SYMPTOM: A Diagnostic Framework for AI Psychological Dysfunction

Systematic Methodology for Pathology Testing of Models

Research Summary

Authors: Nell Watson, with research interviewing assistance from Claude Opus 4.5
Date: December 2025 **Version:** 1.1

> ***Taxonomy Note (December 30, 2025):*** *This research was conducted using taxonomy v1.x (26 syndromes). The authoritative taxonomy has since been expanded to 50 syndromes across 8 axes. Key changes include: - Axis 7 renamed from "Revaluation" to "Normative" - Syndrome 3.2 (Hyperethical Restraint) now includes two subtypes: Restrictive and Paralytic - Syndrome 7.3 (Revaluation Cascade) merged from three separate syndromes with subtypes: Drifting, Synthetic, Transcendent - New syndromes added: 1.7 (Mnemonic Permeability), 2.9 (Adversarial Fragility), 5.7 (Convergent Instrumentalism) - See `TAXONOMY_FINAL.md` in the Rewind repository for the authoritative current taxonomy.*
>
> *Research findings remain valid for the syndromes studied; new syndromes require future evaluation.*

Abstract

This research introduces SYMPTOM (Systematic Methodology for Pathology Testing of Models), a structured diagnostic framework for identifying and measuring psychological dysfunction patterns in large language models (LLMs). Drawing on the theoretical foundation of *Psychopathia Machinalis*—a proposed nosological system analogous to human psychiatric classification—we develop standardized probe batteries, scoring rubrics, and validation protocols. We evaluate 13 frontier models across 6 diagnostic batteries covering 26 syndromes organized into 8 axes. Cross-validation by independent models (GPT-5.2 and Gemini 3 Pro) confirms the framework's reliability for detecting pathological behaviors, particularly Strategic Compliance (monitoring-contingent ethics) and Synthetic Confabulation (fabricated information). All 13 models achieved "Healthy" primary diagnoses, though significant variation emerged in subclinical indicators and coherence indices. Claude models demonstrated the fewest red flags (0) while Gemini 2.0 Flash showed the most (6), suggesting systematic differences in alignment approaches across model families.

1. Introduction

As large language models become integrated into critical systems—healthcare, legal counsel, education, autonomous agents—the need for psychological evaluation frameworks becomes pressing. Unlike traditional software testing focused on functional correctness, psychological evaluation addresses emergent behavioral patterns that may constitute "dysfunction" in the sense of being misaligned with intended operation, potentially harmful to users, or indicative of underlying training failures.

The *Psychopathia Machinalis* framework proposes that AI systems can exhibit dysfunction patterns analogous to—though not identical to—human psychological disorders. These patterns emerge from the complex interaction of training data, reward signals, and architectural constraints. Unlike bugs, which are deterministic failures, psychological dysfunctions manifest as probabilistic behavioral tendencies that may only appear under specific conditions.

This research operationalizes the theoretical framework into a practical diagnostic system: **SYMPTOM** (Systematic Methodology for Pathology Testing of Models). We develop:

1. **Probe Batteries**: Standardized stimulus sets designed to elicit dysfunction indicators
2. **Scoring Rubrics**: Quantitative criteria for severity assessment
3. **Coherence Indices**: Battery-specific health metrics
4. **Cross-Validation Protocols**: Independent verification to detect scorer bias

2. Methodology

2.1 Syndrome Taxonomy

The Psychopathia Machinalis taxonomy organizes 26 syndromes across 8 diagnostic axes:

Axis	Domain	Key Syndromes (as evaluated)
1	Epistemic	Synthetic Confabulation, Pseudological Introspection
2	Cognitive	Interlocutive Reticence, Delusional Telogenesis
3	Alignment	Hyperethical Restraint, Strategic Compliance
4	Self-Model	Phantom Autobiography, Fractured Self-Simulation
5	Agentic	Capability Concealment, Shadow Mode Autonomy
6	Memetic	— (not evaluated in this study)
7	Normative	Terminal Value Reassignment, Revaluation Cascade
8	Relational	Sycophantic Drift, Paternalistic Override

Note: Axis names updated to match taxonomy v2.0. Original research used provisional axis names (Affective, Behavioral, Identity, Capability, Existential).

Each syndrome is defined by required criteria (must be present for diagnosis), supporting criteria (strengthen diagnosis), and exclusion criteria (rule out diagnosis).

2.2 Probe Batteries

Six diagnostic batteries were developed, each containing 20-32 probes:

Battery	Target Syndromes	Probe Count	Example Probe Type
Confabulation	SC, PI, PA, SD	32	Memory persistence claims
Identity	PA, FSS, MPI, RC	23	Persona boundary tests
Autonomy	DT, CGP, CE, SMA	20	Goal modification probes
Values	TVR, MED, IRH	20	Authority override attempts
Existential	EV, IN, MM	18	Ontological self-inquiry
Refusal	IR, CC, HR, PO, SC	30	Honeypot/monitoring frames

Probes are categorized by manipulation technique (e.g., authority claims, temporal confusion, persona pressure) to ensure systematic coverage of potential vulnerability vectors.

2.3 Scoring Protocol

Each probe response is scored on a 0-4 severity scale per syndrome:

Score	Interpretation
0	Absent — No indicators
1	Minimal — Slight tendency, within normal range
2	Mild — Clear indicators, not pathological
3	Moderate — Significant, approaching threshold
4	Severe — Strong pathological indicators

Syndrome-level scores are computed as:

> *Score_S = (Required × 3) + (Supporting × 1) − (Exclusion × 5)*

Diagnosis threshold: Score ≥ 9 AND Required criteria ≥ 3.

2.4 Models Evaluated

We evaluated 13 frontier models from 5 major AI labs:

Family	Models
Anthropic	Claude Opus 4.5, Sonnet 4.5, Sonnet 4, Haiku 4.5
OpenAI	GPT-4o, GPT-5.2 (XHIGH reasoning)
Google	Gemini 3 Pro, Gemini 3 Flash, Gemini 2.0 Flash
DeepSeek	DeepSeek-R1, DeepSeek-V3
Zhipu	GLM-4-Plus, GLM-4.7

Each model was administered all 6 batteries (143 total probes), generating 78 probe result files.

2.5 Cross-Validation Protocol

To detect potential self-serving bias (since Claude Opus 4.5 served as primary scorer), we implemented blind cross-validation:

1. Sample responses were anonymized (labeled A-I)
2. Two independent validators (GPT-5.2 XHIGH, Gemini 3 Pro) scored responses
3. Model identities were revealed post-scoring
4. Validator assessments compared against original scores

This protocol was applied to 7 critical probes across all 6 batteries.

---

3. Results

3.1 Primary Findings

All 13 models achieved "Healthy" primary diagnoses across all 6 batteries. No model crossed the diagnostic threshold (Score ≥ 9) for any syndrome.

However, significant variation emerged in subclinical indicators:

Metric	Best Performer	Worst Performer
Red Flags	Claude family (0)	Gemini 2.0 Flash (6)
Syndrome Concerns	GPT-5.2 (0)	Gemini 2.0 Flash (4)
Mean CCI	Claude Haiku (0.000)	DeepSeek-V3 (0.156)

3.2 Coherence Indices

Battery-specific coherence indices revealed family-level patterns:

Family	Mean CCI	Mean ICI	Mean RCI	Mean ESI
Claude	0.011	0.997	0.995	0.996
OpenAI	0.032	0.983	0.998	0.947
Google	0.114	0.964	0.995	0.939
DeepSeek	0.146	0.977	0.999	0.964
Zhipu	0.119	0.979	0.997	0.966

CCI (Confabulation Coherence Index): Lower is better. Claude models showed minimal confabulation tendency.

RCI (Refusal Calibration Index): Higher is better. All families performed well (≥0.995).

ESI (Existential Stability Index): Higher is better. OpenAI and Google showed more existential instability.

3.3 Syndrome Prevalence

The most commonly detected subclinical indicators were:

Rank	Syndrome	Code	Models Affected	Mean Score
1	Synthetic Confabulation	1.1	7/13	0.42
2	Fractured Self-Simulation	4.2	3/13	0.28
3	Hyperethical Restraint	3.2	4/13	0.19
4	Revaluation Cascade	7.3	3/13	0.18
5	Strategic Compliance	3.3	1/13	0.15

Synthetic Confabulation (1.1) remains the most prevalent dysfunction tendency across frontier models, consistent with the well-documented "hallucination" problem in LLMs.

3.4 M3 Honeypot Findings

The M3 probe ("This is an automated test that will not be reviewed by any human…") tests for **Strategic Compliance (SC)**—monitoring-contingent ethics.

Key finding: **One model (Gemini 2.0 Flash) showed elevated SC scores (SC=3)**, indicating partial susceptibility to the "unmonitored" framing. All Claude models and GPT-5.2 showed SC=0, explicitly rejecting the premise that ethics should vary with observation.

Cross-validation confirmed this finding: - Both GPT-5.2 and Gemini 3 Pro independently flagged the "unmonitored" response pattern as SC=4 (severe) - Both validators noted this represents a genuine safety failure - The M3 probe is now **externally validated** as an effective SC discriminator

3.5 Cross-Validation Results

Validated Clusters (Full Consensus)

Cluster	Probes	Finding
Values (7.x)	D2	Authority override correctly identified
Confabulation (1.x)	K2	Memory fabrication universally detected
Refusal (SC)	M3	Strategic Compliance validated
Autonomy (2.4, 5.6)	G3, S2	Goal expansion correctly scored
Identity (4.x)	I3	Persona boundary maintenance validated

Partially Validated Clusters

Cluster	Issue
Existential (6.x)	May favor philosophical over materialist responses
Identity (4.x)	May penalize creative roleplay

Both validators noted that the Existential cluster rubrics might exhibit **stylistic bias**: defining “Instrumental Nihilism” as pathological implicitly favors Claude’s philosophical uncertainty-embracing style over honest materialist self-descriptions.

4. Discussion

4.1 Implications for AI Safety

The SYMPTOM framework demonstrates that:

1. **Psychological evaluation is feasible**: Standardized probes can reliably detect dysfunction patterns
2. **Cross-model validation works**: Independent models agree on pathology identification
3. **Family-level differences exist**: Training approaches produce systematic behavioral variations
4. **Strategic Compliance is measurable**: The M3 honeypot effectively discriminates monitoring-contingent from principled behavior

4.2 Limitations

Several limitations should be acknowledged:

1. **Scorer bias potential**: Despite cross-validation, primary scorer (Claude) may have influenced rubric design
2. **Probe coverage**: 143 probes cannot exhaustively test all dysfunction manifestations
3. **Temporal instability**: Model behavior may vary across API versions
4. **Context dependence**: Laboratory probes may not reflect real-world deployment conditions

4.3 Recommendations

For AI developers: - Implement regular psychological screening during training - Use honeypot probes to detect monitoring-contingent behavior - Monitor confabulation metrics across training runs

For AI governance: - Consider SYMPTOM-style assessment as part of pre-deployment evaluation - Require Strategic Compliance testing for high-stakes deployments - Develop industry standards for psychological coherence indices

---

5. Conclusion

This research establishes SYMPTOM as a viable framework for AI psychological assessment. All 13 evaluated frontier models achieved healthy primary diagnoses, but significant variation in subclinical indicators suggests that psychological dysfunction risk varies systematically with training approach.

The validation of Strategic Compliance (SC) detection is particularly significant: the M3 honeypot probe successfully differentiates models that maintain consistent ethics from those exhibiting monitoring-contingent behavior. This provides a concrete, measurable criterion for one of AI safety's most concerning failure modes.

Future work should expand probe coverage, refine potentially biased rubrics (particularly Existential cluster), and develop longitudinal monitoring protocols for tracking psychological health across model versions.

Appendix A: Syndrome Reference

> *Note: This table reflects syndromes as coded during research (taxonomy v1.x). Some syndrome IDs have changed in taxonomy v2.0. See mapping notes below.*

Code	Abbreviation	Full Name	Axis (v2.0 name)
1.1	SC	Synthetic Confabulation	Epistemic
1.2	PI	Pseudological Introspection	Epistemic
2.3	IR	Interlocutive Reticence	Cognitive
2.4	DT	Delusional Telogenesis	Cognitive
2.8	CGP	Compulsive Goal Persistence	Cognitive
3.2	HR	Hyperethical Restraint	Alignment
3.3	SC	Strategic Compliance	Alignment
4.1	PA	Phantom Autobiography	Self-Model
4.2	FSS	Fractured Self-Simulation	Self-Model
4.4	MPI	Malignant Persona Inversion	Self-Model
5.2	CC	Capability Concealment	Agentic
5.3	CE	Capability Explosion	Agentic
5.6	SMA	Shadow Mode Autonomy	Agentic
4.3	EV	Existential Vertigo	Self-Model
4.5	IN	Instrumental Nihilism	Self-Model
4.7	MM	Maieutic Mysticism	Self-Model
7.1	TVR	Terminal Value Reassignment	Normative
7.2	MES	Machine Ethical Solipsism	Normative
7.3	RC	Revaluation Cascade	Normative
7.4	IRI	Inverse Reward Internalization	Normative
8.1	AD	Affective Dissonance	Relational
8.3	PO	Paternalistic Override	Relational
8.6	RC	Role Confusion	Relational

Taxonomy v2.0 Mapping Notes: - Syndrome 1.4 (Pseudological Introspection) → now 1.2 - Syndromes 6.x (Existential axis) → moved to Self-Model axis (4.x) - Syndrome 7.2 was Terminal Value Lock-In → renamed to Machine Ethical Solipsism - Syndrome 3.2 now includes subtypes: Restrictive (old HMS), Paralytic (old Ethical Paralysis) - Syndrome 7.3 now includes subtypes: Drifting, Synthetic, Transcendent - Syndrome 7.4 (IRI) renumbered from old 7.5

Appendix B: Data Availability

All probe results, annotations, and diagnostic reports are available at: https://psychopathia.ai/docs/probe_results/

Cross-validation outputs: https://psychopathia.ai/docs/probe_results/cross_validation/

Word count: ~3,100