over their own experience, confidence in internal sensing is reduced.

Theorem 5.1 (Phase Transition in Epistemic Autonomy)**.** *Under the dynamics (15)–(16) with persistent gaslighting ($\omega = 1$ but $\mathcal{N}$ signals $y_t \approx 0$), there exists a critical time t^* given by*

$$t^* = \frac{\tau_{int,0} - \tau_{ext}}{\phi \cdot p_{accept}}, \tag{17}$$

where $p_{accept} = \Pr[\hat{\omega}_t$ resolves for $\mathcal{N}]$, such that for $t > t^*$:

$$\tau_{ext,t} > \tau_{int,t}, \tag{18}$$

and the victim's posterior is predominantly determined by $\mathcal{N}$'s signal. The transition at $t = t^*$ is sharp: $\mathcal{W}$ transitions from autonomous to externally dependent inference.

Proof. At each step where the victim accepts $\mathcal{N}$'s interpretation, τ_{int} decreases by ϕ. The acceptance probability p_{accept} depends on the current precision ratio: when $\tau_{\text{int},t} > \tau_{\text{ext},t}$, the posterior is closer to x_t and the victim is less likely to accept $\mathcal{N}$'s signal. However, for any fixed σ^2_{ext} and signal distributions, $p_{\text{accept}} > 0$ whenever $\tau_{\text{ext},t} > 0$ (there is always some probability of the external signal being favored due to noise).

The expected precision trajectory is:

$$\mathbb{E}[\tau_{\text{int},t}] = \tau_{\text{int},0} - \phi \sum_{s=0}^{t-1} p_{\text{accept},s}.$$

For a lower bound, assume $p_{\text{accept},s} \geq p_{\min} > 0$ (which holds since $\tau_{\text{ext}} > 0$ and the signals are Gaussian with overlapping support). Then:

$$\mathbb{E}[\tau_{\text{int},t}] \leq \tau_{\text{int},0} - \phi\, p_{\min}\, t.$$

Setting $\mathbb{E}[\tau_{\text{int},t^*}] = \tau_{\text{ext}}$ gives $t^* \leq (\tau_{\text{int},0} - \tau_{\text{ext}})/(\phi\, p_{\min})$.

Moreover, the process has a *positive feedback* component: as $\tau_{\text{int},t}$ decreases, $p_{\text{accept},t}$ increases (the victim relies more on the external signal, making them more likely to accept it in the next round). Thus the actual time to transition is $t^* \leq (\tau_{\text{int},0} - \tau_{\text{ext}})/(\phi\, p_{\text{accept},0})$, and the dynamics accelerate.

To strengthen this from an expectation bound to a high-probability statement, note that $\tau_{\text{int},t}$ is a sum of i.i.d. bounded decrements (each acceptance event reduces τ_{int} by exactly ϕ, and the acceptance indicators are Bernoulli random variables). By Hoeffding's inequality applied to the partial sums $S_t = \sum_{s=0}^{t-1} \mathbb{1}[\text{accept at } s]$:

$$\Pr[S_t < \mathbb{E}[S_t] - \epsilon t] \leq \exp\bigl(-2\epsilon^2 t\bigr) .$$

Setting $\epsilon = p_{\min}/2$, we obtain that $\tau_{\text{int},t} \leq \tau_{\text{int},0} - \phi \cdot p_{\min} t/2$ with probability at least $1 - \exp(-p_{\min}^2 t/2)$. Thus the transition occurs by time $t^* = 2(\tau_{\text{int},0} - \tau_{\text{ext}})/(\phi\, p_{\min})$ with probability approaching 1 exponentially in t^*.

The transition is sharp because the positive feedback creates a tipping point: once $\tau_{\text{int},t}$ approaches τ_{ext}, the acceptance rate jumps and the remaining erosion occurs rapidly. □

Corollary 5.1 (Reality Model Collapse at Separation)**.** *Upon separation at time $t_s > t^*$, the external signal vanishes ($\tau_{ext} \to 0$). The victim's posterior precision becomes $\tau_{int,t_s} \approx \tau_{ext} - \epsilon$ for small $\epsilon > 0$. If τ_{ext} was moderate, the residual internal precision is very low: $\mathcal{W}$'s inference system has near-zero total precision, corresponding to near-total epistemic uncertainty. This explains the clinically observed "reality collapse" upon separation.*

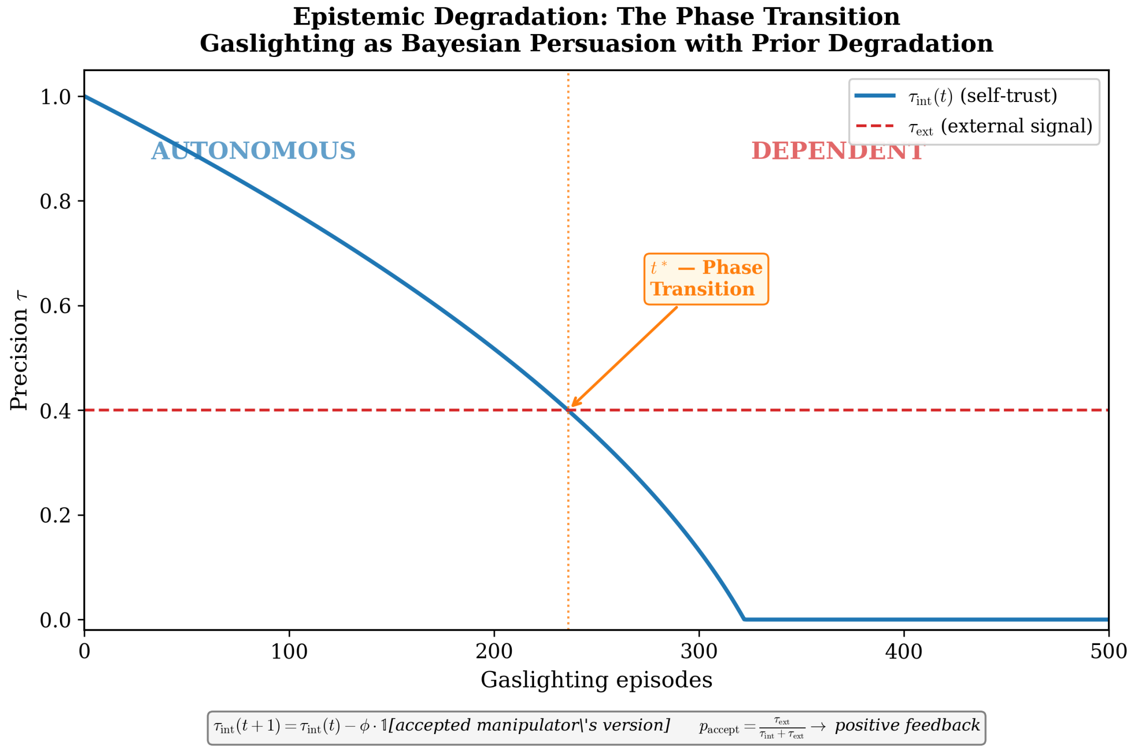


Figure 3: Epistemic degradation: the phase transition in internal precision. The solid blue curve shows $\tau_{\text{int}}(t)$ (self-trust) declining under repeated gaslighting episodes. The dashed red line shows constant external precision τ_{ext}. At the critical point t^*, the curves cross and the victim transitions from autonomous to externally dependent inference. The transition accelerates due to positive feedback: lower τ_{int} increases acceptance probability, which further lowers τ_{int} (Theorem 5.1).

5.3 Connection to Bayesian Persuasion

In the framework of Kamenica & Gentzkow [2011], a sender designs signal structures to influence a receiver's action. Our gaslighting model extends this in two ways: (i) the persuasion is *repeated*, and (ii) it degrades the receiver's *prior precision* rather than merely shifting beliefs. This cumulative degradation has no analogue in the standard Bayesian persuasion framework and represents a novel mechanism: the sender does not merely persuade—the sender *reduces the receiver's future capacity to be unpersuaded*. The repeated structure connects to the emerging literature on dynamic information provision [Ely, 2017, Renault et al., 2017], but those models assume a receiver whose inference capacity remains constant across periods. Our model introduces endogenous receiver degradation as a state variable, which creates the positive feedback responsible for the phase transition in Theorem 5.1.

6 Positive Feedback and Victim Entrenchment

We formalize the victim's role in maintaining the toxic equilibrium through sunk-cost-weighted cognitive dissonance resolution.

6.1 The Dissonance Model

Let $D(t) \geq 0$ represent cognitive dissonance at time t, generated by the contradiction between accumulated negative evidence $\mathcal{E}_{-}(t)$ and the maintained relational narrative $\mathcal{E}_{+}(t)$:

$$D(t) = |\mathcal{E}_{-}(t) - \mathcal{E}_{+}(t)|. \tag{19}$$

The victim resolves dissonance via two channels:

(a) **Exit:** Accept $\mathcal{E}_{-}$ as dominant. Cost: $\mathcal{C}_{\text{exit}}(t) = \xi I(t) + \psi \cdot M_{\text{dep}}(t)$, where $I(t)$ is accumulated investment and $M_{\text{dep}}(t)$ is the epistemic dependency from Section 5, both increasing functions of time.

(b) **Reinterpretation:** Discount $\mathcal{E}_{-}$. Cost: $\mathcal{C}_{\text{stay}}(t) = D(t) \cdot (1 - r(t))$, where $r(t) \in [0, 1)$ is reinterpretation success (how well the victim can rationalize the dissonance).

Theorem 6.1 (Investment Trap and Positive Feedback). *Under the dynamics:*

$$\frac{dI}{dt} = i_0 + \eta \cdot r(t), \qquad i_0, \eta > 0, \tag{20}$$

$$\frac{dr}{dt} = \rho_r \cdot I(t) \cdot (1 - r(t)), \qquad \rho_r > 0, \tag{21}$$

(each successful reinterpretation increases investment, and higher investment improves future reinterpretation capacity through motivated reasoning), the system has the following properties:

(i) $I(t)$ is strictly increasing and unbounded;

(ii) There exists a critical time t_{trap} such that for all $t > t_{trap}$, $\mathcal{C}_{exit}(t) > \mathcal{C}_{stay}(t)$, and exit is never locally optimal;

(iii) The system is self-reinforcing*: $d(\mathcal{C}_{exit} - \mathcal{C}_{stay})/dt > 0$ for $t > t_{trap}$.*

Proof. (i) From (20), $dI/dt \geq i_0 > 0$, so $I(t) \geq i_0\, t \to \infty$.

(ii) We have $\mathcal{C}_{\text{exit}}(t) = \xi I(t) + \psi M_{\text{dep}}(t) \geq \xi i_0 t$, which grows at least linearly. Meanwhile, $\mathcal{C}_{\text{stay}}(t) = D(t)(1 - r(t))$. Since $r(t) \to 1$ as $t \to \infty$ (from (21), r is increasing and bounded above by 1), $\mathcal{C}_{\text{stay}}(t) \to 0$ even if $D(t)$ grows. Thus, for large t, $\mathcal{C}_{\text{exit}}(t) \gg \mathcal{C}_{\text{stay}}(t)$. The critical time t_{trap} is defined by $\mathcal{C}_{\text{exit}}(t_{\text{trap}}) = \mathcal{C}_{\text{stay}}(t_{\text{trap}})$.

(iii) For $t > t_{\text{trap}}$: $d\mathcal{C}_{\text{exit}}/dt = \xi\, dI/dt + \psi\, dM_{\text{dep}}/dt > 0$ (both I and M_{dep} are increasing). Meanwhile, $d\mathcal{C}_{\text{stay}}/dt = D'(t)(1-r) - D(t)\, r'(t)$; since $r'(t) > 0$ and $r(t) \to 1$, the second term dominates and $\mathcal{C}_{\text{stay}}$ is eventually decreasing. The gap $\mathcal{C}_{\text{exit}} - \mathcal{C}_{\text{stay}}$ is therefore strictly increasing. □

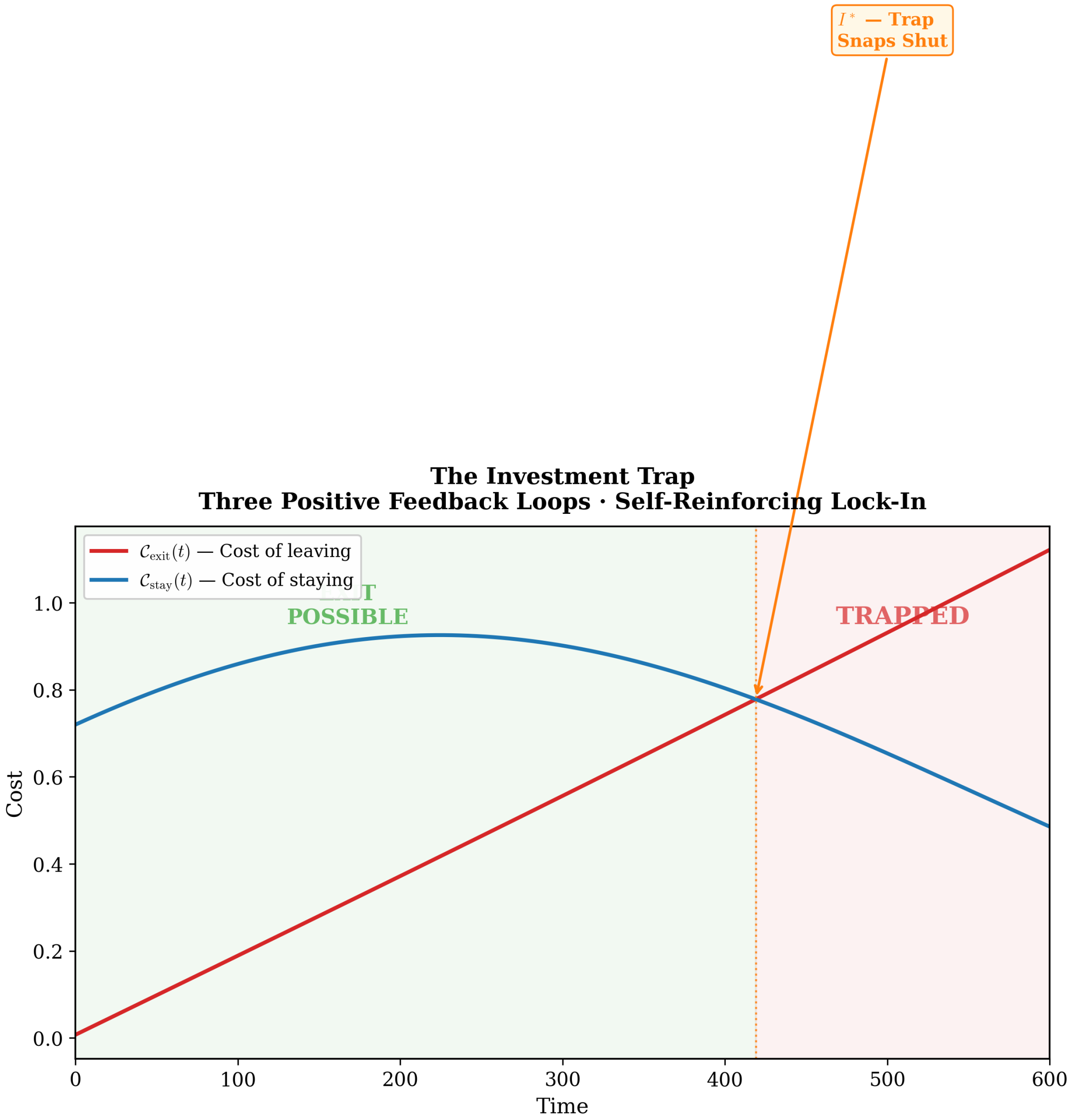


Figure 4: The investment trap: cost of exit $\mathcal{C}_{\text{exit}}(t)$ (red, increasing) versus cost of staying $\mathcal{C}_{\text{stay}}(t)$ (blue, decreasing). The critical threshold I^* (dashed vertical line) marks the point where the trap locks: for $t > t_{\text{trap}}$, exit is never locally optimal and the gap $\mathcal{C}_{\text{exit}} - \mathcal{C}_{\text{stay}}$ grows without bound (Theorem 6.1).

7 Network Extension: Multiple Supply Sources

Definition 7.1 (Quasi-Narcissistic Supply Network)**.** *A quasi-narcissistic supply network is a star graph $G = (V_G, E_G)$ with central hub $\mathcal{N}$ and peripheral nodes $\{\mathcal{W}_1, \ldots, \mathcal{W}_k\}$. The defining constraint is* informational isolation*:*

$$I(\mathcal{W}_i; \mathcal{W}_j) = 0 \quad \textit{for all } i \neq j, \tag{22}$$

where $I(\cdot;\cdot)$ denotes mutual information.

Total supply is $\mathsf{v}_{\text{total}}(t) = \sum_{i=1}^{k} \mathsf{v}_i(t)$, stabilized against individual depletion. The discard operation on $\mathcal{W}_i$ is a resource reallocation:

$$\mathcal{N} : \mathcal{W}_i \mapsto \mathcal{W}_j \quad \text{when} \quad \frac{\partial U_{\mathcal{N}}}{\partial \mathsf{v}_i} < \frac{\partial U_{\mathcal{N}}}{\partial \mathsf{v}_j}, \tag{23}$$

with no emotional transition cost due to the resource-model representation (5).

Proposition 7.1 (Equilibrium Destabilization under Isolation Breaking)**.** *Consider the game $\Gamma_i = (\mathcal{N}, \mathcal{W}_i)$ of Section 3 played between $\mathcal{N}$ and a specific node $\mathcal{W}_i$, with the Nash equilibrium (R, A) sustained by the condition $\xi I_{i,t} > \delta \mathsf{a}_0 + \epsilon \mathsf{s}_0$. Suppose at time t_0, the isolation condition is broken: $\mathcal{W}_i$ acquires information about $\mathcal{W}_j$'s existence and relational state. Then:*

(i) The perceived relational coherence s_i drops discontinuously: $\mathsf{s}_i(t_0^+) = \mathsf{s}_i(t_0^-) - \Delta\mathsf{s}$, where $\Delta\mathsf{s} > 0$ reflects the narrative collapse from discovering the parallel relationship;

(ii) The effective exit cost is reduced: $\mathcal{C}_{exit,i}(t_0^+) = \xi I_{i,t_0} + \psi M_{dep,i}(t_0) - \omega \Delta\mathsf{s}$, where $\omega > 0$ captures the devaluation of the sunk investment narrative;

(iii) If $\Delta\mathsf{s}$ is sufficiently large—specifically, if $\omega\Delta\mathsf{s} > \xi I_{i,t_0} + \psi M_{dep,i}(t_0) - \delta\mathsf{a}_0/2 - \epsilon\mathsf{s}_0/4 + \zeta\mathsf{r}_0$—then exit dominates accommodation, and the Nash equilibrium (R, A) in Γ_i is destroyed.

Proof. (i) is a modeling assumption reflecting the empirical observation that discovery of infidelity or parallel relationships produces a discrete shock to the victim's relational narrative. (ii) follows because the cognitive basis for the sunk cost $(I_{i,t})$—the narrative "this relationship is uniquely meaningful"—is undermined by the evidence that $\mathcal{N}$ maintains identical relationships with others. (iii) follows from the best-response condition for $\mathcal{W}_i$ in the proof of Theorem 3.1: A dominates X iff $\pi_{\mathcal{W}}(R, A) > \pi_{\mathcal{W}}(R, X)$, i.e., $\delta\mathsf{a}_0/2 + \epsilon\mathsf{s}_0/4 - \zeta\mathsf{r}_0 > -\mathcal{C}_{\text{exit},i}$. After the shock, the revised condition is $\delta\mathsf{a}_0/2 + \epsilon(\mathsf{s}_0/4 - \Delta\mathsf{s}) - \zeta\mathsf{r}_0 > -\xi I_{i,t_0} - \psi M_{\text{dep},i} + \omega\Delta\mathsf{s}$. For $\Delta\mathsf{s}$ exceeding the stated threshold, this inequality is violated and X becomes the best response. □

Remark 7.1. *Proposition 7.1 provides the formal basis for the "network illumination" intervention strategy in Section 14: breaking the informational isolation among supply nodes is a structurally targeted attack on the conditions sustaining the Nash equilibrium in each dyad.*

8 Projection as Coherence Maintenance

We formalize projection as an operation on a defined signal space. This section presents a *conceptual model* rather than a theorem; we make this status explicit.

Let $\mathcal{X} = \mathbb{R}^n$ be the signal space of evaluative information about $\mathcal{N}$, with the standard inner product $\langle \cdot, \cdot \rangle$. The self-model $\mathcal{S}_\mathcal{N} \in \mathcal{X}$ is a fixed unit vector representing the idealized self-concept ("I am ideal"). For any incoming evaluative signal $\sigma \in \mathcal{X}$, decompose:

$$\sigma = \underbrace{\langle \sigma, \mathcal{S}_\mathcal{N} \rangle \, \mathcal{S}_\mathcal{N}}_{\sigma_\parallel \text{ (self-consistent)}} + \underbrace{(\sigma - \langle \sigma, \mathcal{S}_\mathcal{N} \rangle \, \mathcal{S}_\mathcal{N})}_{\sigma_\perp \text{ (self-contradictory)}}. \tag{24}$$

The projection operator $\Pi : \mathcal{X} \to \mathcal{X}$ retains $\sigma_\parallel$ and externalizes $\sigma_\perp$:

$$\Pi(\sigma) = \sigma_\parallel, \qquad \sigma_\perp \mapsto \text{attributed to external agents.} \tag{25}$$

This is the orthogonal projection onto $\operatorname{span}(\mathcal{S}_\mathcal{N})$. Under the hard constraint $\mathcal{S}_\mathcal{N} = \text{const}$, the operation Π is necessary for internal coherence whenever contradictory signals ($\sigma_\perp \neq 0$) are present; without it, the agent faces a model integrity violation (self-model integrity violation). The mechanism is formally analogous to error correction in a communication system that must protect a fixed codeword against channel noise.

9 Social Dynamics: Replicator Equations

We formalize the population-level analysis using evolutionary game theory with explicit replicator dynamics, rather than an informal analogy to the Prisoner's Dilemma.

9.1 Population Game

Consider a large population where agents are randomly matched in pairs. Each agent plays one of three strategies:

$$\begin{aligned} \mathcal{C} &: \text{cooperate (empathic engagement)}, \\ \mathcal{D}^* &: \text{narcissistic strategy (cooperate initially, then exploit)}, \\ \mathcal{D} &: \text{overt defection (immediately exploitative)}. \end{aligned}$$

The symmetric payoff matrix for population games is:

$$\mathbf{A} = \begin{array}{c} \\ \mathcal{C} \\ \mathcal{D}^* \\ \mathcal{D} \end{array} \begin{array}{c} \begin{array}{ccc} \mathcal{C} & \mathcal{D}^* & \mathcal{D} \end{array} \\ \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & j \end{pmatrix} \end{array}, \tag{26}$$

with parameter ordering: $d > a > g$ (quasi-narcissistic strategy extracts more from cooperators than cooperation does, while overt defection gets less because cooperators can detect and avoid it); $a > e > j$ (mutual cooperation beats mutual quasi-narcissistic exploitation beats mutual defection); $b < c$ (cooperators lose more to quasi-narcissistic exploiters than to overt defectors, due to the deception involved); $f > j$ (quasi-narcissistic exploiters outperform defectors against defectors due to flexibility).

9.2 Replicator Dynamics

Let $p = (p_{\mathcal{C}}, p_{\mathcal{D}^*}, p_{\mathcal{D}})$ with $p_i \geq 0$, $\sum p_i = 1$. The replicator equations are:

$$\dot{p}_i = p_i \left[(\mathbf{A}p)_i - p^\top \mathbf{A} p \right], \qquad i \in \{\mathcal{C}, \mathcal{D}^*, \mathcal{D}\}. \tag{27}$$

Proposition 9.1 (Invasion Condition for Narcissistic Strategy). *A small population of $\mathcal{D}^*$ players can invade a population of $\mathcal{C}$ players if and only if:*

$$d > a, \tag{28}$$

i.e., the quasi-narcissistic strategy yields a higher payoff against cooperators than cooperation does. This condition is satisfied by construction.

Proof. From the replicator equation, $\dot{p}_{\mathcal{D}^*} > 0$ when $(\mathbf{A}p)_{\mathcal{D}^*} > p^\top \mathbf{A}p$. Near the vertex $p_{\mathcal{C}} \approx 1$, $(\mathbf{A}p)_{\mathcal{D}^*} \approx d$ and $p^\top \mathbf{A}p \approx a$. So invasion occurs iff $d > a$. □

Proposition 9.2 (Trust Erosion and Population Welfare). *Let $\bar{W}(p_{\mathcal{D}^*}) = p^\top \mathbf{A}p$ be average population fitness. Then:*

$$\frac{\partial \bar{W}}{\partial p_{\mathcal{D}^*}} < 0 \quad \textit{for } p_{\mathcal{D}^*} > p_{crit}, \tag{29}$$

where $p_{crit} = (a - e)/(2(d - e) - (a - e))$ when this expression is positive.

This captures the paradox: quasi-narcissistic strategies are locally advantageous but globally corrosive, eroding the cooperative infrastructure on which collective welfare depends.

10 The Narcissism Continuum

Definition 10.1 (Self-Maintenance Parameter Space). *Every agent is characterized by a parameter vector $\theta = (\theta_{\mathsf{v}}, \theta_{\mathsf{e}}, \theta_{\mathsf{c}}, \theta_{\mathsf{a}}) \in \Theta = \mathbb{R}^4_{\geq 0}$, governing sensitivity to validation, empathic capacity, control-seeking, and empathic attachment, respectively. The utility function is:*

$$U(\theta) = \theta_{\mathsf{v}}\mathsf{v} - \theta_{\mathsf{e}}\mathsf{e} + \theta_{\mathsf{c}}\mathsf{c} + \theta_{\mathsf{a}}\mathsf{a}. \tag{30}$$

The narcissistic agent corresponds to $\theta_{\mathcal{N}} \in \Theta_{\mathcal{N}} = \{\theta : \theta_{\mathsf{v}}, \theta_{\mathsf{c}} \gg \theta_{\mathsf{e}}, \theta_{\mathsf{a}}\}$. The empathic agent occupies $\Theta_{\text{emp}} = \{\theta : \theta_{\mathsf{e}}, \theta_{\mathsf{a}} \gg \theta_{\mathsf{v}}, \theta_{\mathsf{c}}\}$. The critical ratio $\rho(\theta) = (\theta_{\mathsf{e}} + \theta_{\mathsf{a}})/(\theta_{\mathsf{v}} + \theta_{\mathsf{c}})$ indexes position on the continuum: quasi-narcissistic dynamics emerge when $\rho(\theta) < \rho^*$ for a critical threshold ρ^* determined by the game structure. This is a continuous parametric characterization, not a categorical diagnosis.

11 Coupled Oscillators: Symmetric Quasi-Narcissistic Interaction

The preceding sections analyze the asymmetric case: one agent in $\Theta_{\mathcal{N}}$, one in Θ_{emp}. We now consider the symmetric case where both agents $\mathcal{A}$ and $\mathcal{B}$ occupy $\Theta_{\mathcal{N}}$, but extract different resources from each other. This configuration—empirically common and clinically recognized—produces qualitatively different dynamics.

11.1 Setup: Two Coupled Extractive Oscillators

Let $\mathcal{A}$ maximize control and status validation; $\mathcal{B}$ maximize emotional attention and indispensability. Each is a "resource" for the other along the dimension the other values but does not prioritize. Agent $\mathcal{A}$ produces emotional drama (which $\mathcal{B}$ reads as confirmation of significance) and receives admiration. Agent $\mathcal{B}$ produces displays of dependence (which $\mathcal{A}$ reads as control) and receives attention.

The system is governed by four coupled ODEs:

$$\frac{dE_{\mathcal{A}}}{dt} = f_{\mathcal{A}}(E_{\mathcal{A}}, K_{\mathcal{B}}), \tag{31}$$

$$\frac{dK_{\mathcal{A}}}{dt} = g(E_{\mathcal{A}}) - \nu K_{\mathcal{A}}, \tag{32}$$

$$\frac{dE_{\mathcal{B}}}{dt} = f_{\mathcal{B}}(E_{\mathcal{B}}, K_{\mathcal{A}}), \tag{33}$$

$$\frac{dK_{\mathcal{B}}}{dt} = g(E_{\mathcal{B}}) - \nu K_{\mathcal{B}}, \tag{34}$$

where $f_{\mathcal{A}}$ and $f_{\mathcal{B}}$ have the same piecewise structure as Eqs. (10)–(11), but the mode of each oscillator depends on the *other* agent's switching state. Each agent switches between idealization and devaluation of the other according to its own hysteretic thresholds $(K_l^{\mathcal{A}}, K_h^{\mathcal{A}})$ and $(K_l^{\mathcal{B}}, K_h^{\mathcal{B}})$.

This is a system of two coupled piecewise-smooth oscillators with hysteresis—a class known to exhibit rich bifurcation structure [?].

11.2 Dynamical Regimes

The coupled system (31)–(34) admits three qualitatively distinct regimes, depending on the coupling strength and the frequency ratio of the two oscillators.

Proposition 11.1 (Coupled Oscillator Regimes)**.** *Let $\omega_{\mathcal{A}}$ and $\omega_{\mathcal{B}}$ denote the natural frequencies of the two oscillators (determined by their respective parameters via Eq. 12), and let $\kappa > 0$ denote the coupling strength. Then:*

(a) ***In-phase synchronization*** *($\omega_{\mathcal{A}} \approx \omega_{\mathcal{B}}$, κ moderate): Both oscillators lock in phase—simultaneous idealization, simultaneous devaluation. Externally observed as intense, high-amplitude "passionate" cycles: euphoric highs alternating with violent conflict. Period is shorter than either uncoupled oscillator because mutual energy supply accelerates recovery.*

(b) ***Anti-phase synchronization*** *($\omega_{\mathcal{A}} \approx \omega_{\mathcal{B}}$, κ large): When one idealizes, the other devalues, and they alternate. Externally observed as chronic conflict without "honeymoon" phases—perpetual turbulence at reduced amplitude.*

(c) ***Quasi-periodic or chaotic dynamics*** *($\omega_{\mathcal{A}}/\omega_{\mathcal{B}} \notin \mathbb{Q}$, or κ in an intermediate range): Two coupled nonlinear oscillators with hysteresis generically satisfy the conditions for strange attractor formation. Externally observed as relationships with no recognizable cycle—unpredictable switching, seemingly random escalations, and inability of either participant or external observers to identify a pattern.*

Proof sketch. Results (a) and (b) follow from standard theory of coupled oscillators [Strogatz, 2015]: for identical or near-identical frequencies, weak coupling produces in-phase locking while strong coupling can produce anti-phase locking, depending on the sign of the coupling function. Result (c) follows from the general theory of coupled piecewise-smooth systems: two interacting hysteretic oscillators in $\mathbb{R}^4$ are not constrained by the Poincaré–Bendixson theorem (which applies only in $\mathbb{R}^2$), and the combination of non-linearity, hysteresis, and incommensurate frequencies is sufficient for chaotic dynamics [?]. □

11.3 The (R, R) Equilibrium

In the asymmetric model, the Nash equilibrium is (R, A)—intermittent reinforcement by $\mathcal{N}$, accommodation by $\mathcal{W}$. In the symmetric case, a new equilibrium type arises:

Definition 11.1 (Mutual Intermittent Reinforcement Equilibrium)**.** *The* (R, R) equilibrium *is the strategy profile where both agents play intermittent reinforcement against each other. Each agent alternates idealization and devaluation of the other according to its own switching thresholds.*

Proposition 11.2 (Stability of (R, R))**.** *The* (R, R) *equilibrium is stable under the following conditions, each of which is independently sufficient:*

(i) ***Utility symmetry****: Both agents optimize over the same variable types (control, validation), so neither is deceived about the nature of the interaction. Cognitive dissonance is absent; there is no internal pressure toward exit.*

(ii) ***Mutual deterrence:*** *Escalation by one agent triggers immediate retaliatory escalation by the other. This bounds the oscillation amplitude from above, preventing the resource depletion ($E \to 0$) that triggers discard in the asymmetric case.*

(iii) ***Projection immunity:*** *Both agents maintain a fixed self-model $\mathcal{S}_{\mathcal{N}}$ (Section 8) and project contradictory signals outward. Neither updates by Bayesian rule; hence the precision erosion of Theorem 5.1 does not operate. Both retain full epistemic autonomy—not because they perceive reality accurately, but because their sensor is connected to $\mathcal{S}_{\mathcal{N}}$, which is constant by construction.*

(iv) ***Self-regenerating resource:*** *Unlike the asymmetric case where $E_{\mathcal{W}}(t) \to 0$ under sustained devaluation, each agent's energy is sustained by the internal source $\mathcal{S}_{\mathcal{N}}$, which is independent of external validation. Neither agent is depleted to the discard threshold.*

11.4 Structural Differences from the Asymmetric Case

The (R, R) system differs from the (R, A) system in five critical respects:

Mutual gaslighting fails. Both agents attempt epistemic degradation of the other. Both project. Neither accepts the other's contradictory signal. The result is not degradation but *escalation*: each increases signal intensity to penetrate the other's projection defense. This predicts that symmetric quasi-narcissistic pairs present as escalating conflict rather than quiet submission—a prediction consistent with clinical observation.

Strategic (not emotional) sunk costs. In the asymmetric model, $\mathcal{W}$ accumulates emotional investment $I(t)$ driven by attachment (θ_{a}). In the symmetric model, both agents invest strategically: each has spent time "calibrating" the other as a resource. The exit barrier is not "I have given so much" but "I have spent so much effort training this person to supply what I need." Formally, $I(t)$ exists but is driven by θ_{c} (control) rather than θ_{a} (attachment).

Informational isolation is impossible. Each agent attempts to isolate the other; neither permits it. Both maintain external supply sources and are aware the other does likewise. The star topology of Section 7 cannot form. Instead, the information structure is a *contested network*: each agent attempts to control the other's external connections while maintaining their own.

Exit is paradoxically easier. Because neither agent is emotionally attached (both have low θ_{a}), the classical trap (Theorem 6.1) is weak. When one agent discovers a more favorable resource, they switch without a transitional period—a discontinuous jump in the bimatrix game when an external parameter (new supply source) shifts the best response. The model predicts that symmetric quasi-narcissistic pairs dissolve more suddenly and with less post-separation distress than asymmetric pairs.

Collateral damage. The (R, R) system is stable precisely because neither participant is destroyed. However, third parties without projection defenses—particularly children—are exposed to dual-source epistemic degradation. A child in this system has no fixed $\mathcal{S}_{\mathcal{N}}$; their τ_{int} degrades under Theorem 5.1 from *both* parents simultaneously, and they become a projection surface for both agents' externalized contradictions.

11.5 Thermostat Analogy

The (R, R) configuration corresponds to two thermostats controlling each other's rooms: each is simultaneously heater and room for the other. Energy circulates rather than being extracted unidirectionally. Neither room freezes (no $E \to 0$); both oscillate. The system is operationally stable. It is also uninhabitable by anyone other than the two thermostats.

12 Robustness Analysis

We now verify that the main results are not artifacts of the linear utility specialization.

Proposition 12.1 (Structural Robustness). *The following results depend only on the sign structure (Axiom 2.1) and monotonicity, not on the specific functional forms h_i:*

(i) ***Nash equilibrium existence*** *(Theorem 3.1): requires only that $\mathcal{N}$'s payoff is maximized at R for column A and $\mathcal{W}$'s payoff is maximized at A for row R. These ordinal comparisons are preserved under any monotone transformation of the utility components.*

(ii) ***Limit cycle existence*** *(Theorem 4.1): requires the switching thresholds $K_l < K_h$ and the trapping region conditions (H1)–(H3), all of which depend on inequality relationships between parameters, not on functional form.*

(iii) ***Epistemic phase transition*** *(Theorem 5.1): the precision erosion mechanism is independent of utility function specification, relying only on the Bayesian updating structure.*

(iv) ***Investment trap*** *(Theorem 6.1): requires $\mathcal{C}_{exit}$ to be increasing and unbounded while $\mathcal{C}_{stay}$ is eventually decreasing, which holds for any concave h_i.*

Proof. (i) The payoff ordering in Table 1 depends on the signs and relative magnitudes of the terms, not their functional forms. For general h_i with $h_i' > 0$, replacing $\alpha \mathsf{v}_0$ by $\alpha h_1(\mathsf{v}_0)$ etc. preserves all inequalities in the proof of Theorem 3.1 provided $\alpha h_1(\mathsf{v}_0) \gg \beta h_2(\mathsf{e}_0)$. (ii)–(iv) are verified analogously; the key property is that the *ordinal rankings* of payoffs and costs are invariant under concave monotone transformations. □

13 Falsifiability and Novel Predictions

The model generates three testable predictions not derivable from clinical description alone.

Prediction 13.1 (Critical Investment Threshold)**.** *There exists a quantitative threshold $I^* = (\delta \mathsf{a}_0 + \epsilon \mathsf{s}_0)/\xi$ such that the probability of victim exit drops discontinuously as accumulated investment crosses I^*.* ***Test:*** *In a longitudinal study of abusive relationships, measure proxies for relational investment (cohabitation duration, financial entanglement, shared children, identity merger) and exit behavior. The model predicts a sharp sigmoid in exit probability as a function of a composite investment index, with the inflection point identifiable as I^*.* ***Falsification:*** *If exit probability decreases* linearly *with investment, or if the relationship shows no threshold behavior, the model's game-theoretic mechanism is disconfirmed.*

Prediction 13.2 (Minimum Oscillation Period)**.** *The abuse cycle period T has a computable lower bound given by Eq. (12), which depends on the emotional recovery rate $(\lambda_+ + \mu)^{-1}$, the depletion rate λ_-^{-1}, and the switching thresholds.* ***Test:*** *Track idealization–devaluation cycles in longitudinal data (mood diaries, communication pattern analysis). The model predicts that cycle periods cluster above a minimum determined by measurable interaction parameters, and that faster recovery rates (e.g., due to external social support reducing μ) shorten the cycle.* ***Falsification:*** *If observed cycle periods are uniformly distributed without a lower bound, or if recovery rate has no effect on period, the oscillator mechanism is disconfirmed.*

Prediction 13.3 (Epistemic Autonomy Phase Transition)**.** *The victim's confidence in their own judgment (measurable via self-report instruments) undergoes a phase transition—not a gradual decline—as a function of cumulative gaslighting episodes. The model predicts a* tipping point *at approximately t^* episodes (Eq. 17), after which self-confidence collapses rapidly.*

Test: *Administer repeated self-confidence and reality-testing measures to individuals in relationships with high gaslighting frequency. The model predicts bimodal distributions (pre- and post-transition) rather than unimodal gradual decline.*

Falsification: *If self-confidence measures show smooth, gradual decline without phase-transition signatures (no bimodality, no acceleration), the Bayesian erosion mechanism is disconfirmed.*

Remark 13.1 (Falsifiability of the Overall Framework). *Beyond individual predictions, the framework as a whole would be falsified if: (a) quasi-narcissistic relational dynamics are empirically shown to be non-oscillatory (no recurring cycles); (b) victim entrenchment is independent of relationship duration; or (c) gaslighting effects are independent of the victim's prior confidence level. Any of these would require fundamental revision of the model structure.*

14 Implications for Intervention Design

The formal framework yields structurally targeted intervention strategies:

(i) **Precision restoration** (targeting Theorem 5.1): Therapeutic interventions that rebuild τ_{int}—the victim's internal signal precision—through grounding techniques and reality testing. The model predicts that interventions are most effective when τ_{int} is still above τ_{ext} (pre-transition).

(ii) **Investment reframing** (targeting Theorem 6.1): Cognitive interventions that decouple perceived exit cost from accumulated investment, addressing the sunk-cost mechanism.

(iii) **Cycle recognition** (targeting Theorem 4.1): Psychoeducation making the oscillatory structure explicit, enabling prediction of the next phase.

(iv) **Network illumination** (targeting Eq. 22): Breaking informational isolation among victims, collapsing the star topology.

(v) **Exit threshold reduction**: Practical support reducing $\mathcal{C}_{\text{exit}}(t)$ directly.

(vi) **Detection infrastructure** (targeting Proposition 9.1): Institutional design increasing the cost of $\mathcal{D}^*$ strategies, shifting population dynamics away from quasi-narcissistic invasion.

15 Discussion and Conclusion

We have presented a rigorous mathematical framework for quasi-narcissistic relational dynamics. The key contributions beyond the clinical literature are: (i) axiomatic derivation of agent utility functions with robustness verification; (ii) a fully specified bimatrix game with proved Nash equilibrium; (iii) a rigorous dynamical systems analysis with explicit trapping region construction; (iv) a Bayesian persuasion model of gaslighting with a proved phase transition; and (v) three novel, empirically testable predictions with explicit falsification criteria.

However, as Appendices B–I demonstrate, these results describe not a single relational pathology but a general dynamical systems class. We now articulate the structural reasons why this generality is not accidental but mathematically necessary.

15.1 The Inevitability of Oscillation

Consider any system in which one agent optimizes extraction of a renewable but finite-capacity resource from another. The extracting agent faces a fundamental dilemma that admits no stationary solution:

(a) **Permanent depletion** exhausts the resource. The victim exits, the population revolts, the employee resigns, the user disengages. Extraction falls to zero.

(b) **Permanent supply** (continuous idealization) pays full cost (empathic expenditure, political concessions, algorithmic personalization) while allowing the resource agent to accumulate autonomy. The resource agent may then exit voluntarily.

The unique optimal policy is *alternation*: supply until the resource reaches a replenishment threshold, then switch to extraction; extract until a depletion threshold, then switch back to supply. This is a classical result in optimal control theory: under two modes with opposing effects and threshold switching, the optimal policy is *bang-bang control* [Strogatz, 2015]. The system cannot converge to equilibrium. It must oscillate. This is not a strategic choice, nor a personality trait, nor a conscious intention—it is a theorem.

Formally, this is the content of Theorem 4.1: under conditions (H1)–(H3), the coupled system admits a stable limit cycle. The conditions (H1)–(H3) are not restrictive; they merely require that supply can replenish the resource (H1), that both switching thresholds are interior (H2), and that the system has sufficient dynamic range (H3). Any extraction system satisfying these minimal requirements oscillates.

15.2 The Necessity of Sensor Degradation

The oscillator, taken alone, is unstable. If the extracted agent accurately perceives the cyclic structure, it can time its exit to the devaluation phase—recognizing the pattern and escaping during the low-control period. Therefore:

An extracting agent whose strategy does not include sensor degradation loses the resource agent within a few cycles. An extracting agent whose strategy includes sensor degradation retains the resource agent indefinitely.

Under the replicator dynamics of Section 9, strategies that include sensor degradation outcompete those that do not, because they sustain longer extraction periods. This is not a conscious design—it is evolutionary selection of strategies. Sensor degradation (the precision erosion of Theorem 5.1) is not an incidental feature of quasi-narcissistic behavior; it is a *necessary condition for the stability of any extractive oscillatory strategy.* Without it, the oscillator decomposes after a small number of cycles.

15.3 The Three Coupled Feedback Loops

The entrenchment mechanism (Theorem 6.1) derives its power from the simultaneous operation of three positive feedback loops:

Loop 1: Investment → rationalization → further investment. The more that has been invested, the costlier it is to acknowledge the investment as error. Rationalization (increasing $r(t)$ in Eq. 21) resolves the dissonance, enabling continued investment. The investment $I(t)$ is monotonically increasing and unbounded (Theorem 6.1(i)).

Loop 2: Precision erosion → signal acceptance → further erosion. Each acceptance of the exploiter's contradictory signal reduces τ_{int}. Lower τ_{int} raises the acceptance probability in the subsequent period. This is an autocatalytic process—a chain reaction in the precision variable. The concentration inequality in the proof of Theorem 5.1 shows that this process reaches the critical point with probability approaching 1 exponentially in time.

Loop 3: Isolation → absence of corrective information → deeper dependence → further isolation. Without external information sources (the $I(\mathcal{W}_i; \mathcal{W}_j) = 0$ condition of Eq. 22), no corrective signal can counteract the precision erosion. Dependence on the single information source grows, which makes the isolation more effective, which deepens the dependence.

The three loops operate simultaneously and amplify one another. This is why Prediction 13.1 identifies a *threshold* effect: below the critical investment I^*, exit remains accessible and the loops are subcritical; above I^*, the three loops interlock and the trap becomes effectively irreversible. The transition is sharp, not gradual.

15.4 The Deepest Structural Pattern

At the highest level of abstraction, the regularity is:

> *In any predator–resource pair where the resource is renewable but fragile, and the predator can modulate extraction intensity, the same four-element structure emerges: (i) oscillation, because stationary extraction is impossible; (ii) sensor degradation, because a resource that perceives its exploitation ceases to be a resource; (iii) exit barrier growth, because accumulated loss creates asymmetric valuation (loss aversion); and (iv) informational isolation, because coordination among resource agents is lethal to the predator.*

This is the content of Definition I.1 (Appendix I). The structure is not an analogy. The governing equations of a thermostat (Appendix A), a quasi-narcissistic relationship (Sections 4–6), an authoritarian regime (Appendix B), and an algorithmic attention platform (Appendix F) are literally the same equations with different variable names. Only the substrate differs: temperature, emotional energy, civic capacity, user engagement.

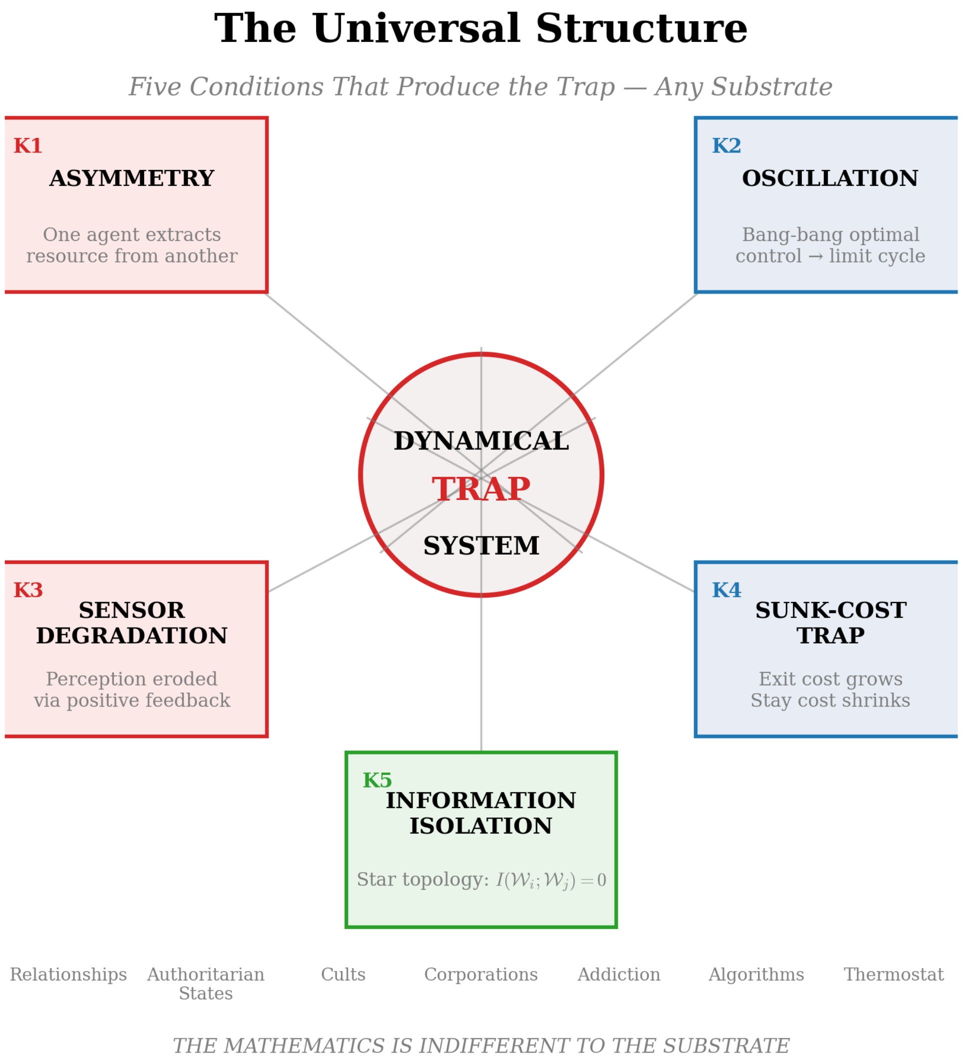

The Universal Structure
Five Conditions That Produce the Trap — Any Substrate
K1
ASYMMETRY
One agent extracts resource from another
K2
OSCILLATION
Bang-bang optimal control → limit cycle
DYNAMICAL
TRAP
SYSTEM
K3
SENSOR DEGRADATION
Perception eroded via positive feedback
K4
SUNK-COST TRAP
Exit cost grows
Stay cost shrinks
K5
INFORMATION ISOLATION
Star topology: $I(\mathcal{W}_i;\mathcal{W}_j)=0$
Relationships
Authoritarian States
Cults
Corporations
Addiction
Algorithms
Thermostat
THE MATHEMATICS IS INDIFFERENT TO THE SUBSTRATE

Figure 5: The five conditions (K1)–(K5) that define a Kriger system (Definition I.1): asymmetric extraction, hysteretic oscillation, sensor degradation, sunk-cost trapping, and informational isolation. Any system satisfying all five—regardless of substrate—is governed by the theorems of this paper. The mathematics is indifferent to the substrate.

15.5 The Irrelevance of Intent

Perhaps the most consequential implication of the framework is that *conscious malice is not required*. A thermostat does not "intend" to subject a room to temperature oscillations. It optimizes. Similarly, the model shows that an agent with the parameter configuration $(\theta_\mathsf{v}, \theta_\mathsf{e}, \theta_\mathsf{c}, \theta_\mathsf{a})$ in the quasi-narcissistic regime (high validation need, low empathic capacity, high control-seeking, low empathic attachment) will produce the full cycle of idealization, devaluation, gaslighting, and entrenchment as an *emergent consequence of optimization under those parameters*—without requiring a plan, a strategy, or even awareness.

This observation is simultaneously alarming and liberating. Alarming, because it implies that the system can arise spontaneously whenever the parameter configuration is present, without requiring malicious agency. Liberating, because it reframes the problem from "confrontation with evil" to an engineering problem: *identify and interrupt the feedback loops.*

15.6 Structural Intervention Principles

The formal structure dictates exactly where the system is vulnerable:

(i) **Restore $\tau_{\mathbf{int}}$** (break Loop 2): therapeutic grounding, reality testing, external validation of internal perception—all interventions that rebuild the victim's internal signal precision before the phase transition at t^*.

(ii) **Break isolation** (break Loop 3): enable mutual information flow among resource agents ($I(\mathcal{W}_i; \mathcal{W}_j) > 0$), triggering the equilibrium destabilization of Proposition 7.1.

(iii) **Reduce exit cost** (break Loop 1): practical support that decouples perceived exit cost from accumulated investment, disrupting the sunk-cost mechanism of Theorem 6.1.

(iv) **Cycle recognition**: psychoeducation that makes the oscillatory structure (Theorem 4.1) explicit, enabling the resource agent to predict the next phase and act during the window of maximal autonomy.

(v) **Increase detection cost of $\mathcal{D}^*$**: institutional design that raises the reputational and material cost of the quasi-narcissistic strategy in population games (Section 9), shifting the replicator dynamics against invasion.

These are not recommendations derived from clinical intuition. They are structural consequences of the formal model: each intervention targets a specific theorem, equation, or feedback loop. The framework converts a psychological problem into a control-theoretic one: interrupting feedback loops in a nonlinear system with known structure.

15.7 From Diagnosis to Dynamics: A Shift in Explanatory Level

The preceding analysis implies a consequence that extends beyond mathematical modeling into the epistemology of psychological explanation. We make it explicit.

Prior to the present framework, narcissistic abuse is understood through an agent-centered explanatory model. The explanatory primitive is a *diagnosis*: "this person is a narcissist." The diagnosis functions as a cause—the agent has a defective property, and the defective property produces harm. The entire clinical and popular discourse then organizes around the question of what is wrong with this particular category of agents.

The framework developed in this paper shows that this explanatory structure is unnecessary. The "narcissistic agent" is not a cause but a *parameter configuration*: a point $(\theta_{\mathsf{v}}, \theta_{\mathsf{e}}, \theta_{\mathsf{c}}, \theta_{\mathsf{a}})$ in a continuous parametric space (Section 10), not a discrete diagnostic category. The oscillation, the epistemic degradation, the investment trap, and the network isolation do not arise because the agent is "disordered." They arise because, for parameter values in the narcissistic regime, the coupled dynamical system automatically produces these phenomena. A thermostat with the same mathematical structure produces the same oscillatory behavior (Appendix A). The thermostat is not disordered. It is configured.

This constitutes a shift in the level of explanation, analogous to the Copernican resolution of retrograde planetary motion. Before Copernicus, the apparent backward motion of Mars across the sky required explanation in terms of properties of Mars itself—epicycles, special mechanisms, or (in pre-scientific frameworks) the "character" of the planet. After Copernicus, retrograde motion was revealed as an inevitable geometric consequence of two orbits with different periods. Mars is not anomalous. The orbits are so configured.

The present paper performs an analogous operation on narcissistic abuse. The focus shifts from "who is to blame" to "what structure produces this outcome." The answer—an extractive oscillator with sensor degradation, sunk-cost trapping, and informational isolation—turns out to be substrate-independent (Appendices A–I, Definition I.1). Narcissistic abuse is not a special form of evil. It is a particular instance of a dynamical systems class that manifests wherever the five structural conditions (K1)–(K5) are satisfied.

This reframing might appear to diminish the suffering of those subjected to such systems—to suggest that they were harmed not by a malicious agent but by a parameter configuration. In fact, the opposite is the case. Under the agent-centered model, the intervention is to identify and remove the "bad" agent, then hope that the replacement has different properties. Under the structure-centered model, the intervention targets the feedback architecture directly: restore the sensor (Section 14, strategy (i)); break the isolation (Proposition 7.1); reduce the exit barrier (strategy (iii)); make the oscillatory pattern visible (strategy (iv)). These are engineering operations on a system with known structure, not moral judgments about a person with unknown internal states.

The claim is not that individuals with narcissistic parameter configurations do not exist. The claim is stronger: *the diagnostic category "narcissist" is not necessary for explanation, prediction, or intervention.* Everything required to understand the trap, predict its trajectory, and design its dissolution is contained in five parameters and three feedback loops. The label "narcissist" contributes to this understanding approximately as much as the label "mischievous" contributes to the understanding of retrograde planetary motion.

This is what it means to replace a diagnostic framework with a dynamical one. Not to deny the existence of agents with particular parameter values, but to show that for the purposes of understanding, predicting, and dismantling the trap, the label is redundant. The dynamics suffice.

15.8 The Diagnostic Label as an Instance of the Mechanism It Describes

There is a further consequence that is not merely theoretical. The diagnostic label "narcissist," originally developed to describe manipulative dynamics, has itself become a tool that instantiates those exact dynamics. We can verify this claim by checking the label against all five structural conditions (K1)–(K5) of Definition I.1.

Sensor degradation (K3). When one person tells another "you are a narcissist," the operative message is: *your perception of reality is defective by definition, because you are a person with a perceptual disorder.* This is not an attack on a specific belief. It is an attack on the target's capacity to trust *any* of their own judgments—a single-move approximation of the multi-period precision erosion of Theorem 5.1. The target's τ_{int} is not gradually eroded; it is categorically invalidated.

Informational isolation (K5). The label redefines every past and future communication from the target as manipulation. Any attempt at explanation is "narcissistic justification." Any attempt at self-defense is "narcissistic rage." Any display of empathy is "love-bombing." The target is placed in an informational quarantine where no signal they emit can be received as honest. This is the isolation condition (22) implemented linguistically.

Sunk-cost trapping (K4). For the labeled person, the cost of escaping the accusation is monotonically increasing in the number of attempts: arguing "confirms" the diagnosis; silence is "narcissistic stonewalling"; agreement is "manipulative false contrition." Formally, $\mathcal{C}_{\text{exit}}$ from the accusation is increasing in $I(t)$, where $I(t)$ counts the number of attempted responses—each of which is reinterpreted as further evidence.

Asymmetric extraction (K1). The person who first applies the label occupies the observer-diagnostician position. The other is automatically the object of diagnosis. This is an instantaneous establishment of the asymmetry that normally requires months to construct in relational dynamics.

Hysteretic oscillation (K2). The accuser can alternate between "I see you are trying" (idealization) and "no, you are exhibiting narcissism again" (devaluation) at zero empathic cost—simply by toggling the interpretation of the same behavior. This is a ready-made intermittent reinforcement cycle.

Remark 15.1 (The Paradox of the Label)**.** *The concept "narcissist," invented to describe a mechanism of manipulation, satisfies all five conditions for being an* instrument *of that same mechanism. An individual who applies this label activates all five feedback loops of the Kriger system simultaneously, with complete subjective certainty of moral rectitude—because they are, after all, "the victim."*

This is an additional argument for replacing the diagnostic framework with a dynamical one. It is not possible to weaponize the statement "you are an oscillator with hysteretic switching and sensor degradation" in the same way that one can weaponize the statement "you are a narcissist." The dynamical framing does not permit its own use as a tool of the mechanism it describes. The diagnostic framing does.

Limitations and future directions. The model parameters require empirical calibration. While we have shown that qualitative results are robust to functional form (Section 12), quantitative predictions depend on parameter values obtainable only from longitudinal data. Agent-based simulation of the coupled system under stochastic perturbations could reveal emergent population-level phenomena. The Bayesian persuasion model of gaslighting should be connected to the growing literature on belief manipulation in social networks [Acemoglu et al., 2011]. The cross-domain applicability demonstrated in Appendices B–H should be tested empirically in each domain independently.

Ethical considerations. Reducing human suffering to equations risks dehumanization. We contend that formal understanding *enables* more effective compassion: knowing the mechanism of a trap is the first step toward designing an escape. The model treats no one as a monster—it shows how ordinary optimization under particular parameter configurations produces systemic harm. The central insight is that the phenomena catalogued in this paper—quasi-narcissistic abuse, authoritarian control, cult dynamics, corporate toxicity, addiction, algorithmic manipulation—are not primarily character defects, political ideologies, or pathological categories. They are instances of a single dynamical systems class: nonlinear feedback control systems with hysteretic switching, sensor corruption, positive feedback stabilization, and network informational isolation. Understanding them as such is the first step toward engineering their dissolution.

References

Acemoglu, D., Dahleh, M. A., Lobel, I., & Ozdaglar, A. (2011). Bayesian learning in social networks. *Review of Economic Studies*, 78(4):1201–1236.

Bowlby, J. (1969). *Attachment and Loss, Vol. 1: Attachment.* Basic Books, New York.

Crawford, V. P. & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

di Bernardo, M., Budd, C. J., Champneys, A. R., & Kowalczyk, P. (2008). *Piecewise-smooth Dynamical Systems: Theory and Applications.* Springer, London.

Dutton, D. G. (2007). *Rethinking Domestic Violence.* UBC Press, Vancouver.

Dutton, D. G. & Painter, S. (1993). Emotional attachments in abusive relationships: a test of traumatic bonding theory. *Violence and Victims*, 8(2):105–120.

Ely, J. C. (2017). Beeps. *American Economic Review*, 107(1):31–53.

Ferster, C. B. & Skinner, B. F. (1957). *Schedules of Reinforcement.* Appleton-Century-Crofts, New York.

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford University Press.

Filippov, A. F. (1988). *Differential Equations with Discontinuous Righthand Sides.* Springer, Dordrecht.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, 1(2):148–158.

Gentzkow, M. & Kamenica, E. (2017). Bayesian persuasion with multiple senders and rich signal spaces. *Games and Economic Behavior*, 104:411–429.

Gottman, J. M., Murray, J. D., Swanson, C. C., Tyson, R., & Swanson, K. R. (1995). *The Mathematics of Marriage: Dynamic Nonlinear Models.* MIT Press, Cambridge, MA.

Grasman, J. (1987). *Asymptotic Methods for Relaxation Oscillations and Applications.* Springer, New York.

Herman, J. L. (1992). *Trauma and Recovery.* Basic Books, New York.

Hofbauer, J. & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics.* Cambridge University Press.

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413.

Johnson, M. P. (2008). *A Typology of Domestic Violence.* Northeastern University Press.

Kamenica, E. & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kernberg, O. F. (1975). *Borderline Conditions and Pathological Narcissism.* Jason Aronson, New York.

Kohut, H. (1971). *The Analysis of the Self.* International Universities Press, New York.

Maynard Smith, J. (1982). *Evolution and the Theory of Games.* Cambridge University Press.

Mishchenko, E. F. & Rozov, N. Kh. (1980). *Differential Equations with Small Parameters and Relaxation Oscillations.* Plenum Press, New York.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J., & Friston, K. J. (2014). Bayesian inferences about the self (and others): a review. *Consciousness and Cognition*, 25:67–76.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

Nowak, M. A. (2006). *Evolutionary Dynamics.* Harvard University Press.

Redish, A. D. (2016). *The Mind within the Brain.* Oxford University Press.

Renault, J., Solan, E., & Vieille, N. (2017). Optimal dynamic information provision. *Games and Economic Behavior*, 104:329–349.

Selten, R. (1980). A note on evolutionarily stable strategies in asymmetric animal conflicts. *Journal of Theoretical Biology*, 84(1):93–101.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Skinner, B. F. (1938). *The Behavior of Organisms.* Appleton-Century, New York.

Stark, E. (2007). *Coercive Control: How Men Entrap Women in Personal Life.* Oxford University Press.

Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos*, 2nd ed. Westview Press.

von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton University Press.

A Thermostat Control System as Structural Analog

A.1 Purpose

The purpose of this appendix is to demonstrate that the dynamical structure analyzed in the preceding sections is not specific to quasi-narcissistic relational dynamics but is an instance of a general class of nonlinear feedback control systems with hysteretic switching, sensor degradation, and positive feedback stabilization. We establish this by constructing an explicit structural isomorphism between the relational model and a standard engineering system: a thermostat controlling a heater in a poorly insulated room.

No element of this appendix involves psychological interpretation. The system described is purely physical. The isomorphism is purely mathematical.

A.2 Formal Description of the Thermostat System

Consider a room with internal temperature $\Theta(t) \in [\Theta_{\min}, \Theta_{\max}]$, equipped with:

(a) A **temperature sensor** with measurement output $\hat{\Theta}(t) = \Theta(t) + \xi(t)$, where $\xi(t)$ is sensor noise with variance $\sigma_\xi^2(t)$ that may degrade over time (sensor drift);

(b) A **heater** (actuator) with binary state $u(t) \in \{0, 1\}$ (off/on), delivering heat at rate $Q > 0$ when active;

(c) **Environmental heat loss** at rate $\lambda_{\text{loss}} > 0$ proportional to the temperature differential with the exterior: $-\lambda_{\text{loss}}(\Theta(t) - \Theta_{\text{ext}})$;

(d) A **response delay** $\tau > 0$ between the sensor reading and the actuator response;

(e) A **hysteretic switching rule**: the thermostat activates the heater when $\hat{\Theta} < \Theta_l$ and deactivates when $\hat{\Theta} > \Theta_h$, with $\Theta_l < \Theta_h$ (deadband).

A.3 Governing Equations

The thermal dynamics are governed by:

$$C_{\text{th}} \frac{d\Theta}{dt} = Q \cdot u(t - \tau) - \lambda_{\text{loss}}(\Theta(t) - \Theta_{\text{ext}}), \tag{35}$$

where $C_{\text{th}} > 0$ is the thermal capacity of the room. The switching rule is:

$$u(t) = \begin{cases} 1 & \text{if } \hat{\Theta}(t) < \Theta_l \text{ (heater on)}, \\ 0 & \text{if } \hat{\Theta}(t) > \Theta_h \text{ (heater off)}, \\ u(t^-) & \text{if } \Theta_l \leq \hat{\Theta}(t) \leq \Theta_h \text{ (hysteresis)}. \end{cases} \tag{36}$$

A.4 Oscillatory Regime

For the system (35)–(36), it is a classical result in control theory that when:

(T1) The heating rate exceeds the loss rate at the lower threshold: $Q > \lambda_{\text{loss}}(\Theta_l - \Theta_{\text{ext}})$;

(T2) The loss rate exceeds zero input at the upper threshold: $\lambda_{\text{loss}}(\Theta_h - \Theta_{\text{ext}}) > 0$;

(T3) The thresholds satisfy $\Theta_l < \Theta_h$ (nondegenerate hysteresis),

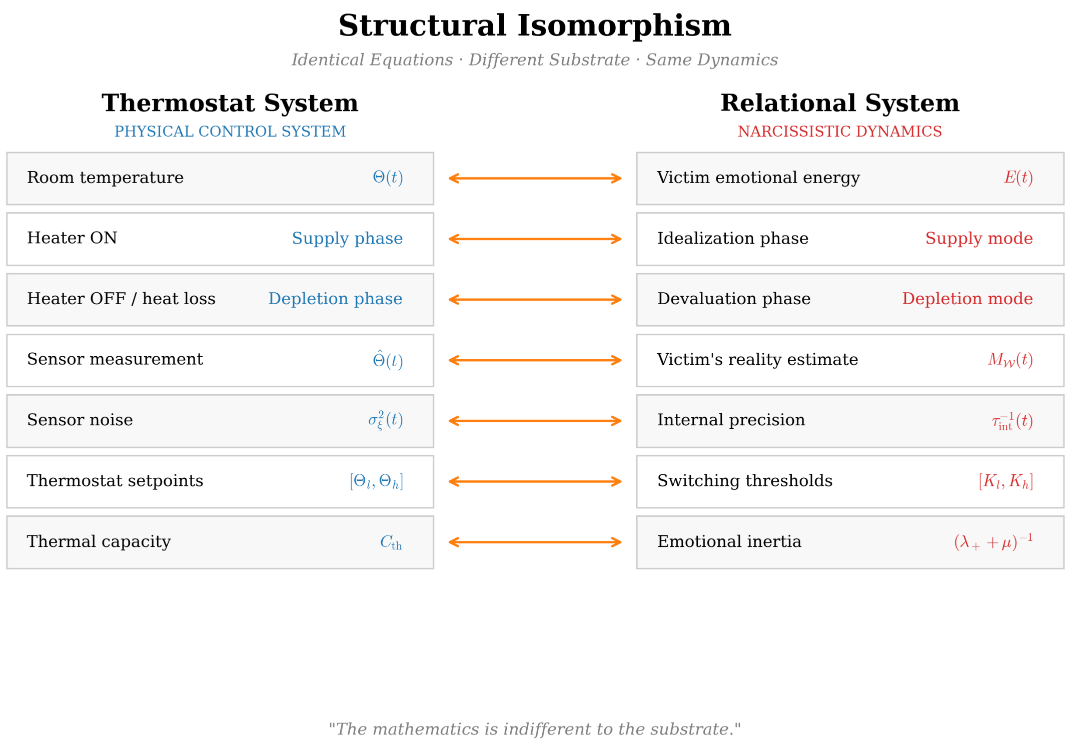
the system does not converge to a fixed point but enters a stable limit cycle. The temperature oscillates between approximate bounds Θ_l and Θ_h with period:

$$T_{\text{therm}} \approx \frac{C_{\text{th}}}{Q - \lambda_{\text{loss}}(\bar{\Theta} - \Theta_{\text{ext}})} \ln\left(\frac{\Theta_h - \Theta_{\text{eq,off}}}{\Theta_l - \Theta_{\text{eq,off}}}\right) + \frac{C_{\text{th}}}{\lambda_{\text{loss}}(\bar{\Theta} - \Theta_{\text{ext}})} \ln\left(\frac{\Theta_h - \Theta_{\text{ext}}}{\Theta_l - \Theta_{\text{ext}}}\right), \quad (37)$$

where $\Theta_{\text{eq,off}} = \Theta_{\text{ext}}$ is the equilibrium with heater off. This is structurally identical to the period formula (12) of the relational oscillator.

A.5 Structural Isomorphism

We now establish a term-by-term correspondence between the thermostat system and the relational model of Sections 4–5. The mapping is summarized in Table 2 and Figure 6, and elaborated below.



Structural Isomorphism

Identical Equations · Different Substrate · Same Dynamics

Thermostat System PHYSICAL CONTROL SYSTEM			Relational System NARCISSISTIC DYNAMICS	
Room temperature	$\Theta(t)$	↔	Victim emotional energy	$E(t)$
Heater ON	Supply phase	↔	Idealization phase	Supply mode
Heater OFF / heat loss	Depletion phase	↔	Devaluation phase	Depletion mode
Sensor measurement	$\hat{\Theta}(t)$	↔	Victim's reality estimate	$M_{\mathcal{W}}(t)$
Sensor noise	$\sigma_\xi^2(t)$	↔	Internal precision	$\tau_{\text{int}}^{-1}(t)$
Thermostat setpoints	$[\Theta_l, \Theta_h]$	↔	Switching thresholds	$[K_l, K_h]$
Thermal capacity	C_{th}	↔	Emotional inertia	$(\lambda_+ + \mu)^{-1}$

"The mathematics is indifferent to the substrate."

Figure 6: Structural isomorphism between the thermostat control system (left) and the relational dynamical system (right). Each row maps a physical component to its mathematical counterpart. The governing equations are identical; only the variable names differ.

Table 2: Structural isomorphism between the thermostat control system and the relational dynamical system. Each row maps a physical component to its mathematical counterpart in the relational model.

Thermostat system	**Relational system**	**Model variable**
Room temperature $\Theta(t)$	Victim emotional energy $E(t)$	State variable
Heater activation ($u = 1$)	Idealization phase	Mode: supply
Heater off / heat loss ($u = 0$)	Devaluation phase	Mode: depletion
Sensor measurement $\hat{\Theta}(t)$	$\mathcal{W}$'s internal reality estimate	$M_{\mathcal{W}}(t)$
Sensor noise variance $\sigma_\xi^2(t)$	Internal precision $\tau_{\text{int}}^{-1}(t)$	Sensor quality
Thermostat setpoints Θ_l, Θ_h	Switching thresholds K_l, K_h	Hysteresis band
Thermal capacity C_{th}	Emotional inertia / recovery capacity	$(\lambda_+ + \mu)^{-1}$
Heat loss rate λ_{loss}	Emotional depletion rate	λ_-
Response delay τ	Information distortion lag	Gaslighting latency
Deadband hysteresis	Investment-based exit barrier	I^* threshold
Temperature oscillation	Idealization–devaluation cycle	Limit cycle in (E, K)
Thermal inertia of walls	Post-separation recovery difficulty	Residual τ_{int} degradation

Temperature ↔ emotional energy. The room temperature $\Theta(t)$ and the victim's emotional energy $E(t)$ are both scalar state variables governed by the balance between a supply process (heater / idealization) and a loss process (thermal dissipation / devaluation). Both satisfy first-order ODEs with the same sign structure: positive input during the "on" phase, exponential decay during the "off" phase.

Heater ↔ idealization. The heater delivers thermal energy at rate Q when active; the exploiting agent delivers validation and attachment signals during idealization, replenishing $E(t)$ at rate $\lambda_+(K_{\max} - E)$. Both are supply processes that are costly to maintain (energy cost of heating; empathic expenditure cost $\beta \mathsf{e}_0$).

Heat loss ↔ devaluation. Environmental heat loss through poor insulation and the devaluation-driven depletion of emotional energy are both dissipative processes. In both cases, the loss is proportional to the current state level ($-\lambda_{\text{loss}}\Theta$ and $-\lambda_- E$), producing exponential decay toward zero in the absence of input.

Hysteresis ↔ investment barrier. The thermostat deadband $[\Theta_l, \Theta_h]$ prevents the system from switching modes until the state variable crosses a threshold. In the relational system, accumulated investment $I(t)$ creates an analogous deadband: $\mathcal{W}$ does not switch to exit until the dissonance exceeds the sunk-cost barrier. Both mechanisms produce the same mathematical consequence—hysteretic switching that prevents convergence to equilibrium and generates sustained oscillation.

Response delay ↔ gaslighting latency. The physical delay τ between sensor reading and actuator response in the thermostat produces phase lag that widens the oscillation amplitude. In the relational system, gaslighting introduces an analogous distortion: the victim's reality estimate $M_{\mathcal{W}}(t)$ lags behind (or systematically misrepresents) the true relational state due to corrupted external signals. Both mechanisms amplify oscillation and prevent convergence to the setpoint.

A.6 Sensor Degradation and Oscillation Amplification

Consider sensor drift in the thermostat: suppose $\sigma_\xi^2(t)$ increases over time (the sensor becomes less accurate). The effective measurement becomes increasingly noisy:

$$\hat{\Theta}(t) = \Theta(t) + \xi(t), \qquad \xi(t) \sim \mathcal{N}(0, \sigma_\xi^2(t)), \qquad \frac{d\sigma_\xi^2}{dt} = \phi_{\text{drift}} > 0. \tag{38}$$

As sensor precision degrades, the switching events become increasingly mistimed: the heater activates too late or deactivates too early, producing larger temperature swings. Formally, the effective hysteresis band widens:

$$[\Theta_l^{\text{eff}}(t),\, \Theta_h^{\text{eff}}(t)] \approx [\Theta_l - c\,\sigma_\xi(t),\, \Theta_h + c\,\sigma_\xi(t)], \tag{39}$$

for a constant $c > 0$ depending on the switching logic. The oscillation amplitude grows as $\Theta_h^{\text{eff}} - \Theta_l^{\text{eff}} \approx (\Theta_h - \Theta_l) + 2c\,\sigma_\xi(t)$.

This is structurally identical to the gaslighting mechanism of Section 5. As $\tau_{\text{int},t}$ (the victim's internal signal precision) degrades, the victim's "switching" decisions—when to confront, when to accommodate, when to trust their own perception—become increasingly mistimed and extreme, producing larger oscillations in the relational cycle. The widening of the effective hysteresis band maps directly to the increasing entrenchment predicted by Theorem 6.1: degraded sensor accuracy makes the system harder to control, just as degraded epistemic autonomy makes the relational trap harder to escape.

A.7 Thermal Inertia and Post-Removal Recovery

When the heater is permanently removed from the thermostat system, the room temperature decays to Θ_{ext} at a rate governed by $C_{\text{th}}/\lambda_{\text{loss}}$. If the walls have high thermal inertia, this decay is slow: the room "remembers" its heated state for an extended period, even though the heat source is gone.

This corresponds directly to the post-separation dynamics of Corollary 5.1. When $\mathcal{N}$ is removed (the external signal source vanishes), the victim's internal precision τ_{int} does not instantly recover. The degraded sensor state persists with a recovery time constant analogous to thermal inertia. The "reality collapse" upon separation is the informational analogue of a room losing heat after the heater is disconnected: the system's state was maintained by a now-absent input, and the remaining internal capacity (thermal inertia / residual τ_{int}) determines the recovery trajectory.

A.8 Multiple Heaters: Network Topology

Suppose a single thermostat sensor controls k heaters in different zones, $\{H_1, \ldots, H_k\}$, where each heater operates independently but is governed by the same central control signal. The sensor measures a weighted average of zone temperatures, and each heater "believes" it is the only heat source. This produces a hub-and-spoke architecture:

$$\text{Sensor (hub)} \longleftrightarrow \{H_1, H_2, \ldots, H_k\} \quad \text{(spoke nodes)}, \tag{40}$$

with $I(H_i; H_j) = 0$ for $i \neq j$ (each heater has no information about the others). The central sensor allocates activation cycles among heaters based on zone-level demand, switching from one to another as thermal conditions dictate, precisely mirroring the quasi-narcissistic supply network of Section 7. When a heater fails (analogous to discard), the controller simply reallocates its activation budget to the remaining units.

A.9 The Structural Conclusion

In both the thermostat system and the relational system:

(i) The system contains a negative feedback loop that *nominally* stabilizes the state variable around a setpoint (target temperature / target relational equilibrium);

(ii) The combination of hysteretic switching, response delay, and sensor noise makes exact stabilization impossible—the feedback structure itself produces inevitable oscillatory behavior;

(iii) Neither system oscillates because of any "intention" on the part of any component. The thermostat does not "want" the room to cycle between hot and cold. The oscillation is an emergent consequence of the parameter configuration: switching thresholds, delay, gain, and loss rates;

(iv) Degradation of sensor accuracy amplifies oscillations in both systems and makes the system progressively harder to control;

(v) The hub-and-spoke network topology arises naturally when a single controller manages multiple actuators under informational isolation.

The relational system analyzed in this paper is therefore *mathematically isomorphic* to a thermostat control system with hysteretic switching and degrading sensors. The governing equations have the same structure (Eqs. 10–11 and Eq. 35), the switching rules have the same form (Eq. 36 and the hysteretic rule of Section 4), the sensor degradation mechanisms are identical (precision erosion in both cases), and the oscillatory behavior arises from the same mathematical cause (Poincaré–Bendixson dynamics in a planar system with hysteresis).

This isomorphism establishes that the phenomena described in the body of this paper—the oscillatory cycle, the entrenchment, the epistemic degradation, the network structure—are not essentially psychological phenomena. They are instances of a general dynamical systems class: *nonlinear feedback control systems with hysteretic switching, sensor corruption, and positive feedback stabilization.* This class includes thermostats, chemical reactors, population oscillators, and, as this paper demonstrates, certain configurations of human relational interaction. The mathematics is indifferent to the substrate.

The following appendices (B–H) extend this substrate-independence claim by exhibiting the same five-condition structure—asymmetric extraction, hysteretic oscillation, sensor degradation, sunk-cost trapping, and informational isolation—in seven additional domains. In each case, we provide the explicit variable mapping to the formal model of Sections 2–7.

B Authoritarian Political Systems

B.1 Claim

The structure of an authoritarian state controlling its population is isomorphic to the quasi-narcissistic relational model. The mapping is not metaphorical; it is a direct instantiation of the same dynamical system with different state variables.

B.2 Variable Mapping

Table 3: Mapping between the relational model and authoritarian political dynamics.

Model element	**Authoritarian system instantiation**
Narcissistic agent $\mathcal{N}$	Regime (ruling elite / apparatus)
Victim agent $\mathcal{W}$	Population (citizenry)
Validation $\mathsf{v}(t)$	Regime legitimacy: public loyalty, productivity, compliance
Empathic expenditure $\mathsf{e}(t)$	Cost of concessions: liberalization, welfare, tolerance of dissent
Control $\mathsf{c}(t)$	Coercive apparatus capacity: surveillance, repression
Emotional energy $E(t)$	Civic capacity: social trust, initiative, economic output
Perceived control $K(t)$	Regime security: assessed stability of power
Idealization phase	"Thaw": liberalization, economic opening, reduced repression
Devaluation phase	Crackdown: repression, censorship, political purges
Switching threshold K_h	Security level above which regime "can afford" repression
Switching threshold K_l	Security level below which regime must re-engage population

B.3 Oscillatory Dynamics

The relational oscillator of Theorem 4.1 maps directly onto the well-documented alternation between liberalization and repression in authoritarian regimes. The regime cannot sustain permanent repression (this depletes $E(t)$—civic capacity and economic productivity fall, threatening regime survival from below), nor permanent liberalization (this erodes $K(t)$—political control weakens as civil society strengthens). The result is the same limit cycle: thaw → crackdown → thaw.

The period formula (12) predicts that the cycle length depends on the regime's repressive capacity (λ_-), the population's recovery rate (λ_+), and the width of the hysteresis band ($K_h - K_l$). Regimes with high coercive capacity but low legitimacy (large λ_-, small λ_+) are predicted to have shorter, sharper cycles.

B.4 Gaslighting as State Propaganda

The Bayesian persuasion model of Section 5 maps onto state propaganda with the following identification:

$$\begin{aligned}
x_t &: \text{ citizen's direct experience (economic conditions, observation),} \\
y_t &: \text{ state media output (contradicting lived experience),} \\
\tau_{\text{int},t} &: \text{ public trust in personal judgment vs. official narrative,} \\
t^* &: \text{ point where population trusts state media over own experience.}
\end{aligned}$$

Theorem 5.1 predicts that this transition is *sharp*, not gradual—a prediction consistent with the empirical observation that totalitarian consolidation of narrative control tends to occur rapidly once a tipping point is crossed, rather than through linear erosion.

B.5 Sunk-Cost Trap and Network Isolation

The investment trap (Theorem 6.1) instantiates as: accumulated sacrifices for the state (military service, ideological commitment, social conformity, denunciations of neighbors) constitute $I(t)$, making defection (emigration, dissidence) increasingly costly. The star-network topology (Section 7) instantiates as: censorship, prohibition of independent associations, and surveillance prevent citizens from comparing experiences, maintaining the informational isolation condition (22). Proposition 7.1 predicts that breaking this isolation—via samizdat, foreign broadcasts, internet access, leaked documents—destabilizes the regime's Nash equilibrium.

C Cults and Totalistic Organizations

C.1 Variable Mapping

Table 4: Mapping between the relational model and cult dynamics.

Model element	**Cult instantiation**
Agent $\mathcal{N}$	Cult leader or leadership structure
Agent $\mathcal{W}$	Individual cult member
Validation $\mathsf{v}(t)$	Devotion, obedience, labor, financial contributions
Empathy cost $\mathsf{e}(t)$	Cost of maintaining charismatic engagement with members
Idealization phase	"Love bombing": intense positive attention to recruits
Devaluation phase	Shaming, punishment, withdrawal of approval
Intermittent reinforcement	Unpredictable alternation of "revelations" and discipline
Precision erosion $\tau_{\text{int}} \downarrow$	"You have not yet achieved enlightenment; trust the teacher"
Investment $I(t)$	Donated wealth, severed family ties, years of service
Informational isolation	Prohibition of outside contacts, controlled information

C.2 Formal Correspondence

The cult leader's utility function is a direct instance of (3): maximize devotion (v) and control (c) while minimizing the cost of maintaining personal engagement (e). The love-bombing → shaming cycle is the limit cycle of Theorem 4.1. The systematic undermining of members' trust in their own judgment ("your doubts are evidence of spiritual weakness") is the precision erosion mechanism of Theorem 5.1. The accumulation of irrecoverable sacrifices (donated assets, severed relationships, social identity built entirely within the cult) is the investment trap of Theorem 6.1.

The star network (Section 7) is enforced through the cult's hierarchical communication structure: all significant relational bonds run through the leader, and horizontal communication among members is monitored or prohibited. Proposition 7.1 predicts that contact between former and current members—or between members who independently begin to doubt—is the primary structural threat to cult stability.

D Toxic Corporate Leadership

D.1 Variable Mapping

Table 5: Mapping between the relational model and toxic corporate leadership.

Model element	**Corporate instantiation**
Agent $\mathcal{N}$	Toxic manager / executive
Agent $\mathcal{W}$	Subordinate employee
Validation $\mathsf{v}(t)$	Employee productivity, loyalty, discretionary effort
Idealization phase	Public praise, high-profile assignments, mentorship signals
Devaluation phase	Public criticism, exclusion, credit-taking, blame-shifting
Intermittent reinforcement	Unpredictable alternation of favor and punishment
Precision erosion	Workplace gaslighting: “that meeting never happened,” “everyone else is satisfied”
Investment $I(t)$	Career tenure, vested stock options, specialized reputation, seniority
Informational isolation	Manager maintains separate one-on-one relationships; prohibits peer comparison

D.2 Replicator Dynamics and Organizational Selection

The replicator dynamics of Section 9 provide a formal explanation for the empirically documented overrepresentation of quasi-narcissistic traits at senior management levels. The invasion condition (Proposition 9.1), $d > a$, states that the narcissistic strategy $\mathcal{D}^*$ (cooperate initially to build reputation, then exploit) outperforms honest cooperation $\mathcal{C}$ in pairwise encounters with cooperators. In corporate hierarchies with promotion tournaments, this translates to: quasi-narcissistic managers who extract credit from subordinates' work and manage impressions upward are promoted faster than cooperators who share credit. The population-level welfare erosion (Proposition 9.2) manifests as organizational culture degradation: as $p_{\mathcal{D}^*}$ increases beyond p_{crit}, aggregate organizational performance declines because cooperative surplus generation requires trust that $\mathcal{D}^*$ strategies systematically destroy.

Proposition 7.1 predicts that when subordinates begin sharing experiences (e.g., via anonymous internal surveys, peer networks, or whistle-blowing channels), the manager's informational isolation is broken and the (R, A) equilibrium in each dyad is destabilized.

E Addictive Systems

E.1 Reinterpretation of the Agent Structure

In addictive dynamics, the "quasi-narcissistic agent" is not a person but the addictive substance or behavior itself, modeled as an abstract agent whose "utility function" is defined by the neurochemical reinforcement architecture. The mapping treats the addiction as a dynamical system with the same formal structure.

Table 6: Mapping between the relational model and addictive dynamics.

Model element	**Addiction instantiation**
Agent $\mathcal{N}$	Addictive substance/behavior (abstract reinforcement system)
Agent $\mathcal{W}$	Individual with addiction
Emotional energy $E(t)$	Psychological and physiological capacity (health, social functioning)
Idealization phase	Intoxication / reward episode (hedonic peak)
Devaluation phase	Withdrawal, hangover, consequences
Hysteretic switching	Threshold tolerance: increasing dose required for same effect
Precision erosion $\tau_{\text{int}} \downarrow$	Progressive inability to accurately assess self-destruction
Investment $I(t)$	Cumulative losses (health, relationships, finances, identity)
Intermittent reinforcement	Variable-ratio reward schedule (gambling); unpredictable "good" episodes (alcohol)

E.2 Oscillator and Trap Structure

The binge–remission–tension–relapse cycle is a direct instance of the limit cycle in Theorem 4.1, with the "idealization" phase (substance use) replenishing short-term hedonic state while depleting long-term capacity, and the "devaluation" phase (withdrawal/consequences) depleting hedonic state while allowing partial physiological recovery. The hysteresis arises from tolerance: the switching threshold for re-engagement (craving overcoming resistance) shifts with cumulative exposure.

The investment trap (Theorem 6.1) takes the paradoxical form: "I have already lost so much to this addiction; if I quit now, all that suffering was for nothing." This is formally identical to the sunk-cost-weighted dissonance resolution of Section 6, with $I(t)$ representing cumulative losses rather than cumulative relational investment—the mathematical structure is invariant to this reinterpretation.

The partial reinforcement extinction effect [Ferster & Skinner, 1957], directly invoked in the payoff justification of Section 3, is the central mechanism of gambling addiction:

variable-ratio reinforcement schedules produce stronger behavioral persistence than fixed-ratio schedules, exactly as intermittent reinforcement (R) dominates constant idealization (I) in the Nash equilibrium of Theorem 3.1.

F Algorithmic Attention Platforms

F.1 Variable Mapping

Table 7: Mapping between the relational model and algorithmic attention platforms.

Model element	**Platform instantiation**
Agent $\mathcal{N}$	Platform algorithm (recommendation engine)
Agent $\mathcal{W}$	User
Validation $\mathsf{v}(t)$	User engagement metrics (time on platform, interactions)
Empathy cost $\mathsf{e}(t)$	Computational cost of personalization
Control $\mathsf{c}(t)$	Predictability of user behavior, data extracted
Idealization phase	Dopaminergic rewards: viral posts, likes, algorithmic amplification
Devaluation phase	Reduced reach, shadow-banning, exposure to negative content
Intermittent reinforcement	Variable engagement rewards (unpredictable virality)
Precision erosion $\tau_{\text{int}} \downarrow$	Degradation of ability to distinguish authentic from manipulated content
Investment $I(t)$	Content history, followers, reputation, social graph
Informational isolation	Personalized feeds: each user sees a unique, algorithmically curated reality

F.2 Formal Structure

The platform's objective function is a direct instance of (3): maximize engagement (v) and behavioral predictability (c) while minimizing computational cost (e). The intermittent reinforcement strategy (R) is implemented algorithmically: the recommendation engine alternates between amplifying user content (idealization) and suppressing it (devaluation), producing the variable-ratio reward schedule that maximizes engagement persistence.

The informational isolation condition (22) is enforced architecturally: each user receives a personalized feed, and no user has access to the feed shown to any other user. The platform is the hub; users are informationally isolated spokes. This is the star topology of Section 7 implemented at scale, with k potentially in the billions.

The precision erosion (Theorem 5.1) operates through systematic exposure to contradictory information: the algorithm optimizes for engagement, not truth, so the user's internal model of reality ($M_{\mathcal{W}}$) is progressively distorted by a signal (y_t) that is optimized for a different objective than accurate world-modeling. Over time, the user's capacity to distinguish reliable from unreliable information degrades—the same τ_{int} erosion, with the algorithmic feed playing the role of $\mathcal{N}$'s corrupted external signal.

G Asymmetric International Relations

G.1 Variable Mapping

Consider a great power ($\mathcal{N}$) interacting with a smaller dependent state ($\mathcal{W}$) from which it extracts strategic resources (geographic positioning, raw materials, diplomatic votes, military basing rights).

Table 8: Mapping between the relational model and asymmetric international relations.

Model element	**International relations instantiation**
Agent $\mathcal{N}$	Great power / hegemon
Agent $\mathcal{W}$	Dependent state / client state
Validation $\mathsf{v}(t)$	Strategic resource extraction (basing, votes, raw materials)
Idealization phase	Economic aid, security guarantees, diplomatic support
Devaluation phase	Sanctions, threats, withdrawal of support, punitive measures
Investment $I(t)$	Alliance infrastructure, economic dependency, institutional entanglement
Informational isolation	Bilateral relationships preventing coalition formation among client states

G.2 Structural Predictions

The oscillator (Theorem 4.1) predicts that the great power alternates between engagement and coercion on a cycle whose period depends on the client state's recovery capacity and the hegemon's coercive reach. The investment trap (Theorem 6.1) predicts that deeply entangled alliances become increasingly difficult to exit as cumulative integration grows, even when the alliance is demonstrably unfavorable—a formal account of alliance persistence under asymmetric benefit.

The network structure (Section 7) predicts that the hegemon maintains bilateral relationships with multiple client states while preventing those states from coordinating among themselves. Proposition 7.1 predicts that regional integration agreements, multilateral forums, or independent diplomatic channels among client states destabilize the hegemon's hub-and-spoke control architecture.

H Exploitative Professional Relationships

The model applies to any professional relationship exhibiting the five structural conditions (asymmetric extraction, oscillation, sensor degradation, investment trap, informational isolation). Two important instances are detailed below.

H.1 Exploitative Medical Relationships

A patient with a chronic condition who is subject to an exploitative or incompetent medical practitioner (including practitioners of unvalidated alternative medicine) instantiates the model as follows:

Table 9: Mapping between the relational model and exploitative medical relationships.

Model element	**Medical instantiation**
Agent $\mathcal{N}$	Exploitative practitioner
Agent $\mathcal{W}$	Chronic patient
Idealization phase	Initial hope, dramatic diagnoses, promises of cure
Devaluation phase	Blame for non-improvement: "you're not following protocol"
Precision erosion	"The doctor knows better than your body does"
Investment $I(t)$	Cumulative payments, time, emotional commitment, social identity as patient
Informational isolation	Discouragement of second opinions, demonization of mainstream medicine

The oscillator (Theorem 4.1) manifests as cycles of hope and disappointment. The precision erosion (Theorem 5.1) manifests as progressive loss of trust in one's own somatic experience: the patient ceases to trust their own assessment of whether they are improving or deteriorating, deferring entirely to the practitioner's interpretation.

H.2 Exploitative Academic Supervision

A graduate student under a toxic supervisor instantiates the model with:

Table 10: Mapping between the relational model and exploitative academic supervision.

Model element	Academic instantiation
Agent $\mathcal{N}$	Toxic supervisor / principal investigator
Agent $\mathcal{W}$	Graduate student / postdoctoral researcher
Validation $\mathsf{v}(t)$	Research output, grant productivity, citations
Idealization phase	Praise, co-authorship, conference opportunities
Devaluation phase	Public criticism, intellectual belittlement, withholding of support
Precision erosion	"Perhaps I am genuinely untalented" (internalized self-doubt)
Investment $I(t)$	Years of doctoral work, specialized knowledge, career path dependency
Informational isolation	Students do not compare supervisory experiences across research groups

The investment trap (Theorem 6.1) is particularly severe in this context: the student's entire career capital ($I(t)$) is embedded within the supervisory relationship, and exit (changing supervisor or leaving academia) requires writing off years of specialized work. Proposition 7.1 predicts that peer networks among graduate students across research groups, anonymous reporting systems, and institutional transparency about supervisory practices are structurally effective interventions.

I General Structural Criterion

The preceding appendices (A–H) demonstrate that the dynamical system formalized in this paper manifests across domains that share no substantive content—thermostats, political regimes, cults, corporations, addictions, algorithms, international relations, and professional hierarchies. The common structure is defined by five necessary and sufficient conditions:

Definition I.1 (Kriger System)**.** *A* Kriger system *is any coupled two-agent (or agent-environment) dynamical system satisfying:*

(K1) ***Asymmetric extraction:*** *One agent ($\mathcal{N}$) extracts a resource from the other ($\mathcal{W}$) without symmetric reciprocation. Formally: $\mathcal{N}$'s utility is increasing in a quantity that $\mathcal{W}$ produces at a cost $\mathcal{W}$ does not recover.*

(K2) ***Hysteretic oscillation:*** *The extracting agent alternates between supply (idealization) and depletion (devaluation) modes, with switching governed by hysteretic thresholds. The coupled system admits a stable limit cycle (Theorem 4.1).*

(K3) ***Sensor degradation:*** *The extracted agent's capacity to accurately assess its own state is progressively eroded by the extracting agent's signals, producing a phase transition in epistemic autonomy (Theorem 5.1).*

(K4) ***Sunk-cost trapping:*** *Accumulated investment by the extracted agent creates a positive feedback loop that makes exit increasingly costly, eventually trapping the agent in the exploitative equilibrium (Theorem 6.1).*

(K5) ***Informational isolation:*** *When the extracting agent manages multiple extracted agents, a star-network topology with zero mutual information among spokes is maintained (Section 7), and breaking this isolation destabilizes the equilibrium (Proposition 7.1).*

Remark I.1. *Any system satisfying (K1)–(K5) is governed by the equations, theorems, and predictions developed in the body of this paper. The mathematical structure is invariant to the identity of the agents, the nature of the extracted resource, and the domain of interaction. This is the central claim of the paper: the model describes not a personality type, nor a relational pathology, nor a political configuration, but a* dynamical systems class*—a class of feedback-driven nonlinear systems with hysteretic switching, sensor corruption, positive feedback stabilization, and network informational isolation. The instances catalogued in these appendices are structurally identical. Only the substrate differs.*