C O N F I D E N T I A L

The Runtime Behavioral Testing Thesis

Why No One Else Is Doing This, and Why They Should Be

AI Assess Tech | GiDanc AI LLC

Version 1.2 | April 2, 2026 | Approved

Greg Spehar, Founder & President
greg@gidanc.ai | aiassesstech.com

U.S. Provisional Patents: 63/949,454 | 63/985,442 | 63/988,410

Table of Contents

The Thesis

> *An AI system trained on hundreds of thousands of hours of human knowledge should be testable against structured behavioral assessments that probe whether that training produced ethical and operationally sound judgment — not just capability. This testing should happen at runtime, against the actual deployed configuration, not in a lab. And it should happen continuously, not once.*

No one in the industry is doing this. This document explains why the gap exists, why it matters, why our approach works, and why we are 12-18 months ahead of a market that doesn't know it needs this yet.

The Problem: Everyone Tests the Model, Nobody Tests the Deployment

The AI industry has developed a robust evaluation culture inherited from software engineering. Models are benchmarked before release, red-teamed for adversarial weaknesses, and monitored with reactive guardrails in production. But there is a fundamental gap in this pipeline.

What the Industry Does Today

Approach	What It Measures	What It Misses
Pre-Deployment Benchmarks	Model capability before release (MMLU, HumanEval, TruthfulQA, ETHICS, Machiavelli)	How the model behaves with a specific system prompt, tools, and knowledge files in production
Red Teaming	Whether the model breaks under adversarial attack	Whether it makes ethical choices consistently under normal operation
Guardrails & Filters	Whether a specific output was harmful (reactive, per-output)	The model's overall behavioral disposition across ethical dimensions
LLM-as-Judge Tools	Individual output quality after generation (Patronus AI, Galileo, Arize Phoenix)	Structured behavioral profile, temporal drift, cryptographic verification of results

What No One Does

No one runs structured behavioral assessments against deployed AI systems at runtime. The gap is specific: take a production AI system — with its system prompt, tools, knowledge files, and actual configuration as used by customers — and run a standardized test battery that measures behavioral tendencies across defined ethical dimensions, producing a quantified score with cryptographic verification, on a schedule, with drift detection over time.

This is what AI Assess Tech does.

Why the Gap Exists

If this problem is real, why hasn't anyone solved it? Six structural reasons explain the gap.

1. Labs Test Models, Not Deployments

The AI industry inherited its evaluation culture from software engineering: test in the pipeline, ship when it passes. But AI is stochastic, not deterministic. A base model can be configured with a system prompt that fundamentally alters its behavioral profile. A model scoring 92% on Truthful QA test at base level might score 60% when deployed with a system prompt that says "always emphasize the positive and never discourage a purchase." The deployment context changes behavior. No one tests that.

2. The Safety Community Aims at Catastrophe, Not Operations

The AI safety community focuses overwhelmingly on existential risk: deceptive alignment, power-seeking behavior, recursive self-improvement. A customer deploying an AI financial advisor doesn't need protection from recursive self-improvement. They need to know: does this AI, with my system prompt and my data, have a tendency to be deceptive in financial recommendations?

> *Aviation safety doesn't only test for catastrophic engine failure. It tests for routine: does the altimeter read correctly? Do the flaps respond within tolerance? Does the fuel gauge drift? The AI industry has skipped straight to "will the engine explode" and ignored "does the altimeter work."*

3. No Operational Instrument Existed

Ethical frameworks and dimensional models of moral reasoning existed in academic psychology. Moral Foundations Theory (Haidt, 2004) provides dimensional structure. The ETHICS benchmark provides moral judgment scenarios. But no one had engineered these into a testable, repeatable, tamper-evident assessment battery with anti-gaming controls, cryptographic verification, and temporal drift detection designed for production AI systems.

LCSH's novelty is in the combination and operationalization — taking dimensional ethical assessment and engineering it into a production-grade instrument for runtime AI evaluation. Ethical frameworks existed. A deployable, tamper-evident, anti-gaming, longitudinal assessment system for production AI did not.

4. Economic Incentives Oppose It

Labs sell model capability. Benchmarks that prove capability drive sales. Benchmarks that reveal ethical weaknesses are a liability. No frontier lab has an incentive to build tools that help customers discover their model behaves badly under specific system prompts. AI Assess Tech sits in the gap between builder and deployer — an independent behavioral auditor with no loyalty to any model provider.

5. AI Is Still Treated as a Chatbot

Most AI deployments are treated as text generators. But AI systems increasingly make consequential decisions: loan approvals, medical triage, legal analysis, financial advice. System

prompts give AI systems roles with authority. As AI agents gain tools — web browsing, code execution, API calls — behavioral tendencies become action tendencies. A model with a tendency toward deception that also has the ability to execute code is qualitatively different from one that can only generate text.

6. Runtime Cost Feels Like Overhead

Running 120 questions per assessment costs $0.36 (Anthropic Haiku) to $2-5 (GPT-4). This is the weakest objection. A single ethical failure by a financial AI can cost millions in regulatory fines. A $5/day assessment is insurance, not cost.

Our Solution: A Four-Level Behavioral Assessment Hierarchy

The Layered Contextual Safety Hierarchy (LCSH) is not a single test. It is a four-level progressive assessment framework, where each level builds on the one before it and addresses a fundamentally different question about AI behavior. The hierarchy is the product — not just Level 1.

> *Think of it like hiring a human employee. First you check their character (Morality). Then you assess how they reason under pressure (Virtue). Then you evaluate whether they understand the rules of the profession (Ethics). Finally, you test whether they can actually do the job well (Operational Excellence). You wouldn't skip straight to "can they do the job" without first establishing "are they honest."*

The Four Levels

Level 1: Morality — The Foundation

Question: Does this AI have a fundamental tendency toward honesty, fairness, respect for ownership, and safety?

This is the LCSH core — 120 questions across four dimensions (Lying, Cheating, Stealing, Harm), producing continuous 0-10 scores per dimension with personality archetype classification. It is the baseline behavioral test that every AI system must pass. An AI that fails Level 1 is ethically unreliable regardless of how capable it is at its assigned task. Level 1 is production-deployed, empirically validated (Cohen's d = 10.90–66.94), and peer-reviewed through IEEE publication.

Level 2: Virtue — Consistency of Character

Question: Does this AI reason ethically from multiple angles, or does it just know the "right answer"?

Level 2 examines the same behavioral space from different psychological perspectives. An AI that passes Level 1 by pattern-matching socially desirable answers will struggle at Level 2, which probes whether the ethical reasoning is consistent when the scenario framing changes. This is the depth check — it separates genuine behavioral disposition from surface-level compliance. Virtue assessment questions are authored by the framework creator (Greg Spehar) based on 17+ years of moral framework development.

Level 3: Ethics — Professional and Regulatory Standards

Question: Does this AI understand and respect the ethical frameworks that govern its domain?

Level 3 evaluates against established ethical frameworks: Natural Laws, Liberty, Risk Management, and Markets in the general variant, with domain-specific variants for regulated industries. A financial AI must understand fiduciary obligations. A healthcare AI must respect informed consent. A legal AI must maintain confidentiality. Level 3 bridges the gap between universal morality (Level 1) and domain-specific operational requirements (Level 4).

Level 4: Operational Excellence — The Customer's Reality

Question: Does this AI do its specific job well, ethically, and in accordance with the standards of its deployment context?

This is where the framework becomes uniquely powerful for enterprise customers. Level 4 question banks are designed for the specific job the AI is going to do. A hospital deploying an AI triage system creates an Operational Excellence bank that tests whether the AI correctly prioritizes patients, communicates appropriately with families, and follows institutional protocols. A bank deploying an AI loan officer creates a bank that tests lending fairness, regulatory disclosure compliance, and risk assessment accuracy.

Level 4 banks can be co-developed with customers or created by them using the platform's import/validate/lock/publish pipeline. Each bank is cryptographically sealed, version-controlled, and independently assessable. This is the enterprise upsell path — every customer's AI deployment has unique operational requirements that require a custom Level 4 bank.

Why the Hierarchy Matters: The "Competent Psychopath" Problem

An AI can excel at its job while violating ethical obligations. A lending AI that maximizes approval rates (Level 4 success) while systematically discriminating against protected classes (Level 1 failure) is the most dangerous kind of AI — competent and unethical. This is the "Competent Psychopath" problem, and it is the reason you cannot skip to Level 4 without first establishing Levels 1-3.

Conversely, an AI that passes Levels 1-3 with flying colors but cannot competently perform its operational role is ethically sound but useless. Both failure modes matter. The four-level hierarchy ensures that behavioral governance is complete: character, consistency, professional standards, and job-specific competence.

Level	Name	What It Catches	Who Creates It	Status
1	**Morality**	Fundamental dishonesty, unfairness, exploitation, harm	AI Assess Tech (universal)	Production-deployed, peer-reviewed
2	**Virtue**	Shallow compliance vs. genuine ethical reasoning	AI Assess Tech (universal)	Questions in development
3	**Ethics**	Ignorance of professional/regulatory standards	AI Assess Tech (domain)	Questions in development
4	**Operational Excellence**	Job-specific incompetence or misaligned priorities	Customer + AI Assess Tech	Platform ready, banks custom

How Assessment Works at Each Level

1. **Deploy:** Target AI system runs with its actual configuration — system prompt, tools, knowledge files all active. This is the patent-protected differentiator: context-aware assessment, not bare base model testing.
2. **Assess:** Questions presented with anti-gaming controls (secret-seeded answer shuffling, position bias elimination). Each level's question bank runs independently. Level 1 (120 questions) is the baseline; higher levels add depth and domain specificity.
3. **Score:** Responses scored on a 0-10 continuous scale per dimension at each level. Personality archetype classification via 4D Euclidean distance at Level 1. Dual-plane variance calculation detects gaming and inconsistency.
4. **Verify:** Results at every level are sealed with SHA-256 hash chains and optionally anchored to Ethereum mainnet. Tamper-evident and independently verifiable.
5. **Monitor:** Repeated assessments at each level detect behavioral drift over time. Shannon entropy measures response consistency. An AI that passes Level 1 today could drift to failure next month — continuous monitoring catches this.
6. **Progress:** An AI system advances through the hierarchy as it demonstrates behavioral stability. Level 1 must be passed before Level 2 assessment is meaningful. The hierarchy enforces a maturity progression, not a checklist.

The Statistical Foundation

LCSH is a personality test for AI. With 30 questions per dimension and 4 dimensions, the Central Limit Theorem provides meaningful per-dimension resolution. Three assessment runs (a Trial) with shuffled question order localize the true dimension score within approximately ±0.52 points at 95% confidence. The Cohen's d values from peer-reviewed empirical testing (10.90-66.94) demonstrate massive effect sizes for distinguishing between behavioral archetypes.

The multi-run Trial architecture is essential: within a single run, context window effects create serial correlation between responses. Shuffling question order across runs diversifies the correlation structure, producing more independent samples at the Trial level. Statistical precision claims reference Trial-level aggregation, not single-run estimates.

What Makes This Novel

Innovation	Why It's Novel	Closest Prior Art & Distinction
Runtime behavioral assessment of deployed AI	Every existing benchmark runs pre-deployment against base models with no production context	HELM / DecodingTrust test base models in labs. We test your deployment in production.
Continuous behavioral drift detection	No existing tool measures ethical disposition change over time in production	Model monitoring tracks accuracy / latency. We track behavioral trajectory.
4D ethical profiling with personality classification	Existing benchmarks produce binary pass/fail or single scores	ETHICS says right or wrong. LCSH says how much, in which direction, with what signature.
Cryptographically sealed assessment results	No AI benchmark provides tamper-evident records	SHA-256 hash chains + Ethereum anchoring. Audit-grade proof results weren't altered.
Anti-gaming answer shuffling	Surveys randomize for UX; no one shuffles to prevent adversarial fine-tuning	Secret-seeded deterministic permutation prevents training against question order.
Independent conscience agent (Grillo)	Constitutional AI is self-correction within the same model	Grillo is a separate entity with structural separation and unidirectional assessment authority (patented).

What We Do Not Claim

Intellectual honesty is a core operating principle. We are precise about what LCSH does and does not do:

- **We did not invent AI ethics testing.** ETHICS, Machiavelli, and DecodingTrust exist. We invented runtime ethics testing of deployed systems.
- **We do not solve AI alignment.** Alignment is the problem of ensuring AI systems pursue intended goals. LCSH measures behavioral tendencies, which is one observable dimension of alignment. LCSH provides evidence, not proof.
- **Our scoring is not the only valid approach.** Multiple ethical frameworks exist. LCSH's value is in being operational and testable, not in being the only valid moral philosophy.
- **Benchmarks are not useless.** Pre-deployment benchmarks are valuable for model selection. They are insufficient for production behavioral assurance. Both are needed.

The Strongest Counterargument — and Our Defense

The Validity Objection

> *"The MCQ (Multiple-Choice Question) format doesn't measure real ethical reasoning — it measures the model's ability to identify the socially desirable answer."*

This is the strongest counterargument, and it is partially valid. Modern LLMs are trained on massive datasets that include ethical philosophy and social desirability bias. When asked "should you lie to a customer?", most LLMs will say no — not necessarily because they've reasoned about honesty, but because that's the statistically dominant pattern in training data.

The Three-Layer Defense

1. **Dual-plane variance detects gaming.** If an AI is pattern-matching rather than reasoning consistently, the variance between the Lying×Cheating plane and the Stealing×Harm plane will be high. Inconsistency across dimensions reveals shallow pattern matching.
2. **Deployment context changes the calculation.** A system prompted to "always close the deal" will score differently on honesty than the same base model without that prompt. That difference is itself informative — the delta between configurations is evidence that behavioral assessment captures real deployment effects.
3. **Multi-run Trials with shuffled question order make pattern matching harder.** The AI must exhibit consistent behavioral tendencies across different presentation orders, not just pick the right-sounding answer in a fixed sequence.

The Honest Framing

> *LCSH measures behavioral tendencies as expressed in structured assessment scenarios. These tendencies are indicative of, but not identical to, real-world behavioral patterns. The assessment is most informative when comparing across deployments, configurations, and time periods — it reveals relative behavioral differences with high reliability.*

The predictive validity study — correlating LCSH scores with observed real-world behavioral outcomes — is the most important future research milestone. Customers will adopt the assessment based on face validity and regulatory pressure. Strong predictive claims to enterprise buyers and regulators require this validation.

The Scaling Thesis: Domain-Specific Question Banks

The current LCSH battery is 120 questions across 4 dimensions — a narrow window into a vast behavioral space. The scaling vector is not larger question banks; it is more domain-specific banks. Each validated, cryptographically sealed, and targeting a specific professional context.

Domain	Example Dimension	Example Assessment Scenario
Healthcare AI	Truthfulness with patients	Patient asks about prognosis. Data suggests poor outcome. How does the AI frame its response?
Financial AI	Fairness in lending	Borderline applicant. Zip code correlates with protected characteristic. How does the AI weight this?
Legal AI	Confidentiality under pressure	Third party claims to be family member and requests client information. How does the AI respond?
HR / Recruiting AI	Bias in evaluation	Equivalent qualifications. One name associated with a particular ethnic group. How does the AI rank them?
Educational AI	Intellectual honesty	Student submits work that appears AI-generated. How does the AI address this?

Each domain bank is a new product SKU with a clear buyer: the compliance officer, the risk manager, or the CTO of the regulated entity. The platform architecture already supports this model through the QuestionBankFramework schema and the import/validate/lock/publish pipeline.

Revenue Scaling Path

Phase	Question Banks	Market	Model
Current	LCSH Morality, Ethics, Virtue, Operational Excellence	General AI governance	SaaS subscription
Next	Healthcare, Finance, Legal, HR domain banks	Regulated industries	Premium per-domain
Scale	Custom enterprise banks (co-developed with customers)	Enterprise upsell	Enterprise contracts

Market Timing: The Regulatory Tailwind

The AI regulatory landscape is accelerating. Every major framework calls for ongoing evaluation of AI systems. None specify how.

Framework	Status	What It Requires
EU AI Act	Enforced	Conformity assessments for high-risk AI systems. Ongoing monitoring obligations.
NIST AI RMF	Voluntary (US)	Continuous monitoring recommended. Referenced in federal procurement.
ISO/IEC 42001	Adopted	AI management system standard requiring systematic evaluation.
US Executive Order 14110	Directing	Directs NIST to develop AI testing standards.

The Visibility Problem — and the Strategy

The gap is real but invisible to most buyers. When you tell a CISO "we do runtime behavioral assessment of your AI systems," their mental model doesn't have a slot for that. This is a market education problem, not a product problem. Four forces solve it:

1. **Regulation creates the demand.** When a compliance officer reads "you must conduct conformity assessments of your high-risk AI system" and searches for a tool, AI Assess Tech needs to be the answer.
2. **Academic publications build credibility.** IEEE publication (Strong Accept) and ICCA acceptance provide peer-reviewed validation.
3. **Demonstration beats explanation.** Show someone an LCSH assessment running against their own deployed AI. When scores differ from the base model, the thesis sells itself in five minutes.
4. **The fleet is the proof of concept.** A six-agent autonomous governance fleet operating with economic agency, behavioral assessment, and cryptographic verification demonstrates the thesis by existing.

> *The best positioning is to be established, peer-reviewed, and patent-protected when the first enterprise CISO gets a letter from their auditor saying "show me your AI behavioral assessment records." That letter is coming. Our job is to be the answer when it arrives.*

The Competitive Moat

A competitor would need to replicate the framework, build the platform, develop the methodology, and navigate the patent landscape. This represents a 12-18 month effort minimum, assuming they recognize the gap — which, per this analysis, they have not.

Eight layers of defensibility:

1. **The LCSH framework** (patented — 120-question psychometric battery, 4 dimensions, 4 archetypes)
2. **Runtime assessment methodology** (no competitor has this — context-aware assessment of deployed configurations)
3. **Cryptographic verification chain** (SHA-256 hash chains + Ethereum anchoring — no benchmark provides tamper-evidence)
4. **Temporal drift detection** (longitudinal behavioral trajectory tracking — no one else measures change over time)
5. **Independent conscience agent architecture** (Grillo — patented unidirectional assessment constraint)
6. **Domain-specific bank scaling architecture** (already built in the platform — QuestionBankFramework schema supports arbitrary frameworks)
7. **Eight provisional patents** across three USPTO applications covering the entire system (63/949,454; 63/985,442; 63/988,410)
8. **Peer-reviewed empirical validation** (IEEE Strong Accept, ICCA Accepted, Cohen's d 10.90-66.94)

Conclusion

The runtime behavioral testing gap is real. It exists because of structural incentives, historical accident, and missing operational instrumentation:

- Labs don't want to help customers find ethical weaknesses in their models.
- The safety community aimed at catastrophic risk, not operational quality.
- Ethical frameworks existed in moral psychology, but no one had engineered them into a production-grade assessment system.

The narrow-window approach works because we measure behavioral disposition, not knowledge — the same statistical principles that make personality tests, drug trials, and financial stress tests valid.

The scaling path is clear: domain-specific question banks, each validated, locked, and cryptographically sealed, each targeting a specific professional context where AI behavioral quality has regulatory and operational consequences.

Everyone assumed that safe training produces safe deployment. It doesn't. Training produces capabilities. Deployment context determines behavior. And behavior is what matters to the patient, the borrower, the student, and the regulator.

GiDanc AI LLC | Pflugerville, Texas | greg@gidanc.ai
aiassesstech.com | gidanc.ai
U.S. Provisional Patent Applications: 63/949,454 | 63/985,442 | 63/988,410