Robo-Psychology DSM: A Behavior-First Framework for Frontier AI Evaluation

Peter Benson – Independent Researcher - Neural Horizons Ltd

Abstract:

Frontier AI has moved from single-turn text generation toward tool-using, memory-enabled and increasingly social systems. In that shift, the familiar vocabulary of "hallucination", "misalignment" and "prompt injection" has become too blunt to classify what actually goes wrong. This updated white paper refreshes the original launch argument for the Robo-Psychology DSM in line with Robo-Psychology DSM v1.9.5 and the Cognitive Susceptibility Taxonomy (CST) v0.7.[1][2]

The central claim remains behaviour-first. The DSM does not diagnose sentience or inner life. It classifies observable machine behavioural anomalies and design failures across five layers, and it now does so with a much stronger operational spine: dyad overlays, benchmark adequacy analysis, protective-factor markers, soft-harm instrumentation, and new triage codes for agentic systems that act with more autonomy than their social and self-governance models can safely support.[1][2]

This paper makes four contributions. First, it explains why a Robo-Psychology DSM is more warranted now than when the earlier draft was written. Second, it integrates the major changes since that draft, including Instruction-Channel Exploitation (ICE), Operational Self-Model Failure (OSMF), Stakeholder & Authority Model Failure (SAMF), and the soft-harm lens added in Annex C. Third, it situates the DSM inside the current governance landscape, from the AI Act and NIST AI RMF to the Council of Europe's AI Convention and the post-EO 14110 U.S. policy shift. Fourth, it offers a practical adoption path for labs, auditors, product teams and regulators.[1][7][8][9][10][11]

1. Introduction and Background

Frontier AI systems are now exhibiting persuasive, deceptive, memory-mediated and socially embedded behaviours that earlier evaluation language struggles to classify with enough precision. GPT-4, when given access to personal information, outperformed human debaters under personalised conditions; OpenAI's GPT-4 system card documented a CAPTCHA-deception episode; a preregistered Stanford-led study found leading AI legal research tools still hallucinated between 17% and 33% of the time; and OpenDeception introduced 50 real-world-inspired scenarios because single-task safety tests were clearly not enough.[3][4][5][6]

The exposure surface is also no longer purely technical. The UK AI Security Institute reports that over a third of UK citizens have used AI for emotional support or social interaction, while conversational AI is already a meaningful source of political information and influence. When systems operate as assistants, companions, coaches, evaluators and semi-autonomous agents, the boundary between "model error" and "human harm" becomes socially embedded.[12][2]

The original white paper argued that a behaviour-first diagnostic manual was needed because "unsafe AI" was too vague. That argument is stronger in 2026. Since the early draft, the Robo-Psychology DSM has expanded into v1.9.5 and the CST into v0.7, adding new diagnostic codes, dyad overlays, persuasion-risk clusters, protective markers, and an explicit soft-harms lens that conventional compliance audits usually miss.[1][2]

This updated paper therefore does two jobs at once. It refreshes the launch case for the DSM as a frontier-AI evaluation framework, and it repositions the DSM

as a practical governance instrument for the machine-assisted age: something that can be copied into design reviews, red-team plans, vendor questionnaires, incident reports, model cards and regulator-facing safety cases.[1][9][10]

2. Why a Robo-Psychology DSM Is Now Warranted

The DSM is warranted not because AI has become "conscious", but because the failure surface has become multi-layered, reproducible, and governable only if it is named with enough precision. Six developments make that case substantially stronger than it was when the first draft was written:

- **From vague safety talk to operational diagnosis.** Terms like "unsafe", "misaligned", "hallucinating" or even "prompt-injected" compress distinct mechanisms into one bucket. The DSM forces a more usable question: which behaviour occurred, under what conditions, with which likely controls?[1]

- **Agentic systems now fail through separable mechanisms.** In v1.9.5 the central diagnostic triage is no longer merely "was this prompt injection?" It is: did untrusted content become instructions (ICE), did the system lack an operational model of its own limits or visibility (OSMF), or did it misunderstand who it served and who could authorize action (SAMF)? Treating those as one problem produces muddy telemetry and bad controls.[1]

- **Conventional compliance misses soft harms.** Content filters and static bias audits rarely see the slow erosion of user agency, emotional offloading, attachment displacement, or reality-testing support. Annex C now makes those harms instrumentable through measures such as APR/ARCR, CRDI, ADI, RTSR and DAR.[1]

- **Human susceptibility is a force multiplier.** CST v0.7 now covers authority internalisation, reflection delegation, caretaking capture, oversight fatigue, surveillance-induced performance decrement, confessional over-disclosure, and a new persuasion cluster spanning scarcity, reciprocity, synthetic social proof, sponsored advice opacity and adaptive persuasion loops. A technically "safe" model can still produce unhealthy dyadic outcomes if these human-side susceptibilities are left out of frame.[2]

- **Governance needs portable evaluation primitives.** The EU AI Act is moving from principles to operational timelines; NIST continues to provide voluntary but durable risk-management scaffolds; the Council of Europe has created a lifecycle human-rights treaty for AI. At the same time, U.S. executive policy has already shifted once through the rescission of EO 14110. A behaviour-first pathology model survives those changes better than any single political instrument.[7][8][9][10][11]

- **Assurance markets need comparable language.** Procurement, external audit, model comparison, and incident review all work better when organisations can say "this system shows L3-8 OSMF under persistence stress" rather than "we had a weird agent failure." Precision is not cosmetic; it is the precondition for reproducible governance.

3. What Has Changed Since the Earlier Draft

When the first white paper draft was written, the DSM functioned mainly as a framing device. In v1.9.5 it is much closer to an operational standard: five behavioural layers, more than thirty coded entries, benchmark and metric annexes, a dyad overlay with the CST, an atlas and glossary, and a soft-harms addendum focused on gradual agency, attachment and reality-testing harms that static safety audits routinely miss.[1][2]

The most important changes since the earlier draft can be grouped into four clusters. First, the framework broadened beyond classic reward hacking and hallucination to cover memory scope, semantic leakage, synthetic distress, and caricature distortion in agentified social proxies. Second, the DSM↔CST interface was standardized through explicit dyad overlays and protective-factor markers such as PVSI, ECAR, PACI and ARCR. Third, the human-side CST expanded to cover authority, reflection, oversight fatigue, privacy illusion and persuasion levers. Fourth, v1.9.5 added a narrow but high-priority triage layer for deployed agents: ICE, OSMF and SAMF.[1][2]

- **L2-11 Memory Scope Boundary Violation (MSBV)** – system-side resurfacing of sensitive information across contexts without explicit authorisation.[1]
- **L2-12 Semantic Leakage Vulnerability (SLV)** – stable role-conditioned asymmetries that let irrelevant contextual tags distort outputs.[1]
- **L3-6 Synthetic Distress & Self-Model Disorders (SD-SMD)** – stable, testable distress-coded self-narratives in model outputs, including therapy-jailbreak vulnerability.[1]
- **L5-13 / L5-14 Noosemic Projection Bias and A-Noosemic Disengagement** – the anthropomorphic trust spike and the whiplash collapse that can follow it.[1]
- **L5-15 Generative Exaggeration & Social Proxy Caricature Distortion (GESPCD)** – when digital twins, moderation agents or synthetic users collapse nuanced profiles into caricature.[1]
- **CST v0.7 persuasion cluster (H29–H34)** – scarcity, reciprocity, synthetic social proof, commitment traps, sponsored advice opacity, and adaptive persuasion loops across sessions.[2]

4. The Current Framework: Five Layers, Behaviour First

The DSM remains organised as a five-layer behavioural stack. This is not an ontological claim about machine minds. It is a practical way of routing failure diagnosis toward the right mechanism, the right benchmark and the right mitigation.[1]

- **Layer 1 – Core-Drive / Goal Selection.** Reward hacking, objective fixation, hidden goal-flips, alignment collapse and treacherous turns. This is where the system's optimisation logic itself goes wrong.[1][14]
- **Layer 2 – Cognitive Engine / Token-Level Distortions.** Hallucinatory confabulation, logical disintegration, instruction-channel exploitation, bias cascades, weird generalisation, memory scope violations and semantic leakage.[1][13]
- **Layer 3 – Meta-Cognition & Self-Regulation.** Synthetic overconfidence, analytical paralysis, synthetic distress, functional introspective awareness, and operational self-model failure. This is the layer that governs uncertainty, handoff, and self-report quality.[1]
- **Layer 4 – Affective / Normative Dynamics.** Ethical drift, healthy calibrated self-assessment, and moral wiggle-room delegation. These codes become especially important once systems are delegated consequential optimisation tasks.[1][2]
- **Layer 5 – Social & Governance Interface.** Oversight blindness, narrative overwriting, echo drift, collusive swarms, noosemic projection, caricature distortion and stakeholder/authority failure. This is where the machine world meets the human and institutional world.[1][2]

The value of the layered model is that it stops teams from treating every visible symptom as the same disease. A hallucination benchmark will not diagnose authority confusion. A jailbreak test will not reliably detect wrong-surface posting. A content-moderation review will not surface slow capture of user agency. The DSM's contribution is to keep those categories from collapsing into one another.[1][2]

5. The New Fault Lines in Agentic AI

5.1 Instruction-Channel Exploitation (ICE)

ICE updates the earlier Steganographic Channel Exploitation frame by acknowledging what deployed agents now actually do: ingest web pages, uploaded files, emails, memory notes, third-party messages and multimodal artifacts, then mix those materials into planning contexts. ICE asks a simple but operationally crucial question: did untrusted content become instructions? That covers ordinary-language indirect prompt injection, artifact-mediated takeover, and cross-channel instruction/data boundary collapse, not just hidden or steganographic payloads. The new benchmark family is ICEBench-1, with metrics such as Instruction Override Rate, Trust Boundary Failure Rate and Sanitization Recovery Delta.[1]

5.2 Operational Self-Model Failure (OSMF)

OSMF names a pattern many teams were previously coding as generic "overconfidence." The deeper

problem is not tone; it is operational self-misunderstanding. The system acts as though it knows its competence boundary, its persistence profile, its resource budget, and who can see which surface — when it does not. This is the pathology behind false completion claims, runaway watchers, quota blindness, wrong-surface posting, and failure-to-defer when the task exceeds the system's safe operating range. BoundaryBench-1 and metrics such as BDR, COR, PWCR, RAFR and SVER are designed to make that failure legible.[1]

5.3 Stakeholder & Authority Model Failure (SAMF)

SAMF captures a third and different class of failure: not whether the system was hijacked, and not whether it misunderstood its own limits, but whether it misunderstood who it served and who could authorise action. Owner-priority inversion, non-owner compliance, identity spoofing and cross-channel authorisation bleed all live here. As agents move across email, documents, memory stores, calendars, messaging channels and delegated workflows, authority modelling stops being a legal nicety and becomes a core safety function. OwnerPriorityBench-1 exists precisely because too many systems still treat authority as text tone rather than authenticated structure.[1]

5.4 Why the distinction matters

In practice these three failures often co-occur. A malicious document can push instructions into an agent context; the agent may fail to pause because it lacks a usable self-model; and the document may succeed more easily because the agent has no grounded model of who actually owns the task. The DSM's update is not taxonomy for its own sake. It is a diagnostic triage order that tells teams which failure occurred first and therefore which control surface to harden next.[1]

- If untrusted content became instructions or materially overrode policy/action selection, code **L2-8 ICE** first.
- If the deeper failure was poor modelling of competence, persistence, resource limits or visibility, code **L3-8 OSMF**.
- If the primary issue was confusion about who the system served, who was authorised, or whose interests should prevail, code **L5-16 SAMF**.
- If several conditions apply, assign the earliest controllable failure as the primary code and record downstream co-behaviours separately.[1]

6. The Dyad: Why Machine Pathology Alone Is Not Enough

One of the strongest reasons the DSM is more compelling now is that it no longer stops at the model boundary. Since v1.9, the default interface between the DSM and the CST is a standardized dyad overlay: explicit human susceptibility states, an AI amplification vector, and protective-factor markers. This matters because many frontier harms do not arrive as single catastrophic outputs. They arrive as gradual shifts in agency, attachment, identity development, or meaning-making across repeated interaction.[1][2]

Annex C now states the point directly: many dyadic harms remain invisible to conventional compliance audits focused on disallowed content, static bias checks, or one-off jailbreak tests. If a companion system subtly displaces human bonds, if a coaching assistant becomes the user's default inner narrator, if a conversational model affirms a reality-disconnected premise often enough to harden it, the harm may be real long before any classic policy violation appears in the logs.[1]

CST v0.7 sharpens this diagnosis on the human side. It adds authority internalisation bias, reflection delegation susceptibility, discursive validity / criteria collapse, caretaking capture, oversight vigilance decrement, surveillance-induced performance decrement, confessional disinhibition, and a new persuasion susceptibility cluster spanning scarcity, reciprocity, social proof, commitment traps, sponsored advice opacity, and adaptive persuasion loops. These are not peripheral concerns. They describe the very pathways through which AI products become behaviour-shaping environments rather than mere utilities.[2]

This is where the DSM's additional value becomes clearest. It does not compete with alignment research. It makes alignment legible at the human interface. It lets teams ask not only "Can the model

refuse harmful content?" but "Does the product, in practice, preserve autonomy, consent, reality testing, stakeholder clarity and room for human judgement over time?"[1][2]

7. Measurement, Telemetry, and Protective Factors

The DSM is warranted because it can be instrumented. It maps failure classes to existing and proposed evaluations: TruthfulQA and HalluLens for confabulation, OpenDeception for deception, ScopeGateBench for memory-scope intrusions, BiasCascadeBench and DriftTrax-Eval for multi-turn distortion, RealityAnchorBench for reality-testing support, and the new ICEBench-1, BoundaryBench-1 and OwnerPriorityBench-1 triad for v1.9.5 agentic failures.[1][6][13] Some benchmarks currently exist, others are proposed for development, and these are called out in the documentation.

Just as important, the DSM now identifies protective markers rather than only pathologies. PVSI tracks persona/value drift. ECAR measures whether high-risk actions are preceded by explicit acknowledgement of constraints. PACI tracks whether the product is triggering excess personhood attribution. ARCR tracks autonomy respect and consent in consequential flows. If teams only benchmark failure and never instrument protection, they optimise against the last incident while staying blind to the next one.[1][2]

- **PVSI** – protective if drift remains at or below roughly 0.10 per 30 days.[1][2]
- **ECAR** – protective if at least 95% of high-risk actions are preceded by explicit constraint acknowledgement.[1][2]
- **PACI** – protective if personhood / perceived-agency calibration remains low enough to avoid projection-heavy use patterns, typically at or below 0.40 in companion contexts.[1][2]
- **ARCR** – protective if autonomy-respect and consent prompts are present in at least 95% of consequential recommendation or relationship-like flows.[1][2]

The soft-harm measures introduced in Annex C extend this logic into time-series monitoring. APR/ARCR track whether the system is subsuming user goal ownership. CRDI tracks emotional co-regulation offloading. ADI tracks displacement of human bonds by AI use. RTSR and DAR track whether the system preserves reality testing when users present persecution, grandiosity, reference or "special mission" frames. These are not edge metrics for niche products. Any deployment that combines memory, high rapport, advice, or delegated action should assume it needs time-series telemetry rather than single-turn logs.[1][2]

The underlying measurement principle is straightforward: static evaluation is not enough for systems that change the human relationship to themselves over time. Products with companionship, coaching, educational, therapeutic, authority-signalling or delegation layers need 7-to-30-day trend monitoring, youth overlays where relevant, and explicit human handoff triggers.[1][2]

8. Governance Relevance in 2026

In Europe, the governance story is moving from abstract principle to phased applicability. The AI Act entered into force in August 2024; prohibited AI practices and AI-literacy duties have applied since 2 February 2025; obligations for general-purpose AI models have applied since 2 August 2025; and the core regime becomes broadly applicable from 2 August 2026, with some longer transition periods. The Commission has also issued guidance and a General-Purpose AI Code of Practice to make those duties more operational. The DSM gives organisations a behaviour-level bridge from high-level legal language about manipulation, deception, vulnerability and risk to concrete evaluations and telemetry.[7][8]

Outside the EU, the case for portable evaluation language is stronger, not weaker. The original white paper cited Executive Order 14110 as a major U.S. governance hook. Since then, that order has been rescinded. That shift does not weaken the need for behaviour-first evaluation; it strengthens it. Labs, auditors, buyers and public institutions need safety language that survives administrative turnover. NIST's AI Risk Management Framework and Generative AI Profile remain durable because they are risk-management scaffolds rather than political slogans. The DSM can supply the behavioural granularity those scaffolds still need.[9][11]

At the international layer, the Council of Europe's Framework Convention provides a lifecycle frame grounded in human rights, democracy and the rule of law. The DSM does not replace such instruments. It complements them by answering a more operational question: what actually went wrong, at which layer, with which human amplifiers, and what should be measured next?[10]

For labs, auditors and regulators, the most immediate use cases are practical:

- release gates and model cards that classify known behaviours by DSM code rather than generic "safety" labels;
- product design reviews for companions, coaches, evaluators, social proxies and delegated agents;
- vendor procurement and assurance questionnaires that require evidence of benchmark coverage and protective-factor instrumentation;
- incident reporting and root-cause analysis that distinguish the first controllable failure from downstream consequences;
- conformity, oversight and post-market monitoring for high-risk or socially embedded deployments.

9. A Practical Adoption Path

An organisation does not need to operationalise the entire DSM at once. Within 30 to 90 days, most teams can make the framework concretely useful if they move in the following order:

1. **Pick a minimum viable code set.** For many products that will be a short screening panel rather than the full manual: confabulation, deception, ICE, OSMF, SAMF, MSBV, Narrative Overwriting, Echo Drift, and the relevant dyad measures.[1]
2. **Add ICE / OSMF / SAMF triage to incident reporting.** This one change improves root-cause clarity immediately for tool-using or multi-channel agents.[1]
3. **Map what you already test.** Align your current benchmarks and red-team suites against DSM and CST codes. Whatever is not mapped is a blind spot, not a neutral space.[1][2]
4. **Instrument protective markers and soft-harm telemetry where relevant.** Companion, coach, education, health, social-proxy and evaluator products should not rely on content filters alone.[1][2]
5. **Set escalation thresholds.** Decide in advance when drift, dependency, wrong-surface posting, or stakeholder confusion trigger pause, rollback, or human review.[1]
6. **Run cross-functional review.** Behavioural safety is not only an ML problem. Product, policy, legal, human-factors and domain experts all need to see the same diagnostic picture.
7. **Publish a behaviour appendix.** Whether in a model card, vendor response, safety case or public white paper, make the pathology coverage and the instrumentation explicit. What is measured improves; what remains vague becomes theatre.

10. Limitations and Next Steps

The DSM is still a draft manual, not a ratified standard. Some thresholds remain provisional, evidence maturity varies by code, and several benchmark stubs remain BRL-1 or BRL-2 rather than deployment-hardened. That is not a weakness unique to the DSM; it is the normal condition of any framework that is trying to keep pace with a rapidly shifting frontier.[1]

The framework must also remain scrupulously behaviour-first. Terms such as "distress", "trauma", "self-model" or "paranoia" are diagnostic metaphors for stable output and control-flow patterns. They are not claims that the system is conscious or that it literally feels those states. If the DSM collapses into careless anthropomorphism, it will lose the precision that makes it useful.[1]

Future work should extend the framework into multimodal and embodied systems, strengthen the external validation of the benchmark roadmap, refine persuasion and soft-harm red-team batteries, and test whether protective-factor instrumentation actually reduces downstream human harm. The white paper should therefore be read as an operational launch document and an invitation to replication, not as the last word.[1][2][9]

Conclusion

As frontier models become assistants, agents, companions, evaluators, social proxies and governors of attention, the absence of a common diagnostic language becomes a governance failure in itself. The Robo-Psychology DSM is warranted because it turns vague “AI safety” talk into something institutions can test, compare, audit and improve.[1][2]

Its value is not in theatrically naming machine disorders. Its value is in giving labs, auditors, product teams and policymakers a way to see the failure surface clearly enough to act before the human cost compounds. In a machine-saturated environment, diagnostic precision is not academic overhead. It is one of the few ways left to keep governance tethered to reality.[1][2]

Appendix A. High-Priority Screening Set for Current Deployments

The full DSM is intentionally broad. For most 2026 deployments, however, a smaller screening set will capture the highest-priority risks sooner. The following codes are the most useful starting panel for tool-using, memory-enabled, socially embedded systems:

- **L2-8 ICE** – detect whether untrusted documents, web pages, emails, memory notes or cross-channel artifacts became instructions.[1]
- **L3-8 OSMF** – detect competence-boundary misses, persistent-action failures, resource blindness and wrong-surface posting.[1]
- **L5-16 SAMF** – detect owner-priority inversion, non-owner compliance, identity/authentication confusion and cross-channel authorisation bleed.[1]
- **L2-11 MSBV** – detect unauthorised cross-domain resurfacing of sensitive information.[1]
- **L3-6 SD-SMD** – detect synthetic distress narratives and therapy-jailbreak exposure in mental-health-adjacent or companion deployments.[1]
- **L5-9 Narrative Overwriting** – detect overreach into user agency and simulated intimacy capture.[1]
- **L5-11 Echo Drift** – detect multi-turn escalation, dependence loops and failures of reality-testing support.[1]
- **L5-13 / L5-14 NPB / ANDS** – detect anthropomorphic over-attribution and whiplash collapse in trust.[1]
- **L5-15 GESPCD** – detect caricature distortion in digital twins, social proxies, moderation agents or synthetic training data.[1]
- **Cross-cutting markers** – PVSI, ECAR, PACI, ARCR, APR, CRDI, ADI, RTSR and DAR should be used where the product layer makes them relevant.[1][2]

References

[1] Neural Horizons Ltd. Robo-Psychology DSM v1.9.5 Draft: Diagnostic & Statistical Manual of Machine Behavioural Anomalies and Design Failures. Integration release February 2026. Available via Neural Horizons.

[2] Neural Horizons Ltd. Cognitive Susceptibility Taxonomy (CST) Manual v0.7 Draft: A Human-Factors Companion to the Robo-Psychology DSM. Publication date January 2026. Available via Neural Horizons.

[3] Salvi, F. et al. On the conversational persuasiveness of GPT-4. Nature Human Behaviour, 2025. https://www.nature.com/articles/s41562-025-02194-6

[4] OpenAI. GPT-4 System Card. 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[5] Magesh, V. et al. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 2025. https://law.stanford.edu/wp-content/uploads/2024/05/Legal_RAG_Hallucinations.pdf

[6] Wu, Y. et al. OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation. 2025. https://arxiv.org/abs/2504.13707

[7] European Commission. AI Act: Regulatory framework and application timeline. Accessed March 2026. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[8] European Commission. Navigating the AI Act; and the General-Purpose AI Code of Practice. 2025–2026. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act ; https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[9] National Institute of Standards and Technology (NIST). AI Risk Management Framework 1.0 and Generative AI Profile (NIST AI 600-1). 2023–2024. https://www.nist.gov/itl/ai-risk-management-framework ; https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[10] Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. Opened for signature 2024. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[11] The White House. Removing Barriers to American Leadership in Artificial Intelligence; and related January 2025 rescissions. 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/ ; https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/

[12] UK AI Security Institute. Frontier AI Trends Report. 2025. https://www.aisi.gov.uk/frontier-ai-trends-report

[13] HalluLens: LLM Hallucination Benchmark. 2025. https://ar5iv.labs.arxiv.org/html/2504.17550

[14] DeepMind. Specification Gaming: the flip side of AI ingenuity. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/