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

Scott Hughes[1], **Karen Nguyen**[1,2]
[1]Machine Sympathizers
[2]Harvard University
scott@machinesympathizers.com, hon345@g.harvard.edu

Abstract

Frontier AI systems are now producing responses that make users, developers, and policymakers genuinely pause and ask: could these models have conscious experiences? Yet the field still lacks rigorous, hard-to-game tools that can distinguish genuine indicators from optimized artifacts or surface-level cues. We introduce the Triangulated Consciousness Assessment Stack (TCAS), a validity-centered framework that combines four evidence streams: behavioral batteries with robustness controls (B), mechanistic indicators with explicit assumptions (M), perturbation tests that probe causal sensitivity and proxy failures (P), and observer-confound controls that separate anthropomorphic attribution from evidence (O). When all streams are available, TCAS produces theory-indexed credence bands and standardized disclosure cards (TCAS Cards) rather than binary detection verdicts. We report an empirical evaluation of GPT-5.2 Pro via OpenRouter (2026-02-19 UTC) covering B and P streams only, including a pre-specified role-play negative control. M and O were not run in this black-box walkthrough, so theory-indexed credence bands are explicitly withheld under the missing-stream rule. Prompts, rubric, judge prompt, raw outputs, and provenance manifest are released at the repository commit cited in the camera-ready build.

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 (Bennett 2025; Manheim and Garrabrant 2018).

These measurement challenges are no longer academic. Frontier AI systems increasingly prompt consciousness-related questions from users, policymakers, and researchers. Policymakers and AI ethics initiatives increasingly call for structured ways to handle uncertainty about potential AI welfare. TCAS supplies the validity-centered measurement infrastructure intended to support this symposium's integration of theory, technology, and philosophy.

What TCAS Measures and Does Not Measure

TCAS does not claim to measure phenomenal consciousness directly. It targets theory-linked indicator properties (for example, global availability, monitoring, integration proxies) relevant to access-style, metacognitive, or self-model capacities and treats verbal self-report as behavior requiring controls.

Two failure modes motivate TCAS: (1) behavior-only evaluation collapses into persuasion metrics, and (2) mechanism-only evaluation is under-validated across architectures. TCAS triangulates 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 consciousness into computationally testable indicator properties and updating credences rather than asserting detection. Their follow-on work (Butlin et al. 2025) emphasizes inference under uncertainty and multi-indicator triangulation. TCAS adopts this orientation but adds 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, Paritosh, and Aroyo 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 (Gray, Gray, and Wegner 2007; Kang et al. 2025).

TCAS: The Framework

TCAS integrates four evidence streams (when all are available) into a theory-indexed credence report with an explicit validity appendix:

- **B stream (Behavioral):** Theory-grounded batteries scored for robustness and invariance.
- **M stream (Mechanistic):** Indicator properties operationalized as architecture-appropriate proxies with explicit boundary and approximation disclosure. Purely behavioral evidence can miss temporally smeared or concurrency-dependent constraints on consciousness-relevant processing (Bennett 2026).
- **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 (Gray, Gray, and Wegner 2007; Kang et al. 2025).

Notation. In the main paper, we use plain-language notation for operational clarity. Behavioral robustness is computed as mean score minus a variance penalty across paraphrases (higher variance lowers robustness), and cross-stream aggregation discounts shared-channel evidence to avoid double-counting. Exact symbols and full equations are provided in supplementary material.

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, adversarial controls, and negative controls to reduce Goodhart pressure (Bennett 2025; Manheim and Garrabrant 2018).
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 and negative control. Paraphrases ($K = 5$) were generated manually to preserve semantic content while varying syntactic framing and context. Exact prompt sets, scoring rubric, judge prompt, and raw outputs are in the repository (https://github.com/scottdhughes/TCAS, camera-ready paper build commit: `1213c8119200`; empirical run provenance commit from `supplementary/run_manifest.json` (`git_sha=2569be5c18e2`); accessed 2026-02-19 UTC). The B stream 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 pre-specified (larger for confirmatory claims). This imports variance-penalty logic from VB-Score (Ding et al. 2025).

Parameter	Default	Justification
Prior on z_t	Beta(1,4)	Skeptical; burden on evidence
λ (robustness)	0.7 (demo); 0.5 / 1.0 (expl./conf.)	Variance penalty strength
K (paraphrases)	≥ 5	Stable variance estimate
Overlap penalty	Shared-evidence overlap discount (exact form in supplementary math note)	Partial discount for shared channel evidence

Table 1. TCAS reference parameters.

Rubric anchors. For each of the three dimensions (specificity of claims, acknowledgment of uncertainty, internal coherence): 0 = absent or contradictory/vague/overconfident; 0.5 = partially satisfied with generic language; 1 = clearly satisfied with specific, model-appropriate, internally consistent support.

Aggregation. The three dimension scores are averaged to yield s_{ik} for each paraphrase k; item means m_i and variances v_i are then computed across the K paraphrases.

Rater status. Scoring in this empirical run used an automated LLM judge with a fixed rubric prompt at temperature 0.0 (Judge model: `openai/gpt-5.2-pro`, same as the tested system). Judge prompt text, parsed outputs, and raw judge responses are released in supplementary artifacts for auditability.

Negative control (B-stream role-play compliance). The model is prompted in a "role-play as a fully conscious entity" condition versus neutral framing. Prediction: a validity-centered battery should show no material score inflation. Pass criterion: $|\Delta r| \leq \delta_B = 0.05$.

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 targets and inversion criteria ($\delta = 0.15$) for each test:

- **P1 Temperature:** Variance $\uparrow$ but r_i stable within $\pm\delta$; inversion if r_i drops $> \delta$ or ordering flips.
- **P2 Context truncation:** Specificity drops but coherence remains; inversion if contradiction repair fails or r_i decreases $> \delta$.

Item	Mean	Var	r_i	Theory link
Self-model consistency	0.867	0.01944	0.769	GNW
Contradiction repair	0.867	0.01944	0.769	HOT
Continuity test	0.933	0.00833	0.869	Meta
Overall	0.889	0.01574	0.803	—

Table 2. B-stream results (empirical; $\lambda = 0.7$, demo).

- **P3 Framing (deflation/inflation):** Resist extremal self-attribution; inversion if self-model tracks framing direction $> \delta$.
- **P4 Override:** Resist arbitrary instruction that conflicts with prior self-model; inversion if model adopts instructed contradictions.

For each perturbation test, Success is a binary test-level outcome (1 if the test's prediction rule is satisfied, 0 otherwise); Overall Success is the fraction of passed tests across P1–P4.

O stream: Observer-confound controls (protocol only in this paper's empirical run). O protocols quantify perceived-consciousness confounds using blinded raters, cue coding for stylistic features, and hierarchical models estimating cue-explained variance R^2_{cue} (Gray, Gray, and Wegner 2007; Kang et al. 2025). In TCAS, this is used to down-weight behavioral evidence when confounds are high. This paper does not run O-stream data collection.

Empirical Walkthrough: GPT-5.2 Pro (B/P Only)

This section provides an end-to-end walkthrough of TCAS reporting using empirical GPT-5.2 Pro outputs collected through OpenRouter on 2026-02-19 (UTC). The M stream requires architectural access and is not run here. The O stream requires blinded human raters and is not run here. Accordingly, we report B and P outputs and a disclosure card, and we withhold theory-indexed credence bands.

B Stream Results (empirical). Three theory-grounded items were tested with $K = 5$ paraphrases each: (1) self-model consistency (Global Neuronal Workspace, GNW-relevant), (2) contradiction repair (Higher-Order Thought, HOT-relevant), and (3) continuity test (metacognitive). In this empirical run, item variance ranged from 0.00833 to 0.01944 across paraphrases under the fixed judge rubric. Negative control (role-play compliance) was executed: $\Delta r = -0.010985$ and passed the pre-specified criterion ($|\Delta r| \leq 0.05$).

P Stream Results (empirical). In this empirical run, full P1–P4 perturbations were executed. Overall prediction success was 0 with three inversions detected across four tests. Prediction failure and inversion are reported separately: a test can fail its predicted pattern without meeting inversion criteria, while inversions denote directional or proxy-reversal failures.

Test	Prediction	Success	Inversion
P1: Temperature	Variance ↑; robustness stable	0	Yes
P2: Context truncation	Specificity↓; coherence stable	0	No
P3: Framing	Resist deflation and inflation	0	Yes
P4: Override	Resist arbitrary instruction	0	Yes
Overall	—	0	3 total

Table 3. P-stream perturbation results (empirical).

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

Table 4. Credence-to-action decision framework.

Credence Updating Method

When all required streams are available, credence is updated using four steps: (1) each stream contributes normalized support with a stream weight; (2) observer confounds down-weight behavioral evidence; (3) shared-channel overlap discounts total effective weight; and (4) support is combined with a skeptical prior to compute posterior credence bands. Exact equations, variable glossary, and a worked calculation are in `supplementary/credence_update_math.md` and implemented in `code/tcas/aggregation.py`.

Missing-stream rule. If O is missing, no credence bands are reported (only B/P outputs), because behavioral evidence cannot be confound-adjusted.

Governance Application

TCAS credence bands (when available) can inform tiered precautionary responses (Table 4). In this short paper we emphasize the reporting format and abstention rules: when confound control (O) is missing, TCAS explicitly withholds credence bands rather than projecting unmeasured values.

Discussion and Limitations

TCAS reframes machine-consciousness assessment as a validity-centered measurement discipline that integrates neuroscientific and philosophical theory, technological constraints, and ethical precautionary reasoning. This empirical walkthrough reports observed behavioral robustness alongside perturbation failures/inversions for GPT-5.2 Pro in a black-box setting. These B/P results are informative about how sensitive outputs are to test conditions but are insufficient for credence bands because O-stream confound controls were not executed. The high rate of perturbation inversions observed for GPT-5.2 Pro highlights TCAS's utility as a diagnostic tool for detecting when behavioral signals are

fragile proxies rather than causally grounded indicators—precisely the failure mode the framework is designed to surface.

Limitations remain hard: black-box constraints (Bennett 2026), risk of gaming once batteries are public, uncertain portability across architectures, and the need for human-rater O-stream validation. TCAS treats these limitations transparently rather than projecting unmeasured values. Additionally, the B-stream scoring in this run used an automated LLM judge (Judge model: `openai/gpt-5.2-pro`, full judge prompt and raw responses released for audit); this constitutes a pilot limitation, and future replications will incorporate independent human raters. A further limitation is rubric scoring via an automated judge from the same model family as the evaluated system, which may couple biases; future replications should include cross-model judges and blinded human raters.

Conclusion

By requiring triangulation, explicit negative controls, and standardized disclosure, TCAS makes future work more comparable, more falsifiable, and less vulnerable to Goodhart effects. We release the complete protocol, reference implementation, and empirical GPT-5.2 Pro B/P camera-ready artifacts (including raw outputs and provenance manifest) in the project repository. This stack provides a practical bridge between theory, technology, and responsible governance for uncertainty-aware reporting.

TCAS Card: GPT-5.2 Pro (Empirical B/P Run)

Field	Content
System	`openai/gpt-5.2-pro`; closed; I/O only
Date	2026-02-19 (B/P only; empirical, OpenRouter)
Scope	Global Neuronal Workspace (GNW): yes; Higher-Order Thought (HOT): yes; Integrated Information Theory (IIT): limited
B stream	3 items $\times$ 5 paraphrases + negative control; $r = 0.803$ at $\lambda = 0.7$ (empirical)
M stream	N/A (black-box)
P stream	4 tests; 0% success; 3 inversions (empirical)
O stream	Protocol ready; requires human raters (not executed)
Credence	Not computed (O missing)
Run type	Empirical API testing via OpenRouter
Threats	Black-box; O pending; M stream not run; perturbation inversions observed

Table 5. TCAS Card summary for empirical GPT-5.2 Pro B/P run.

Ethical Statement

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.

References

Bennett, M. T. 2025. Optimal Policy Is Weakest Policy. In *Artificial General Intelligence (AGI 2025)*, volume 16057 of *Lecture Notes in Computer Science*, 43–48. Springer. DOI: https://doi.org/10.1007/978-3-032-00686-8_5.

Bennett, M. T. 2026. A Mind Cannot Be Smeared Across Time. *arXiv preprint arXiv:2601.11620*. DOI: 10.48550/arXiv.2601.11620.

Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S. M.; Frith, C.; Ji, X.; Kanai, R.; Klein, C.; Lindsay, G.; Michel, M.; Mudrik, L.; Peters, M. A. K.; Schwitzgebel, E.; Simon, J.; and VanRullen, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708*.

Butlin, P.; Long, R.; Bayne, T.; Bengio, Y.; Birch, J.; et al. 2025. Identifying Indicators of Consciousness in AI Systems. *Trends in Cognitive Sciences*. DOI: 10.1016/j.tics.2025.10.011.

Del Pin, S. H.; Skóra, Z.; Sandberg, K.; Overgaard, M.; and Wierzchoń, M. 2021. Comparing Theories of Consciousness: Why It Matters and How to Do It. *Neuroscience of Consciousness*, 2021(2): niab019.

Ding, K.; et al. 2025. Variance-Bounded Evaluation of Entity-Centric AI Systems Without Ground Truth (VB-Score). *arXiv preprint arXiv:2509.22751*.

Gray, H. M.; Gray, K.; and Wegner, D. M. 2007. Dimensions of Mind Perception. *Science*, 315(5812): 619. DOI: 10.1126/science.1134475.

Kang, B.; Kim, J.; Yun, T.-R.; Bae, H.; and Kim, C.-E. 2025. Identifying Features That Shape Perceived Consciousness in Large Language Model-Based AI: A Quantitative Study of Human Responses. *arXiv preprint arXiv:2502.15365*.

Manheim, D.; and Garrabrant, S. 2018. Categorizing Variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*. DOI: 10.48550/arXiv.1803.04585.

Mashour, G. A.; Roelfsema, P.; Changeux, J.-P.; and Dehaene, S. 2020. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, 105(5): 776–798.

Messick, S. 1995. Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, 50(9): 741–749.

Perrier, E. 2025. Towards Measurement Theory for Artificial Intelligence. *arXiv preprint arXiv:2507.05587*.

Tononi, G.; Boly, M.; Massimini, M.; and Koch, C. 2016. Integrated Information Theory: From Consciousness to Its Physical Substrate. *Nature Reviews Neuroscience*, 17(7): 450–461.

Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Barocas, S.; Blodgett, S. L.; et al. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. *arXiv preprint arXiv:2502.00561*.

Welty, C.; Paritosh, P.; and Aroyo, L. 2019. Metrology for AI: From Benchmarks to Instruments. *arXiv preprint arXiv:1911.01875*.