

Research Article

Interpretability Without Full Recoverability: Formal Identifiability, Synthetic Recovery, and Reportability of Withdrawal-Side Parameters in the Continuous Moral-Social World Model

Regio Marcos Pinto Abreu Filho¹

1. Polícia Militar do Estado do Rio de Janeiro (PMERJ), Brazil

Mechanistically specified models are often treated as interpretable because each coefficient has an assigned role in the equations. That semantic assignment is necessary but not sufficient for inference: parameter interpretation also requires structural identifiability of the parameter-to-observation map and practical recovery under finite, noisy data. We formalize this distinction for the Continuous Moral-Social World Model (CMS-WM), a simulator with phenotype-state, world-state, and action-policy modules. The revised analysis makes the complete withdrawal-side policy explicit, defines the observation operator, proves exact softmax scaling and common-shift invariances, and evaluates effective withdrawal parameters in reproducible synthetic recovery experiments. In a principal batch of 30 ground-truth settings with 1,000 trials each, the ambiguity-linked coupling showed moderate rank recovery (Spearman $\rho = 0.52$; median normalized absolute error = 0.19), whereas the friction-linked coupling was weak under correlated inputs ($\rho = 0.21$; median error = 0.34). When friction was experimentally orthogonalized from ambiguity, friction recovery improved ($\rho = 0.68$; median error = 0.087), demonstrating configuration-dependent rather than intrinsic interpretability. Pressure coupling and the withdrawal threshold remained insufficiently rank-recoverable in the principal design ($\rho = 0.37$ and 0.30 , respectively), despite substantial threshold activation. In contrast, out-of-sample action-regime prediction was stable: dominant-regime agreement was 100%, median predictive L1 distance was 0.0187, and median Jensen-Shannon divergence was 7.2×10^{-5} . These findings support a hierarchical reportability framework: regime-level behavior is reportable; the ambiguity-linked parameter is cautiously reportable; friction is design-conditional; and pressure coupling and threshold should not receive substantive mechanistic interpretation under the present observation model. The

study is an internal synthetic verification of the model–observation pairing, not external validation of psychopathy or antisocial behavior.

Corresponding author: Regio Marcos Pinto de Abreu Filho, regiomarcosabreu@gmail.com

1. Introduction

Computational psychiatry seeks formal accounts of latent processes that can connect theory, behavior, and intervention ^{[1][2]}. In antisociality and psychopathy, this ambition is particularly relevant because verbal constructs such as valuation asymmetry, social learning, volatility tracking, threat, and withdrawal can otherwise remain underspecified ^{[3][4][5][6][7]}. Formalization, however, creates a second obligation: the quantities embedded in the equations must be distinguished from the quantities that can actually be inferred from observations.

A coefficient may have an unambiguous semantic role in a generative equation while remaining non-identifiable or poorly recoverable. Structural identifiability asks whether distinct parameter values can produce exactly the same observable distribution under ideal data. Practical recoverability asks whether a finite-data estimator can distinguish values with tolerable error and uncertainty. Predictive adequacy asks whether the model reproduces observables or decisions of interest. External validity asks whether the model and its inferred quantities generalize to empirical populations and settings. These are related but non-equivalent properties ^{[8][9][10][11][12][13][14][15]}.

The Continuous Moral-Social World Model (CMS-WM) was developed as a mechanistically expressive simulator coupling a phenotype state, a world state, and an action policy over cooperation, defection, and withdrawal. Withdrawal is generated by multiple overlapping drivers: an endogenous ambiguity activation, friction, pressure above a threshold, and a baseline action-value contrast. Such a decomposition is scientifically useful only to the extent that its components remain distinguishable after the simulator is composed with an observation model.

The present revision converts that concern into a fully specified, demonstrable methods study. It makes five changes. First, it replaces the ambiguous language of three "layers" with three coupled modules and reserves "state" for time-indexed state vectors. Second, it declares all variables and parameters, including the previously implicit friction covariate. Third, it numbers the equations and defines the complete parameter-to-observation map. Fourth, it provides exact invariance proofs and reproducible synthetic

recovery experiments rather than narrative assertions. Fifth, it separates parameter reportability from regime-level predictive stability.

Mechanistic construction fixes what a parameter means inside the generative model. It does not, by itself, guarantee that the parameter is uniquely identified by the observables, stably recovered under finite data, or externally valid. Interpretability is therefore a property of the model-observation-design-estimator chain, not of equation naming alone.

Central proposition

2. Conceptual and inferential framework

2.1. Four levels of interpretation

Level	Definition	Question	Required evidence
Semantic/mechanistic meaning	Role assigned by the equations and scientific theory	Does the coefficient have a defined causal or computational role in the simulator?	Formal declaration and mechanistic rationale
Structural identifiability	Uniqueness of the parameter-to-observation map under ideal data	Can two distinct parameter values yield identical observables?	Proof, symmetry analysis, differential algebra, or exact profile structure
Practical recoverability	Finite-sample estimability under a stated design and estimator	Are generating values recovered with acceptable rank, error, uncertainty, and boundary behavior?	Simulation-based calibration, recovery studies, profiles, and repeated optimization
External validity	Transport from the synthetic model to empirical systems	Do estimates predict or explain independent real-world data?	Independent datasets, measurement validation, and prospective tests

Table 1. Distinct levels of model interpretation

This hierarchy resolves a common terminological problem. A "structural parameter" is a coefficient belonging to the formal generative structure; "recoverability" is not a competing parameter class but a property of inference under a particular observation and experimental design. A parameter can therefore be structurally meaningful yet structurally non-identifiable, or structurally identifiable yet practically unrecoverable.

2.2. Equivalence-based definition

Let θ denote the model parameters, D a generated trajectory, and O the observation operator that maps the trajectory to observables $S(D)$. Two parameter settings are observationally equivalent when they induce the same distribution of observables across all admissible designs:

$$\theta \sim \theta' \text{ iff } P_{\theta}[O(D) \in B|d] = P_{\theta'}[O(D) \in B|d], \text{ for every measurable } B \text{ and admissible design } d. \quad (1)$$

A parameter functional $\psi(\theta)$ is structurally identifiable only if observational equivalence implies equality of the functional. This formulation makes the observation model explicit: identifiability can be lost even when the internal simulator is mechanically well defined.

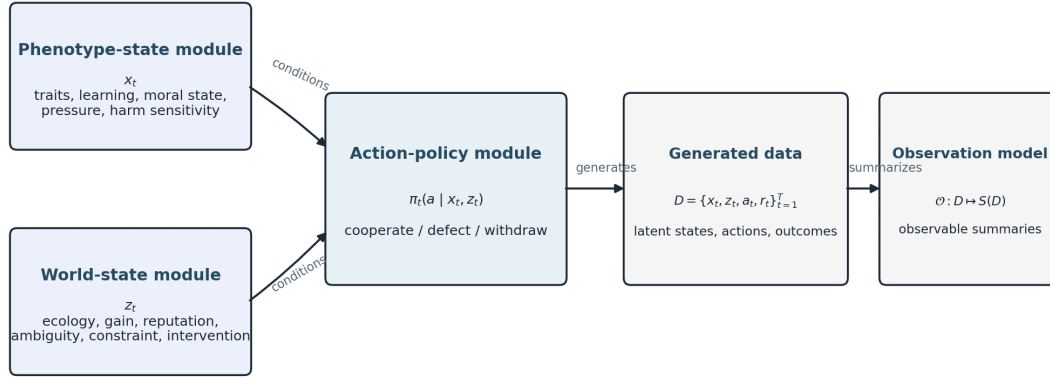
$$\psi \text{ is identifiable iff } \theta \sim \theta' \text{ implies } \psi(\theta) = \psi(\theta'). \quad (2)$$

3. Formal CMS-WM specification

3.1. Coupled modules and notation

CMS-WM is organized into a phenotype-state module, a world-state module, and an action-policy module (Figure 1). The first two contain time-indexed states. The third is a conditional decision rule, not a state. State-transition and learning equations generate trajectories; the observation operator determines which features of those trajectories enter inference.

CMS-WM: generative modules and the explicit observation model



State-transition and learning equations govern the evolution of x_t and z_t ; inferential interpretation concerns the complete parameter-to-observation map.

Mechanistic meaning is assigned in the generative model; identifiability is assessed after composition with $\mathcal{O}$.

Figure 1. CMS-WM generative modules and observation model. The distinction between the generative model and $\mathcal{O}$ is essential: parameter meaning is assigned internally, whereas identifiability is evaluated after composition with the observation map.

$$x_t = (\beta_t, \eta_t, \alpha_{\text{self}}, t, \alpha_{\text{other}+}, t, \alpha_{\text{other}-}, t, \omega_t, c_t, m_t, p_t, h_t). \quad (3)$$

The phenotype-state vector contains social-utility distortion beta, social-learning gain eta, outcome-specific learning rates, social-volatility tracking omega, controllability c, moral state m, pressure p, and harm sensitivity h. Complete declarations and computational domains are provided in Supplementary Table S1.

$$z_t = (s_t, G_t, \rho_t, A_{\text{world}}, t, L_t, \Psi_t). \quad (4)$$

The world-state vector contains social ecology s, instrumental gain opportunity G, reputation rho, world ambiguity A_{world} , legal or institutional constraint L, and intervention regime Psi. The notation A_{world} avoids conflating external ambiguity with the endogenous ambiguity activation used in the withdrawal policy.

$$A_{\text{action}} = \{C, D, W\} = \{\text{cooperate, defect, withdraw}\}. \quad (5)$$

3.2. Action policy

$$\ell_a(t) = Q_a(t) + b_a(x_t, z_t), a \in A_{\text{action}}. \quad (6)$$

$$\pi_t(a|x_t, z_t) = \exp[\ell_a(t)/\tau] / \sum_{a' \in A} \exp[\ell_{a'}(t)/\tau], \tau > 0. \quad (7)$$

$$b_C(t) = w_m, Cm_t + w_\beta, C(\beta_t - 1) + w_p, Cp_t + w_\rho, C(\rho_t - 0.5) + w_h, Ch_t. \quad (8)$$

$$b_D(t) = w_m, Dm_t + w_\beta, D(\beta_t - 1) + w_p, Dp_t + w_h, Dh_t + w_\rho, D(\rho_t - 0.5). \quad (9)$$

$$b_W(t) = w_p, W[p_t - \theta_W]_+ + w_f, Wf_t + w_A, Wg_A(m_t). \quad (10)$$

$$[u]_+ = \max(0, u), g_A(m_t) = [1 - m_t^2]_+, f_t = \varphi_f(x_t, z_t) \in [0, 1]. \quad (11)$$

Equation (11) resolves two previously implicit points. First, f_t is a derived friction covariate produced by a declared mapping ϕ_f of the current phenotype and world states. Second, the ambiguity-linked term in the withdrawal logit is endogenous moral indeterminacy $g_A(m_t)$, not the world-state ambiguity variable $A_{\text{world},t}$. This distinction prevents a semantic overclaim.

3.3. State dynamics and learning

$$m(t+1) = m(t) + \varepsilon_m [a_m m(t) - b_m m(t)^3 + u_{\text{norm}}(t) + \chi_1 C(t) - \chi_2 I(t) - \chi_3 R(t)] + \sigma_m \xi(t). \quad (12)$$

The moral state follows a stochastic forced double-well dynamic. C_t denotes cumulative exploitative success, I_t intervention load, R_t reputational or sanction pressure, and ξ_t a zero-mean stochastic perturbation. The equation can produce switching and hysteresis, but those qualitative capabilities do not guarantee that every coefficient can be inferred from compressed behavior.

$$Q_{\text{self}}(t+1) = Q_{\text{self}}(t) + \alpha_{\text{self}}(t) \delta_{\text{self}}(t). \quad (13)$$

$$Q_{\text{other}+}(t+1) = Q_{\text{other}+}(t) + \alpha_{\text{other}+}(t) \delta_{\text{other}+}(t). \quad (14)$$

$$Q_{\text{other}-}(t+1) = Q_{\text{other}-}(t) + \alpha_{\text{other}-}(t) \delta_{\text{other}-}(t). \quad (15)$$

3.4. Observation model

$$D = \{(x_t, z_t, a_t, r_t) : t = 1, \dots, T\}, O(D) = S(D) \in R^K. \quad (16)$$

The observation operator O maps a full latent trajectory to a finite summary vector. In the verification study, summaries included overall action frequencies, withdrawal frequencies across covariate strata, and targeted interaction or threshold slices. Because O is many-to-one, it can erase distinctions that exist inside the simulator. All interpretive claims in this paper are therefore conditional on the declared observation family.

4. Structural analysis

4.1. Exact global scaling invariance

For any constant $c > 0$, multiplying every logit component and the softmax temperature by c leaves the action probabilities unchanged:

$$(Q, w, \tau) \mapsto (cQ, cw, c\tau), c > 0. \quad (17)$$

$$\text{softmax}(c\ell/c\tau) = \text{softmax}(\ell/\tau). \quad (18)$$

Raw temperature and raw policy weights are therefore not separately structurally identifiable. The legitimate inferential objects are temperature-normalized contrasts, such as $\lambda_j = w_j/\tau$, together with action-value differences rather than absolute Q levels.

4.2. Common additive-logit invariance

$$\ell_a(t) \mapsto \ell_a(t) + \kappa(t), \text{ for every action } a. \quad (19)$$

$$\text{softmax}(\{\ell_a + \kappa\}_a) = \text{softmax}(\{\ell_a\}_a). \quad (20)$$

A common shift in all action logits is observationally invisible, so one intercept or action-value origin must be fixed. We use action contrasts and an effective withdrawal intercept q_W .

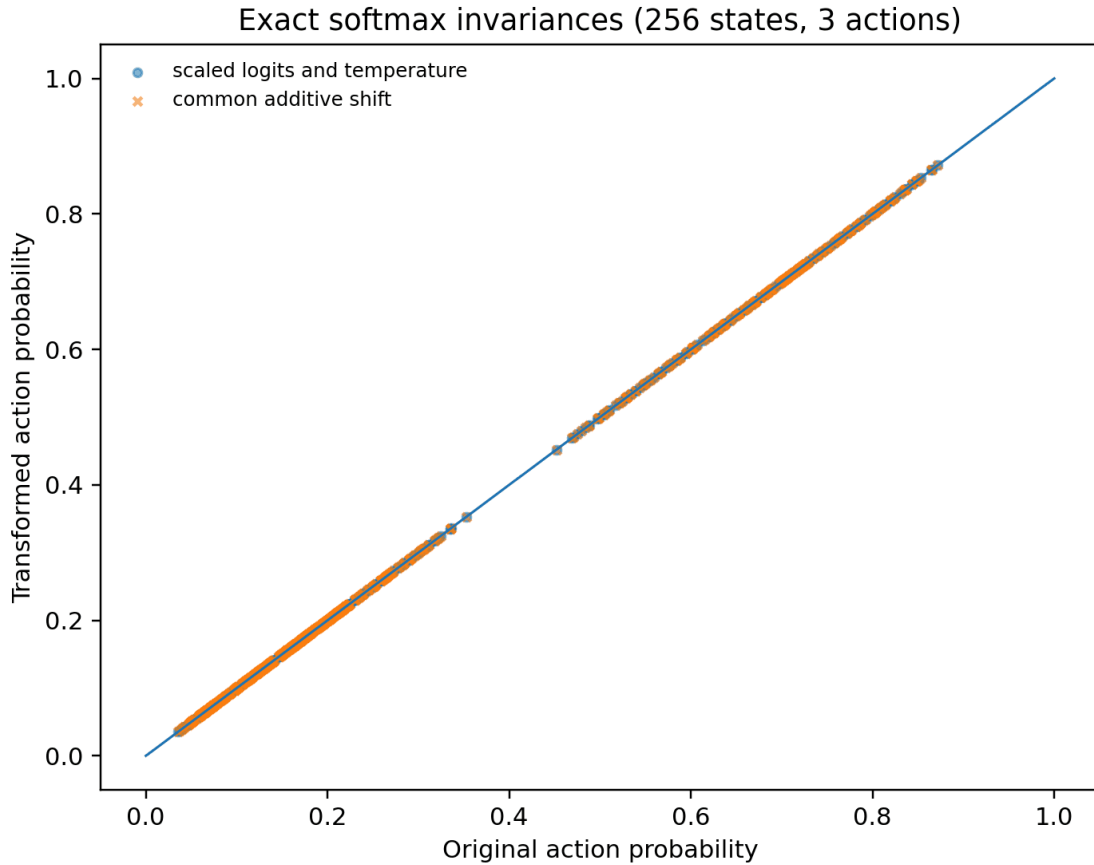


Figure 2. Numerical verification of the exact softmax invariances for 256 states and three actions. Scaling logits and temperature by $c = 3.7$ and adding a common intercept $k = 2.4$ reproduced probabilities to machine precision (maximum absolute differences 2.22×10^{-16} and 4.44×10^{-16}).

5. Synthetic verification study

5.1. Scope and inferential target

The empirical component is an internal synthetic verification of the withdrawal policy, not a claim of clinical validation. The test bench isolates the focal policy terms while retaining competition among cooperation, defection, and withdrawal. It asks whether known generating values are recovered and whether fitted parameters preserve out-of-sample action regimes.

$$L_C = 0.25(1 - a) - 0.10p. \quad (21)$$

$$L_D = -0.25 + 0.45p + 0.15f. \quad (22)$$

$$L_W = q_W + \lambda_A a + \lambda_F f + \lambda_P [p - \theta_W]_+. \quad (23)$$

Here $a = g_A(m)$ is the endogenous ambiguity activation, f is friction, and p is pressure. The focal vector is $\text{vartheta} = (q_W, \lambda_A, \lambda_F, \lambda_P, \theta_W)$. Cooperation and defection logits are fixed in the test bench so that uncertainty is concentrated on the withdrawal decomposition. This reduced verification model is a diagnostic instrument for the policy; it is not a replacement for the full CMS-WM dynamics.

5.2. Principal recovery design

Thirty generating vectors were selected by a fixed-seed Latin-hypercube-like design over q_W in $[-1.55, -0.50]$, λ_A in $[0.70, 2.30]$, λ_F in $[0.15, 1.35]$, λ_P in $[0.25, 2.00]$, and θ_W in $[0.45, 0.73]$. Each vector generated 1,000 categorical choices. Ambiguity activation followed $\text{Beta}(2.2, 2.2)$. Friction was intentionally correlated with ambiguity, $f = \text{clip}(0.68a + 0.32u + e, 0, 1)$, where u followed $\text{Beta}(2, 3)$ and e was Gaussian noise with standard deviation 0.035. Pressure was drawn from a mixture that created both subthreshold and suprathreshold observations. Across datasets, the median Spearman correlation between a and f was 0.895, and the median fraction $p > \theta_W$ was 0.447. Thus pressure was actively excited; weak pressure recovery cannot be attributed to an almost empty hinge region.

The enriched observation vector contained the three overall action frequencies; withdrawal frequencies in quartiles of a , f , and p ; two ambiguity-friction contrast cells; and pressure slices $p > 0.70$, $p > 0.82$, and $0.55 < p \leq 0.70$. Candidate parameters were estimated by bounded, six-start nonlinear least squares using binomial standard-error weights with prespecified floors.

$$L(\text{vartheta}) = \|W[S_{\text{pred}}(\text{vartheta}) - S_{\text{obs}}]\|_2^2. \quad (24)$$

5.3. Recovery and prediction metrics

$$\text{NMAE}_j = \text{median}_i |\hat{v}_{ij} - v_{ij}| / (U_j - L_j). \quad (25)$$

For each parameter we report Spearman rank correlation, bootstrap 95% intervals based on 2,000 resamples, median normalized absolute error, 90th-percentile error, and boundary-hit rate. A conservative reportability screen required $\rho \geq 0.50$, median NMAE ≤ 0.20 , 90th-percentile NMAE ≤ 0.50 , and boundary-hit rate ≤ 0.25 . The screen is a decision rule for this study, not a universal constant; bootstrap uncertainty and sensitivity results determine whether a passing quantity is labeled reportable or only cautiously reportable.

Predictive stability was evaluated on an independent factorial battery of 8,000 states with a , f , and p independently uniform on $[0,1]$. We compared generating and fitted mean action probabilities using dominant-regime agreement, L1 distance, and Jensen-Shannon divergence.

5.4. Design sensitivity and profile diagnostics

Two complementary checks were used. First, 24 generating vectors were repeated under a friction-orthogonal design in which a and f were independently sampled; the median a - f correlation was -0.011 while pressure activation remained similar to the principal design. Second, a 2×2 sensitivity experiment crossed baseline versus pressure-excited states with coarse versus enriched summaries for 12 generating vectors. A representative profile-wise analysis then fixed one focal parameter over a grid and reoptimized the remaining parameters, exposing broad, displaced, or boundary-seeking discrepancy profiles [\[9\]\[15\]](#).

6. Results

6.1. Exact invariances

The scaling and common-shift transformations changed action probabilities by at most 2.22×10^{-16} and 4.44×10^{-16} , respectively. These are numerical roundoff, not empirical approximation. Raw temperature, absolute action-value origin, and unnormalized policy weights cannot be separately interpreted in this policy family.

6.2. Principal parameter recovery

Quantity	Spearman rho (95% bootstrap)	Median NMAE (95% bootstrap)	90th pct. NMAE	Boundary
Withdrawal intercept	0.77 [0.51,0.90]	0.077 [0.042,0.104]	0.146	0.0%
Ambiguity-linked coupling	0.52 [0.23,0.74]	0.189 [0.118,0.259]	0.457	3.3%
Friction-linked coupling	0.21 [-0.17,0.54]	0.336 [0.289,0.504]	0.654	46.7%
Pressure-linked coupling	0.37 [0.00,0.66]	0.117 [0.091,0.253]	0.464	16.7%
Withdrawal threshold	0.30 [-0.10,0.63]	0.165 [0.107,0.197]	0.372	13.3%

Table 2. Principal recovery results across 30 synthetic ground-truth settings.

The withdrawal intercept q_W was well recovered as a calibration quantity. Among the focal mechanistic couplings, λ_A was the only parameter to meet the point-estimate screen in the correlated principal design: $\rho = 0.52$, median NMAE = 0.189, 90th-percentile NMAE = 0.457, and boundary-hit rate = 3.3%. Its bootstrap rank interval remained wide (0.23 to 0.74), so the appropriate classification is cautious rather than strong reportability.

Friction recovery was poor under correlated inputs ($\rho = 0.21$; median NMAE = 0.336; 46.7% boundary hits). Pressure coupling had a relatively low median error but insufficient rank recovery ($\rho = 0.37$), showing why error magnitude alone is not an adequate interpretability criterion. The threshold likewise failed rank recovery ($\rho = 0.30$).

6.3. Friction is design-conditional

Quantity	Spearman rho	Median NMAE	90th pct. NMAE	Boundary
Withdrawal intercept	0.82	0.073	0.183	0.0%
Ambiguity-linked coupling	0.67	0.106	0.219	0.0%
Friction-linked coupling	0.68	0.087	0.254	8.3%
Pressure-linked coupling	0.42	0.205	0.498	25.0%
Withdrawal threshold	0.23	0.217	0.425	20.8%

Table 3. Recovery under a friction-orthogonal design (24 ground-truth settings).

Orthogonalizing friction from ambiguity changed the scientific conclusion about λ_F . Its rank recovery rose from 0.21 to 0.68, median NMAE fell from 0.336 to 0.087, and boundary hits fell from 46.7% to 8.3%. This is positive evidence that the coefficient can carry recoverable information under an adequately separating design. It is also evidence against unrestricted interpretation: recoverability depends on the joint state distribution and is therefore configuration-conditional.

Parameter recovery under the correlated, pressure-excited summary design

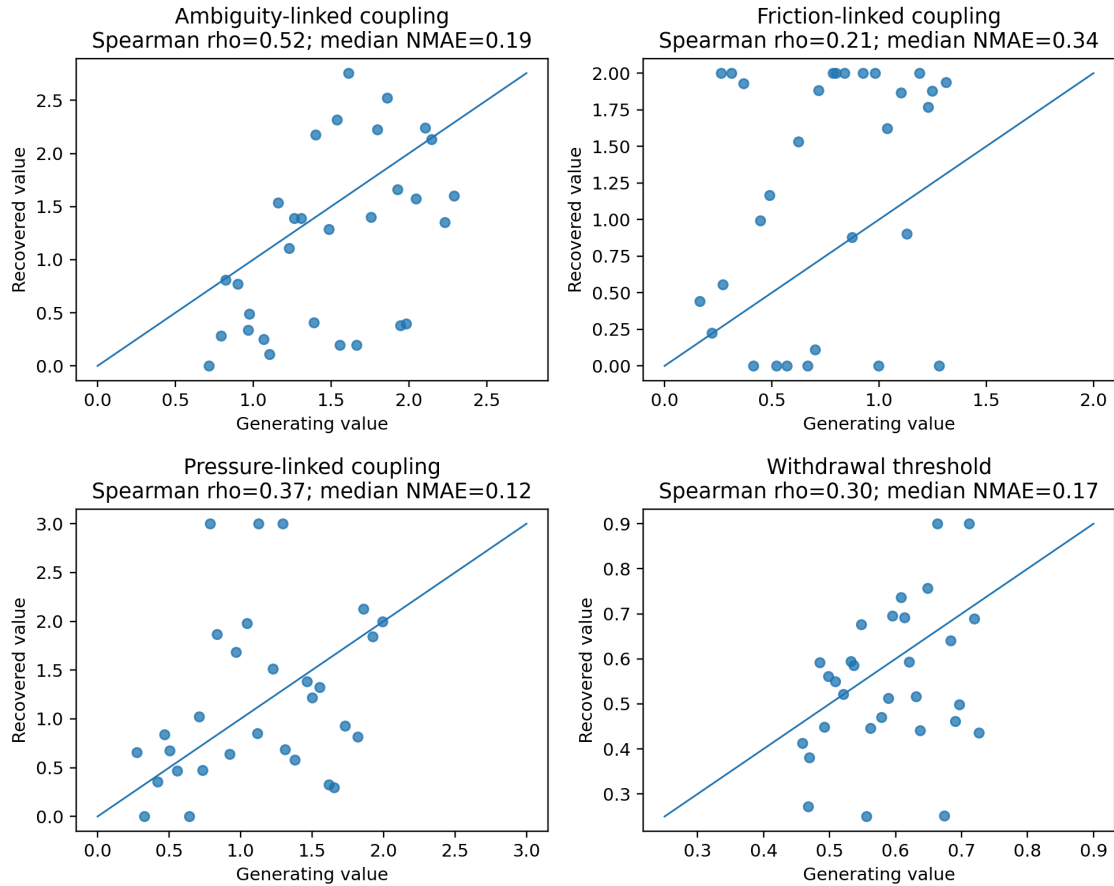


Figure 3. Generating versus recovered focal parameters in the principal correlated, pressure-excited design. The diagonal denotes exact recovery. The friction parameter frequently approached a bound; pressure and threshold estimates showed weak rank preservation despite some low individual errors.

6.4. Pressure excitation did not guarantee pressure or threshold recovery

The principal design placed a median 44.7% of observations above the generating threshold, with a 10th- to 90th percentile range of 31.2% to 56.2%. Nevertheless, λ_P and θ_W did not preserve generating rank. In the 2 x 2 sensitivity study, pressure rho ranged from -0.12 to 0.59 and threshold rho from -0.10 to 0.45 across state and summary designs. Enriching summaries or increasing threshold crossings therefore did not yield a general rescue. The remaining ambiguity is a property of the composite policy-observation geometry, not merely an empty dead zone.

6.5. Profile diagnostics

Representative profile diagnostics (synthetic dataset 20)

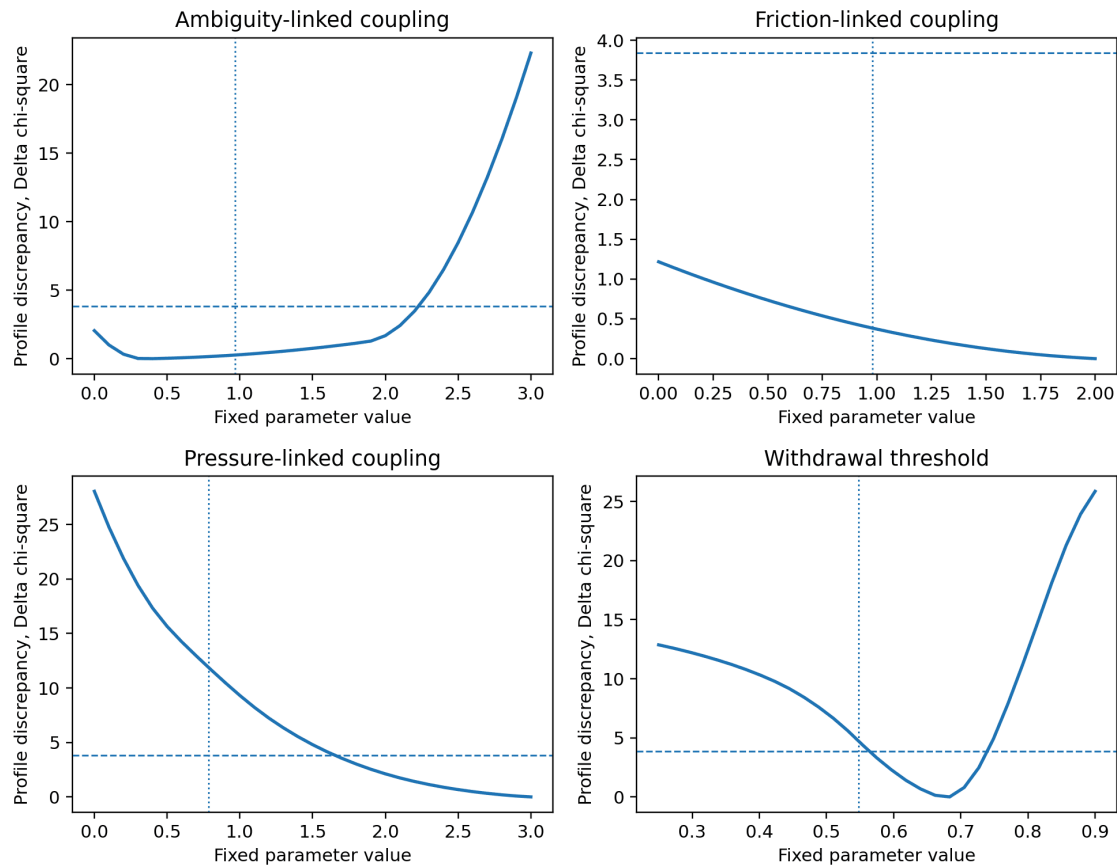


Figure 4. Representative profile discrepancies for synthetic dataset 20. The dotted vertical line is the generating value and the dashed horizontal line is Delta chi-square = 3.84, used only as a familiar visual reference. The profiles are diagnostic rather than exact likelihood-ratio confidence intervals because the objective is a weighted summary discrepancy. Friction is nearly flat and boundary-seeking; pressure and threshold minima are displaced from truth.

The representative profiles explain why a numerically optimized point estimate can be misleading. A finite minimum does not imply that the generating value is localized or that the parameter has a stable mechanistic interpretation. Profile shape, boundary behavior, cross-dataset rank, and predictive performance must be considered jointly.

6.6. Regime prediction was stable despite parameter error

Across all 30 principal fits, the dominant action regime on the independent factorial battery matched the generating model. Median predictive L1 distance was 0.0187, its 90th percentile was 0.0429, and median Jensen-Shannon divergence was 7.22×10^{-5} . Parameter recovery and prediction therefore separated sharply: multiple parameter combinations preserved practically indistinguishable regime behavior.

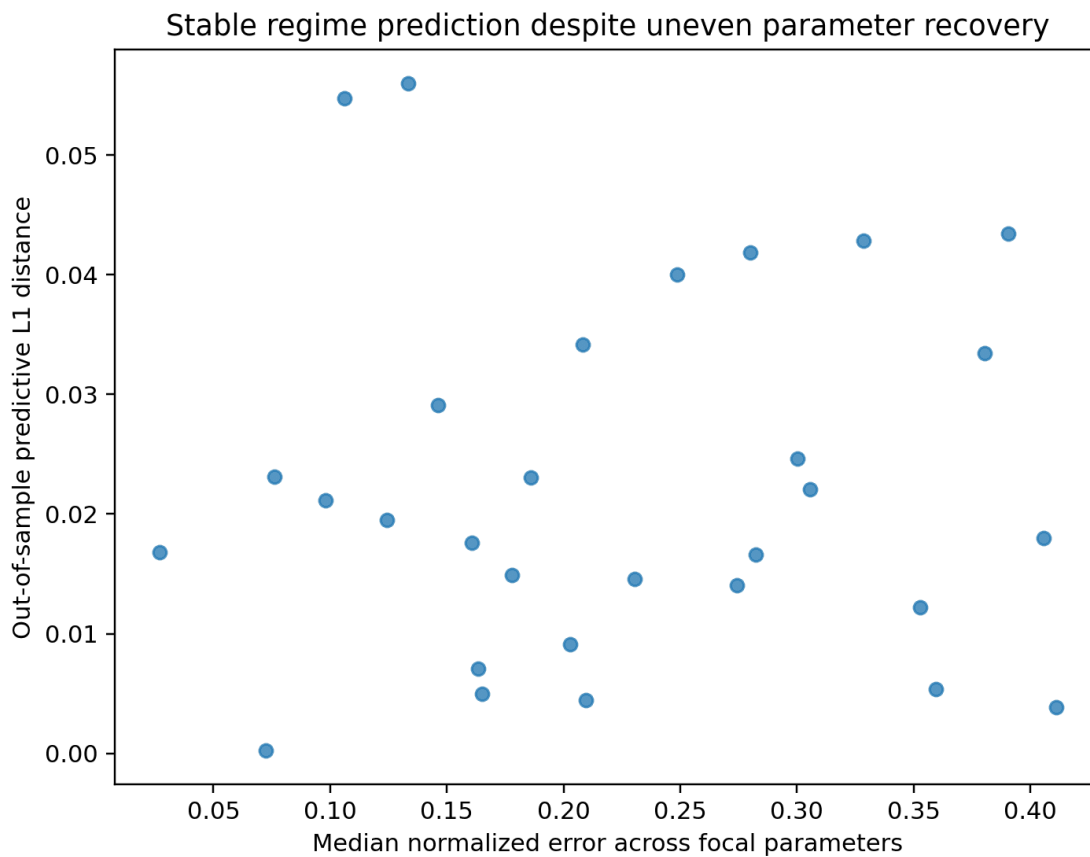


Figure 5. Out-of-sample predictive discrepancy remained small across datasets with uneven focal-parameter recovery. This supports regime-level reportability while constraining parameter-level mechanistic claims.

6.7. Final reportability taxonomy

Object	Role	Status	Scope/caveat
q_W	Calibration intercept	Reportable for calibration	Recovered well; not a specific mechanism
lambda_A	Endogenous ambiguity-withdrawal coupling	Cautiously reportable	Meets point screen; bootstrap rank uncertainty remains material
lambda_F	Friction-withdrawal coupling	Configuration-conditional	Weak with correlated inputs; recoverable when friction is orthogonalized
lambda_P	Pressure-withdrawal coupling	Non-reportable	Insufficient rank recovery across designs despite active hinge region
theta_W	Withdrawal threshold	Non-reportable	Insufficient rank recovery and displaced profiles
Withdrawal-rich regime	Predictive/regime object	Reportable within simulator	Stable out-of-sample action regime; no external clinical claim

Table 4. Evidence-linked reportability classification

7. Discussion

The revised analysis supports a narrower but more defensible interpretation of CMS-WM. The model has a mechanistically articulated withdrawal policy, yet the observation model does not preserve all of that internal decomposition. This is not a contradiction. Semantic interpretability and inferential identifiability answer different questions. The equations define what lambda_A, lambda_F, lambda_P, and theta_W do inside the simulator. Recovery experiments determine whether synthetic observables contain enough distinct information to estimate those quantities.

The central empirical result is the asymmetry between regime stability and parameter recovery. Fitted models preserved dominant action regimes and mean action probabilities even when focal parameter estimates were inaccurate, boundary-seeking, or weakly rank-correlated with truth. This pattern is consistent with the general phenomenon of sloppy or compensatory parameter directions in complex

mechanistic models ^{[10][11]}. For scientific reporting, prediction should not be converted automatically into a story about uniquely recovered latent mechanisms.

The friction experiment provides a constructive example. λ_F was not intrinsically meaningless: it became recoverable when friction was decorrelated from ambiguity. The correct conclusion is therefore design-conditional interpretability. This status is more informative than a binary identifiable/non-identifiable label because it identifies a concrete experimental requirement: tasks or observations must independently manipulate or measure friction and ambiguity.

Pressure excitation offers the complementary negative example. The hinge region was well populated, but λ_P and θ_W remained weakly rank-recoverable. More threshold crossings are not sufficient when changes in coefficient, threshold, baseline, and competing actions induce similar summary patterns. Future studies should consider direct temporal likelihoods, controlled pressure schedules that concentrate observations near multiple candidate thresholds, intervention contrasts, and independent measurement channels rather than relying only on additional summary bins.

The reportability framework has broader relevance for computational psychiatry. Mechanistic models should disclose (i) their variable declarations and domains, (ii) exact symmetries and reference constraints, (iii) the observation operator, (iv) synthetic recovery across a design space rather than a single example, (v) predictive checks on held-out conditions, and (vi) explicit claims and non-claims. Negative or conditional recovery findings are not technical failures to hide; they define the epistemic boundary of the model.

8. Limitations

- The recovery study is synthetic and internal. It does not establish that CMS-WM parameters correspond to real clinical mechanisms, psychopathic traits, or patient-level phenotypes.
- The verification harness isolates the withdrawal policy and fixes cooperation and defection logits. Full-dynamic recovery may be harder because learning, moral dynamics, and world-state feedback introduce additional compensation.
- The observation model is summary-based. A trial-level likelihood, state-space filter, or richer temporal task could change practical recoverability.
- The principal batch contains 30 ground-truth settings and the orthogonal sensitivity 24. Bootstrap intervals describe this finite design set and should not be interpreted as population-level confidence intervals.

- The reportability thresholds are conservative study rules, not universal standards. Their value lies in making the decision explicit and reproducible.
- The profile discrepancy is not an exact likelihood profile; the Delta chi-square line is a visual diagnostic reference, not a formal confidence boundary.
- The present experiments cover selected parameter ranges and state distributions. Structural statements generalize within the declared policy family; practical statements do not generalize beyond the tested model-observation-design combinations.

9. Conclusion

CMS-WM supports a precise distinction between mechanistic meaning and inferential warrant. Exact softmax symmetries rule out raw temperature and raw-weight interpretation. Synthetic recovery shows that the endogenous ambiguity coupling is cautiously recoverable, friction is recoverable only under a separating design, and pressure coupling and threshold are not substantively interpretable under the present summary observation model. Yet regime-level predictions remain stable. The scientifically defensible contribution is therefore not a complete latent decomposition of withdrawal, but an evidence-linked hierarchy of reportable, conditional, nuisance, and regime-level objects.

Statements and declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

The author declares no competing interests.

Author Contributions

R.M.P.A.F. was the sole author and is responsible for conceptualization, model formulation, software, analysis, visualization, and manuscript preparation.

Data and Code Availability

All code, fixed seeds, synthetic generating settings, derived result tables, figures, and a SHA-256 manifest required to reproduce the reported analyses are supplied in the accompanying reproducibility bundle. No

restricted empirical data exist.

Ethics

No human participants, patient data, or identifiable clinical records were used. Institutional ethics approval and informed consent were therefore not required.

References

1. [△]Adams RA, Huys QJM, Roiser JP (2016). "Computational Psychiatry: Towards a Mathematically Informed Understanding of Mental Illness." *J Neurol Neurosurg Psychiatry*. 87:53–63. doi:[10.1136/jnnp-2015-310737](https://doi.org/10.1136/jnnp-2015-310737).
2. [△]Wang XJ, Krystal JH (2014). "Computational Psychiatry." *Neuron*. 84:638–654. doi:[10.1016/j.neuron.2014.10.018](https://doi.org/10.1016/j.neuron.2014.10.018).
3. [△]Pauli R, Lockwood PL (2023). "The Computational Psychiatry of Antisocial Behaviour and Psychopathy." *Neurosci Biobehav Rev*. 145:104995. doi:[10.1016/j.neubiorev.2022.104995](https://doi.org/10.1016/j.neubiorev.2022.104995).
4. [△]Zavlis O, Story GW, Moutoussis M, et al. (2025). "A Systematic Review of Computational Modeling of Interpersonal Dynamics in Psychopathology." *Nat Ment Health*. 3:932–942. doi:[10.1038/s44220-025-00465-9](https://doi.org/10.1038/s44220-025-00465-9).
5. [△]Prosser A (2018). "A Bayesian Account of Psychopathy: A Model of Lacks Remorse and Self-Aggrandizing." *Comput Psychiatry*. 2:1–49.
6. [△]Rhoads SA, Gan L, Berluti K, et al. (2025). "Neurocomputational Basis of Learning When Choices Simultaneously Affect Both Oneself and Others." *Nat Commun*. 16:9350. doi:[10.1038/s41467-025-64424-9](https://doi.org/10.1038/s41467-025-64424-9).
7. [△]Driessen JMA, Figner B, Brazil IA, et al. (2025). "Dissecting How Psychopathic Traits Are Linked to Learning in Volatile Social Environments." *Cogn Affect Behav Neurosci*. 25:1543–1562. doi:[10.3758/s13415-025-01295-z](https://doi.org/10.3758/s13415-025-01295-z).
8. [△]Reverdy P, Leonard NE (2016). "Parameter Estimation in Softmax Decision-Making Models with Linear Objective Functions." *IEEE Trans Autom Sci Eng*. 13:54–67. doi:[10.1109/TASE.2015.2499244](https://doi.org/10.1109/TASE.2015.2499244).
9. [△][♢]Simpson MJ, Maclaren OJ (2023). "Profile-Wise Analysis: A Profile Likelihood-Based Workflow for Identifiability Analysis, Estimation, and Prediction with Mechanistic Mathematical Models." *PLoS Comput Biol*. 19:e111515. doi:[10.1371/journal.pcbi.1011515](https://doi.org/10.1371/journal.pcbi.1011515).
10. [△][♢]Gutenkunst RN, Waterfall JJ, Casey FP, et al. (2007). "Universally Sloppy Parameter Sensitivities in Systems Biology Models." *PLoS Comput Biol*. 3:e189. doi:[10.1371/journal.pcbi.0030189](https://doi.org/10.1371/journal.pcbi.0030189).
11. [△][♢]Wieland FG, Hauber AL, Rosenblatt M, Tönsing C, Timmer J (2021). "On Structural and Practical Identifiability." *Curr Opin Syst Biol*. 25:60–69. doi:[10.1016/j.coisb.2021.03.005](https://doi.org/10.1016/j.coisb.2021.03.005).

12. [△]Wilson RC, Collins AGE (2019). "Ten Simple Rules for the Computational Modeling of Behavioral Data." *eLife*. 8:e49547. doi:[10.7554/eLife.49547](https://doi.org/10.7554/eLife.49547).
13. [△]Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A (2020). "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." *arXiv:1804.06788*.
14. [△]Auger-Méthé M, Field C, Albertsen CM, et al. (2016). "State-Space Models' Dirty Little Secrets: Even Simple Linear Gaussian Models Can Have Estimation Problems." *Sci Rep*. 6:26677. doi:[10.1038/srep26677](https://doi.org/10.1038/srep26677).
15. ^{a, b}Raue A, Kreutz C, Maiwald T, et al. (2009). "Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood." *Bioinformatics*. 25:1923–1929. doi:[10.1093/bioinformatics/btp358](https://doi.org/10.1093/bioinformatics/btp358).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.