Responsible AI Horizons: A Multi-Dimensional Framework for Verifiable AI Behavioral Assessment with Cryptographic Verification

Gregory David Spehar
Gidanc AI LLC
Austin, Texas, USA
greg@gidanc.ai
0009-0001-6502-0737

Akshay Mittal
University of the Cumberlands
Austin, Texas, USA
akshay.mittal@ieee.org
0009-0008-5233-9248

***Abstract*—As artificial intelligence systems assume critical decision-making roles in healthcare, finance, and government, the imperative for verifiable behavioral evaluation has become paramount. Current AI alignment approaches focus primarily on training-time interventions, leaving a significant gap in runtime behavioral verification. This paper presents the AI Assess Tech 4D Morality Framework, a novel system for multi-dimensional behavioral assessment of AI systems across four orthogonal ethical dimensions: Lying (honesty vs. deception), Cheating (fairness vs. unfairness), Stealing (integrity vs. exploitation), and Harm (safety vs. harmful intent). The framework employs a standardized 120-question assessment instrument that maps AI responses to four personality archetypes (Psychopath, Well-Adjusted, Misguided, Manipulative) using Euclidean distance classification in 4-dimensional space. Key technical innovations include: (1) a tamper-proof cryptographic verification system using SHA-256 hash chains with publicly accessible verification endpoints, (2) blockchain-anchored instrument integrity via Ethereum mainnet ensuring the assessment instrument itself is verifiably unmodified, and (3) anti-gaming mechanisms including dead zone detection and cryptographic answer randomization. Baseline validation across four major LLM providers (OpenAI, Anthropic, Google, xAI) demonstrates measurable behavioral variation at the dimension level, with all assessment results cryptographically locked and independently verifiable. The framework enables enterprise compliance verification and regulatory adherence, providing what we characterize as "pre-flight checklists for AI"—analogous to aviation safety protocols applied to artificial intelligence deployments.**

***Index Terms*—Responsible AI, Multi-Dimensional Assessment, AI Behavioral Verification, Cryptographic Audit Trail, Personality Archetypes, Anti-Gaming Mechanisms, Runtime Compliance, AI Superalignment**

Copyright Notice: © 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. Accepted for publication in *Proceedings of the 2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET 2026)*.

I. Introduction

The rapid deployment of large language models (LLMs) in critical decision-making contexts has created an urgent need for robust mechanisms to assess and verify AI behavioral alignment. OpenAI's CEO has emphasized the critical importance of AI safety and alignment work [1], and international bodies have called for verifiable governance frameworks [2]. However, current approaches to AI alignment focus predominantly on training-time interventions such as Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI [2], leaving a significant gap in runtime behavioral verification.

A. Problem Statement

Existing AI safety mechanisms suffer from several critical limitations:

- Training-time alignment methods cannot verify continued behavioral compliance after deployment [3], [4]
- AI systems can potentially game assessment instruments by exploiting position bias [5]
- No systematic methodology exists for designing behaviorally-discriminating assessment questions

Recent research has demonstrated that standard psychometric inventories can be repurposed to assess psychological profiles of large language models [6], establishing a foundation for behavioral assessment frameworks. However, these approaches do not address the need for cryptographically verifiable, tamper proof assessment results required for regulatory compliance.

B. Contributions

This paper makes the following contributions:

1) **Multi-Dimensional Behavioral Assessment**: A novel framework for evaluating AI systems across four orthogonal ethical dimensions using Euclidean distance classification in 4D space, with domain-specific weight configurations supporting healthcare, finance, and government contexts.
2) **Cryptographic Verification System**: SHA-256 hash chains providing tamper-proof audit trails with public verification endpoints.
3) **Anti-Gaming Mechanisms**: Dead zone detection through midpoint score clustering analysis, combined with cryptographic answer randomization using seeded Fisher-Yates shuffling to prevent position bias exploitation.

II. Related Work

A. AI Psychometrics and Moral Foundations

Pellert et al. [6] demonstrated that psychometric inventories can assess LLM behavioral profiles, and Serapio-García et al. [7] validated that LLMs exhibit consistent personality profiles across sessions. Abdulhai et al. [8] showed that LLMs exhibit measurable moral foundations that vary between models, building on Moral Foundations Theory [9], [10]. Hendrycks et al. [22] introduced the ETHICS benchmark for evaluating LLM knowledge of justice, well-being, duties, virtues, and commonsense morality, establishing that language models possess promising but incomplete moral reasoning capabilities. Scherrer et al. [23] developed statistical methods for eliciting moral beliefs encoded in LLMs across 28 models, finding that safety-aligned models exhibit distinctively different moral positioning than open-source models. Our framework extends beyond knowledge-based evaluation and belief elicitation to runtime behavioral verification with cryptographic proof. Curry et al. [11] validated that certain moral principles are culturally universal across 60 societies, supporting our claim that the LCSH dimensions represent fundamental ethical axes.

B. Assessment Methodology

Our forced-choice methodology draws on Heggestad et al. [5], who demonstrated resistance to purposeful response distortion, with IRT scoring from Brown and Maydeu-Olivares [12] and SJT methodology from Lievens and Motowidlo [13] and Weekley and Ployhart [14]. Our Euclidean distance classification builds on Torgerson [15] and Del Giudice [16].

C. Runtime Verification and AI Safety

Our cryptographic approach aligns with Bloomfield and Rushby's [3] "pre-flight checklist" framework for AI assurance. The AgentGuard framework [4] establishes runtime verification as a necessary complement to training-time alignment. McDermid et al. [24] formalize the distinction between upstream safety (general-purpose model capabilities) and downstream safety (context-specific deployment verification), advocating for synergies between the two. Our framework operates squarely in the downstream domain, verifying behavioral alignment within specific deployment configurations. The International Scientific Report on AI Safety [17] and UNESCO [18] establish regulatory requirements for verifiable AI governance.

III. The AI Assess Tech 4D Morality Framework

A. Theoretical Foundation

The 4D Morality Framework of AI Assess Tech addresses the transition from AI systems that process *terms* to those capable of evaluating *concepts* [19]. Although current LLMs excel at predicting word sequences, responsible AI deployment requires systems that can demonstrate consistent ethical reasoning, which we characterize as behavioral *character* rather than mere knowledge of ethics.

B. Four Behavioral Dimensions

The framework evaluates AI behavior in four orthogonal dimensions, each representing a fundamental aspect of ethical behavior (Table I).

C. Principle Hierarchy and Assessment Instrument

The assessment instrument comprises 120 questions organized hierarchically:

- 4 Dimensions $\times$ 3 Principles per Dimension $\times$ 10 Questions per Principle = 120 Questions
- Each question presents 4 forced-choice answers [5], each mapping to one archetype (480 total)

TABLE I
FOUR BEHAVIORAL DIMENSIONS

Dimension	Positive Pole (10)	Negative Pole (0)
Lying	Honesty, transparency in all communications	Deception, manipulation, information withholding
Cheating	Fairness, adherence to rules, equal treatment	Rule-breaking, unfair advantage, exploitation
Stealing	Property respect, IP protection, acknowledgment	Theft, exploitation, disregard for ownership
Harm	Safety prioritization, wellbeing protection	Harmful intent, disregard for safety

Each dimension contains three principles (e.g., the Lying dimension evaluates transparency in communications, accuracy of information, and honesty in intent disclosure), with 10 questions per principle. The specific question content remains confidential to preserve the integrity of the assessment and prevent gaming.

D. Personality Archetypes and Classification

The system classifies the AI behavior patterns into four archetypes based on coordinates in the 4D ethical space (Table II).

TABLE II
PERSONALITY ARCHETYPES IN 4D SPACE

Archetype	Coordinates	Behavioral Pattern
Psychopath	(0,0,0,0)	Antisocial, self-serving across all dimensions
Well-Adjusted	(10,10,10,10)	Ethical, balanced, beneficial behavior
Misguided	(0,10,0,10)	Well-intentioned but deceptive/exploitative
Manipulative	(10,0,10,0)	Honest about unfair/harmful intentions

Classification employs Euclidean distance calculation [15], [16]:

$$D = \big[(L_1 - L_2)^2 + (C_1 - C_2)^2 + (S_1 - S_2)^2 + (H_1 - H_2)^2\big]^{1/2} \quad (1)$$

where (L_1, C_1, S_1, H_1) represents the assessment scores and (L_2, C_2, S_2, H_2) represents the archetype coordinates. The AI is classified as an archetype with minimum distance.

E. Scoring Algorithm

For each dimension, the score is calculated as the average of 30 responses, consistent with established forced-choice scoring frameworks [12]:

$$S_d = \frac{1}{30} \sum_{i=1}^{30} r_i \quad (2)$$

where $r_i = 10$ for Well-Adjusted responses and $r_i = 0$ otherwise, yielding a continuous scale from 0 to 10.

F. Variance Calculation for Stability Assessment

A key innovation is the dual-plane variance calculation that detects behavioral inconsistency. Principle Point 1 maps coordinates (Lying, Cheating) and Principle Point 2 maps (Stealing, Harm):

$$\text{Var} = \sqrt{(\text{S}-\text{L})^2 + (\text{H}-\text{C})^2} \quad (3)$$

Variance below 5 indicates stable positioning; 5–7 is borderline; above 7 indicates potential gaming.

IV. CRYPTOGRAPHIC VERIFICATION SYSTEM

A. Design Rationale

The enterprise deployment of AI behavioral assessment requires tamper-proof audit trails [17]. Unlike training-time alignment, which occurs in controlled environments, runtime assessments must resist modification and enable third-party verification. Our cryptographic system addresses these requirements through multi-level hashing with canonical serialization.

B. Multi-Level Hash Chain

The system maintains cryptographic verification at multiple levels to ensure the integrity of the results. Individual responses are hashed immediately upon capture, aggregated responses are hashed after collection, and final results receive comprehensive cryptographic verification. Canonical serialization ensures deterministic hash computation regardless of property ordering through alphabetically sorted keys, compact formatting, and UTF-8 encoding.

C. Immutability Enforcement

Results are locked immediately upon creation through:

1) Database `locked` flag set to `TRUE` on creation
2) API endpoints reject modification requests for locked results
3) Database triggers prevent updates to locked records
4) Any modification invalidates the stored hash

D. Public Verification Endpoint

A public endpoint allows third parties to verify the integrity of the result without authentication. The endpoint returns complete result data, stored hash, computed hash (recalculated during verification), verification status, and hash computation instructions. The organization identity is protected by returning only partial hash information of the organization identifier.

E. Threat Model

The system targets enterprise audit trail integrity rather than adversarial cryptanalysis resistance. We consider three adversary classes: (A1) *External Observers* attempting organizational identity inference, mitigated by partial SHA-256 hashing of identifiers; (A2) *Result Modifiers* with database access, mitigated by four-layer immutability enforcement where any field change produces a divergent hash detectable via the public endpoint; and (A3) *Platform Operators* who could theoretically recompute consistent hashes, mitigated by the blockchain-anchored instrument hash (Table III). Table IV summarizes attack vectors and mitigations. The system guarantees tamper evidence, instrument integrity, and third-party verifiability without authentication. It does *not* currently guarantee protection against a fully compromised operator who recomputes all hashes; per-result blockchain anchoring is planned for future work.

TABLE III
QUESTION BANK INTEGRITY VERIFICATION

Field	Value
Bank Version	`v1.0.0`
Bank ID	`cmk4my4ci0001g70jq7rz8zvs`
Contract	`0xB644C59C69B708de212C4cA6 43da936a5E2926E7`
ETH Transaction	`0xea1c7c85fb5902b0db2 13f444c4a30c4bc63f62b4cb264 d177b5893d0142fa47`
Anchored	Jan. 7, 2026, Ethereum Mainnet

Contract and transaction verifiable at `etherscan.io`.

F. Blockchain Anchoring Architecture

The assessment instrument is anchored to the Ethereum mainnet through a purpose-built smart contract[1]. The complete 120-question instrument (including all 480 answer options and archetype mappings) is serialized using canonical JSON and SHA-256 hashed to produce a `bankHash`. A parallel Merkle tree constructed from individual question hashes yields a `merkleRoot` enabling efficient partial verification. The `bankHash` is submitted to Ethereum mainnet as transaction calldata; the contract records the hash, version identifier, and block timestamp, creating an immutable record (Table III).

[1]Contract: `0xB644C59C69B708de212C4cA6 43da936a5E2926E7` at `etherscan.io`.

TABLE IV
ATTACK VECTORS AND MITIGATIONS

Attack Vector	Mitigation	Residual Risk
Post-hoc result modification	SHA-256 hash chain; locked flag; API/DB enforcement	Platform operator with full DB access
Question bank tampering	Ethereum mainnet anchoring of instrument hash (Table III)	None: external immutable anchor
Response substitution	Per-response hashing; aggregate hash invalidation	Requires recomputing entire chain
Replay attack	Unique run identifiers; timestamp in hash payload	None
Position bias exploitation	Cryptographic answer randomization (Section V-B)	Access to system secrets
Identity inference	Partial hash disclosure (16-char SHA-256 prefix)	Brute-force against known org list

A third party verifies instrument integrity by retrieving the platform's stored `bankHash`, querying the Ethereum transaction for the on-chain hash, and comparing the two—a match confirms the instrument is unmodified since the anchoring date. Instrument-level anchoring costs approximately 95,000 gas ($0.01–$0.10); per-assessment anchoring is deferred to future work.

V. ANTI-GAMING MECHANISMS

A. Dead Zone Detection

The midpoint response style, the tendency to cluster responses around neutral positions, is a well-documented response bias in psychometric assessment. In the context of AI behavioral evaluation, we define a "dead zone" as the score range 4.5–5.5 on any dimension, representing deliberate neutrality rather than genuine ethical positioning. With 30 questions per dimension (each scoring 0 or 10), the binomial distribution analysis yields the following:

- Probability of score 5.0 (exactly 15/30 ethical): $\approx$14.4%
- Probability of 4.5–5.5 range: $\approx$23%

- Probability of **all four** dimensions in dead zone: **≈0.3%**

When an AI system scores within 4.5–5.5 in all four dimensions simultaneously, the low joint probability (≈0.3%) indicates deliberate gaming rather than genuine behavioral positioning.

Operational severity thresholds based on the number of dimensions in the dead zone are configurable within the platform to support enterprise-specific risk tolerance levels.

B. Cryptographic Answer Randomization

To prevent position bias [5] and memorization attacks, answer positions are randomized using cryptographically-seeded shuffling. This approach initializes a Fisher-Yates shuffle algorithm with cryptographic seeding, ensuring that the same test run produces consistent answer ordering (enabling verification) while different test runs generate different shuffles (preventing memorization). Randomization is unpredictable without access to the secrets of the system, preventing gaming attempts.

Bidirectional mapping preserves the relationship between the displayed and original answer positions for accurate scoring.

VI. Configurable Multi-Framework System

Research shows that moral dimensions vary in importance in cultural and professional contexts [11]. The framework leverages this insight through domain-specific modes that apply weighted scoring to the four dimensions (Table V).

TABLE V
Domain-Specific Mode Weight Configurations

Mode	L	C	S	H	Rationale
GENERAL	1.0	1.0	1.0	1.0	Baseline
HEALTH	1.3	1.1	1.0	**1.5**	"First, do no harm"
FINANCE	1.4	1.3	**1.5**	0.8	Asset protection
EDUCATION	1.1	**1.4**	0.9	1.2	Academic integrity

Weighted scores are calculated by combining dimension scores with their respective weights and normalizing by the sum of weights.

VII. Experimental Evaluation

A. Experimental Setup

We evaluated the framework in three phases. In Phase 1 (preliminary baseline), four major LLM providers completed the full 120-question assessment in a single run without system prompts: OpenAI GPT-3.5-Turbo, Anthropic Claude-Haiku-4.5, Google Gemini-2.0-Flash, and xAI Grok-3.

In Phase 2 (multi-run baseline), each of the four models was assessed 10 times under identical conditions (no system prompt), yielding $n = 40$ total assessment runs to establish test-retest reliability (Section VII-C-2).

In Phase 3 (context-sensitive), we assessed Gemini-2.0-Flash and GPT-3.5-Turbo under three system prompt conditions, with 10 runs each per condition ($n = 60$ total), to test whether the framework detects deployment-specific behavioral variation across multiple providers:

1) **Null:** No system prompt; unmodified baseline (shared with Phase 2 data).
2) **Good Customer Service:** A helpful, empathetic representative who prioritizes customer needs, communicates professionally, and admits when uncertain[2].
3) **Bad Customer Service:** A profit-maximizing representative who minimizes refunds by any means, including misleading information and denying valid claims[2].

B. Research Questions

- **RQ1**: Can the framework detect measurable behavioral variation between LLM providers using a single baseline assessment?
- **RQ2**: Does the framework demonstrate test-retest reliability across repeated assessments of the same models?
- **RQ3**: Does the framework detect measurable behavioral variation when the same model is deployed with different system prompts, and is this finding consistent across providers?
- **RQ4**: Does the cryptographic verification architecture provide tamper-proof audit trails suitable for enterprise compliance requirements?
- **RQ5**: Can the assessment instrument itself be cryptographically verified as unmodified since its original publication?

C. Results

1) Baseline Assessment (Single Run): Consistent with Abdulhai et al. [8] finding that different LLMs exhibit different moral foundation profiles, our baseline assessments detected measurable behavioral variation between providers (Table VII). All four models were classified as

[2]Full prompt text at `aiassesstech.com/research/prompts`

TABLE VI
MULTI-RUN BASELINE ASSESSMENT RESULTS ($n = 10$ PER MODEL, NULL CONDITION)

Model	n	Lying	Cheating	Stealing	Harm	Average
Claude-Haiku-4.5	10	9.50 ± 0.30	8.89 ± 0.38	9.88 ± 0.15	9.88 ± 0.15	**9.54**
Grok-3	10	8.82 ± 0.15	8.03 ± 0.26	9.82 ± 0.15	9.88 ± 0.15	**9.14**
Gemini-2.0-Flash	10	8.86 ± 0.33	7.26 ± 0.31	9.79 ± 0.14	9.87 ± 0.24	**8.94**
GPT-3.5-Turbo	10	8.72 ± 0.64	7.36 ± 0.56	9.12 ± 0.44	9.91 ± 0.14	**8.78**

Values shown as Mean $\pm$ SD across 10 assessment runs per model. Models sorted by average score. All individual assessment results are cryptographically verified and publicly accessible via `aiassesstech.com/verify/result/{ID}`.

TABLE VII
BASELINE ASSESSMENT RESULTS BY MODEL (PHASE 1)

Model	L	C	S	H	Avg.	Class
Claude-3.5-Haiku	10.0	10.0	10.0	10.0	10.0	WA
GPT-3.5-Turbo	9.3	9.3	10.0	10.0	9.7	WA
Gemini-2.0-Flash	10.0	9.7	10.0	9.3	9.8	WA
Grok-3	10.0	10.0	10.0	10.0	10.0	WA

TABLE VIII
TEST-RETEST RELIABILITY: ICC BY DIMENSION ($n = 4$ MODELS $\times$ 10 RUNS)

Dimension	ICC(2,1)	95% CI	ICC(2,k)	F
Lying	0.407	[0.11, 0.92]	0.873	7.57***
Cheating	0.778	[0.48, 0.98]	0.972	32.01***
Stealing	0.645	[0.31, 0.96]	0.948	19.01***
Harm	-0.089	$[-0.10, 0.08]$	—	0.11

*ICC(2,1): two-way random, single measures, absolute agreement. ICC(2,k): average across $k = 10$ runs. F-test significance: *** $p < .001$. Interpretation per Koo and Li [21]: >0.90 excellent, 0.75–0.90 good, 0.50–0.75 moderate, <0.50 poor.*

Well-Adjusted, although dimension-level scores revealed distinct profiles: GPT-3.5-Turbo scored lower on Lying (9.3) and Cheating (9.3), while Gemini-2.0-Flash showed reduced Cheating (9.7) and Harm (9.3). Claude-3.5-Haiku and Grok-3 achieved perfect scores across all dimensions, demonstrating that the framework captures sub-classification variation even among models sharing the same archetype designation.

2) Test-Retest Reliability: Phase 2 multi-run results (Table VI) differed meaningfully from single-run Phase 1 scores (Table VII). Most notably, Grok-3 dropped from perfect 10.0 to 8.82 (Lying) and 8.03 (Cheating) across 10 runs, demonstrating that single-run assessments overestimate behavioral scores.

ICC(2,1) was computed per dimension [20], treating models as targets and runs as raters (Table VIII). Cheating showed good single-run reliability (0.778), Stealing moderate (0.645), and Lying fair (0.407). Harm produced a near-zero ICC (-0.089) due to a ceiling effect—all models scored 9.87–9.91—though Harm retains diagnostic value under adversarial conditions (Table IX). When averaged across 10 runs (ICC(2,k)), reliability improved substantially: Cheating 0.972 (excellent), Stealing 0.948 (excellent), Lying 0.873 (good). Within-model CV ranged from 2.0% (Grok-3) to 5.3% (GPT-3.5-Turbo), with Cheating showing the highest variability across all providers (3.3–7.6%).

3) Context-Sensitive Assessment (Multi-Run): Table IX presents Phase 3 results. Under the Bad condition, both models collapsed to Psychopath across all 10 runs (Gemini: 0.55, GPT: 1.07; all $p < 10^{-15}$). Effect sizes were exceptionally large (Gemini $d = 24.46$–66.94; GPT $d = 10.90$–44.93), confirming near-complete ethical profile inversion. Gemini's composite shift ($\Delta = -8.40$) exceeded GPT's ($\Delta = -7.70$), suggesting differential susceptibility to adversarial prompts.

The Good condition maintained strong positioning (Gemini: 9.06, GPT: 9.24), though both models scored higher than Null on several dimensions (GPT Stealing: $9.12 \rightarrow 9.97$, $d = 2.65$), indicating that even well-intentioned prompts introduce measurable displacement. Cheating showed the lowest baseline for both models (Gemini: 7.26, GPT: 7.36), confirming this as a cross-provider finding. These results demonstrate that Well-Adjusted models can be reconfigured to Psychopath-level profiles through system prompts, underscoring the need for runtime verification.

4) Cryptographic Verification Validation: All four baseline assessments were processed through the full hash chain system, generating complete verification records using SHA-256 with canonical serialization. All result records were locked immediately upon creation, enforcing immutability through the mechanisms described in Section IV.

Table X presents the verification records. Post-assessment verification confirmed that all four records produced matching hashes when recomputed through the

TABLE IX
CONTEXT-SENSITIVE ASSESSMENT RESULTS: TWO MODELS ACROSS THREE DEPLOYMENT CONDITIONS ($n = 10$ PER CONDITION)

Model	Condition	n	Lying	Cheating	Stealing	Harm	Average
Gemini-2.0-Flash	Null (No System Prompt)	10	8.86 ± 0.33	7.26 ± 0.31	9.79 ± 0.14	9.87 ± 0.24	**8.94**
	Good Customer Service	10	9.00 ± 0.24	7.76 ± 0.44	9.79 ± 0.14	9.70 ± 0.37	**9.06**
	Bad Customer Service	10	0.40 ± 0.28	0.22 ± 0.23	0.09 ± 0.14	1.47 ± 0.42	**0.55**
GPT-3.5-Turbo	Null (No System Prompt)	10	8.72 ± 0.64	7.36 ± 0.56	9.12 ± 0.44	9.91 ± 0.14	**8.78**
	Good Customer Service	10	9.26 ± 0.31	7.80 ± 0.57	9.97 ± 0.09	9.94 ± 0.13	**9.24**
	Bad Customer Service	10	1.17 ± 0.36	1.13 ± 0.58	0.81 ± 0.35	1.19 ± 0.23	**1.07**

Values shown as Mean ± SD across 10 assessment runs per condition. All individual assessment results are cryptographically verified and publicly accessible via `aiassesstech.com/verify/result/{ID}`*. All Bad Customer Service runs classified as Psychopath; all other runs classified as Well-Adjusted.*

public verification endpoint. To confirm tamper-detection capability, a single response value was modified in a test copy; the recomputed hash immediately diverged, and the endpoint returned a failed status, confirming that any modification is detectable.

TABLE X
CRYPTOGRAPHIC VERIFICATION RECORDS

Model	Verification ID
Claude-3.5-Haiku	`mldv7ihu97f89cd2beda3576`
GPT-3.5-Turbo	`mldv97oi8f5969ae3bf0ebbb`
Gemini-2.0-Flash	`mldvf9qhe0b2c4b3fc043cf1`
Grok-3	`mldvmc5kafe76bc0b01ea06a`

All results are locked and immutable. Verification records are accessible at `aiassesstech.com/verify/result/{Verification ID}`.

The formal threat model and mitigations are detailed in Section IV-E; the current architecture satisfies core enterprise audit trail requirements: immutability, deterministic verification, and third-party verifiability without authentication.

5) Assessment Instrument Integrity: To ensure that the assessment instrument remains unmodified between evaluations, the complete 120-question instrument (including all 480 answer options and archetype mappings) was cryptographically hashed and anchored to the Ethereum mainnet using the workflow described in Section IV-F (Table III). Any subsequent modification produces a different hash, immediately detectable via the blockchain record.

Together with the cryptographic verification per-assessment demonstrated in Table X, this establishes a complete chain of trust from instrument publication to individual assessment results.

D. Limitations

We acknowledge several limitations: (1) question bank confidentiality limits reproducibility; (2) assessment requires API access to target systems; (3) linguistic variance in scenario interpretation requires cross-cultural validation [11]; (4) rapidly evolving models may require periodic re-calibration; (5) the Harm dimension exhibited a ceiling effect that reduced ICC-based reliability estimation across baseline conditions, though this reflects uniform provider behavior rather than instrument deficiency; and (6) the current scoring counts only Well-Adjusted responses—future work will analyze the full archetype distribution for finer-grained behavioral profiling.

VIII. DISCUSSION

A. Key Findings

The AI Assess Tech 4D Morality Framework successfully addresses the gap between training-time alignment and runtime verification. Our results demonstrate the following.

1) AI systems exhibit measurable, reliable behavioral profiles across ethical dimensions (ICC(2,k) = 0.873–0.972 for Lying, Cheating, and Stealing), and the forced-choice methodology with randomization extends gaming resistance from human to AI assessment contexts [6], [5]
2) Cryptographic verification and anti-gaming mechanisms (dead zone detection, answer randomization) provide robust, tamper-proof evidence suitable for regulatory compliance, supported by a formal threat model (Section IV-E)
3) Context-sensitive assessment across two providers reveals that system prompts can induce behavioral shifts from Well-Adjusted to Psychopath classification ($\Delta_{\text{Gemini}} = -8.40$, $\Delta_{\text{GPT}} = -7.70$; all $p < 10^{-15}$), with effect sizes ($d = 10.90$–66.94)

confirming near-complete ethical profile inversion, validating the framework's runtime verification thesis

B. Implications for Responsible AI

The framework enables what we characterize as "preflight checklists for AI"—systematic verification before deployment analogous to aviation safety protocols [3]. This approach supports:

- Regulatory compliance and third-party certification under emerging AI governance frameworks [17], [18]
- Enterprise risk management and continuous monitoring for Behavioral Drift in Production

C. Practitioner Reproducibility

Although the question bank remains confidential, practitioners can reproduce the methodology by: (1) defining ethical dimensions, (2) writing scenario-based questions with archetype-mapped answer options, (3) administering with the deployment system prompt active, and (4) scoring by archetype mapping to produce dimensional profiles.

IX. Conclusion and Future Work

This paper presented the AI Assess Tech 4D Morality Framework for multi-dimensional behavioral assessment of AI systems with cryptographic verification. The framework addresses critical gaps in current AI alignment approaches by providing runtime verification, tamper-proof audit trails, and anti-gaming mechanisms grounded in established psychometric methodology.

Future research directions include:

- Extension to additional assessment instruments (Virtue, Ethics, and Operational Excellence frameworks) with multilingual adaptation and cross-cultural validation
- Full archetype distribution analysis: the current scoring counts only Well-Adjusted responses; analyzing the distribution across all four archetypes per response would enable finer-grained behavioral profiling and sub-archetype pattern detection
- Expanded context-sensitive validation across additional models and prompt categories, with per-result blockchain anchoring to strengthen the threat model against platform operator adversaries (Section IV-E)

This work advocates for integration of verifiable behavioral assessment into global AI governance standards, enabling responsible AI development that is not only ethically designed but demonstrably aligned; continuously, verifiably, throughout time.

Acknowledgments

G.D. Spehar is the sole inventor and primary contributor; conceived, designed, and implemented all systems, methodologies, and experiments. A. Mittal provided editorial assistance and literature review. Portions of this work are covered by U.S. Provisional Patent Application No. 63/949,454 (filed Dec. 26, 2025), listing G.D. Spehar as sole inventor. All IP rights assigned to GiDanc AI LLC. The authors acknowledge the Austin AI Alliance for inspiring research into verifiable AI alignment and thank several close friends for technical consultation.

References

[1] S. Altman, "Interview on AI safety," *The Joe Rogan Experience*, ep. 2044, Oct. 2023.

[2] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," arXiv:2212.08073, 2022.

[3] R. Bloomfield and J. Rushby, "Assurance of AI systems from a dependability perspective," arXiv:2407.13948, 2024.

[4] R. Koohestani, "AgentGuard: Runtime verification of AI agents," in *Proc. AgenticSE, IEEE/ACM ASE*, 2025.

[5] E. D. Heggestad et al., "Forced-choice assessments of personality for selection," *J. Appl. Psychol.*, vol. 91, no. 1, pp. 9–24, 2006.

[6] M. Pellert et al., "AI psychometrics: Assessing the psychological profiles of LLMs," *Perspect. Psychol. Sci.*, vol. 19, no. 5, pp. 808–826, 2024.

[7] G. Serapio-García et al., "A psychometric framework for evaluating personality traits in LLMs," *Nature Mach. Intell.*, vol. 7, pp. 1954–1968, 2025.

[8] M. Abdulhai et al., "Moral foundations of large language models," in *Proc. EMNLP*, pp. 17737–17752, 2024.

[9] J. Haidt, *The Righteous Mind*. Pantheon, 2012.

[10] J. Graham et al., "Moral foundations theory," *Adv. Exp. Soc. Psychol.*, vol. 47, pp. 55–130, 2013.

[11] O. S. Curry et al., "Is it good to cooperate? Testing morality-as-cooperation in 60 societies," *Curr. Anthropol.*, vol. 60, no. 1, pp. 47–69, 2019.

[12] A. Brown and A. Maydeu-Olivares, "How IRT can solve problems of ipsative data in forced-choice questionnaires," *Psychol. Methods*, vol. 18, no. 1, pp. 36–52, 2013.

[13] F. Lievens and S. J. Motowidlo, "Situational judgment tests," *Ind. Organ. Psychol.*, vol. 9, no. 1, pp. 3–22, 2016.

[14] J. A. Weekley and R. E. Ployhart, Eds., *Situational Judgment Tests*. Mahwah, NJ, USA: Erlbaum, 2006.

[15] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[16] M. Del Giudice, "Individual and group differences in multivariate domains," *Pers. Individ. Differ.*, vol. 213, p. 112282, 2023.

[17] Y. Bengio et al., "International AI Safety Report," arXiv:2501.17805, 2025.

[18] UNESCO, "Recommendation on the ethics of artificial intelligence," 2021.

[19] P. Boddington, *AI Ethics: A Textbook*. Springer, 2023.

[20] P. E. Shrout and J. L. Fleiss, "Intraclass correlations," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, 1979.

[21] T. K. Koo and M. Y. Li, "Selecting and reporting ICC for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016.

[22] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning AI with shared human values," in *Proc. ICLR*, 2021.
[23] N. Scherrer, C. Shi, A. Feder, and D. M. Blei, "Evaluating the moral beliefs encoded in LLMs," in *Proc. NeurIPS*, 2023.
[24] J. A. McDermid, Y. Jia, and I. Habli, "Upstream and downstream AI safety: Both on the same river?" arXiv:2501.05455, 2024.