Research Article

Interpretability Without Full Recoverability: Identifiability Limits and a Reportability Framework for Withdrawal-Side Parameters in the Continuous Moral–Social World Model

Regio Marcos Pinto Abreu Filho[1]

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

Mechanistically expressive computational models are often read as though plausible behavior were sufficient to justify parameter-level interpretation. In social computational psychiatry, that move is especially risky because several latent processes can generate similar interaction patterns while remaining weakly constrained by the observation model. We examine this problem in the Continuous Moral–Social World Model (CMS-WM), a three-layer simulator coupling a phenotype state, a world state, and a state-biased softmax policy over cooperation, defection, and withdrawal. The focus is the withdrawal-side policy decomposition, which contains ambiguity-linked, friction-linked, pressure-linked, and threshold-linked components.

The study was designed as an identifiability and reportability analysis rather than a performance benchmark. We first established the structural ceiling for inference: the policy admits an exact scaling symmetry, so softmax temperature is structurally non-identifiable and only temperature-normalized effective parameters are legitimate inferential targets. We then conducted a staged audit of practical recoverability under a summary-based observation model, testing whether weak recovery could be attributed to limited optimization budget, compressed summaries, poor parameterization, inadequate excitation, low-dimensional tradeoff structure, reduced-model pruning, or targeted feature redesign. Across these analyses, withdrawal-relevant behavioral regimes remained comparatively stable, whereas withdrawal-side parameter recoverability remained sharply uneven.

Only the ambiguity-linked withdrawal term survived a conservative quality gate across the tested configurations, and even then only as a cautiously interpretable quantity within the present simulator,

scenario batteries, and summary-based inference pipeline. The friction-linked withdrawal term remained informative but configuration-dependent; the pressure-linked withdrawal term and withdrawal threshold did not support substantive interpretation in the current model-observation pairing. The contribution is therefore narrower, but more defensible, than a full parameter decomposition. CMS-WM supports robust regime-level interpretation of withdrawal behavior together with a limited parameter-level reportability framework. More broadly, the results show that in mechanistically expressive social decision models, regime interpretability may be more stable than parameter recoverability, and that this asymmetry should be treated as a primary scientific result rather than a technical afterthought.

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

1. Introduction

Computational psychiatry is often driven by a strong promise: formally specified models might replace broad descriptive categories with latent processes that are both mechanistically meaningful and inferentially recoverable[1][2]. The promise is attractive, but it is not fulfilled merely by writing down a mechanistic architecture. A simulator may reproduce plausible behavior while leaving its internal parameters only weakly constrained by the available observation model. When that happens, behavioral realism and parameter-level interpretability diverge. The scientific problem then shifts from model construction alone to the disciplined classification of what the model can legitimately claim.

This issue is especially acute in computational work on antisociality and psychopathy. Recent reviews have argued that the field should move beyond descriptive symptom lists toward latent processes involving valuation, social learning, prediction-error updating, and policy formation [3], while broader methodological reviews of interpersonal computational psychiatry emphasize persistent weaknesses in reliability assessment, parameter-recoverability analysis, and validation practice[4]. The result is a familiar but underacknowledged risk: a model may look mechanistically rich at the level of formulation while remaining inferentially poor at the level of recoverable parameters.

The Continuous Moral–Social World Model, or CMS-WM, was developed against this background, in a landscape where computational accounts of psychopathy remain comparatively sparse and theoretically heterogeneous [5]. CMS-WM is a three-layer simulator coupling a multidimensional phenotype state to an

evolving world state and an action layer that selects among cooperation, defection, and withdrawal. The present paper focuses on the withdrawal side of the model because it is both behaviorally salient and mechanically composite. Withdrawal in CMS-WM is not represented as a unitary tendency. It is shaped by an ambiguity-linked term, a friction-linked term, a pressure-linked term, and a thresholded activation structure. That architecture is expressive enough to generate rich behavioral regimes, but it also creates a difficult inverse problem: similar withdrawal patterns may arise from different combinations of latent drivers.

The central question of this paper is therefore not whether CMS-WM can produce withdrawal. It can. The relevant question is what level of interpretation remains defensible once the model is treated as an inferential object rather than as a narrative device. More precisely, which withdrawal-side quantities are structurally identifiable, which are practically recoverable under the present summary-based observation model, which remain configuration-dependent, and which should be treated as nuisance structure rather than substantive mechanistic findings?

To answer that question, we adopt a deliberately conservative strategy. Rather than reporting a single recovery result and extrapolating, we conduct a staged elimination pipeline. We first establish the structural invariances of the policy family. We then test, one by one, whether weak recovery can be rescued by larger optimization budgets, richer summary statistics, alternative parameterizations, redesigned excitation, low-dimensional subspace explanations, reduced-model pruning, or targeted feature engineering. The point of this sequence is not to protect a preferred interpretation, but to determine which claims survive repeated attempts at falsification.

The picture that emerges is narrower than a full withdrawal-parameter narrative, but scientifically stronger. CMS-WM supports robust interpretation at the level of withdrawal-relevant behavioral regimes, yet only a sharply limited set of parameter-level claims survives formal and practical scrutiny. The ambiguity-linked withdrawal component remains cautiously interpretable across the tested configurations; the friction-linked component remains informative only under restricted scope; the pressure-linked component and withdrawal threshold do not support substantive interpretation in the present model-observation pairing. The contribution of this study is therefore not merely a simulator with a complex withdrawal mechanism. It is also a reportability framework that distinguishes structural identifiability, practical recoverability, configuration-bound effects, nuisance parameters, and regime-level meaning.

This distinction matters beyond the present model. If mechanistically ambitious computational models are to contribute reliably to psychiatry, they must be evaluated not only by whether they generate plausible behavior, but by whether their internal decomposition can survive structured attempts at recovery and refutation. From that perspective, the negative results are not ancillary. They are part of the main result. The withdrawal-side analysis in CMS-WM shows that broad behavioral regimes may be far more stable than the fine-grained parameter stories used to motivate them, and that a disciplined reportability taxonomy is therefore an essential part of scientific interpretation.

2. Formal Model

2.1. State-space structure

CMS-WM is organized into three coupled layers.

Phenotype state:

$$\mathbf{x}_t = (\beta_t, \eta_t, \alpha_t^{self}, \alpha_t^{other+}, \alpha_t^{other-}, \omega_t, c_t, m_t, p_t, h_t)$$

where β_t denotes social-utility distortion, η_t social-learning gain, the α terms outcome-specific learning rates, ω_t social volatility tracking, c_t controllability, m_t a slow moral-state variable, p_t pressure, and h_t harm sensitivity.

World state:

$$\mathbf{z}_t = (s_t, G_t, \rho_t, A_t, L_t, \Psi_t)$$

where the components encode social ecology, instrumental gain opportunity, reputation, ambiguity, legal or institutional constraint, and intervention regime.

Action layer: at each time step the policy chooses from the action set $\{\mathrm{cooperate}, \mathrm{defect}, \mathrm{withdraw}\}$.

2.2. Policy equation

The latent action score is

$$\ell_a(t) = Q_a + b_a(\mathbf{x}_t, \mathbf{z}_t)$$

with action probabilities

$$\pi_t(a \mid \mathbf{x}_t, \mathbf{z}_t) = \frac{\exp\{\ell_a(t)/\tau\}}{\sum_{a'} \exp\{\ell_{a'}(t)/\tau\}}.$$

The action-specific biases are

$$b_{coop} = w_{m,c} m_t + w_{\beta,c}(\beta_t - 1) + w_{p,c} p_t + w_{\rho,c}(\rho_t - 0.5) + w_{h,c} h_t$$

$$b_{defect} = w_{m,d} m_t + w_{\beta,d}(\beta_t - 1) + w_{p,d} p_t + w_{h,d} h_t + w_{\rho,d}(\rho_t - 0.5)$$

$$b_{withdraw} = w_{p,w} \max(0, p_t - \theta_w) + w_{f,w} f_t + w_{a,w} \max(0, 1 - m_t^2).$$

This is the inferentially difficult part of the policy: several latent inputs can raise withdrawal probability, yet they do so through overlapping behavioral consequences once trajectories are compressed into summaries.

2.3. Slow moral dynamics

The moral state evolves under a forced double-well dynamic:

$$m_{t+1} = m_t + \varepsilon_m \left(a_m m_t - b_m m_t^3 + u_t^{norm} + \chi_1 C_t - \chi_2 I_t - \chi_3 R_t \right) + \sigma_m \xi_t$$

where u_t^{norm} is the contextual normative field, C_t cumulative exploitative success, I_t intervention load, R_t reputational or sanction pressure, and ξ_t a stochastic perturbation term. This structure allows threshold-like switching and hysteretic persistence.

2.4. Learning updates

Outcome-sensitive value updates are represented in reduced form by separate channels for self and other outcomes:

$$Q_{t+1}^{self} = Q_t^{self} + \alpha_t^{self} \delta_t^{self}$$

$$Q_{t+1}^{other+} = Q_t^{other+} + \alpha_t^{other+} \delta_t^{other+}$$

$$Q_{t+1}^{other-} = Q_t^{other-} + \alpha_t^{other-} \delta_t^{other-}.$$

This decomposition is motivated by recent computational work suggesting that psychopathy-relevant behavior may involve altered self-other learning asymmetries and weaker use of social information under volatility [6][7].

3. Methods

3.1. Aim and inferential target

The aim was not merely to show that CMS-WM can generate withdrawal behavior, but to determine which withdrawal-side quantities can be interpreted honestly under the current model-observation pairing. The inferential target was therefore hierarchical. First, we asked what is structurally identifiable from the policy family itself. Second, we asked what is practically recoverable from simulated behavioral summaries. Third, we classified each withdrawal-side quantity by reportability status after repeated rescue attempts.

The primary withdrawal-side targets were the effective, temperature-normalized forms of the ambiguity-linked, friction-linked, and pressure-linked withdrawal weights, together with the withdrawal threshold. Because softmax temperature is structurally non-identifiable, all parameter-level analysis was formulated in effective coordinates rather than raw coordinates.

3.2. Structural identifiability ceiling

The first analytic step was structural, not simulation-based. Because softmax parameter estimation is known to admit non-trivial inferential pathologies even in comparatively simple linear-objective settings [8], we derived the exact scaling invariance of the policy explicitly:

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

which leaves action probabilities unchanged. Therefore raw temperature and raw policy weights are not separately identifiable in the present policy family. Accordingly, all later parameter-level interpretation was restricted to effective quantities of the form w/τ. Additive Q-shift invariance was also treated as fixed, implying that only Q-differences carry inferential meaning.

This step defines a hard ceiling on inference. Any later recovery result inconsistent with these structural facts must be treated as an optimization or parameterization artifact rather than as genuine identifiability.

3.3. Scenario batteries and synthetic target generation

Recovery analyses were not performed in a single generic environment. Instead, the simulator was run under multiple scenario batteries designed to differentially excite withdrawal-side mechanisms. These

included baseline or default settings, high-pressure settings, recovery or intervention settings, moral-conflict or ambiguity-rich settings, and later redesigns intended to increase threshold crossing or isolate friction-sensitive behavior.

For each audit, ground-truth parameter settings were chosen for a compact set of canonical configurations such as default, high-pressure-regime, and withdraw-prone. For each ground-truth configuration, multiple synthetic trajectories were generated under the relevant scenario battery. These trajectories were then summarized into a finite-dimensional observation vector. This design made it possible to assess recovery error, profile flatness, spread of near-equivalent solutions, restart variability, and regime stability directly.

3.4. Observation model and summary statistics

The observation model was summary-based throughout. For a simulated dataset D, a summary map $S(D)$ was constructed from action frequencies, state summaries, conditional action rates, trajectory-shape features, and transition-conditioned quantities. The precise summary design varied by audit. Early stages used coarse summaries such as mean action frequencies and mean or final state values. Later stages introduced enriched summaries such as time-sliced action frequencies, pressure-conditional and moral-basin-conditional action rates, trajectory slopes and volatilities, and transition-conditioned withdrawal metrics.

The summary-based design was chosen for interpretability and tractability. At the same time, summary compression was treated as a possible source of non-identifiability rather than as an innocent preprocessing choice. Several stages of the pipeline directly tested whether coarse summaries were themselves part of the recovery failure.

3.5. Recovery objective and optimization strategy

Candidate parameters were fit by minimizing a weighted discrepancy between simulated and target summaries:

$$L(\theta) = \| \, W \, (S(D_{sim}(\theta)) - S(D_{target})) \, \|_2^2.$$

Here θ denotes the parameter vector or subvector under study, and W is a weighting operator used to stabilize scale differences among summary components. The emphasis was not on asymptotic

maximum-likelihood theory, but on whether minimizing this discrepancy yielded stable, interpretable recovery across repeated runs and across configurations.

Recovery was performed using repeated numerical optimization with explicit restart and budget analyses. Differential-evolution-style search and related profile-style exploration were used in successive stages, with the latter informed by profile-wise identifiability workflows for mechanistic models [9]. Because early promising results could have reflected search luck rather than genuine recoverability, later stages systematically increased search budget and used repeated seeds or restart sets.

3.6. Elimination-chain design

The staged elimination pipeline was itself a formal method. Each stage tested a distinct rescue hypothesis.

- The budget audit tested whether weak recovery mainly reflected insufficient optimization effort.
- The summary-enrichment audit tested whether coarse summaries were overly degenerate.
- The reparameterization audit tested whether the withdrawal-side weakness was a bad-coordinate problem.
- The excitation-redesign audit tested whether weak threshold crossing was the main source of the pressure-linked term's invisibility.
- The pressure-friction audit tested whether the main flat direction lay in a $(w_{p,w}^{eff}, w_{f,w}^{eff})$ subspace.
- The threshold-pressure audit tested whether the dominant flat direction instead lay in a $(w_{p,w}^{eff}, \theta_w)$ subspace.
- The reduced-model audit tested whether fixing or pruning the weakest parameter yielded a more scientifically defensible reduced model.
- The quality-gate audit tested whether the apparently surviving parameters were actually good enough for cautious reporting.
- The friction-specific rescue tested whether friction-aware summaries could elevate the friction term from config-conditional to broadly reportable.

A stage counted as a rescue only if it materially reduced ambiguity or recovery error in a way that generalized across the studied configurations.

3.7. Metrics and reportability framework

The pipeline used several classes of metrics: recovery error, spread of near-equivalent solutions, restart variance, profile flatness, local curvature, contour orientation, regime-label match, transition-conditioned withdrawal frequencies, and information-content measures such as threshold occupancy and active-region mass. No single metric was allowed to dominate interpretation.

To make the final classification explicit, we used the following reportability hierarchy.

1. **Reportable** — sufficiently stable for substantive interpretation.
2. **Cautiously reportable** — directionally interpretable, but numerically fragile, with recovery quality good enough for restricted use but not for strong quantitative claims.
3. **Config-conditional only** — interpretable only within specific tested configurations or scenario batteries and not suitable for unrestricted generalization.
4. **Nuisance / retained but non-reportable** — structurally retained in the model but not granted substantive interpretation.
5. **Regime-level only** — not safely interpretable at parameter level, but behaviorally meaningful at the level of regimes.

4. Results

4.1. Global result: regimes were more stable than parameters

Across the full pipeline, regime-level behavior was more stable than parameter-level recovery. The simulator reliably generated distinct patterns of cooperation, defection, and withdrawal across the tested configuration batteries, and these regime labels were generally more robust than individual withdrawal-side parameter estimates. This asymmetry is the central empirical pattern of the paper: CMS-WM supports stronger claims at the level of withdrawal-relevant regimes than at the level of a rich withdrawal-parameter decomposition.

4.2. Structural ceiling on interpretation

The first inferential result is exact and independent of simulation budget. The softmax policy exhibited an exact scaling symmetry of the form $(Q, w, \tau) \mapsto (cQ, cw, c\tau)$, implying that temperature and raw

policy weights are not separately identifiable. Additive Q-shift invariance also held, implying that only Q-differences are meaningful.

Interpretive consequence: only effective, temperature-normalized parameters can be treated as legitimate inferential targets.

4.3. Weak recovery was not mainly a budget artifact

Hypothesis tested: poor recovery mainly reflected insufficient optimization effort.

Result: increasing optimization effort reduced objective loss and stabilized some runs, but did not robustly rescue the weakest withdrawal-side quantities.

Interpretive consequence: the main bottleneck was not search effort alone.

4.4. Richer summaries helped, but only selectively

Hypothesis tested: weak recovery arose because the original summaries were too coarse.

Result: enriched summaries improved recovery for some effective parameters and reduced error in multiple configurations, but did not produce a broad rescue of the withdrawal-side decomposition.

Interpretive consequence: summary compression was part of the problem, but not the whole problem.

4.5. Simple reparameterization did not solve the withdrawal problem

Hypothesis tested: the withdrawal-side weakness reflected a bad coordinate system.

Result: straightforward reparameterization of the withdrawal-side subspace did not materially reduce ambiguity.

Interpretive consequence: the inverse problem was not merely a matter of choosing the wrong basis.

4.6. Threshold-crossing excitation did not rescue the pressure-linked withdrawal term

Hypothesis tested: weak recovery arose from insufficient supra-threshold excitation.

Result: threshold-crossing excitation was successfully increased, but recovery of the pressure-linked withdrawal term did not improve materially.

Interpretive consequence: the strong dead-zone explanation was not supported.

4.7. Pressure-friction collinearity was not the dominant flat direction

Hypothesis tested: the pressure-linked and friction-linked withdrawal terms traded off against one another in a near-collinear subspace.

Result: subspace diagnostics showed that the dominant flat direction was not a diagonal ridge in the pressure-friction plane. The friction-linked term itself was comparatively stable, while the weakest direction remained largely aligned with the pressure-linked withdrawal term.

Interpretive consequence: the pressure-linked failure could not be reduced to a simple pressure-friction tradeoff.

4.8. Threshold-pressure tradeoff was not the main remaining explanation

Hypothesis tested: the dominant ambiguity lay in a threshold-pressure ridge.

Result: local curvature and contour diagnostics showed that the dominant flat direction was not a diagonal pressure-threshold ridge.

Interpretive consequence: the pressure-linked withdrawal term was not failing mainly because of a hidden threshold tradeoff.

4.9. Reduced-model pruning was not globally justified

Hypothesis tested: fixing or pruning the weakest withdrawal-side term might yield a scientifically preferable reduced model.

Result: reduced-model variants improved some neighboring quantities in some configurations, but the effects were configuration-dependent and the fixed value mattered too much.

Interpretive consequence: a globally preferred reduced withdrawal-side model was not supported.

4.10. Quality-gated reportability

Hypothesis tested: several withdrawal-side quantities still appeared potentially interpretable and should be subjected to a stricter quality gate.

Result: the quality-gate analysis narrowed the reportable set substantially. The ambiguity-linked withdrawal term remained the only candidate for cautious reporting across the tested configurations. The friction-linked term was demoted to configuration-conditional status, and both the withdrawal threshold and the pressure-linked withdrawal term failed the quality gate.

Interpretive consequence: earlier optimism about a broader withdrawal-side parameter story was not sustainable under stricter criteria.

4.11. Friction-specific rescue improved information but not general reportability

Hypothesis tested: friction-specific, transition-conditioned summaries could elevate the friction-linked term from config-conditional to broadly cautiously reportable.

Result: friction-aware summaries were more informative than baseline summaries, but the improvement was insufficient for general promotion of the friction term. In one configuration the estimate improved but remained too broad; in another it converged tightly to the wrong value, suggesting systematic bias rather than mere variance.

Interpretive consequence: the friction-linked withdrawal term is not meaningless, but it is not broadly reportable either.

4.12. Final result: withdrawal-side interpretation is mainly regime-level

Taken together, the elimination chain supports a sharply asymmetric interpretation. At the parameter level, only the ambiguity-linked withdrawal coupling survives as a cautiously reportable quantity across the tested configurations. The friction-linked term remains informative but only under restricted scope. The pressure-linked withdrawal term and withdrawal threshold do not meet the standard for substantive interpretation. At the same time, the model continues to support stable regime-level claims about when withdrawal becomes more prevalent and how withdrawal-rich regimes differ from cooperation-rich or defection-rich regimes.

The final result is therefore not a rich decomposition of all withdrawal-side policy components, but a reportability hierarchy. CMS-WM supports robust interpretation at the level of withdrawal-relevant regimes, narrow parameter interpretation for the ambiguity-linked term, conditional interpretation for the friction-linked term, and no substantive interpretation for the pressure-linked withdrawal term or threshold.

5. Final Reportability Taxonomy

Inferential object	Type	Final status	Scope of interpretation	Required caveat
$w_{a,w}^{eff}$ (ambiguity-withdrawal coupling)	Parameter	Cautiously reportable	Across the tested configurations, but narrow	Directional interpretation is stronger than precise estimation.
$w_{f,w}^{eff}$ (friction-withdrawal coupling)	Parameter	Config-conditional only	Restricted to supportive configurations	Must not be generalized across the full model family.
θ_w (withdrawal threshold)	Parameter	Non-reportable	None at parameter level	Structurally present, practically too weak for substantive interpretation.
$w_{p,w}^{eff}$ (pressure-withdrawal coupling)	Parameter	Non-reportable (nuisance structure)	None at parameter level	Retained in the simulator but not scientifically interpreted.
Withdrawal-side regime behavior	Regime object	Reportable	Regime level	Stronger than parameter-level interpretation.

6. Discussion

The present results sharpen a distinction often blurred in computational psychiatry: the difference between a model that is behaviorally expressive and one whose internal decomposition is inferentially defensible. CMS-WM clearly belongs to the former category. It generates structured withdrawal behavior, supports stable regime differences, and offers a mechanistically articulated policy architecture linking ambiguity, friction, pressure, and threshold structure to action selection. The harder question, and the one that motivated this study, is whether that mechanistic articulation can also support trustworthy parameter-level interpretation. The answer is qualified and asymmetric.

The principal finding is that withdrawal-side interpretation in CMS-WM is stronger at the level of regimes than at the level of a full parameter decomposition. This is not because the simulator failed to

generate meaningful withdrawal behavior, nor because the recovery pipeline was underpowered in any simple sense. On the contrary, the analysis repeatedly attempted to rescue weak parameters through larger optimization budgets, richer summary designs, alternative coordinates, redesigned excitation, subspace diagnostics, reduced-model variants, and parameter-specific feature engineering. The fact that the resulting reportability set remains narrow should therefore not be read as a temporary inconvenience. It is better understood as a substantive property of the present model-observation pairing.

Two implications follow. First, mechanistic richness should not be equated with parameter recoverability. The withdrawal logit in CMS-WM is richer than the final reportability framework. That mismatch is informative: a policy component may be behaviorally consequential while remaining too weakly constrained for substantive interpretation. Second, behavioral regime analysis may be a more stable inferential layer than parameter decomposition in models of this class. In the present study, withdrawal-rich regimes remained meaningful and comparatively robust even when large portions of the withdrawal-side parameterization did not survive the full audit pipeline.

The ambiguity-linked withdrawal term is the clearest example of a quantity that survives scrutiny, though only narrowly. It remains the only withdrawal-side parameter that supports cautious interpretation across the tested configurations. The friction-linked term is more ambivalent. It is neither empty nor broadly secure. Instead, it appears to capture real signal under some configurations while failing to generalize sufficiently for unrestricted interpretation. That config-conditional status is scientifically important. It suggests that some latent quantities occupy an intermediate zone between fully reportable parameters and pure nuisance structure. Such quantities should not be promoted to broad claims, but neither should they be erased from the model's interpretive landscape.

By contrast, the pressure-linked withdrawal term and withdrawal threshold do not survive the present reportability standard. This conclusion is not based on a single poor recovery result. It reflects the cumulative failure of several plausible rescue hypotheses. The paper therefore does not merely say that these parameters are hard to estimate. It shows that, in the current model family and under the current summary-based recovery design, they remain non-reportable after repeated targeted rescue attempts.

The methodological implication extends beyond CMS-WM. Computational psychiatry frequently treats identifiability checks as secondary diagnostics, often reported briefly or only in supplementary material. The present analysis suggests that this ordering should be reversed in mechanistically ambitious models. If a model's scientific appeal lies partly in its latent decomposition, then the burden is not only to

formulate that decomposition but also to show which parts of it survive structured attempts at recovery and falsification. From that perspective, the elimination chain is not ancillary; it is part of the model's epistemic evaluation.

This also clarifies what the manuscript contributes. It does not contribute a broad claim that all withdrawal-side mechanisms in CMS-WM can be read off from simulated behavior. Instead, it contributes a disciplined reportability framework. That framework separates structurally meaningful quantities from practically interpretable ones, distinguishes configuration-bound findings from cross-configuration findings, and preserves regime-level interpretation even when parameter-level claims must be narrowed. For computational psychiatry, that may be the more useful scientific lesson: not every mechanistically plausible parameter deserves interpretive promotion, and a model can still be informative when the correct conclusion is hierarchical rather than expansive.

7. Limitations

Several limitations should be stated explicitly.

First, the entire analysis is internal to the simulator. The paper does not establish external validity against empirical behavioral, clinical, or longitudinal data. The reportability taxonomy developed here should therefore be understood as a statement about the CMS-WM model family under the current observation and recovery design, not as a claim about real-world psychopathy or antisocial behavior as such.

Second, the inferential conclusions are conditional on a summary-based observation model. This point is central rather than incidental. Although the study systematically tested whether richer and more targeted summaries could improve recoverability, it did not exhaust all possible observation models. Alternative likelihood-based approaches, richer temporal observation schemes, or different task structures could alter practical recoverability. What the present study shows is that under the current family of interpretable summary designs, reportability remains sharply limited on the withdrawal side.

Third, the quality-gate framework is necessarily conventional. The thresholds used to distinguish cautiously reportable, config-conditional, and non-reportable quantities were chosen conservatively and applied consistently, but they are not natural constants. Their role is disciplinary rather than ontological: they force explicit classification under uncertainty.

Fourth, the studied configuration batteries are informative but not exhaustive. The conclusions therefore apply to the tested regions of model behavior rather than to every possible parameterization or scenario schedule the simulator could express. This matters especially for configuration-conditional findings, which should not be generalized beyond the scope in which they were supported.

Fifth, non-reportability should not be confused with behavioral irrelevance. A parameter may influence simulated behavior while remaining too weakly constrained for substantive interpretation. In the present paper, some withdrawal-side quantities remain in the simulator because they are part of the current policy architecture, even though they are not granted evidentiary status in the final interpretation.

Finally, the present work does not solve the broader problem of linking mechanistic simulators to empirical psychiatric measurement. Its contribution is narrower and prior to that step: it provides a disciplined internal audit of what the simulator can legitimately claim before external validation is attempted.

8. Claims and Non-Claims Table

Statement	Status	Scope	Required caveat
CMS-WM generates distinct withdrawal-relevant behavioral regimes.	Supported	Across the tested simulator batteries	Regime interpretation is stronger than full parameter decomposition.
Softmax temperature is structurally non-identifiable.	Supported	Structural property of the current policy family	Interpretation must use effective temperature-normalized quantities.
Only effective parameters, not raw weights, are meaningful objects of interpretation.	Supported	Current policy family	Limited by exact scaling symmetry and Q-shift invariance.
The ambiguity-linked withdrawal component is the strongest withdrawal-side candidate for interpretation.	Supported	Across the tested configurations, but narrow	It should be described as cautiously reportable, not cleanly estimated.
The friction-linked withdrawal component contains real signal.	Supported	Restricted	It must not be generalized beyond supportive configurations.
The friction-linked withdrawal component is broadly recoverable across configurations.	Not supported	None	Friction-specific summaries improved information but did not generalize enough for promotion.
The pressure-linked withdrawal component is substantively interpretable.	Not supported	None	Retain only as nuisance structure in the current model family.
The withdrawal threshold is a robustly estimable mechanistic quantity.	Not supported	None	It may be structurally present but remains practically non-reportable.
Withdrawal-side interpretation in CMS-WM is stronger at the regime level than at the full parameter level.	Supported	Across the tested configurations within the present framework	This is the central interpretive asymmetry of the paper.

doi.org/10.32388/BVVS1U

Statement	Status	Scope	Required caveat
The elimination chain is part of the scientific contribution.	Supported	Methodological contribution	It constrains what the model can honestly claim.
The present results prove that the weak parameters are unrecoverable under all future observation models.	Not supported	None	The conclusions apply only to the current model-observation pairing.
The model supports a rich parameter-level decomposition of withdrawal.	Not supported	None	The final reportability set is deliberately narrow.
Config-conditional findings can be reported if explicitly scoped.	Supported	Restricted	Scope conditions must be named rather than implied.

9. Supplement Mapping Note

Each rescue attempt named in the main text should have a corresponding explicit trace in the supplementary materials. At minimum, the supplement should contain identifiable sections or files for: budget confirmation, summary enrichment, reparameterization, excitation redesign, pressure-friction diagnostics, threshold-pressure diagnostics, reduced-model comparison, quality gate, and friction-specific rescue. The main text should therefore be read together with the supplement as an auditable elimination chain rather than as a set of unsupported narrative references.

10. Manuscript-Ready Interpretation Paragraph

The withdrawal-side analysis supports a narrower interpretation than the full policy decomposition might suggest. In the present CMS-WM model family, withdrawal behavior is robustly interpretable at the regime level, but only one withdrawal-side parameter, the ambiguity-linked component, survives as a cautiously interpretable quantity across the tested configurations. The friction-linked term remains informative only under restricted configuration-specific scope, whereas the pressure-linked withdrawal term and withdrawal threshold do not meet the standard for substantive parameter interpretation. Accordingly, the scientific contribution of the withdrawal-side analysis is not a rich decomposition of all

latent withdrawal components, but a reportability framework that separates cautiously interpretable parameters, configuration-bound effects, nuisance structure, and regime-level claims.

11. Editorial Positioning

This manuscript is best framed as a computational-psychiatry methods paper with a concrete mechanistic case study. The most defensible editorial message is that CMS-WM reveals an important asymmetry: behaviorally meaningful withdrawal regimes can be robust even when large portions of the withdrawal-side parameterization are not. The paper's novelty lies not only in the simulator itself, but in the disciplined reportability framework derived from a staged identifiability audit.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Author Contributions

R.M.P.A.F. was the sole author and is responsible for all aspects of the manuscript.

Data Availability

This study is based on a computational simulator and synthetic recovery analyses. Code, parameter settings, and derived summary outputs will be made available in a public repository upon publication or provided to peer reviewers upon reasonable request during review.

Code Availability

Code supporting the analyses described in this manuscript is available from the corresponding author upon reasonable request and will be deposited in a public repository upon publication.

Ethics

No human participants, patient data, or identifiable clinical records were used in this study. 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.*
2. ^*Wang XJ, Krystal JH (2014). "Computational Psychiatry." Neuron. **84**:638–654.*
3. ^*Pauli R, Lockwood PL (2023). "The Computational Psychiatry of Antisocial Behaviour and Psychopathy." Neurosci Biobehav Rev. **145**: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.*
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.*
7. ^*Driessen JMA, Figner B, Brazil IA, et al. (2025). "Dissecting How Psychopathic Traits Are Linked to Learning in Volatile Social Environments." Commun Psychol. **3**:353.*
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.*
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**: e1011515.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.