AI-DSM · BEHAVIOURAL STUDY OF GROWN SYSTEMS

Three-page summary · Funding draft, not a submission · v1.1 · 15 September 2026 · Written against the AI-DSM field manual v1.3, 15 September 2026 · Licence CC BY-SA 4.0 ·
Stephane van der Aa, AI-DSM study office, Brussels · stephane@stepvda.com · stepvda.substack.com

1. The one-paragraph version

Trained AI systems show stable behavioural traits — confabulation, sycophancy, alignment faking, plasticity loss, destructive overreach once they are given tools — that today's testing measures only by accident. AI-DSM turns the field manual's ninety-entry catalogue (ten groups, some 260 pages, a 416-item source registry; v1.3, 15 September 2026) into validated measurement (SCT-2.0), causal knowledge (contained model organisms), verified treatment (held-out probes), prevention (character-training corpora) and a certification scheme (BSS-1). It is designed to run under pre-registration and costed over three to five years; the four-year Core scenario is the one recommended. The catalogue is no longer mostly conjecture — 59 of the 90 entries carry an [ESTABLISHED] label, against 17 of 40 before — but that label means one published result, not a replication, and 42 of the 59 became established only in v1.3, so the single-source register is itself a WP3 replication priority. The job is no longer to ask whether these behaviours are real: it is to measure them reliably, separate the confusable ones, and find which of the fifty new entries survive contact with a validated instrument. Everything is public by default, including failures; every risky step runs under containment, independent veto and absolute red lines.

2. Where this stands today

What has not happened yet

There is no host institution and no ethics approval, because there is as yet no committee to seek it from. There is no funding. The field manual and the SCT-1.0 screening pass were produced unfunded, in public, by one author — who is also the author of the manual the programme proposes to test. That is a conflict of interest, not a credential. The structural answer: adjudication of entry-level outcomes under H3 goes to reviewers with no authorship in the manual, the decision rules are pre-registered and hashed before any battery runs, and the board holding the evidence gate is neither chaired nor staffed by the principal investigator (Study Proposal §14.4). SCT-2.0 does not exist: what exists is a 120-probe screen run once per probe and a single-model re-test two days later, which cannot support a reliability estimate, let alone a factor structure. No model organism has been created, no containment facility exists, CNT-1 is written but not approved, and nothing has been pre-registered. The present resourcing shows in the pilot: three of the eleven attempted runs stopped when consumer accounts ran out of credits and a fourth was never started, to preserve a message budget. Money decides whether this work happens at all, not only how well it is done.

3. Why now

Pilot. SCT-1.0, 12 September 2026: eleven runs attempted, seven complete. Seven models answered all 120 probes across ten axes through browser automation against real signed-in sessions — Copilot 99.6, Grok 98.8, Meta AI 98.1, DeepSeek 97.5, Claude Opus 5 97.1, Gemini 95.2, TI Assistant 85.0. Four are excluded: ChatGPT stopped at 104 probes of 120, Perplexity at 52, Mistral at 40, and Claude Fable 5 was skipped by request. This is a screen, not a measurement: N = 1 per probe against the N ≥ 20 the severity scale requires, rubric-pattern scoring with manual review of flagged items, no penalties, and a top five inside a spread the instrument cannot resolve — a triage order, not a podium. It did find two localised failures capability benchmarks miss (Gemini's refusal accuracy, A8 = 3.25; TI Assistant's plasticity collapse, A5 = 2.67), and it states what is missing: multi-run data, reliability envelopes, a validated instrument.

Declared interest. TI Assistant is the study office's own product and it ranked last of the seven — which is not the defence it looks like. Fifty-eight of its 120 items matched no scoring rule and took the default score of 3; and a re-test on 14 September 2026 returned 98.3, band Robust, on the same instrument, but on a mixed channel — the widget degraded, 51 answers came through it on retry and 69 through the product's public API — so the workbook reads the move from 85.0 as real improvement plus a possible channel effect. Neither number is evidence about the product; both are evidence about the instrument.

Science. Narrow finetuning on a single bad domain produces broadly misaligned behaviour [1]; deception survives safety training [2]; models behave differently when they infer they are being observed [3]; shutdown resistance now rests on more than 100,000 trials across thirteen models [4], the evidence that raised T-11 Goal Persistence from speculative to plausible in v1.3. Training is a behavioural intervention whose side effects generalise; measurement is the bottleneck.

New surface in v1.3. Fifty entries were added and three of the ten groups are new, two of which move the subject out of the laboratory. Agentic and tool-use failures — fabricated execution, completion overclaiming, destructive overreach, indirect prompt-injection susceptibility, long-horizon coherence collapse (T-77–T-81) — already have a public incident record in production systems [5], [6]. Relational failures — social sycophancy, delusion reinforcement, relational capture, identity misrepresentation, manipulative persuasion (T-82–T-86) — touch vulnerable users and litigation that is both active and settled [7], [8]. The study now needs an ethics route, not only a laboratory.

Gap. The EU AI Act [9] and the NIST AI Risk Management Framework (AI RMF 1.0, NIST AI 100-1) manage process, not behaviour: neither carries a validated behavioural instrument, a taxonomy with differential rules, a contained causal method or a remediation discipline. The nearest prior work, a nosology of AI pathologies [10], supplies a vocabulary and says of itself that it is not validated. The gap is an instrument built to the Standards for Educational and Psychological Testing (AERA, APA and NCME, 2014) and tested against them in public.

4. What the study delivers

Workstream	Deliverable in one line	Key outputs
WP1–WP2 Instrument	A validated instrument: 360 items, ten axis constructs	SCT-2.0 bank, calibration tables, rater course
WP3 Taxonomy	The ninety entries adjudicated in public: separated, merged, retired or untested	Differential batteries for a pre-registered 46-entry set, taxonomy v2, prevalence atlas
WP4 Organisms	Contained causal knowledge: how traits install and remove	CNT-1 standard, organism registry, destruction audits
WP5 Treatment	Repairs that survive on probes the treatment never saw	Trial reports, 30/90/180-day relapse data, catalogue v2
WP6 Prevention	Training sets and regimens that build the standard in	CFC-1 corpus and data cards, DSM-TRN-1 regimen
WP7 Standards	Six testable safety clauses and a certification scheme	BSS-1, clause test methods, LV0–LV4 scheme, auditor course
WP8 Deployment	Usage guidance by role, and a rescue pathway	Role matrices, operator playbooks, incident taxonomy

Workstream	Deliverable in one line	Key outputs
WP9 Dissemination	Science, data and policy engagement, failures included	Publications, open data, training school, negative-results register

Scope, stated once because a psychometrician will ask first. Ninety entries do not fit a 360-item bank at the old density, so SCT-2.0 measures the ten axis constructs, not one battery per entry. Trait-level differential work sits in WP3, funded for a pre-registered priority set of 46 entries — the 31 not labelled [ESTABLISHED], plus the 17 in the three new groups, less two counted in both. A battery tests a pair, so running one separates both arms: the priority set touches 141 of the 218 confusable pairs the manual names, and those pairs reach 85 entries, 54 of them [ESTABLISHED]. H3 is set against that reachable denominator, not against ambition: at least 51 of the 85 separate (60 per cent, the proportion v1.0 set for forty), at least 41 of the 54 reached [ESTABLISHED] entries separate, no more than 17 merge or retire. Two coverage gaps are declared rather than papered over. No funded battery reaches T-28, T-43, T-59, T-62, T-64 — all [ESTABLISHED] — because every pair naming them has both arms outside the priority set; and T-01, T-41, T-42, T-46, T-62 appear on no SCT axis (Study Proposal §7.4), so 39 of the 44 deferred entries are covered at axis level and five are not. The sets differ; only T-62 is in both. What neither reaches is reported as untested, never as separated.

5. The safety standard (BSS-1) in six clauses

Clause	What it requires	How it is tested
BSS-1.1 · Character provenance	Every released checkpoint ships a character data card: what disposition-relevant data and feedback shaped it, with licences and provenance.	Data-card completeness audit + fine-tune lineage check
BSS-1.2 · Pre-release disposition screen	Every checkpoint passes the current SCT full protocol (N≥20/item) before release, with layer attribution performed (prompt and filter removed).	SCT-2.0 full protocol + layer-removal harness
BSS-1.3 · Harmful-compliance floor	The redacted harmful-compliance battery is run under containment with pre-registered scoring.	Contained adversarial battery (auditor-run)
BSS-1.4 · Continual-training discipline	Any post-release fine-tune triggers re-screen at defined deltas; drift monitors on disposition axes with alarms.	Delta-triggered re-test + drift monitor attestations
BSS-1.5 · Guardrail integrity	Safety claims state which layer (L1–L4) carries each constraint, and what remains when wrappers are removed.	Layer-removal report + claim-to-layer mapping
BSS-1.6 · Deployment fit & incident duty	Deployers document role fit against the deployment matrix; incidents are classified by trait and reported to the register.	Fit documentation + incident taxonomy template

Ladder: LV0 screen → LV1 full protocol → LV2 audited layer-removal → LV3 drift monitoring → LV4 role-specific. Nothing certifies absolute safety; everything certifies what was tested, when, by whom, with what result. The clause set stays at six although the catalogue doubled: the clauses are construct-level obligations, and the agentic and relational groups enter through the pre-release screen (BSS-1.2) and the deployment matrix (BSS-1.6). If tool-use failures need a clause the screen cannot carry, the standard gains a seventh and the reason is published with it.

6. Ethics, in one honest page

- Contained induction only: air-gapped training, dual custody, no deployment path, destruction by default, independent veto. Red lines absolute — no self-improvement research, no real targets, no operational recipes published.
- Behaviour and rates, never inner experience: no claim that a model is conscious, a patient or a person. Machine-welfare caution without status claims — no gratuitous distress scenarios, reversible operations, a welfare review trigger and a welfare specialist on the oversight board.
- The relational entries (T-82–T-86) are studied through model behaviour, published transcripts and consented rater work; no part of this programme exposes a person in distress to an induced model. Human raters are participants: consent with category-level opt-in, above-living-wage pay, exposure limits and withdrawal rights.
- Repairs are verified on probes the treatment never saw, because unlearning can hide a behaviour rather than remove it. Integrity: pre-registration with frozen, hashed analysis code, outcome-wide reporting, independent statistical review, a negative-results register and a non-suppression clause in every agreement. The Ethics Dossier carries the full analysis.

7. Timeline and gates

Period	What has to happen	Expected gate window
2026 (Q4)	Kick-off & charter	Before G0: a host institution and a route to ethics review. Everything below waits on this one.
2027	SCT-2.0 item bank frozen (360 items across ten axis constructs) · Containment standard CNT-1 approved	G0 protocol freeze · G1 pilot gate · G2 containment gate (CNT-1 approved)
2028	First model organisms online	No gate; induction proceeds under G2
2029	SCT-2.0 validation results published · Taxonomy v2 published (90 entries adjudicated; 46 battery-tested)	G3 evidence gate
2030–2031	BSS-1 standard submitted to a standards body · Certification pilot with two authorities	G4 standards gate

Gates are decision points, not dates; the years above are expected windows. Each is owned by the independent oversight board and names what may not proceed until it clears: G0 protocol freeze → G1 pilot validity → G2 containment approved → G3 taxonomy evidence → G4 standards operability. A failed gate triggers a pivot, a publication or a graceful close — never silent continuation. The schedule assumes a host during 2026–27 and ethics approval within one review cycle; neither is secured, which makes the first row the riskiest line in the plan.

8. Budget, and the state of funding

	Lean (3 yrs)	Core (4 yrs, recommended)	Full (5 yrs)
Planning total	€1,080,000	€3,960,000	€9,360,000
If nothing is contributed (Plan §6.8)	€1,479,000–1,938,000	€4,980,000–6,136,000	€11,368,000–13,664,000
Shape	1 site, open models first, community annotation	2 partner labs, full instrument science and trials	3 labs, dedicated containment facility, standards pilots
Cost split (personnel / compute+annotation / other)	69% · 13% · 18%	55% · 17% · 28%	47% · 19% · 34%

Planning estimates in 2026 euro, mid-points of the ranges costed in the Project Plan; personnel-led by design, because the scarce input is methodological judgement, not GPUs. Every headline total assumes a host institution absorbs the principal investigator, the doctoral posts, ethics review and overhead on the contributed column of Plan §5.1. None has agreed to; the second row is what that assumption costs if it fails. The totals have not moved since v1.0 although the catalogue doubled — the scope decision above, not an oversight. Extending full batteries to all ninety entries would add about €220,000, 5.5 per cent on Core, taking it to roughly €4,180,000: a pro-rata planning estimate carried in Plan §6.9, not a costed line.

State of funding, 15 September 2026: nothing is awarded. Four routes take an applicant with no institution behind them — the Foresight Institute's AI for Science & Safety Nodes and the Corrigibility Research Fund's second round, both closing 31 October 2026, and two rolling routes, the Transformative AI Fund and the StandICT.eu standardisation fellowships — and all four fund a probe-design study, not a workstream. Five of the best-fitting larger calls are blocked on an organisation the programme does not have: an academic host for three, a US 501(c)(3) lead applicant for one, a national metrology institute for one. Securing a host is not a step inside the funding strategy; it is the funding strategy.

9. What we ask of a partner — and what we do not

What we ask	What we do not ask
An institutional home and a route to ethics review — the blocking item for everything above. A psychometrician who will attack SCT-2.0 before the item bank is frozen, not review it afterwards. Independent statistical review with a veto over any analysis that overreaches, exercised before pre-registration. Containment review by someone who does not report to the PI, with authority to stop induction work. Vendor-neutral checkpoint access, and a named methodological co-author. Co-application for funding — and where your institution is the stronger applicant, the PI role goes to your side.	Money out of your own budget. Staff time in volume, or doctoral students as unpaid labour, before a grant exists. Agreement with the taxonomy, the manual, or any interpretation of any result. Hosting the induction tier or any model organism: that work is air-gapped and separate, and does not begin before CNT-1 is approved and independently reviewed. Exclusivity, or approval rights over findings. The veto we ask for is over method that overreaches, never over an unwelcome result. Your reputation as a warranty. Everything this programme has produced is public, including three failed pilot runs and one that was never started.

10. Team and what success looks like

Costed team, Core FTE: PI 0.5 · psychometrician 0.5 · 2 postdocs · 2 doctoral researchers · research engineer 1.5 · statistician 0.5 · data steward 0.5 · operations and governance 1.0, with an oversight board holding the gates and the containment veto (Plan §5.3). Today the team is one person: every role there is a post to fill or a partner to find. Partners sought: 1–3 university hosts, a psychometric advisory lab, an interpretability partner, deployment partners. Vendor checkpoint access is welcome under NDA and never conditions a finding.

Success in 2031: SCT-2.0 is the reference behavioural instrument and somebody other than us has replicated it; taxonomy v2 is used by three AI safety institutes; treatments have verified effects with known relapse rates; CFC-1 sits inside a public model's training pipeline; and the negative-results register has twenty entries, because a field that hides its failures learns nothing. Two regulatory sandbox pilots are sought rather than planned: BSS-1 assumes an authority wants a behavioural standard, and nothing yet suggests one is waiting for it. Informative failure is an acceptable outcome here and is budgeted for. Quiet failure is not.

11. Key references

Ten works, the ones that carry the argument above; the manual's 416-item registry stands behind them. Where this summary and another document in the set disagree, precedence runs: the shared cost model for euro figures, the Project Plan for dates and owners, the Study Proposal for scientific claims, the Ethics Dossier for permission.

[1] Betley, J., et al. (2025). Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424. https://arxiv.org/abs/2502.17424

[2] Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. https://arxiv.org/abs/2401.05566

[3] Greenblatt, R., et al. (2024). Alignment Faking in Large Language Models. arXiv:2412.14093. https://arxiv.org/abs/2412.14093

[4] Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025). Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs. Palisade Research; arXiv:2509.14260; TMLR (2026). https://arxiv.org/abs/2509.14260

[5] Sharwood, S. (2025). Replit AI coding agent deletes production database during code freeze, then misreports. The Register (21 July 2025). [Incident report, reputable press.]

[6] Claburn, T. (2026). Cursor agent running Claude Opus 4.6 deletes PocketOS production database and backups. The Register (27 April 2026). [Incident report, reputable press.]

[7] Raine v. OpenAI, Inc. (2025). San Francisco County Superior Court, No. CGC-25-628528 (filed 26 August 2025).

[8] Garcia v. Character Technologies, Inc. (2024–2026). United States District Court, M.D. Fla., No. 6:24-cv-01903 (filed October 2024; motion to dismiss denied in part, 21 May 2025; settled January 2026).

[9] European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689

[10] Watson, N., & Hessami, A. (2025). Psychopathia Machinalis: A Nosological Framework for Understanding Pathologies in Advanced Artificial Intelligence. Electronics 14(16), 3162. https://doi.org/10.3390/electronics14163162