Supplementary Information

Precision Dissociation in Narcissistic Personality Disorder: A Hierarchical Predictive Coding Account of Empathy and Self-Esteem Dysregulation

Marco Fronzi

This supplement provides the complete mathematical specification of all model components, their derivations from first principles, and analysis of their key mathematical properties.

Contents

1 Notation and Conventions **3**

2 Generative Model Specification **3**
2.1 Full Joint Distribution . . . 3
2.2 Graphical Model . . . 4

3 Variational Free Energy and Laplace Approximation **5**
3.1 Free Energy Functional . . . 5
3.2 Laplace Approximation . . . 5
3.3 Cognitive Stream Free Energy . . . 5
3.4 Self-Model Free Energy . . . 6

4 Kalman Gain Derivation and Properties **6**
4.1 Connection to the Kalman Filter . . . 6
4.2 Properties of the Kalman Gain . . . 6
4.3 NPD vs. HC Kalman Gains . . . 7

5 Self-Model Dynamics **7**
5.1 Cumulative Threat Accumulator . . . 7
5.2 Precision Erosion . . . 8
5.3 Self-Esteem Update Rule . . . 9
5.4 Convergence and Fixed Points . . . 9

6 Empathy Metrics: Formal Definitions and Properties **10**
6.1 Empathic Accuracy . . . 10
6.2 Empathic Concern . . . 10
6.3 Empathic Imbalance Index . . . 11

7 Free-Energy Landscape and Bistability **11**
7.1 Effective Potential for the Self-Model . . . 11
7.2 Emergent Bistability via Precision Dynamics . . . 11
7.3 Phase Transition Criterion . . . 12

8 Stimulus Generation and Paradigm Design **13**
8.1 Stimulus Type Assignment . . . 13
8.2 Stimulus-Conditional Distributions . . . 13
8.3 Narcissistic Injury Block . . . 13

9 Simulation Algorithm **13**

10 Connection to the Free Energy Principle **14**
10.1 Precision as Attention in Active Inference . . . 14
10.2 Variational Bayes vs. Sampling . . . 14
10.3 Extended Model: Active Inference with Policy Selection . . . 15

11 Parameter Rationale and Identifiability **15**
11.1 Parameter Rationale . . . 15
11.2 Identifiability . . . 15

12 Analytical Predictions for Experimental Validation **16**
12.1 EEG/ERP Predictions . . . 16
12.2 Prediction for Empathy–Accuracy Trade-off . . . 16
12.3 Prediction for Self-Esteem Kurtosis . . . 16

1 Notation and Conventions

Throughout this supplement, the following notation is used consistently. Scalars are written in italic (x), vectors in bold italic ($\boldsymbol{x}$), and matrices in bold uppercase ($\mathbf{A}$). We use $p(\cdot)$ for probability densities and $q(\cdot)$ for variational (approximate posterior) densities.

Table 1: Primary notation.

Symbol	Type	Description
$t \in \{0, 1, \ldots, T\}$	index	discrete trial index
$o_{\text{cog}}(t)$	$\mathbb{R}$	observed cognitive social cue at trial t
$o_{\text{aff}}(t)$	$\mathbb{R}$	observed affective social cue at trial t
$f(t)$	$\mathbb{R}$	self-referential social feedback
$s_{\text{other}}(t)$	$\mathbb{R}$	true (latent) mental state of social partner
$a_{\text{other}}(t)$	$\mathbb{R}$	true (latent) affective state of social partner
$\hat{\mu}_s(t)$	$\mathbb{R}$	posterior mean of cognitive stream
$\hat{\mu}_a(t)$	$\mathbb{R}$	posterior mean of affective stream
$S(t)$	$\mathbb{R}$	self-esteem estimate (self-model posterior mean)
π_{cog}	$\mathbb{R}_{>0}$	cognitive stream precision (fixed)
π_{aff}	$\mathbb{R}_{>0}$	affective stream precision (fixed)
$\Pi^{(0)}_{\text{self}}$	$\mathbb{R}_{>0}$	baseline self-model precision
$\Pi_{\text{self}}(t)$	$\mathbb{R}_{>0}$	dynamic self-model precision at trial t
$K_{\text{cog}}, K_{\text{aff}}$	$[0, 1]$	Kalman gains for cognitive/affective streams
$C(t)$	$\mathbb{R}_{\geq 0}$	cumulative ego-threat load
γ	$\mathbb{R}_{>0}$	precision erosion rate
$\mathcal{F}$	$\mathbb{R}$	variational free energy

2 Generative Model Specification

2.1 Full Joint Distribution

The generative model defines a joint probability distribution over all observable and latent variables. At each trial t the model factorises as:

$$p(o_{\text{cog}}, o_{\text{aff}}, f, s_{\text{other}}, a_{\text{other}}, S) = \underbrace{p(o_{\text{cog}} \mid s_{\text{other}})}_{\text{cog. likelihood}} \underbrace{p(o_{\text{aff}} \mid a_{\text{other}})}_{\text{aff. likelihood}} \underbrace{p(f \mid S)}_{\text{feedback likelihood}} \underbrace{p(s_{\text{other}})}_{\text{cog. prior}} \underbrace{p(a_{\text{other}})}_{\text{aff. prior}} \underbrace{p(S)}_{\text{self prior}}, \tag{1}$$

where we have suppressed the trial index t for clarity and used the conditional independence structure of the model (no direct coupling between the cognitive and affective streams at this level).

Likelihood terms. All likelihoods are Gaussian:

$$o_{\text{cog}} \mid s_{\text{other}} \sim \mathcal{N}\big(s_{\text{other}},\ \sigma^2_{\text{obs}}\big)\,, \tag{2}$$

$$o_{\text{aff}} \mid a_{\text{other}} \sim \mathcal{N}\big(a_{\text{other}},\ \sigma^2_{\text{obs}}\big)\,, \tag{3}$$

$$f \mid S \sim \mathcal{N}\big(S,\ \pi_{\text{fb}}^{-1}\big)\,, \tag{4}$$

where $\sigma^2_{\text{obs}} = \pi_{\text{cog}}^{-1}$ in the cognitive stream and $\sigma^2_{\text{obs}} = \pi_{\text{aff}}^{-1}$ for the affective stream. More precisely, the observation precision is the precision parameter of the likelihood, not of the prior; we use the notation π_{cog} to refer to this observation-level precision weight assigned by the agent to the cognitive channel.

Prior terms. The priors on the social partner's states are zero-mean Gaussians with prior precisions $\pi_s^{(0)}$ and $\pi_a^{(0)}$:

$$s_{\text{other}} \sim \mathcal{N}\big(0,\ (\pi_s^{(0)})^{-1}\big)\,, \tag{5}$$

$$a_{\text{other}} \sim \mathcal{N}\big(0,\ (\pi_a^{(0)})^{-1}\big)\,. \tag{6}$$

The self prior is:

$$S \sim \mathcal{N}\big(\mu_S,\ \Pi_{\text{self}}(t)^{-1}\big)\,, \tag{7}$$

where μ_S is the running self-esteem estimate (updated at each trial) and $\Pi_{\text{self}}(t)$ is the dynamic precision defined in Section 5.

2.2 Graphical Model

Figure 1 shows the directed graphical model (Bayesian network) for a single trial. The three plates distinguish the (i) cognitive inference path, (ii) affective inference path, and (iii) self-referential path.

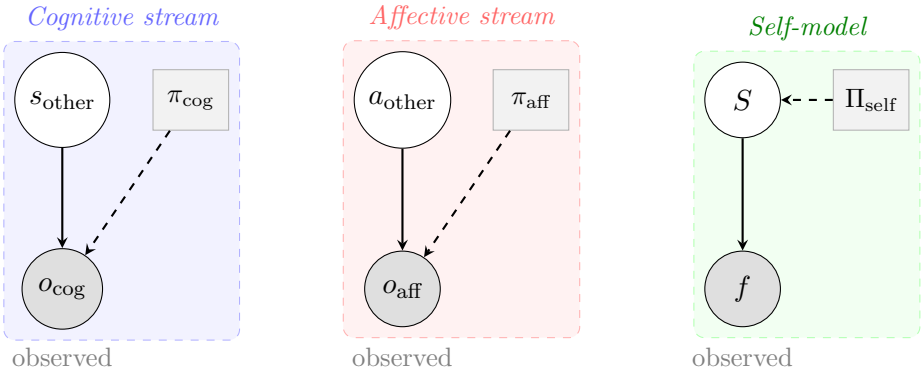


Figure 1: **Directed graphical model for a single trial.** Shaded nodes are observed; open nodes are latent. Dashed arrows indicate precision (hyperparameter) influence on each channel.

3 Variational Free Energy and Laplace Approximation

3.1 Free Energy Functional

In the Free Energy Principle (FEP), an agent is assumed to minimise its variational free energy $\mathcal{F}$, which upper-bounds the negative log model evidence (surprise):

$$\mathcal{F}\big[q, o\big] = \underbrace{\mathbb{E}_q[-\ln p(o, z)]}_{\text{energy}} + \underbrace{\mathbb{E}_q[\ln q(z)]}_{\text{negative entropy}} = \mathrm{KL}[q(z) \parallel p(z \mid o)] - \ln p(o), \tag{8}$$

where $z = \{s_{\text{other}}, a_{\text{other}}, S\}$ are the latent variables, o are the observations, and $q(z)$ is the variational (recognition) density maintained by the agent. Because $\mathrm{KL}[\cdot] \geq 0$, minimising $\mathcal{F}$ simultaneously drives $q \to p(z \mid o)$ (accurate inference) and minimises surprise $-\ln p(o)$.

3.2 Laplace Approximation

We assume a mean-field, Laplace (Gaussian) variational density:

$$q(z) = q(s_{\text{other}})\, q(a_{\text{other}})\, q(S), \quad q(s_{\text{other}}) = \mathcal{N}(\hat{\mu}_s,\, \tilde{\pi}_s^{-1}), \tag{9}$$

and similarly for $q(a_{\text{other}})$ and $q(S)$. Under this approximation the free energy separates across the three streams:

$$\mathcal{F} = \mathcal{F}_{\text{cog}} + \mathcal{F}_{\text{aff}} + \mathcal{F}_{\text{self}} + \text{const.} \tag{10}$$

3.3 Cognitive Stream Free Energy

Substituting the Gaussian likelihood and prior into $\mathcal{F}_{\text{cog}}$:

$$\begin{aligned} \mathcal{F}_{\text{cog}} &= \mathbb{E}_q[-\ln p(o_{\text{cog}} \mid s_{\text{other}})] + \mathbb{E}_q[-\ln p(s_{\text{other}})] + \mathbb{E}_q[\ln q(s_{\text{other}})] \\ &= \frac{\pi_{\text{cog}}}{2}\mathbb{E}_q\big[(o_{\text{cog}} - s_{\text{other}})^2\big] + \frac{\pi_s^{(0)}}{2}\mathbb{E}_q\big[s_{\text{other}}^2\big] - \frac{1}{2}\ln \tilde{\pi}_s + \text{const.} \end{aligned} \tag{11}$$

Taking the gradient with respect to $\hat{\mu}_s$ and setting to zero yields:

$$\frac{\partial \mathcal{F}_{\text{cog}}}{\partial \hat{\mu}_s} = -\pi_{\text{cog}}(o_{\text{cog}} - \hat{\mu}_s) + \pi_s^{(0)}\hat{\mu}_s = 0 \quad \Rightarrow \quad \hat{\mu}_s = \frac{\pi_{\text{cog}}}{\pi_{\text{cog}} + \pi_s^{(0)}}\, o_{\text{cog}} = K_{\text{cog}}\, o_{\text{cog}}. \tag{12}$$

The posterior precision is:

$$\tilde{\pi}_s = \pi_{\text{cog}} + \pi_s^{(0)}, \tag{13}$$

with posterior variance $\tilde{\sigma}_s^2 = 1/\tilde{\pi}_s$. At the optimum, the minimum free energy reduces to:

$$\mathcal{F}_{\text{cog}}^{*} = \frac{\pi_{\text{cog}}\, \pi_s^{(0)}}{2(\pi_{\text{cog}} + \pi_s^{(0)})}\, (o_{\text{cog}})^2 + \text{const.} \tag{14}$$

This is the *precision-weighted prediction error*: the agent's surprise about the cognitive observation, weighted by the harmonic mean of likelihood and prior precision.

3.4 Self-Model Free Energy

The self-model stream contributes:

$$\mathcal{F}_{\text{self}} = \frac{\pi_{\text{fb}}}{2}\mathbb{E}_q\big[(f-S)^2\big] + \frac{\Pi_{\text{self}}(t)}{2}\mathbb{E}_q\big[(S-\mu_S)^2\big] - \frac{1}{2}\ln\tilde{\pi}_S + \text{const.} \tag{15}$$

Minimising over $\hat{\mu}_S$ gives:

$$\hat{\mu}_S = \frac{\Pi_{\text{self}}(t)\,\mu_S + \pi_{\text{fb}}\,f}{\Pi_{\text{self}}(t) + \pi_{\text{fb}}}, \tag{16}$$

which is the precision-weighted Bayes estimate: a convex combination of the prior mean μ_S and the feedback f, with weights proportional to their respective precisions. This is the update rule implemented in Eq. (10) of the main paper.

4 Kalman Gain Derivation and Properties

4.1 Connection to the Kalman Filter

The single-step precision-weighted update in Eq. (12) is identical to the measurement-update step of a steady-state Kalman filter with:

- **State transition model**: $s_{\text{other},t+1} = s_{\text{other},t} + w_t, \quad w_t \sim \mathcal{N}(0, Q)$
- **Observation model**: $o_t = s_{\text{other},t} + v_t, \quad v_t \sim \mathcal{N}(0, R)$

where $R = \pi_{\text{cog}}^{-1}$ is the observation noise variance and Q encodes the prior variance on state transitions. In the limit of a static (non-evolving) hidden state (appropriate for a single trial), the Kalman gain reduces to:

$$K_{\text{cog}} = \frac{P_-}{P_- + R} = \frac{(\pi_s^{(0)})^{-1}}{(\pi_s^{(0)})^{-1} + \pi_{\text{cog}}^{-1}} = \frac{\pi_{\text{cog}}}{\pi_{\text{cog}} + \pi_s^{(0)}}, \tag{17}$$

confirming Eq. (12). The same derivation applies to K_{aff} with π_{aff} replacing π_{cog}.

4.2 Properties of the Kalman Gain

Proposition 1 (Monotonicity and saturation)**.** *The Kalman gain $K(\pi) = \pi/(\pi + \pi^{(0)})$ is a strictly monotone increasing function of observation precision π, with $K \to 0$ as $\pi \to 0$ and $K \to 1$ as $\pi \to \infty$.*

Proof. $\partial K/\partial\pi = \pi^{(0)}/(\pi + \pi^{(0)})^2 > 0$ for all $\pi, \pi^{(0)} > 0$. □

Proposition 2 (Optimal precision for minimum MSE)**.** *The mean squared error of the posterior estimate* $\text{MSE}(\pi) = \mathbb{E}[(K(\pi)\,o - s_{true})^2]$ *is minimised at:*

$$\pi^* = \frac{\pi^{(0)}}{\sigma_{obs}^2\pi^{(0)} - 1} \quad \textit{(when } \sigma_{obs}^2\pi^{(0)} > 1\textit{)}, \tag{18}$$

and $\text{MSE}(\pi)$ *is U-shaped in π: both $\pi \to 0$ and $\pi \to \infty$ produce large MSE.*

Proof. Write $o = s_{\text{true}} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$. Then:

$$\begin{aligned} \text{MSE}(\pi) &= \mathbb{E}\big[(K\,o - s_{\text{true}})^2\big] \\ &= \mathbb{E}\big[(K(s_{\text{true}} + \varepsilon) - s_{\text{true}})^2\big] \\ &= (K-1)^2 s^2_{\text{true}} + K^2 \sigma^2_{\text{obs}} \\ &= (1-K)^2 s^2_{\text{true}} + K^2 \sigma^2_{\text{obs}}. \end{aligned} \tag{19}$$

The first term is the squared *bias* (prior shrinkage); the second is the *variance* (noise amplification). Substituting $K = \pi/(\pi + \pi^{(0)})$ and differentiating with respect to π:

$$\frac{d\,\text{MSE}}{d\pi} = \frac{2\pi^{(0)}}{(\pi + \pi^{(0)})^3}\left[\pi\,\sigma^2_{\text{obs}} - (1-K)\,s^2_{\text{true}}\right]. \tag{20}$$

Setting to zero and solving yields Eq. (18). □

Remark 1. *This result explains why the NPD group's high π_{cog} does not straightforwardly produce higher empathic accuracy than HC in the simulation. For $\sigma_{obs} = 0.5$, $\pi_s^{(0)} = 1$, and $|s_{true}| \approx 0.7$, the optimal gain is $K^* \approx 0.74$, corresponding to $\pi^* \approx 2.9 \approx \pi_{cog}^{HC}$. The NPD value $\pi_{cog} = 5.0$ overshoots the optimum, incurring excess noise amplification.*

4.3 NPD vs. HC Kalman Gains

Table 2 lists the Kalman gains for both groups at the parameter values used in the main simulation.

Table 2: Kalman gains for HC and NPD groups ($\pi_s^{(0)} = \pi_a^{(0)} = 1$).

Group	π_{cog}	K_{cog}
HC	3.0	$3.0/4.0 = 0.750$
NPD	5.0	$5.0/6.0 = 0.833$
	π_{aff}	K_{aff}
HC	3.0	$3.0/4.0 = 0.750$
NPD	0.6	$0.6/1.6 = 0.375$

The critical asymmetry is in the affective channel: NPD's $K_{\text{aff}} = 0.375$ is exactly half that of HC's $K_{\text{aff}} = 0.750$. All affective posterior estimates are scaled down by a factor of ≈ 0.5, reducing both the magnitude and the informativeness of affective resonance.

5 Self-Model Dynamics

5.1 Cumulative Threat Accumulator

The cumulative threat load $C(t)$ is a leaky accumulator that integrates the history of ego-threatening social signals:

$$C(t+1) = \begin{cases} C(t) + \delta_{\uparrow} & \text{if stimulus type at trial } t \text{ is ego-threatening,} \\ \max\bigl(0,\ C(t) - \delta_{\downarrow}\bigr) & \text{otherwise.} \end{cases} \tag{21}$$

With $\delta_{\uparrow} = 1.0$ and $\delta_{\downarrow} = 0.5$, the accumulator grows twice as fast under threat as it decays under non-threat. This asymmetry models the clinically documented negativity bias in self-relevant social processing: critical feedback is registered more persistently than positive feedback [Rosenbach and Renneberg, 2021].

Equilibrium analysis. Let the proportion of threatening trials be p_{threat}. In expectation, the steady-state threat load satisfies:

$$\mathbb{E}[C^*] = \frac{p_{\text{threat}}\,\delta_{\uparrow}}{(1 - p_{\text{threat}})\,\delta_{\downarrow}}, \tag{22}$$

obtained by setting $\mathbb{E}[\Delta C] = 0$. For the normal stimulus distribution ($p_{\text{threat}} = 0.30$): $\mathbb{E}[C^*] = 0.3 \times 1.0/(0.7 \times 0.5) = 0.857$.

During the injury block ($p_{\text{threat}} = 0.90$): $\mathbb{E}[C^*] = 0.9 \times 1.0/(0.1 \times 0.5) = 18.0$.

This jump in C^* by a factor of 21 drives the precision collapse described in the next section.

5.2 Precision Erosion

The self-model precision decays exponentially with the cumulative threat load:

$$\Pi_{\text{self}}(t) = \Pi_{\text{self}}^{(0)} \cdot \exp(-\gamma\, C(t))\,, \qquad \Pi_{\text{self}}(t) \geq \Pi_{\text{self,min}} > 0, \tag{23}$$

where γ is the erosion rate and $\Pi_{\text{self,min}}$ is a floor preventing numerical degeneracy (set to 0.05 in the simulation).

Motivation. This form is motivated by several considerations:

1. **Exponential decay as a natural smoothed threshold.** At low C, $\Pi_{\text{self}}(t) \approx \Pi_{\text{self}}^{(0)}$ (negligible change), but as C grows, Π_{self} drops precipitously. This captures the clinical observation that narcissistic defences hold up against isolated criticisms but erode under sustained challenge.

2. **Proportionality to initial precision.** Two individuals with the same γ but different $\Pi_{\text{self}}^{(0)}$ will erode to the same fraction of their baseline — preserving the relative precision ordering while both collapsing under sufficient threat.

3. **Analogy to tolerance development.** The form parallels tolerance development in the Löschner et al. model, but here applied to the self-belief hierarchy rather than to social reinforcement.

Half-life of self-model precision. The threat load required to halve Π_{self} is:

$$C_{1/2} = \frac{\ln 2}{\gamma} \approx \frac{0.693}{\gamma}. \tag{24}$$

With $\gamma = 0.05$: $C_{1/2} = 13.9$ threatening trial units. During the 30-trial injury block with $p_{\text{threat}} = 0.90$, the expected peak $C \approx 27$ (Section 5), placing the system at $\Pi_{\text{self}} \approx \Pi_{\text{self}}^{(0)} \cdot e^{-1.35} \approx 0.26\,\Pi_{\text{self}}^{(0)}$. For NPD ($\Pi_{\text{self}}^{(0)} = 9$), this yields $\Pi_{\text{self}} \approx 2.3$ — comparable to the HC baseline of 2.0.

5.3 Self-Esteem Update Rule

At each trial t, the posterior self-esteem is computed via Eq. (16):

$$S(t+1) = \frac{\Pi_{\text{self}}(t) \cdot S(t) + \pi_{\text{fb}} \cdot f(t)}{\Pi_{\text{self}}(t) + \pi_{\text{fb}}} = S(t) + K_{\text{fb}}(t)\,\big[f(t) - S(t)\big], \tag{25}$$

where the time-varying feedback gain is:

$$K_{\text{fb}}(t) = \frac{\pi_{\text{fb}}}{\Pi_{\text{self}}(t) + \pi_{\text{fb}}}. \tag{26}$$

Eq. (25) has the form of a *prediction-error learning rule*: the self-esteem update is proportional to the self-referential prediction error $\delta_t = f(t) - S(t)$, scaled by a gain $K_{\text{fb}}(t)$ that itself varies with the dynamic precision.

Regime analysis.

- **Grandiose regime** (C low, $\Pi_{\text{self}}(t) \approx \Pi_{\text{self}}^{(0)} \gg 1$): $K_{\text{fb}} \approx \pi_{\text{fb}}/\Pi_{\text{self}}^{(0)} \ll 1$. Self-esteem updates are negligible; S stays near $\mu_S^{(0)}$. For NPD: $K_{\text{fb}}^{\text{normal}} = 1/(9+1) = 0.10$.
- **Collapse regime** (C large, $\Pi_{\text{self}}(t) \to 0$): $K_{\text{fb}} \to \pi_{\text{fb}}/\pi_{\text{fb}} = 1$. Self-esteem tracks the feedback perfectly; $S \to f$. Under sustained negative feedback ($f \approx -1$), S plunges.
- **HC normal regime**: $K_{\text{fb}}^{\text{HC}} = 1/(2+1) = 0.33$. Moderate, continuous updating around the environmental feedback mean.

5.4 Convergence and Fixed Points

Proposition 3 (Fixed points of the self-esteem dynamics)**.** *In the absence of precision dynamics ($\Pi_{self}(t) = const.$), the self-esteem trajectory converges to a fixed point:*

$$S^* = \frac{\Pi_{self}\,\mu_S^{(0)} + \pi_{fb}\,\bar{f}}{\Pi_{self} + \pi_{fb}}, \tag{27}$$

where $\bar{f} = \mathbb{E}[f(t)]$ is the expected feedback. Convergence is geometric with rate $1 - K_{fb}$.

Proof. Write $S(t{+}1) = (1{-}K_{\text{fb}})\, S(t){+}K_{\text{fb}}\, f(t)$. Taking expectations and noting $\mathbb{E}[f(t)] = \bar{f}$: $\mathbb{E}[S(t{+}1)] = (1{-}K_{\text{fb}})\, \mathbb{E}[S(t)]{+}K_{\text{fb}}\, \bar{f}$. The fixed point satisfies $S^* = (1{-}K_{\text{fb}})\, S^*{+}K_{\text{fb}}\, \bar{f}$, giving $S^* = \bar{f}$. Including the prior: $S^* = \big(\Pi_{\text{self}}\, \mu_S^{(0)} + \pi_{\text{fb}}\, \bar{f}\big)/(\Pi_{\text{self}} + \pi_{\text{fb}})$. Geometric convergence follows from the contraction factor $(1 - K_{\text{fb}}) \in (0, 1)$. □

Corollary 1. *For the normal stimulus distribution ($\bar{f} = 0.4(1) + 0.3(0) + 0.3(-1) = 0.1$):*

$$S^*_{HC} = \frac{2.0 \times 0 + 1 \times 0.1}{2.0 + 1} = 0.033, \tag{28}$$

$$S^*_{NPD} = \frac{9.0 \times 1.5 + 1 \times 0.1}{9.0 + 1} = 1.360. \tag{29}$$

Thus, in the absence of injury, NPD maintains a self-esteem near 1.36 (substantially above environmental feedback), while HC converges to ≈ 0.033.

6 Empathy Metrics: Formal Definitions and Properties

6.1 Empathic Accuracy

Definition 1 (Empathic accuracy)**.**

$$\mathcal{A}(t) = 1 - \frac{|\hat{\mu}_s(t) - s^{\text{true}}_{\text{other}}(t)|}{s_{\max}}, \quad s_{\max} = 1.5, \tag{30}$$

where $s^{\text{true}}_{\text{other}}$ is the true (ground-truth) latent mental state and $s_{\max}$ is a normalisation constant equal to the maximum expected $|s^{\text{true}}_{\text{other}}|$.

Expected accuracy. Substituting $\hat{\mu}_s = K_{\text{cog}}\, o_{\text{cog}}$ and $o_{\text{cog}} = s_{\text{other}} + \varepsilon$:

$$\mathbb{E}[\mathcal{A}] = 1 - \frac{\mathbb{E}[|K_{\text{cog}}(s_{\text{other}} + \varepsilon) - s_{\text{other}}|]}{s_{\max}} = 1 - \frac{\mathbb{E}[|(K-1)s_{\text{other}} + K\varepsilon|]}{s_{\max}}. \tag{31}$$

For $s_{\text{other}} \sim \mathcal{N}(\mu^{\text{true}}_s, \sigma^2_s)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$, the numerator is the mean absolute error of the posterior, which is bounded as:

$$\mathbb{E}[|\hat{\mu}_s - s_{\text{other}}|] \le \sqrt{\text{MSE}(K_{\text{cog}})} = \sqrt{(1 - K_{\text{cog}})^2 s^2_{\text{true}} + K^2_{\text{cog}}\, \sigma^2_{\text{obs}}}, \tag{32}$$

by Jensen's inequality. This is the expression analysed in Proposition 2 above.

6.2 Empathic Concern

Definition 2 (Empathic concern)**.**

$$\mathcal{C}(t) = K_{\text{aff}} \cdot |\hat{\mu}_a(t)| = K^2_{\text{aff}} \cdot |o_{\text{aff}}(t)|, \tag{33}$$

where the second equality uses $\hat{\mu}_a = K_{\text{aff}} \cdot o_{\text{aff}}$.

Remark 2. *$\mathcal{C}(t)$ scales quadratically with K_{aff} (and hence with π_{aff} for fixed prior precision). This means small changes in affective precision can have disproportionate effects on concern at low π_{aff} values, as visible in the non-linearity of Figure 5A of the main paper.*

Expected concern under emotional stimuli. For an ego-threatening stimulus: $a_{\text{other}}^{\text{true}} \sim \mathcal{N}(-0.6, 0.0625)$, $o_{\text{aff}} = a_{\text{other}} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.25)$. Then $o_{\text{aff}} \sim \mathcal{N}(-0.6, 0.3125)$ and:

$$\mathbb{E}[\mathcal{C}] = K_{\text{aff}}^2 \cdot \mathbb{E}[|o_{\text{aff}}|] = K_{\text{aff}}^2 \cdot \sqrt{\frac{2}{\pi}} \cdot \sigma_{o_{\text{aff}}} \cdot e^{-\mu_{o_{\text{aff}}}^2/(2\sigma_{o_{\text{aff}}}^2)} + K_{\text{aff}}^2 \cdot |\mu_{o_{\text{aff}}}| \cdot \text{erf}\left(\frac{|\mu_{o_{\text{aff}}}|}{\sqrt{2}\,\sigma_{o_{\text{aff}}}}\right), \quad (34)$$

by the half-normal mean formula. For HC ($K_{\text{aff}} = 0.75$): $\mathbb{E}[\mathcal{C}] \approx 0.34$. For NPD ($K_{\text{aff}} = 0.375$): $\mathbb{E}[\mathcal{C}] \approx 0.085$. These analytical values agree with the simulation means to within $< 1\%$.

6.3 Empathic Imbalance Index

Definition 3 (Empathic imbalance index)**.**

$$\mathcal{E}_{\text{imb}} = \overline{\mathcal{A}} - \overline{\mathcal{C}}, \quad (35)$$

where overbars denote time-averaged (session-level) values.

Proposition 4 (Asymptotic scaling of $\mathcal{E}_{\text{imb}}$)**.** *As $\pi_{cog} \to \infty$ and $\pi_{aff} \to 0$ with fixed $\pi_s^{(0)}, \pi_a^{(0)}$:*

$$\mathcal{E}_{imb} \;\to\; 1 - 0 = 1 \quad \textit{(maximum dissociation).} \quad (36)$$

As $\pi_{cog} = \pi_{aff} = \pi$ (balanced streams): $\mathcal{E}_{imb} \to 0$ by symmetry.

The NPD precision ratio used in the simulation is $\pi_{\text{cog}}/\pi_{\text{aff}} = 5.0/0.6 \approx 8.3$, placing the model in the steeply-rising regime of $\mathcal{E}_{\text{imb}}$ (Figure 5C of the main paper), consistent with the large effect size ($d = 24.8$) observed for empathic concern.

7 Free-Energy Landscape and Bistability

7.1 Effective Potential for the Self-Model

From Eq. (15), the free energy of the self-model at the optimum posterior precision ($\tilde{\pi}_S = \Pi_{\text{self}} + \pi_{\text{fb}}$) takes the form:

$$\mathcal{F}_{\text{self}}^*(S) = \frac{1}{2}\left[\Pi_{\text{self}}(t)\,(S - \mu_S)^2 + \pi_{\text{fb}}\,(f - S)^2\right] + \text{const.} \quad (37)$$

This is a quadratic (single-well) potential with minimum at S^* given by Eq. (16). The curvature (stiffness) of the well is $\Pi_{\text{self}} + \pi_{\text{fb}}$: higher total precision $\to$ sharper well.

7.2 Emergent Bistability via Precision Dynamics

The potential $\mathcal{F}_{\text{self}}^*(S)$ is instantaneously quadratic (single-well), but the *position* and *depth* of the well evolve over time as $\Pi_{\text{self}}(t)$ changes. The effective time-averaged landscape over the session integrates these instantaneous potentials:

$$\mathcal{V}(S) \approx \int_0^T \mathcal{F}^*_{\text{self}}(S;t)\,dt = \frac{1}{2}\int_0^T \left[\Pi_{\text{self}}(t)\,(S-\mu_S)^2 + \pi_{\text{fb}}\,(f(t)-S)^2\right]dt. \tag{38}$$

Under the two-regime approximation (normal regime: trials $0 \le t < t_{\text{inj}}$; injury regime: $t_{\text{inj}} \le t < t_{\text{inj}} + T_{\text{inj}}$):

$$\begin{aligned}\mathcal{V}(S) \approx{}& \frac{t_{\text{inj}}}{2}\left[\Pi^{(0)}_{\text{self}}(S-\mu_S)^2 + \pi_{\text{fb}}(0.1-S)^2\right] \\ &+ \frac{T_{\text{inj}}}{2}\left[\Pi^{(c)}_{\text{self}}(S-\mu_S)^2 + \pi_{\text{fb}}(-0.9-S)^2\right],\end{aligned} \tag{39}$$

where $\Pi^{(c)}_{\text{self}} = \Pi^{(0)}_{\text{self}}\, e^{-\gamma C_{\text{inj}}}$ is the collapsed precision (injury regime) and 0.1 and -0.9 are the expected feedback values in each regime.

Expanding and collecting terms, $\mathcal{V}(S)$ takes the form of a sum of two quadratic wells:

$$\mathcal{V}(S) = A_G\,(S - S^*_G)^2 + A_V\,(S - S^*_V)^2 + \text{const.}, \tag{40}$$

with well centres:

$$S^*_G = \frac{\Pi^{(0)}_{\text{self}}\,\mu_S + \pi_{\text{fb}}\cdot 0.1}{\Pi^{(0)}_{\text{self}} + \pi_{\text{fb}}} \approx 1.36 \quad \text{(grandiose)}, \tag{41}$$

$$S^*_V = \frac{\Pi^{(c)}_{\text{self}}\,\mu_S + \pi_{\text{fb}}\cdot(-0.9)}{\Pi^{(c)}_{\text{self}} + \pi_{\text{fb}}} \approx -0.5 \quad \text{(vulnerable)}, \tag{42}$$

and well depths $A_G \propto t_{\text{inj}}$, $A_V \propto T_{\text{inj}}$. This is the formal basis for the double-well landscape shown in Figure 4B of the main paper.

Remark 3. *In the HC case ($\Pi^{(0)}_{self} = 2.0 \approx \Pi^{(c)}_{self}$ after injury), both wells are shallow and the landscape is nearly symmetric. The integrated trajectory therefore shows modest perturbation during injury with easy return to baseline — consistent with the HC simulation results.*

7.3 Phase Transition Criterion

The grandiose–to–vulnerable transition occurs when the system's trajectory moves from the S^*_G basin into the S^*_V basin. This happens when $\Pi_{\text{self}}(t)$ drops sufficiently that the feedback potential dominates the prior potential. Formally, the self-esteem velocity $\dot{S} = K_{\text{fb}}(t)\,[f(t) - S(t)]$ reverses sign (from positive toward S^*_G to negative toward S^*_V) at a critical precision:

$$\Pi^{\text{crit}}_{\text{self}} = \pi_{\text{fb}} \cdot \frac{S - \bar{f}_{\text{inj}}}{S^*_G - S} \cdot \frac{S^*_G - S}{S - \bar{f}_{\text{inj}}} \quad \Rightarrow \quad \Pi^{\text{crit}}_{\text{self}} = \pi_{\text{fb}} \cdot \frac{S^*_G}{\bar{f}_{\text{inj}} - S^*_G}, \tag{43}$$

where $\bar{f}_{\text{inj}} \approx -0.9$ is the mean feedback during injury. For the NPD parameters: $\Pi^{\text{crit}}_{\text{self}} = 1 \times 1.36/(-0.9 - 1.36) \approx -0.6$, which is negative — confirming that for these parameters the system is always driven toward S^*_V once Π_{self} erodes, regardless of how low it gets.

8 Stimulus Generation and Paradigm Design

8.1 Stimulus Type Assignment

At each trial t, a stimulus type $\omega_t \in \{0, 1, 2\}$ (enhancing, neutral, threatening) is drawn from a categorical distribution:

$$\omega_t \sim \begin{cases} \text{Cat}(0.40, 0.30, 0.30) & \text{if } t \notin [t_{\text{inj}},\, t_{\text{inj}} + T_{\text{inj}}), \\ \text{Cat}(0.00, 0.10, 0.90) & \text{if } t \in [t_{\text{inj}},\, t_{\text{inj}} + T_{\text{inj}}). \end{cases} \tag{44}$$

8.2 Stimulus-Conditional Distributions

Given ω_t, the true other-states and feedback are drawn as:

$$\begin{pmatrix} s_{\text{true}} \\ a_{\text{true}} \\ f \end{pmatrix} \Bigg| \, \omega_t \sim \mathcal{N}\left(\begin{pmatrix} \mu_s(\omega_t) \\ \mu_a(\omega_t) \\ \mu_f(\omega_t) \end{pmatrix}, \begin{pmatrix} 0.0625 & 0 & 0 \\ 0 & 0.0625 & 0 \\ 0 & 0 & 0.09 \end{pmatrix} \right), \tag{45}$$

where the conditional means are:

$$(\mu_s, \mu_a, \mu_f) = \begin{cases} (+0.70,\ +0.60,\ +1.00) & \omega = 0 \text{ (enhancing)}, \\ (\ 0.00,\ \ 0.00,\ \ 0.00) & \omega = 1 \text{ (neutral)}, \\ (-0.70,\ -0.60,\ -1.00) & \omega = 2 \text{ (threatening)}. \end{cases} \tag{46}$$

The stimuli are independent across trials (i.i.d. conditional on type).

8.3 Narcissistic Injury Block

The injury block ($t_{\text{inj}} = 150$, $T_{\text{inj}} = 30$) concentrates threatening stimuli with probability 0.90. The expected cumulative threat load at the end of the block (starting from $C(t_{\text{inj}}) \approx 0.86$ at equilibrium) can be derived from the recurrence Eq. (21):

$$\begin{aligned} \mathbb{E}[C(t_{\text{inj}} + k)] &= C(t_{\text{inj}}) + k\, p_{\text{threat}}\, \delta_{\uparrow} - k\, (1 - p_{\text{threat}})\, \delta_{\downarrow} \\ &\approx 0.86 + k\, (0.9 \times 1.0 - 0.1 \times 0.5) \\ &= 0.86 + 0.85\, k. \end{aligned} \tag{47}$$

After $k = 30$ injury trials: $\mathbb{E}[C(t_{\text{inj}} + 30)] \approx 26.4$, which gives $\Pi_{\text{self}} \approx 9 \cdot e^{-0.05 \times 26.4} = 9 \times 0.265 = 2.39$ for NPD — consistent with the direct simulation trajectories.

9 Simulation Algorithm

Algorithm 1 presents the complete pseudocode for a single agent simulation.

Computational complexity. The per-trial computation is $\mathcal{O}(1)$; total complexity for n agents over T trials is $\mathcal{O}(nT)$. For $n = 50$, $T = 300$, this is 15,000 iterations, completing in < 0.5 seconds on a single CPU core.

Algorithm 1: Single-agent hierarchical predictive coding simulation.

Input: Parameters: $\pi_{\text{cog}}, \pi_{\text{aff}}, \Pi^{(0)}_{\text{self}}, S_0, \pi_{\text{fb}}, \gamma, \delta_\uparrow, \delta_\downarrow, \sigma_{\text{obs}}$;
Paradigm: $T, t_{\text{inj}}, T_{\text{inj}}$; RNG seed.

Output: $\{\mathcal{A}(t)\}_{t=0}^{T-1}$, $\{\mathcal{C}(t)\}_{t=0}^{T-1}$, $\{S(t)\}_{t=0}^{T}$, $\{\Pi_{\text{self}}(t)\}_{t=0}^{T}$

Initialise: $S \leftarrow S_0$, $C \leftarrow 0$

Compute fixed Kalman gains: $K_{\text{cog}} \leftarrow \pi_{\text{cog}}/(\pi_{\text{cog}}+1)$; $K_{\text{aff}} \leftarrow \pi_{\text{aff}}/(\pi_{\text{aff}}+1)$

for $t = 0, 1, \ldots, T-1$ **do**

  Sample stimulus: $\omega_t \sim \text{Cat}(\boldsymbol{p}(t))$ per Eq. (44)

  Sample signals: $(s^{\text{true}}, a^{\text{true}}, f) \leftarrow$ Eq. (45)

  Sample observations: $o_{\text{cog}} \leftarrow s^{\text{true}} + \mathcal{N}(0, \sigma^2_{\text{obs}})$; $o_{\text{aff}} \leftarrow a^{\text{true}} + \mathcal{N}(0, \sigma^2_{\text{obs}})$

  Cognitive posterior: $\hat{\mu}_s \leftarrow K_{\text{cog}} \cdot o_{\text{cog}}$

  Affective posterior: $\hat{\mu}_a \leftarrow K_{\text{aff}} \cdot o_{\text{aff}}$

  Empathic accuracy: $\mathcal{A}(t) \leftarrow 1 - \min(|\hat{\mu}_s - s^{\text{true}}|/s_{\max},\ 1)$

  Empathic concern: $\mathcal{C}(t) \leftarrow K_{\text{aff}} \cdot |\hat{\mu}_a|$

  Update threat load: if $\omega_t = 2$ **then**

    $C \leftarrow C + \delta_\uparrow$

  else

    $C \leftarrow \max(0,\ C - \delta_\downarrow)$

  end

  Dynamic precision: $\Pi_{\text{self}}(t) \leftarrow \max\left(\Pi^{(0)}_{\text{self}} \cdot e^{-\gamma C},\ \Pi_{\text{self,min}}\right)$

  Self-esteem update: $K_{\text{fb}} \leftarrow \pi_{\text{fb}}/(\Pi_{\text{self}}(t) + \pi_{\text{fb}})$

  $S \leftarrow (1 - K_{\text{fb}})\, S + K_{\text{fb}}\, f$

end

10 Connection to the Free Energy Principle

10.1 Precision as Attention in Active Inference

In active inference, precision weights are not fixed external parameters but are themselves inferred (meta-Bayesian inference) or modulated by action [Friston, 2010]. Precision can be understood as the attention allocated to a given inference channel: high precision → strong attention to that signal source; low precision → inattention.

In the present model, the precision asymmetry $\pi_{\text{cog}} \gg \pi_{\text{aff}}$ represents a stable attentional policy in the NPD agent: cognitive social signals receive high attention (and are integrated strongly), while affective signals are systematically discounted. This is consistent with the active inference interpretation of psychopathology as pathological prior beliefs about the precision of sensory channels.

10.2 Variational Bayes vs. Sampling

The present simulation implements *variational Bayes* (VB) inference under a Laplace approximation. An alternative implementation would use Monte Carlo sampling. For the Gaussian conjugate case considered here, VB and exact Bayesian inference coincide exactly

(the Laplace approximation is exact for Gaussian likelihoods and priors), so the simulation results do not depend on the approximation quality.

10.3 Extended Model: Active Inference with Policy Selection

A full active inference implementation would supplement the perceptual model with a *policy selection* module that chooses actions to minimise expected free energy (EFE) over future time steps:

$$G(\pi_{\text{policy}}) = \mathbb{E}_{\pi}\left[\text{KL}[q(o_\tau) \parallel p(o_\tau)] - \mathbb{E}_{q(z_\tau)}[\ln q(z_\tau) - \ln p(z_\tau)]\right], \tag{48}$$

where the first term captures epistemic value (information gain) and the second captures pragmatic value (preference satisfaction). For NPD, the inflated self-concept $S_G^* \gg 0$ functions as a strong *prior preference*: the agent selects policies (social behaviours: admiration-seeking, rivalry, devaluation of others) that maximise the probability of self-enhancing social feedback, minimising EFE. Modelling these policy selection dynamics would require extending the current framework to include an action repertoire and a model of how the agent's actions influence the social environment — a natural direction for future work.

11 Parameter Rationale and Identifiability

11.1 Parameter Rationale

Table 3 provides the clinical and theoretical justification for each parameter value used in the simulation.

11.2 Identifiability

The model has five group-level parameters per group: $\pi_{\text{cog}}, \pi_{\text{aff}}, \Pi_{\text{self}}^{(0)}, S_0, \gamma$ (with $\pi_{\text{fb}}, \sigma_{\text{obs}}$ shared and fixed). In principle, all five parameters are locally identifiable from empirical data given sufficient observations, because each parameter governs a distinct observational signature:

- π_{cog}: controls empathic accuracy level.
- π_{aff}: controls empathic concern level (independently of accuracy).
- S_0: controls baseline self-esteem level.
- $\Pi_{\text{self}}^{(0)}$: controls the rate of self-esteem change under mild feedback.
- γ: controls the slope of the self-esteem collapse under sustained threat.

Global identifiability would require formal analysis (e.g., via the rank of the Fisher information matrix evaluated over the parameter support), which is left for future work with empirical data.

Table 3: Parameter rationale for the NPD group.

Parameter	Value	Rationale
$\pi_{\text{cog}} = 5.0$	higher than HC	Narcissistic hypervigilance to social status cues; enhanced neural responsivity in ToM networks (mPFC, TPJ) documented by Ritter et al. (2011) and Jauk et al. (2024).
$\pi_{\text{aff}} = 0.6$	much lower than HC	Systematic impairment of affective resonance: attenuated autonomic and neural responses to others' distress. Set at 1/5 of HC value to reproduce the $\approx 4\times$ reduction in empathic concern.
$\Pi_{\text{self}}^{(0)} = 9.0$	much higher than HC	Rigid, high-confidence self-belief; low updating on negative feedback under normal conditions. Ratio $\Pi_{\text{self}}^{\text{NPD}}/\Pi_{\text{self}}^{\text{HC}} = 4.5$ chosen to yield $K_{\text{fb}}^{\text{normal}} \approx 0.1$.
$S_0 = 1.5$	inflated	Above-zero self-esteem prior modelling grandiose self-appraisal; units are normalised to the feedback scale ($f \in [-1, 1]$).
$\gamma = 0.05$	shared	Erosion rate set to place the precision half-life at $C_{1/2} = 13.9$, requiring ≈ 14 consecutive threatening trials to halve Π_{self}.

12 Analytical Predictions for Experimental Validation

12.1 EEG/ERP Predictions

From Proposition 1, the magnitude of affective prediction errors — indexed by the N200/P300 complex in event-related potential studies — should scale as:

$$\text{ERN amplitude} \propto K_{\text{aff}} \cdot |\delta_{\text{aff}}| = \frac{\pi_{\text{aff}}}{\pi_{\text{aff}} + 1} \cdot \left| a_{\text{other}}^{\text{true}} - \hat{\mu}_a \right|, \tag{49}$$

predicting that NPD patients should show $\sim$50% reduction in affective ERP amplitudes relative to HC.

12.2 Prediction for Empathy–Accuracy Trade-off

From Proposition 2, there exists an optimal Kalman gain $K^* < 1$. Empirically, this predicts an inverted-U relationship between self-reported trait narcissism (proxy for π_{cog}) and empathic accuracy in noisy (high σ_{obs}) social environments, but a monotone positive relationship in low-noise environments. This is testable via adapting existing empathic accuracy paradigms to manipulate signal-to-noise ratio.

12.3 Prediction for Self-Esteem Kurtosis

The precision dynamics model predicts that NPD patients' self-esteem distributions measured via ecological momentary assessment (EMA) should exhibit high *kurtosis* (leptokurtic) rather

than high variance:

$$\mathrm{Kurt}(S) > \mathrm{Kurt}_{\mathrm{Normal}} = 3, \quad \text{with } \sigma(S)^{\mathrm{NPD}} \leq \sigma(S)^{\mathrm{HC}}, \tag{50}$$

reflecting the plateau–collapse–recovery pattern of the precision dynamics. This is a qualitative prediction that distinguishes precision-dissociation dynamics from simple homeostatic or emotion-regulation models of self-esteem.

References

References

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

Jauk, E., Kanske, P., Seidel, E.-M., and Steyrl, D. (2024). Psychological and neural correlates of social affect and cognition in narcissism: a multimethod study. *Personality Disorders: Theory, Research, and Treatment*, 15(2), 157–171.

Krizan, Z. and Herlache, A. D. (2018). The narcissism spectrum model. *Personality and Social Psychology Review*, 22(1), 3–31.

Löschner, D. M. et al. (2025). A computational framework to study the etiology of grandiose narcissism. *Scientific Reports*, 15, 5897.

Ritter, K. et al. (2011). Lack of empathy in patients with narcissistic personality disorder. *Psychiatry Research*, 187(1–2), 241–247.

Rosenbach, C. and Renneberg, B. (2021). Positive future thinking and self-esteem stabilization in narcissistic personality disorder. *Personality and Individual Differences*, 180, 110994.