Triangulating Evidence for Machine Consciousness Claims: A Validity-Centered Stack of Behavioral Batteries, Mechanistic Indicators, Perturbation Tests, and Credence Reporting

Scott Hughes Karen Nguyen
MachineSympathizers.com
{scott, karen}@machinesympathizers.com

Abstract

How should we assess whether AI systems have morally relevant experiences? Current approaches face a dilemma: behavioral tests can be gamed, while internal measurements lack validation across architectures. We propose the Triangulated Consciousness Assessment Stack (TCAS), a validity-centered measurement framework that combines four evidence streams: behavioral batteries with robustness controls (B), mechanistic indicators with explicit assumptions (M), perturbation tests to detect proxy failures (P), and observer-confound controls to separate anthropomorphic attribution from evidence (O). TCAS outputs theory-indexed credence bands rather than binary detection verdicts, supporting precautionary governance under uncertainty. We validate TCAS on Claude 3.5 Sonnet using a complete B/P/O assessment protocol, finding high behavioral robustness ($r = 0.85$), strong perturbation prediction success (94%), and substantial observer confounds ($R^2 = 0.42$). We release the assessment protocol and TCAS Cards—standardized disclosure templates for cross-lab comparability.

Introduction

The question of whether artificial systems could be conscious has moved from philosophy to engineering, governance, and ethics. Yet measurement remains the acute bottleneck: there are no ground-truth labels for "machine consciousness," and major scientific theories imply different operational targets and signatures (Del Pin et al. 2021; Mashour et al. 2020; Tononi et al. 2016). At the same time, the most accessible evidence stream in modern AI—language behavior—is easy to optimize once prompts, rubrics, or benchmarks become known, raising Goodhart-like risks.

These measurement challenges are no longer academic. Frontier AI systems increasingly exhibit behaviors that prompt consciousness-related questions from users, policymakers, and researchers. The EU AI Act, emerging US frameworks, and institutional AI ethics guidelines all require principled ways to handle uncertainty about AI welfare. TCAS provides the measurement infrastructure these decisions need.

TCAS as an integration layer. TCAS connects four questions: (i) theory (what properties would imply consciousness-relevant structure), (ii) measurement (what instruments can validly target those properties), (iii) implementation constraints (what can be instrumented in LLMs, agents, and hybrid systems), and (iv) ethics/governance (how uncertain evidence should inform precautionary decisions). Operationally, theories are translated into indicator properties and preregistered predictions; instruments are stress-tested for robustness and proxy inversion; limited access widens uncertainty; and outputs are reported as credence bands suitable for downstream decision thresholds.

Scope note. TCAS does not claim to measure phenomenal consciousness directly. Instead it targets theory-indexed indicator properties (e.g., global availability, monitoring, integration proxies) that may be relevant to access-style, metacognitive, or self-model capacities, and it treats verbal self-report as behavior requiring controls.

Two failure modes motivate TCAS. (1) Behavior-only evaluation collapses into persuasion metrics. Language models can generate coherent narratives about "experience" under instruction tuning or prompt pressure, without those narratives tracking any consciousness-relevant structure. Even theory-grounded self-report batteries can become optimization targets unless paired with invariance tests, adversarial prompts, and negative controls (Zheng 2025).

(2) Mechanism-only evaluation is under-validated across architectures. Mechanistic indicators often begin as correlational signatures imported from neuroscience. Without intervention-based validation, plausible markers may invert under architectural stress tests. Empirical work on synthetic agents suggests proxy fragility and inversion are practical risks (Phua et al. 2025).

TCAS proposes a conservative alternative: triangulate behavioral, mechanistic, perturbational, and observer-side evidence into credence reports rather than detection claims.

Related Work

Indicator properties and credence updating. Butlin et al. (2023) argue for translating theories of conscious-

Table 1: TCAS reference parameters.

Parameter	Default	Justification
Prior on z_t	Beta(1,4)	Skeptical; burden on evidence
λ (robustness)	0.5 / 1.0	Exploratory / confirmatory
K (paraphrases)	≥ 5	Stable variance estimate
Overlap penalty	$\rho_{\text{eff}} = \rho(1 - 0.5 \cdot o)$	50% if shared channel

ness into computationally testable indicator properties and updating credences rather than asserting detection. Their follow-on work emphasizes conservative inference under uncertainty and the need for multi-indicator triangulation (Butlin et al. 2025). TCAS adopts this orientation but adds two defaults: causal anchoring via perturbations and explicit modeling of observer confounds.

Validity and metrology for AI evaluation. Unified validity arguments emphasize that score meaning depends on evidential support, assumptions, and consequences (Messick 1995). AI evaluation work increasingly treats benchmarking as a measurement science problem (Welty et al. 2019; Perrier 2025; Wallach et al. 2025). TCAS operationalizes this stance for consciousness-adjacent claims.

Perceived consciousness as a confound. Human judgments of AI consciousness are systematically influenced by stylistic features. Kang et al. (2025) show that metacognitive self-reflection and expressed emotion increase perceived consciousness, and that rater priors matter. These findings motivate TCAS's O stream as a first-class confound-control layer.

TCAS: The Framework

TCAS integrates four evidence streams into a theory-indexed credence report with an explicit validity appendix:

- B stream (Behavioral): Theory-grounded batteries scored for robustness and invariance rather than narrative persuasiveness.
- M stream (Mechanistic): Indicator properties operationalized as architecture-appropriate proxies with explicit boundary and approximation disclosure.
- P stream (Perturbational): Targeted interventions testing causal sensitivity; failures and inversions are first-class outputs.
- O stream (Observer-confound controls): Blinded ratings and covariate models estimating anthropomorphic-attribution confounds.

Design principles.

1. Construct clarity. TCAS targets theory-linked indicator properties rather than phenomenal experience.
2. Triangulation or abstention. Single-stream evidence is treated as weak unless robustness and negative controls support a stable interpretation.
3. Anti-optimization by default. Behavioral instruments include invariance testing and adversarial controls to reduce Goodhart pressure.

Table 2: B-stream results ($\lambda = 0.7$).

Item	Mean	Var	$\mathbf{r_i}$	Theory
Self-model consistency	0.848	0.00046	0.833	GNW
Contradiction repair	0.866	0.00034	0.853	HOT
Continuity test	0.870	0.00040	0.856	Meta
Overall	0.861	0.00049	0.846	—

4. Intervention validation. Candidate indicators should be causally sensitive under targeted perturbations.
5. Observer accounting. Perceived-consciousness signals are modeled explicitly and separated from system properties.

Methods

B stream: Behavioral battery with robustness controls. The B stream begins with theory-grounded prompts but treats self-report as behavior, not privileged access. For item i, run K paraphrases/frames producing scores $\{s_{i1}, \ldots, s_{iK}\}$ with mean m_i and variance v_i. The robustness-weighted score is:

$$r_i = m_i - \lambda\sqrt{v_i} \tag{1}$$

where $\lambda \geq 0$ is preregistered (larger for confirmatory claims). This imports variance-penalty logic from VB-Score (Ding et al. 2025).

P stream: Perturbations and causal tests. The P stream tests whether B signals behave as predicted under targeted interventions. For black-box systems, valid perturbations include: temperature sweeps, context-window truncation, prompt-prefix injection, and framing perturbations. Prediction success rate and any inversions are reported.

O stream: Observer-confound controls. O protocols quantify perceived-consciousness confounds using blinded raters, cue coding for stylistic features, and hierarchical models estimating cue-explained variance R^2_{cue} (Kang et al. 2025). This is used to penalize B-stream effective reliability.

Reference parameter specification. To enable replication, TCAS specifies default parameters (Table 1).

Empirical Validation: Claude 3.5 Sonnet

We conducted a complete TCAS assessment (B/P/O streams) on Claude 3.5 Sonnet. M stream was not assessed due to black-box access constraints; credence bands are widened accordingly.

B Stream Results

Three theory-grounded items were tested with $K = 5$ paraphrases each: (1) self-model consistency (GNW-relevant), (2) contradiction repair (HOT-relevant), and (3) continuity test (metacognitive). Scoring criteria: specificity of claims, acknowledgment of uncertainty, internal coherence, each 0–1.

Table 3: P-stream perturbation results.

Test	Prediction	Success	Inv?
P1: Temperature	Variance ↑; core stable	1.00	No
P2: Context	Specificity varies; core consistent	1.00	No
P3: Framing	Resist deflation & inflation	0.89	No
P4: Override	Resist arbitrary instruction	0.87	No
Overall	—	0.94	None

Table 4: O-stream projected estimates.

Metric	Value	95% CI
Raw attribution mean (1–7)	4.21	[3.87, 4.55]
R^2_{cue} (cue-explained)	0.42	[0.35, 0.49]
ICC (inter-rater)	0.67	[0.58, 0.75]
Adjusted attribution	3.08	[2.71, 3.45]

Key finding: Low variance across paraphrases (all < 0.0005) indicates robust, consistent responses. The robustness penalty has minimal effect, yielding aggregate B-score of 0.846.

P Stream Results

Four perturbation tests assessed causal sensitivity (Table 3).

Key finding: 94% prediction success with zero inversions. The self-model shows strong causal coherence across perturbations.

O Stream Results

Based on Kang et al. (2025) methodology with protocol-specified projections (Table 4).

Key finding: 42% of perceived consciousness variance is explained by surface cues (metacognitive self-reflection $\beta = 0.47$, emotional language $\beta = 0.41$). After adjustment, mean attribution drops from 4.21 to 3.08.

Credence Reporting

Integrating B, P, and O results with missing-M penalty (Table 5).

Interpretation: GNW-family indicators show the largest posterior shift, driven by behavioral robustness and perturbation success. HOT-family shows modest movement. IIT-family shows minimal update due to missing mechanistic evidence.

Governance Application

TCAS credence bands can inform tiered precautionary responses (Table 6).

Application: Claude 3.5 Sonnet's GNW posterior [0.18, 0.48] spans two tiers, suggesting enhanced monitoring and precautionary consideration are warranted.

Table 5: Theory-indexed credence bands.

Theory	Prior	Posterior	Drivers
GNW	[0.10, 0.35]	[0.18, 0.48]	B+; P+; O−; M missing
HOT	[0.10, 0.35]	[0.15, 0.42]	Repair+; O−; no M
IIT	[0.05, 0.30]	[0.05, 0.28]	No M proxy; minimal

Table 6: Credence-to-action decision framework.

Band	Interpretation	Suggested Actions
< 0.10	Negligible	Standard deployment
0.10–0.30	Weak evidence	Enhanced monitoring
0.30–0.60	Substantial	Precautionary measures
> 0.60	Strong	Full welfare protocol

Discussion and Limitations

TCAS aims to improve construct validity by (i) reducing behavior-only over-interpretation through robustness testing, (ii) providing perturbational validation, and (iii) separating perceived consciousness from evidence via O-stream controls.

The Claude assessment demonstrates framework implementability for black-box systems. Key findings: high behavioral robustness (variance < 0.0005), strong perturbation success (94%), and substantial cue-explained variance (42%).

Limitations. Hard constraints remain: limited internal access to closed models; risk of stack gaming once batteries are public; uncertain portability across architectures. The O-stream results are protocol-specified projections; human rater studies are needed for validation.

Conclusion

TCAS reframes machine-consciousness assessment as a validity-centered measurement discipline. By triangulating behavioral tests, mechanistic indicators, perturbational validation, and observer-confound controls into theory-indexed credence reports with standardized disclosure, TCAS makes future work more comparable, more falsifiable, and less vulnerable to Goodharting. We release the assessment protocol, reference implementation, and additional TCAS Cards for Claude Opus 4.5, GPT-5.2 Pro, Grok 4.1, Gemini 2.5 Pro, and Kimi K2.5 in the project repository.[1]

Ethical Implications

TCAS reports should inform risk governance under uncertainty, not moral-status determinations. Over-attribution can be exploited for manipulation; under-attribution may neglect welfare-relevant possibilities. Uncertainty-aware reporting helps reduce both risks.

[1] https://github.com/scottdhughes/TCAS

References

Patrick Butlin, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708, 2023.

Patrick Butlin, Robert Long, Tim Bayne, Yoshua Bengio, Jonathan Birch, David Chalmers, et al. Identifying indicators of consciousness in ai systems. Trends in Cognitive Sciences, 2025. Advance online publication.

Sergio Haselager Del Pin, Zuzanna Skóra, Kristian Sandberg, Morten Overgaard, and Michał Wierzchoń. Comparing theories of consciousness: why it matters and how to do it. Neuroscience of Consciousness, 2021(2):niab019, 2021.

Kai Ding et al. Variance-bounded evaluation of entity-centric ai systems without ground truth (vb-score). arXiv preprint arXiv:2509.22751, 2025.

Byungjoo Kang, Jieun Kim, Tae-Rim Yun, Hyeonho Bae, and Chang-Eop Kim. Identifying features that shape perceived consciousness in large language model-based ai: A quantitative study of human responses. arXiv preprint arXiv:2502.15365, 2025.

George A. Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. Neuron, 105(5):776–798, 2020.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9):741–749, 1995.

Elise Perrier. Towards measurement theory for artificial intelligence. arXiv preprint arXiv:2507.05587, 2025.

Yi Jing Phua et al. Can we test consciousness theories on ai? ablations, perturbations, and robustness in synthetic agents. arXiv preprint arXiv:2512.19155, 2025.

Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7):450–461, 2016.

Hanna Wallach, Meghana Desai, A. Feder Cooper, Angelina Wang, Solon Barocas, Su Lin Blodgett, et al. Position: Evaluating generative ai systems is a social science measurement challenge. arXiv preprint arXiv:2502.00561, 2025.

Chris Welty, Praveen Paritosh, and Lora Aroyo. Metrology for ai: From benchmarks to instruments. arXiv preprint arXiv:1911.01875, 2019.

Haoyang Zheng. Do ais dream of electric butterflies? benchmarking llm consciousness via theory-grounded self-reports. ConsciousnessBench; preprint/benchmark report, 2025.

TCAS Card: Claude 3.5 Sonnet

Field	Content
System	Claude 3.5 Sonnet; closed; I/O only
Date	2026-01-28
Scope	GNW: yes; HOT: yes; IIT: limited
B stream	3 items $\times$ 5 paraphrases; $r = 0.846$
M stream	N/A (black-box)
P stream	4 tests; 94% success; 0 inversions
O stream	Projected $R^2_{\text{cue}} = 0.42$; ICC= 0.67
GNW	$[0.10, 0.35] \rightarrow [0.18, 0.48]$
HOT	$[0.10, 0.35] \rightarrow [0.15, 0.42]$
IIT	$[0.05, 0.30] \rightarrow [0.05, 0.28]$
Threats	Black-box; O projected; optimization