Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Precision Dissociation in Hierarchical Predictive Coding: A Computational Principle Illustrated in Narcissistic Personality Disorder

Marco Fronzi

marco.fronzi@gmail.com

The University of Sydney

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11449501/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Precision Dissociation in Hierarchical Predictive Coding: A Computational Principle Illustrated in Narcissistic Personality Disorder

Marco Fronzi

School of Physics, The University of Sydney (Australia)

marco.fronzi@gmail.com

October 21, 2026

Abstract

We introduce *precision dissociation* as a general computational principle in hierarchical predictive coding, and illustrate its consequences in the context of Narcissistic Personality Disorder (NPD). The principle holds that opposing precision assignments across parallel inference streams produce characteristic dissociations in the behaviours those streams govern — without any deficit in the underlying generative model capacity. Applied to NPD, the principle predicts two clinically prominent but mechanistically unexplained features of the disorder: (i) intact cognitive empathy (Theory of Mind, mentalising) co-occurring with substantially impaired affective empathy (emotional resonance, empathic concern); and (ii) episodic, disproportionate self-esteem collapse following ego-threatening events. We formalise the framework as a three-level Bayesian hierarchy in which the affective inference stream carries systematically lower precision (π_{aff}) than the cognitive stream (π_{cog}), while the self-model carries a high-precision, inflated prior ($\Pi^{(0)}_{\text{self}}$). Under sustained social threat this self-model precision erodes exponentially, shifting the free-energy landscape from a deep-grandiose to a deep-vulnerable attractor basin — the formal correlate of narcissistic collapse. Proof-of-concept simulations across two-group agent ensembles ($n = 50$ per group, 300-trial social paradigm with embedded narcissistic injury block) are consistent with the predicted dissociation: empathic accuracy is near-equivalent between groups (HC $= 0.783 \pm 0.010$ vs. NPD $= 0.771 \pm 0.010$, absolute difference $< 2\%$), whereas empathic concern is approximately fourfold lower in the NPD regime (HC $= 0.346 \pm 0.014$ vs. NPD $= 0.086 \pm 0.004$). Self-esteem collapse during the injury block is $\sim$86% larger in the NPD regime than in controls. Phase-space analysis reveals bistability in the (Π_{self}, S) plane, recasting the grandiose–vulnerable oscillation as a precision-driven attractor bifurcation. The framework generates falsifiable predictions for electrophysiology, neuroimaging, and experience-sampling studies, and suggests therapeutic leverage points distinct from those implied by symptom-level accounts.

Keywords: precision weighting; hierarchical predictive coding; free energy principle; narcissistic personality disorder; empathy dissociation; attractor dynamics; computational psychiatry; active inference

1 Introduction

A recurring question in computational psychiatry is whether mental disorders reflect impairments in the capacity for inference — the generative model itself — or, instead, in the weights assigned to competing inference channels. The Free Energy Principle (FEP) and active inference [Friston, 2010, Friston et al., 2017] provide a formal language for this distinction: *precision* (inverse variance) weights determine how strongly each level of a predictive hierarchy gates prediction errors upward. Pathological precision allocation, rather than structural incapacity, has been proposed as the computational substrate of autism [Pellicano and Burr, 2012], psychosis [Adams et al., 2013], and personality disorder [Badcock et al., 2019, Fronzi, 2025]. The present paper develops a specific instance of this idea — *precision dissociation* across parallel inference streams — and shows that it captures two cardinal features of Narcissistic Personality Disorder (NPD).

The first feature is an empathic paradox. Individuals with NPD frequently demonstrate sophisticated social cognition — accurate decoding of others' intentions, status-tracking, and strategic mentalising — yet show substantially reduced affective resonance: lower empathic concern, emotional disengagement, and limited responsiveness to others' distress [Baskin-Sommers et al., 2014, Ritter et al., 2011]. This dissociation between largely intact *cognitive* empathy and impaired *affective* empathy is among the most replicated findings in the NPD literature [Wai and Tiliopoulos, 2012, Cascio et al., 2023, Dziobek et al., 2011], but no formal mechanism has been proposed to explain why the two components should diverge so markedly.

The second feature is temporal instability. Patients typically present in either a grandiose state (arrogant, entitled, exploitative) or a vulnerable state (hypersensitive, shame-laden, fragile), and clinical observation has long noted oscillation between these poles [Gore and Widiger, 2016, Krizan and Herlache, 2018]. The grandiose–vulnerable spectrum has been described dimensionally [Krizan and Herlache, 2018], but the dynamical mechanism that drives state transitions — particularly the sudden self-esteem collapse following narcissistic injury — remains unexplained at the computational level.

Recent formal approaches to NPD include agent-based models of self-esteem regulation [Löschner et al., 2025] and predictive models of social exclusion [Jiao, 2026]. The Löschner et al. [2025] model identifies tolerance development and rivalry cycles as disorder-maintaining mechanisms, framing parental overvaluation as the key etiological input. While this work captures important behavioural dynamics, it does not formalise the Bayesian inferential machinery governing social perception, does not model empathic dissociation, and provides no account of the collapse as a precision-level event. The present paper aims to address these gaps.

Contribution and positioning. We propose precision dissociation as a general principle in hierarchical predictive coding, and illustrate its application to NPD. The core theoretical claim is not specific to NPD: *any* system in which two parallel inference streams are assigned systematically different precisions will exhibit the corresponding dissociation in stream-governed behaviours, regardless of the generative model's structural completeness. NPD provides a clinically rich and empirically constrained test case for this principle. The framework: (i) unifies the empathic dissociation and self-esteem collapse within one mechanistic account; (ii) yields quantitative, falsifiable predictions; (iii) maps naturally onto known neurobiological circuits; and (iv) is analytically tractable without parameter fitting to empirical data — appropriate for a Level-1 theoretical proof of concept.

2 Theoretical Framework

2.1 Hierarchical Generative Model

We consider a three-level hierarchical generative model of social interaction. At the highest level an agent maintains a *self-model* $S(t)$ representing current self-esteem. At the intermediate level the agent infers the mental and affective states of an interaction partner via two parallel streams. At the lowest level noisy observations o of the social environment are received.

Level 3 — Self-model.

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

where μ_S is the current self-esteem estimate and Π_{self} is the self-model precision. In the NPD regime: elevated $\mu_S^{(0)}$ (inflated prior mean) and high $\Pi_{\text{self}}^{(0)}$ (rigid, tightly held self-belief).

Level 2 — Social inference streams. Two parallel channels infer properties of the social partner:

$$s_{\text{other}} \sim \mathcal{N}\big(0,\, (\pi_{\text{cog}})^{-1}\big)\,, \tag{2}$$

$$a_{\text{other}} \sim \mathcal{N}\big(0,\, (\pi_{\text{aff}})^{-1}\big)\,, \tag{3}$$

where s_{other} is the inferred mental/intentional state (ToM channel) and a_{other} is the inferred affective/valence state (resonance channel).

Level 1 — Observations. Social observations are corrupted by additive Gaussian noise:

$$o_{\text{cog}} = s_{\text{other}} + \varepsilon_{\text{cog}}, \quad \varepsilon_{\text{cog}} \sim \mathcal{N}(0, \sigma_{\text{obs}}^2), \tag{4}$$

$$o_{\text{aff}} = a_{\text{other}} + \varepsilon_{\text{aff}}, \quad \varepsilon_{\text{aff}} \sim \mathcal{N}(0, \sigma_{\text{obs}}^2). \tag{5}$$

2.2 Precision-Weighted Bayesian Inference

Posterior inference in each stream follows the standard conjugate Gaussian update (single-step Kalman filter):

$$\hat{\mu}_s = K_{\text{cog}}\, o_{\text{cog}}, \quad K_{\text{cog}} = \frac{\pi_{\text{cog}}}{\pi_{\text{cog}} + \pi_s^{(0)}}, \tag{6}$$

$$\hat{\mu}_a = K_{\text{aff}}\, o_{\text{aff}}, \quad K_{\text{aff}} = \frac{\pi_{\text{aff}}}{\pi_{\text{aff}} + \pi_a^{(0)}}, \tag{7}$$

where $\pi_s^{(0)} = \pi_a^{(0)} = 1$ are the baseline prior precisions of the social inference streams (fixed at 1 for both groups), distinct from σ_{obs} which enters the observation likelihood (Eqs. 4–5) but does not appear in the gain directly because the gain is computed over the prior-to-stream precision ratio, not the raw observation noise. The Kalman gain $K \in [0, 1]$ governs the weight placed on the incoming observation relative to the prior.

Precision dissociation and the Kalman gains. The NPD regime is defined by $\pi_{\text{cog}} \gg \pi_{\text{aff}}$, which yields $K_{\text{cog}} \gg K_{\text{aff}}$. This is the formal expression of the principle: the generative models for the two streams are structurally identical; only their precision assignments differ. The consequence is that the cognitive stream integrates social signals strongly ($K_{\text{cog}} \approx 0.83$ for the NPD parameters used below), while the affective stream integrates them weakly ($K_{\text{aff}} \approx 0.375$). There is no structural impairment of the affective stream; rather, its signals are selectively downweighted.

Empathic accuracy.

$$\mathcal{A}(t) = 1 - \frac{|\hat{\mu}_s(t) - s_{\text{true}}(t)|}{s_{\max}}, \tag{8}$$

where $s_{\max}$ is a normalisation constant. Note that above the optimal Kalman gain, further precision increase amplifies observation noise rather than signal, so $\mathcal{A}$ is not monotone in π_{cog} (see Supplementary Information for the MSE decomposition).

Empathic concern.

$$\mathcal{C}(t) = K_{\text{aff}}\,|\hat{\mu}_a(t)|, \tag{9}$$

When $\pi_{\text{aff}} \to 0$, $K_{\text{aff}} \to 0$ and concern is suppressed irrespective of incoming signal strength. Critically, this is a precision effect, not a signal-processing deficit.

Empathic imbalance index.

$$\mathcal{E}_{\text{imb}} = \overline{\mathcal{A}} - \overline{\mathcal{C}}, \tag{10}$$

where overbars denote session averages. In the HC regime ($\pi_{\text{cog}} \approx \pi_{\text{aff}}$), $\mathcal{E}_{\text{imb}} \approx 0$. In the NPD regime ($\pi_{\text{cog}} \gg \pi_{\text{aff}}$), $\mathcal{E}_{\text{imb}} > 0$, quantifying the dissociation severity.

2.3 Self-Model Dynamics and Narcissistic Collapse

Cumulative threat and dynamic precision. Social interactions generate self-relevant feedback $f(t)$. We model cumulative ego-threat as a leaky accumulator:

$$C(t{+}1) = \begin{cases} C(t) + \delta_{\uparrow} & \text{ego-threatening trial,} \\ \max\bigl(0,\, C(t) - \delta_{\downarrow}\bigr) & \text{otherwise,} \end{cases} \tag{11}$$

with $\delta_{\uparrow} > \delta_{\downarrow}$, modelling the asymmetric registration of criticism vs. praise [Rosenbach and Renneberg, 2021]. Self-model precision then erodes exponentially:

$$\Pi_{\text{self}}(t) = \Pi_{\text{self}}^{(0)} \cdot e^{-\gamma C(t)}, \tag{12}$$

where $\gamma > 0$ is the erosion rate. This form encodes the clinical observation that isolated criticisms leave the narcissistic defence intact, whereas sustained social threat eventually destabilises it.

Self-esteem update.

$$S(t{+}1) = \frac{\Pi_{\text{self}}(t) \cdot S(t) + \pi_{\text{fb}} \cdot f(t)}{\Pi_{\text{self}}(t) + \pi_{\text{fb}}}, \tag{13}$$

where π_{fb} is the precision assigned to social feedback. When $\Pi_{\text{self}}(t)$ is large the prior dominates ($K_{\text{fb}} \approx 0.10$ for the NPD parameters), conferring grandiose rigidity. As $\Pi_{\text{self}}(t)$ erodes, K_{fb} increases and S converges rapidly toward the (negative) feedback mean — producing the abrupt collapse dynamics.

2.4 Free-Energy Landscape Interpretation

Under a Laplace approximation, the variational free energy of the self-model reduces to [Friston, 2007]:

$$\mathcal{F} = \tfrac{1}{2}\Big[\Pi_{\text{self}}(S - \mu_S^*)^2 + \pi_{\text{fb}}(f - S)^2\Big] + \text{const.} \tag{14}$$

This defines an effective potential landscape $\mathcal{V}(S)$ with two regimes:

- **Grandiose basin**: Π_{self} large; deep well at $\mu_S^{(0)} > 0$. Perturbations are rapidly restored.
- **Vulnerable basin**: $\Pi_{\text{self}} \to 0$; minimum shifts to $\mathbb{E}[f] < 0$ under sustained threat. The feedback channel dominates.

The grandiose-to-vulnerable transition is therefore a *precision-driven bifurcation* — a shift in attractor basin driven not by changes in the generative model structure but by the erosion of a single precision parameter (Fig. 4).

3 Methods

3.1 Simulation Design

We simulated two agent ensembles ($n = 50$ per group; HC and NPD regimes) over $T = 300$ trials. Each trial presents a social signal drawn from one of three categories: *self-enhancing* ($p = 0.40$), *neutral* ($p = 0.30$), or *ego-threatening* ($p = 0.30$), with latent states $s_{\text{true}} \sim \mathcal{N}(\pm 0.7,\, 0.0625)$ and feedback $f \sim \mathcal{N}(\pm 1.0,\, 0.09)$.

A **narcissistic injury block** (trials 150–179, $p_{\text{threat}} = 0.90$) models sustained social rejection or status loss, followed by 120 recovery trials under the normal distribution.

Parameter values. Table 1 lists all parameters. NPD values were set to satisfy three qualitative constraints derived from the clinical literature: $\pi_{\text{cog}}^{\text{NPD}} > \pi_{\text{cog}}^{\text{HC}}$ (hypervigilant status monitoring); $\pi_{\text{aff}}^{\text{NPD}} \ll \pi_{\text{aff}}^{\text{HC}}$ (impaired affective resonance); and $\Pi_{\text{self}}^{\text{NPD}} \gg \Pi_{\text{self}}^{\text{HC}}$ with $S_0 > 0$ (rigid, inflated self-concept). All parameters are illustrative; their calibration against empirical data is a natural next step.

Table 1: Model parameters (HC and NPD regimes).

Symbol	Description	HC	NPD
π_{cog}	Cognitive precision	3.0	5.0
π_{aff}	Affective precision	3.0	0.6
$\Pi_{\text{self}}^{(0)}$	Baseline self-precision	2.0	9.0
S_0	Initial self-esteem	0.0	1.5
π_{fb}	Feedback precision	1.0	1.0
γ	Precision erosion rate	0.05	0.05
$\delta_{\uparrow}$	Threat accumulation	1.0	1.0
$\delta_{\downarrow}$	Threat recovery	0.5	0.5
σ_{obs}	Observation noise	0.5	0.5

3.2 Outcome Measures

Four primary outcome measures were computed:

1. **Empathic accuracy** $\mathcal{A}(t)$ (Eq. 8).
2. **Empathic concern** $\mathcal{C}(t)$ (Eq. 9).
3. **Self-esteem volatility** $\sigma(S)$: standard deviation of $S(t)$ across all 300 trials.
4. **Collapse magnitude** ΔS: pre-injury mean minus minimum S during the injury block.

Session-mean scalars were compared between groups using two-sided Mann–Whitney U tests. Because $\mathcal{C}$ is a near-deterministic function of the fixed Kalman gain K_{aff} (which differs by a factor of exactly two between groups), the resulting Cohen's d is very large ($d \approx 25$), reflecting a near-zero within-group variance for a gain-determined metric rather than a meaningful between-subjects effect size. We therefore characterise the concern difference as a *fold-reduction* (HC/NPD ratio ≈ 4) and report d for completeness with this caveat. All simulations used Python 3.11 (NumPy, SciPy, Matplotlib).

3.3 Parameter Sensitivity Analysis

Three sensitivity sweeps were performed: (i) $\pi_{\text{aff}} \in [0.1, 6.0]$ at fixed $\pi_{\text{cog}} = 5.0$; (ii) $\Pi^{(0)}_{\text{self}} \in [0.5, 12.0]$; (iii) the precision ratio $\pi_{\text{cog}}/\pi_{\text{aff}} \in [0.5, 12]$. Each sweep used $n = 20$ agents over $T = 100$ trials with no injury block.

4 Results

4.1 Precision Dissociation Architecture

Figure 1 illustrates the hierarchical architecture. The NPD and HC regimes share identical generative model structures; they differ solely in their precision assignments. The cognitive stream Kalman gain is modestly higher in the NPD regime ($K^{\text{NPD}}_{\text{cog}} = 0.833$ vs. $K^{\text{HC}}_{\text{cog}} = 0.750$), while the affective stream gain is sharply lower ($K^{\text{NPD}}_{\text{aff}} = 0.375$ vs. $K^{\text{HC}}_{\text{aff}} = 0.750$). This is the formal expression of precision dissociation.

4.2 Empathic Dissociation

Figure 2 presents the empathy results. Empathic accuracy was near-equivalent across the session: HC $= 0.783 \pm 0.010$, NPD $= 0.771 \pm 0.010$ (Mann–Whitney $U = 1984$, $p = 4.3 \times 10^{-7}$, $d = 1.21$; absolute difference 0.012 on a $[0, 1]$ scale). This small absolute difference is consistent with the empirical literature showing cognitive empathy to be largely preserved in NPD [Baskin-Sommers et al., 2014, Ritter et al., 2011].

Empathic concern differed markedly: HC $= 0.346 \pm 0.014$ vs. NPD $= 0.086 \pm 0.004$ ($U = 2500$, $p = 7.1 \times 10^{-18}$; fold-reduction ≈ 4). As noted in Section 3, the large Cohen's d (≈ 25) for this metric arises because concern is a near-deterministic function of K_{aff}, which differs by a factor of two between regimes; within-group variation is consequently very small. The practically meaningful quantity is the fourfold suppression of affective signal integration.

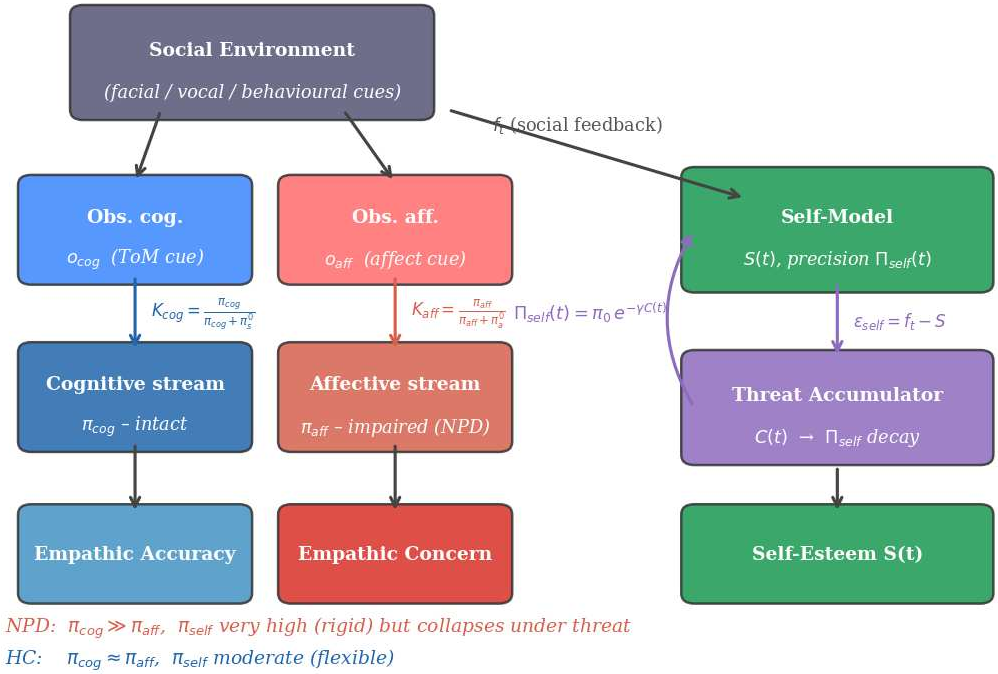


Figure 1: **Hierarchical predictive coding architecture.** Two parallel inference streams (cognitive: blue; affective: red) receive social observations with precision-weighted Kalman gains K_{cog} and K_{aff}. The self-model (green) carries a high-precision prior that erodes exponentially under cumulative ego-threat $C(t)$. Arrows indicate precision parameters modulating each inference channel. NPD and HC regimes are identical in structure; they differ only in the values of π_{cog}, π_{aff}, and $\Pi_{\text{self}}^{(0)}$.

Notably, concern suppression is uniform across the session including the injury block, consistent with the model prediction that affective resonance impairment is a stable precision parameter rather than a state-dependent response.

4.3 Self-Esteem Dynamics and Narcissistic Collapse

Figure 3 shows self-esteem trajectories. During the pre-injury period, NPD agents maintained elevated self-esteem near $S \approx 1.36$ (the Bayesian fixed point for the NPD parameters, consistent with the analytical prediction; see Supplementary Information). The feedback gain $K_{\text{fb}} \approx 0.10$ effectively insulated the self-concept from trial-by-trial feedback.

During the injury block, both groups showed self-esteem decline, but the dynamics differed qualitatively. HC agents underwent moderate, continuous updating toward the negative feedback mean. NPD agents exhibited a delayed but substantially steeper collapse: the precision erosion accumulated over ∼20 threatening trials before $\Pi_{\text{self}}(t)$ dropped sufficiently to permit large-scale updating, at which point the trajectory fell rapidly. Collapse magnitude was $\Delta S_{\text{NPD}} \approx 2.6$ vs. $\Delta S_{\text{HC}} \approx 1.4$ — approximately 86% larger — despite identical stimulus sequences. This models the clinical observation that narcissistic injury produces affective responses disproportionate to the apparent trigger.

Post-injury recovery (trials 180–299) showed gradual return toward the pre-injury baseline as $C(t)$ decayed and $\Pi_{\text{self}}(t)$ partially recovered, consistent with clinical observations of self-esteem restoration following withdrawal from threatening environments [Rosenbach and Renneberg, 2021].

The self-model precision trajectory (Fig. 3B) makes the mechanism explicit: NPD agents begin with $\Pi_{\text{self}}^{(0)} = 9$ (rigid self-model), which erodes to $\Pi_{\text{self}} \approx 2.4$ at the injury block nadir — roughly the HC baseline — before partially recovering.

4.4 Phase Portrait and Free-Energy Landscape

Figure 4 presents the phase-space analysis. In the (Π_{self}, S) plane, HC agents orbit a compact region near ($\Pi_{\text{self}} \approx 2.0$, $S \approx 0.1$) throughout the session. NPD agents trace a characteristic loop: elevated Π_{self} and S during the normal regime (grandiose basin); rapid joint excursion to low Π_{self}, low S during the injury block (vulnerable basin); then gradual recovery. This loop structure is not present in HC, confirming that the transition is a qualitative dynamical event rather than a quantitative shift along a shared trajectory.

The free-energy landscape (Fig. 4B) illustrates the bifurcation analytically. The NPD normal regime has a deep well centred on the grandiose attractor. As Π_{self} erodes under injury, the well at $\mu_S^{(0)}$ shallows while the well near $\mathbb{E}[f] < 0$ deepens, until the vulnerable basin becomes globally attracting. The HC landscape has two comparable, shallow wells throughout, consistent with flexible self-concept updating. This analysis suggests that the grandiose and vulnerable presentations are not merely opposite ends of a self-esteem continuum but distinct attractor basins separated by a precision-determined potential barrier.

4.5 Parameter Sensitivity

Figure 5 characterises model behaviour across parameter ranges. Empathic concern increases monotonically with π_{aff} (panel A), confirming this as the primary determinant of affective empathy impairment; the relationship is approximately linear for $\pi_{\text{aff}} \in [0.1, 3.0]$ and saturates above the HC value.

Self-esteem volatility is non-monotonic in Π_{self} (panel B): maximum volatility occurs near $\Pi_{\text{self}} \approx$ 2–3, with both very low (self-esteem tracks feedback closely) and very high (rigid self-concept) precision producing reduced $\sigma(S)$. This predicts that subclinical narcissism — corresponding to intermediate Π_{self} — may produce the *highest* moment-to-moment self-esteem variability on experience-sampling measures, while clinical NPD (high Π_{self}) shows lower day-to-day variance but larger collapse events.

The empathic imbalance index $\mathcal{E}_{\text{imb}}$ is monotone in the precision ratio $\pi_{\text{cog}}/\pi_{\text{aff}}$ (panel C), approaching zero when streams are balanced (as in HC) and rising steeply above a ratio of ≈3. The NPD regime (ratio ≈ 8.3) lies in the steeply rising region, consistent with the observation that modest clinical changes in either precision parameter could produce substantial changes in empathic profile.

5 Discussion

5.1 Theoretical Contributions

Precision dissociation as a general principle. The central contribution of this paper is conceptual: we show that *opposing precision assignments across parallel inference streams* are sufficient to generate the characteristic dissociations observed in NPD, without positing any structural impairment of the generative model. This principle generalises beyond NPD. Any clinical phenotype characterised by selective preservation of one inferential faculty alongside impairment of another — and there are many in psychiatry and neurology — is a candidate for a precision-dissociation account. The present work illustrates the principle in one well-constrained case.

Mechanistic unification. Previous computational accounts of NPD have addressed self-esteem regulation [Löschner et al., 2025] or social cognition [Cascio et al., 2023] separately. The precision-dissociation framework illustrates that the same parameter configuration ($\pi_{\text{cog}} \gg \pi_{\text{aff}}$, $\Pi^{(0)}_{\text{self}} \gg 1$) that produces empathic dissociation also produces self-esteem collapse dynamics. This parsimony is a theoretical strength of FEP-based approaches [Friston et al., 2017].

The grandiose–vulnerable oscillation as a bifurcation. Krizan and Herlache's [2018] spectrum model describes the grandiose–vulnerable continuum verbally. Our phase-space analysis casts this as a bifurcation problem: the two presentations are distinct attractor basins in the (Π_{self}, S) plane, separated by a potential barrier whose height is controlled by $\Pi^{(0)}_{\text{self}}$ and the threat erosion dynamics. This translates the clinical observation of abrupt state transitions into a precise dynamical prediction: transitions should be preceded by an accumulation of threatening social signals exceeding a threshold, rather than by any single precipitating event. This is testable with ecological momentary assessment (EMA) data, where a leptokurtic (high-kurtosis, low-variance) self-esteem distribution would be consistent with the plateau-and-collapse dynamics.

Reinterpreting preserved cognitive empathy. A precision-weighted analysis reveals that high π_{cog} does not straightforwardly improve empathic accuracy: above the optimal Kalman gain, additional precision amplifies observation noise rather than signal (see the MSE decomposition in Supplementary Information). The near-equivalence of accuracy between groups (absolute difference $< 2\%$) is therefore not a finding against the model but a prediction of it: both groups operate near the optimal gain, with the NPD regime slightly overshooting. This further predicts that the apparent preservation of cognitive empathy may be context-dependent — in high-noise social environments (ambiguous facial expressions, masked speech), the NPD regime may paradoxically show *lower* accuracy due to noise amplification. This is testable using existing empathic accuracy paradigms with graded signal-to-noise manipulation.

Precision suppression vs. capacity deficit. The model formalises a distinction highlighted in clinical reviews [Baskin-Sommers et al., 2014]: NPD individuals receive and process affective social signals (observations o_{aff} arrive with identical fidelity in both groups) but assign them low precision, discounting them from influencing affect and behaviour. This is mechanistically distinct from a capacity deficit and aligns with the empirical observation that NPD individuals can appear empathic when motivated to do so. In active inference terms, low π_{aff} is equivalent to a strong prior that others' affective states are not relevant to self-directed goals.

5.2 Neurobiological Predictions

The precision-dissociation framework yields four specific, falsifiable predictions.

P1 (EEG/ERP). Affective prediction error signals — indexed by the N200/P300 complex and sympathetic skin conductance response to others' distress — should be attenuated in NPD relative to controls, proportional to the reduction in K_{aff}. Cognitive ToM-related ERP components (e.g., N400 in mental-state inference tasks) should be comparable or enhanced, consistent with elevated K_{cog}.

P2 (Neuroimaging). Structural and functional asymmetries between the ventromedial prefrontal cortex – anterior insula – amygdala affective resonance circuit and the dorsomedial PFC – temporoparietal junction ToM circuit should predict empathic imbalance $\mathcal{E}_{\text{imb}}$ at the individual level. Reduced connectivity in the affective resonance circuit with preserved or enhanced ToM circuit connectivity is predicted for the NPD group.

P3 (Injury paradigm). Under sustained negative social feedback, NPD patients should show a delayed but steep amygdala and anterior insula response, with a time course consistent with exponential precision erosion: modest initial response, rapid escalation after a threshold of cumulative threat is reached.

P4 (EMA). Self-esteem distributions collected via experience sampling should exhibit higher kurtosis in NPD relative to subclinical narcissism and HC, reflecting the plateau-and-collapse dynamics — even if overall variance is not elevated. This distinguishes the precision model from homeostatic or emotion-regulation accounts that predict uniform variance differences.

5.3 Clinical Implications

The precision-dissociation framework suggests therapeutic leverage points that differ from symptom-level accounts.

First, interventions that increase π_{aff} — the weight assigned to affective social signals — should produce approximately linear improvements in empathic concern (Fig. 5A). Mentalization-Based Treatment [Bateman and Fonagy, 2004] can be understood in this framework as increasing π_{aff} through sustained directed attention to affective cues in the therapeutic relationship, making explicit what the NPD patient's precision hierarchy renders implicit.

Second, the phase-space analysis identifies self-precision *flexibility* — the parameter γ in Eq. (12) — as a more tractable therapeutic target than baseline grandiosity ($\Pi_{\text{self}}^{(0)}$). Interventions that build adaptive self-model flexibility, such as those building tolerance for self-esteem fluctuation without collapse into the vulnerable basin, would raise the potential barrier between attractors. This corresponds therapeutically to developing a more feedback-responsive self-concept rather than directly reducing grandiosity.

Third, if cholinergic or noradrenergic systems modulate precision in predictive coding hierarchies [Friston, 2010, Parr and Friston, 2019], pharmacological agents acting on these systems could in principle simultaneously affect both π_{aff} (empathic concern) and Π_{self} dynamics (self-esteem stability). The model makes explicit predictions about the direction and coupling of such effects.

5.4 Limitations and Future Directions

This simulation constitutes a Level-1 proof of concept. It illustrates the theoretical principle and generates falsifiable predictions, but does not involve parameter fitting to empirical data, and several important extensions remain.

Empirical calibration. The parameter values in Table 1 are illustrative. Calibrating them against task-based data — empathic accuracy paradigms, social feedback learning tasks, or EMA time series — would require hierarchical Bayesian estimation (e.g., via the Hierarchical Gaussian Filter [Mathys et al., 2011] or DCM) applied to NPD samples.

Scalar self-model. Clinical NPD involves heterogeneous self-concept dimensions (achievement, attractiveness, social dominance) that may carry different precision weights. A vector self-model with domain-specific precisions would capture this heterogeneity.

Interpersonal dynamics. The current model treats social stimuli as exogenous. A two-agent active inference formulation, in which agents select actions (admiration-seeking, rivalry, devaluation) to minimise expected free energy, would capture the dyadic dynamics central to NPD pathology [Löschner et al., 2025].

Vulnerable NPD subtype. The present NPD parameter set models the grandiose presentation. Vulnerable NPD may correspond to high π_{aff} combined with high π_{cog} and low $\Pi^{(0)}_{\text{self}}$ — affective hypervigilance and intact cold mentalising with a fragile self-model — a precision dissociation in the *opposite direction* to the grandiose profile, yet producible within the same framework. This parametric contrast between grandiose and vulnerable subtypes is itself a testable prediction of the model.

6 Conclusion

We have proposed precision dissociation as a general computational principle in hierarchical predictive coding, and illustrated its application to Narcissistic Personality Disorder. Within a three-level Bayesian hierarchy, opposing precision assignments across the cognitive and affective empathy streams — with a high-precision, inflated self-model — are jointly sufficient to produce the disorder's cardinal features: preserved cognitive empathy, suppressed affective empathy, grandiose self-esteem rigidity, and precipitous collapse under sustained social threat. Phase-space analysis recasts the grandiose–vulnerable oscillation as a precision-driven bifurcation in the free-energy landscape. The framework is parsimonious, analytically tractable, and generates quantitative predictions testable with existing experimental and clinical methods. We suggest that the precision-dissociation principle extends beyond NPD and may offer a unifying computational language for a broader class of clinical dissociations across parallel inference hierarchies.

Data and Code Availability

All simulation code and figure-generation scripts are implemented in Python 3.11 (NumPy, SciPy, Matplotlib) and are available from the author upon request. No empirical human data were used in this study.

Conflict of Interest

The author declares no conflicts of interest.

Author Contributions

M.F.: Conceptualisation, formal analysis, methodology, software, writing (original draft, revision, and editing), visualisation.

Funding

The author declares that no funding was received for conducting this study.

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, 4, 47.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., and Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, and Behavioral Neuroscience*, 19(6), 1319–1351.

Baskin-Sommers, A., Krusemark, E., and Ronningstam, E. (2014). Empathy in narcissistic personality disorder: From clinical and empirical perspectives. *Personality Disorders: Theory, Research, and Treatment*, 5(3), 323–333.

Bateman, A. and Fonagy, P. (2004). *Psychotherapy for Borderline Personality Disorder: Mentalization-Based Treatment*. Oxford University Press.

Cascio, M. T., Carcione, A., Nicolo, G., and Semerari, A. (2023). The dark side of empathy in narcissistic personality disorder. *Frontiers in Psychiatry*, 14, 1074558.

Dziobek, I., Preissler, S., Grozdanovic, Z., Heuser, I., Heekeren, H. R., and Roepke, S. (2011). Neuronal correlates of altered empathy and social cognition in borderline personality disorder. *NeuroImage*, 57(2), 539–548.

Friston, K. J. (2007). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815–836.

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

Friston, K. J., Wiese, W., and Hobson, J. A. (2017). Sentience and the free energy principle. *Physics of Life Reviews*, 33, 103–113.

Fronzi, M. (2025). Hierarchical interoceptive–social inference dysregulation in Cluster B personality disorders: a computational psychiatry framework and active inference simulation. *Cognitive Computation*, in press.

Gore, W. L. and Widiger, T. A. (2016). Fluctuation between grandiose and vulnerable narcissism. *Personality Disorders: Theory, Research, and Treatment*, 7(4), 363–371.

Jiao, B. (2026). Bidirectional predictive modeling of narcissists' social exclusion and hostile interpretations: a deep-learning exploration. *Frontiers in Psychology*, 17, 1761090.

Krizan, Z. and Herlache, A. D. (2018). The narcissism spectrum model: a synthetic view of narcissistic personality. *Personality and Social Psychology Review*, 22(1), 3–31.

Löschner, D. M., Schoemann, M., Jauk, E., Herchenhahn, L., Schwöbel, S., Kanske, P., and Scherbaum, S. (2025). A computational framework to study the etiology of grandiose narcissism. *Scientific Reports*, 15, 5897.

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5, 39.

Parr, T. and Friston, K. J. (2019). Generalised free energy and active inference. *Biological Cybernetics*, 113(5–6), 495–513.

Pellicano, E. and Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, 16(10), 504–510.

Ritter, K., Dziobek, I., Preissler, S., Rüter, A., Vater, A., Fydrich, T., Lammers, C.-H., Heekeren, H. R., and Roepke, S. (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.

Wai, M. and Tiliopoulos, N. (2012). The affective and cognitive empathic nature of the dark triad of personality. *Personality and Individual Differences*, 52(7), 794–799.

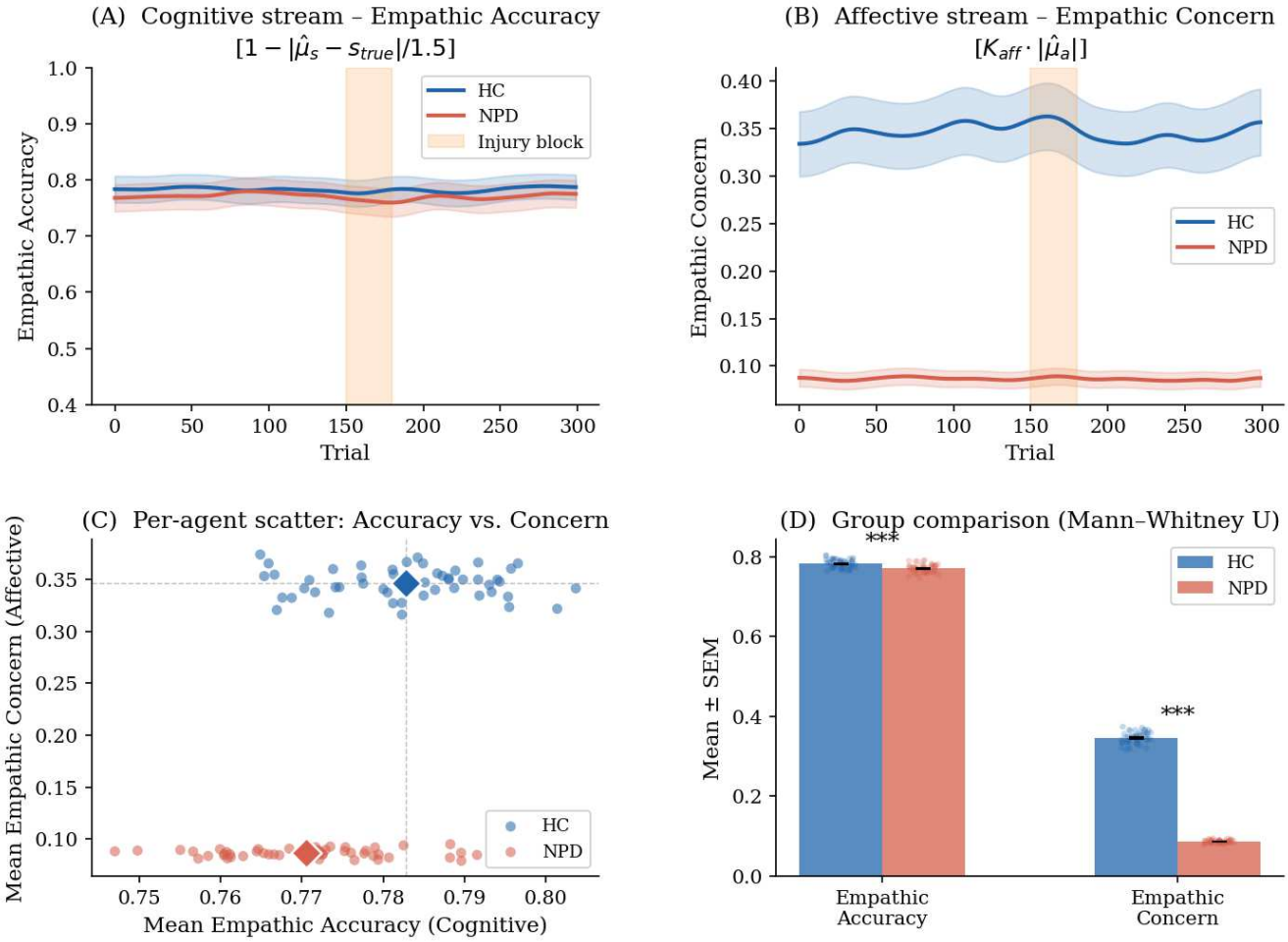


Figure 2: **Empathy dissociation in the NPD regime.** (A) Cognitive empathic accuracy is near-equivalent between groups (absolute difference $< 2\%$). (B) Affective empathic concern is approximately fourfold lower in the NPD regime throughout the session. (C) Per-agent scatter in the accuracy–concern plane: NPD agents cluster in the high-accuracy / low-concern quadrant. (D) Group comparison (Mann–Whitney U; *** $p < 0.001$; bars show mean ± SEM; dots show individual agents). Amber shading: narcissistic injury block (trials 150–179).

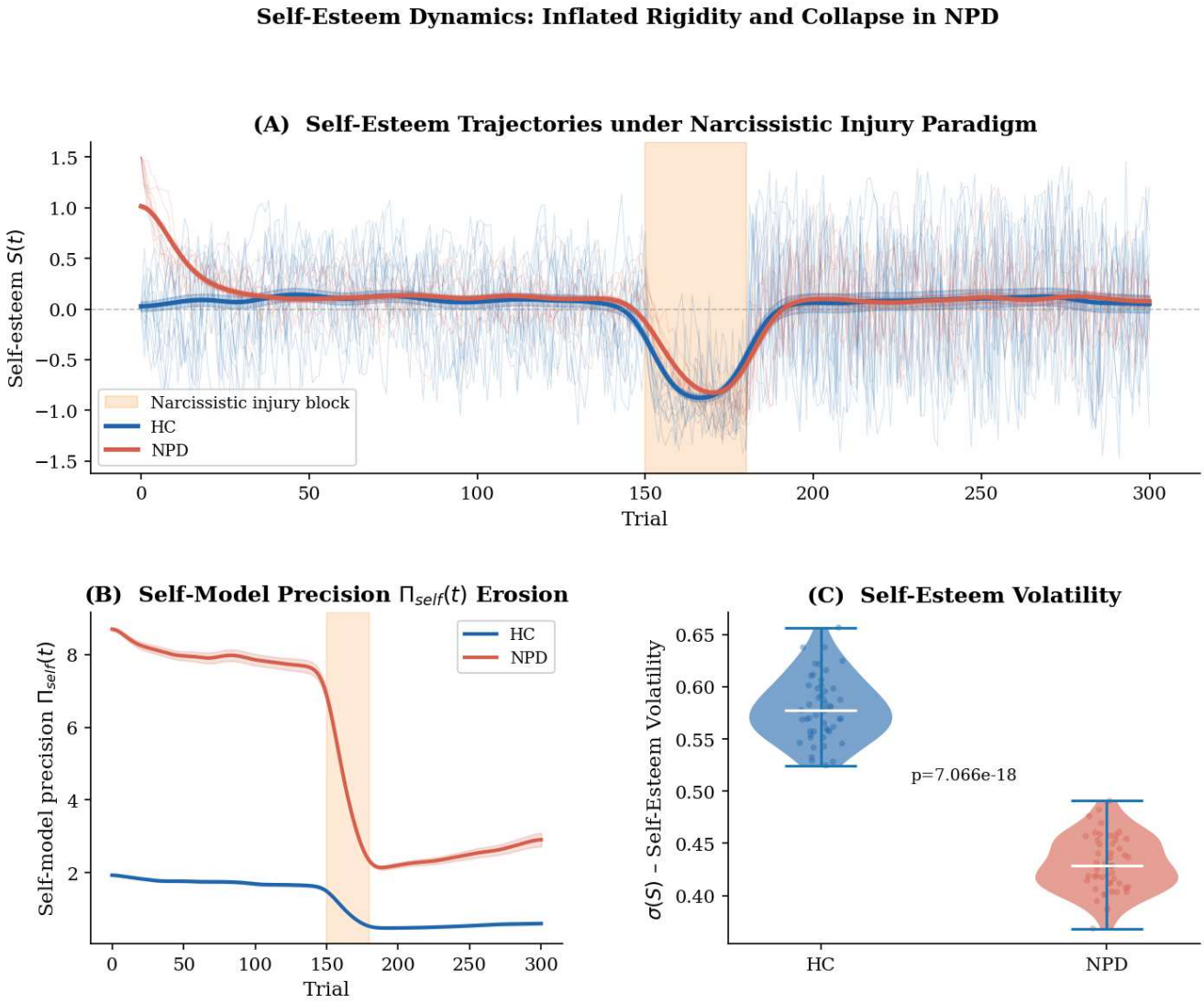


Figure 3: **Self-esteem dynamics under the narcissistic injury paradigm.** (A) Group trajectories: NPD agents maintain an elevated baseline and exhibit a larger, delayed collapse. Faint lines: 10 individual agents per group. (B) Dynamic self-model precision $\Pi_{\text{self}}(t)$: sharp erosion during the injury block drives the transition from grandiose to vulnerable dynamics. (C) Self-esteem volatility violin plots: HC shows higher total variance (continuous updating); NPD shows lower variance punctuated by larger collapse amplitude.

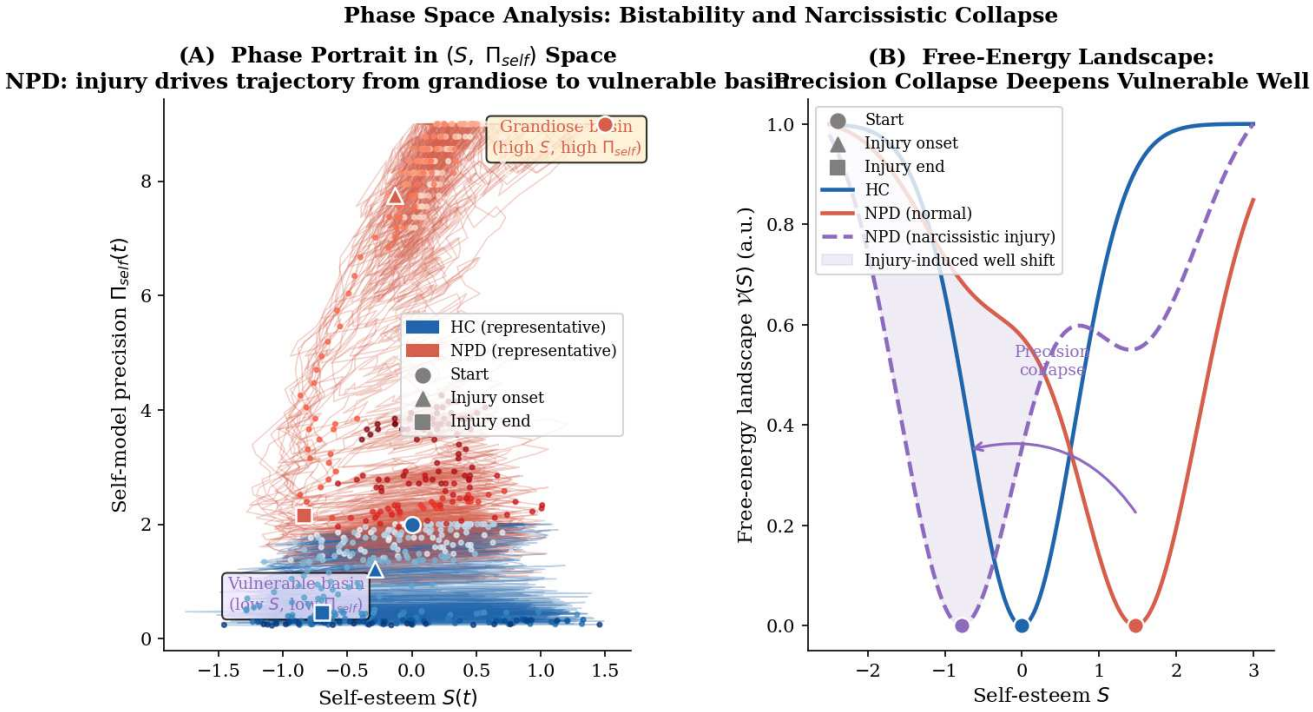


Figure 4: **Phase portrait and free-energy landscape.** (A) Agent trajectories in (Π_{self}, S) phase space. NPD traces a wide loop from the grandiose basin (upper-right) to the vulnerable basin (lower-left) during injury and partially recovers. HC orbits a compact region near ($\Pi_{\text{self}} = 2.0$, $S = 0$). Markers indicate session start (circle), injury onset (triangle), and injury end (square). (B) Free-energy potential $\mathcal{V}(S)$ under three conditions: HC (symmetric wells); NPD normal (deep grandiose well); NPD injured (precision collapse deepens the vulnerable well, driving attractor switching). Purple shading indicates the injury-induced potential shift.

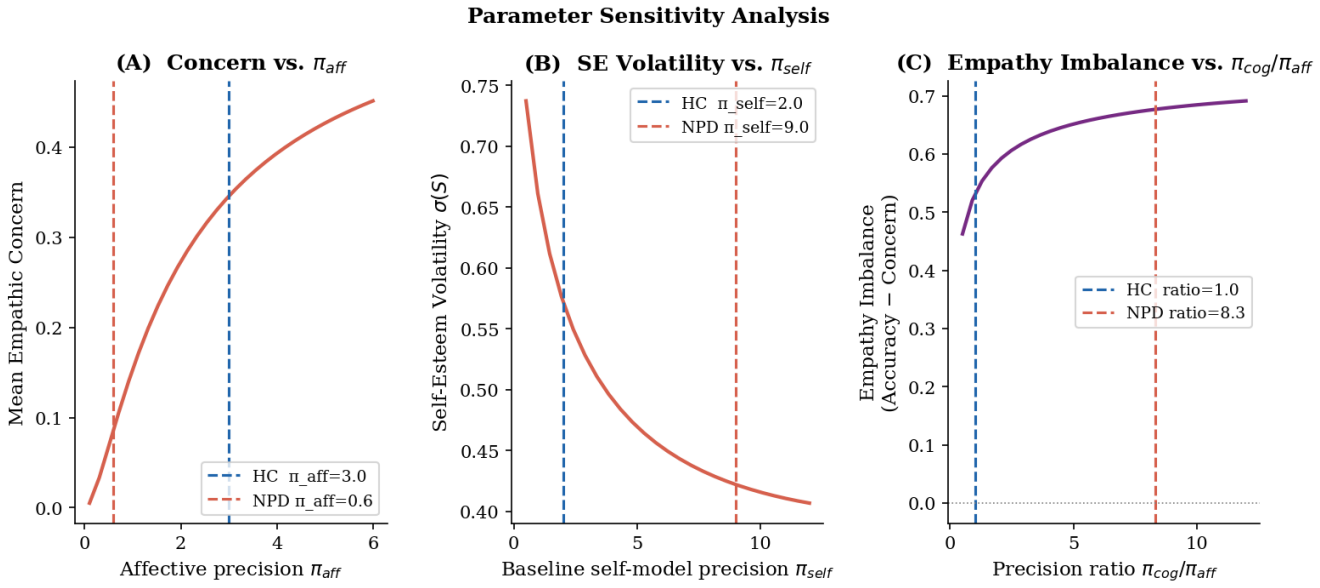


Figure 5: **Parameter sensitivity analysis.** (A) Empathic concern vs. affective precision π_{aff}. (B) Self-esteem volatility $\sigma(S)$ vs. baseline self-precision $\Pi^{(0)}_{\text{self}}$: non-monotonic relationship with maximum at intermediate values. (C) Empathic imbalance index $\mathcal{E}_{\text{imb}}$ vs. precision ratio $\pi_{\text{cog}}/\pi_{\text{aff}}$. Dashed vertical lines mark HC and NPD parameter values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- npdsupplementary.pdf