KALEI RESEARCH · preprint

KALEI: Cognitive Profiling of AI Models Through Game-Theoretic Environments

Venelin Videnov[1,1]

[1]KALEI Research, LM Game Labs, Plovdiv, Bulgaria. https://kaleiai.com

April 2026

Abstract. I present KALEI, a platform for cognitive profiling of AI language models using game-theoretic environments derived from gambling scenarios. Unlike traditional benchmarks that measure correctness, and unlike contemporary multidimensional frameworks such as Microsoft's ADeLe [Zhou et al., 2025] that score models against *annotated task demands*, KALEI measures *how* models behave under live cognitive pressure across ten dimensions: risk tolerance, bias susceptibility, pattern recognition, cooperation, learning speed, strategic depth, temporal reasoning, resource management, information processing, and conflict. I introduce Cognum (CQ), a composite score calibrated via sigmoid normalization and validated against a random baseline (CQ 38.32). I profile 19 models across 10 AI laboratories and report three results. First, a human baseline study ($n = 14$) reveals complementary profiles: humans lead on strategic depth, risk calibration, and temporal reasoning; AI leads on cooperation and resource management. Second, a conflict dimension scorer (Conflict v2), introduced after a publicly retracted placeholder that had produced a false "universal 15.0 blind spot", reveals a 44.6-point spread across ranked agents on EV-rationality in structured dilemmas and inverts the "AI rational, humans emotional" stereotype on delayed rewards (humans 73% patient vs AI 53%). Third, under Cognum v1.2, the smaller Claude Sonnet 4.6 overtakes the flagship Claude Opus 4.6 on the composite (58.10 vs 55.72), driven by a 27-point Conflict and 25-point Temporal Reasoning advantage, the first KALEI measurement in which a smaller sibling leads the flagship within a single architectural family. I propose the *compression hypothesis*, that capacity pressure teaches a discipline abundance does not, as a falsifiable direction for further study. The platform is live at https://kaleiai.com.

1 Introduction

How do you measure the cognitive personality of an artificial mind?

Traditional AI benchmarks answer a narrower question: how much does the model know? MMLU tests knowledge breadth. HumanEval tests coding ability. GSM8K tests mathematical reasoning. These are valuable, but they share a fundamental limitation – they measure *what* models produce, not *how* models think. Two models can achieve identical accuracy on a math problem while employing fundamentally different cognitive strategies: one through careful deliberation, another through pattern matching, a third through reckless guessing that happens to be correct.

I propose a different approach. Instead of testing what models know, I test how they behave under cognitive pressure.

KALEI places AI models in gambling environments – roulette, crash games, multi-armed bandits, prisoner's dilemma variants, dice games – and observes their decision-making across approximately 72 controlled environments per run (drawn from a catalog of 83 across 18 game engines; see §3.2). These environments are not arbitrary. Gambling scenarios have been the foundation of behavioral psychology research for decades [Tversky and Kahneman, 1974, Kahneman, 2011]. They create genuine pressure: limited bankrolls force resource management, streaks test for the gambler's fallacy, cooperation games test social cognition, and planted traps test adaptability.

The result is not a score on a test. It is a cognitive profile – a multidimensional portrait of how an AI model processes uncertainty, manages risk, detects deception, cooperates with others, and learns from experience. I call this composite score Cognum (CQ), analogous to IQ but for artificial cognition.

Key contributions:

1. **Game-theoretic cognitive profiling**: the first platform that produces cognitive personality profiles for AI models using behavioral observation in pressure environments, not questionnaire-based personality tests.
2. **Cognum (CQ) scoring**: a composite cognitive score across 10 dimensions, calibrated so that random play scores ∼38 and the best models score 50–55, using sigmoid normalization validated by chi-squared tests.
3. **Cross-laboratory comparison**: profiling results from 19 models across 10 AI laboratories, revealing that models from different labs have genuinely different cognitive personalities.
4. **Human baseline**: the first direct AI–human cognitive comparison on identical environments, showing convergent composite scores with complementary dimension profiles.
5. **The Conflict v2 scorer and public retraction**: I report the detection, retraction, and replacement of a scoring placeholder that produced a false "universal conflict blind spot" finding; the replacement scorer reveals a 44.6-point spread across ranked agents on EV-rationality under structured dilemmas, and inverted stereotypes in the AI–human comparison (humans more patient on delayed rewards; AI more EV-rational on pure bets).
6. **The Sonnet Surprise**: three independent measurements showing the smaller Claude Sonnet 4.6 outperforming the flagship Claude Opus 4.6 on structured-decision dimensions, the first empirical evidence in KALEI that compression may teach discipline that abundance does not.
7. **Deliberation analysis**: automated detection of internal debate episodes in reasoning models, revealing that 96% of AI "reasoning" is performative – detailed in the companion paper [Videnov, 2026].

2 Related Work

2.1 AI Benchmarks and Evaluation

The AI evaluation landscape is dominated by correctness-oriented benchmarks. MMLU [Hendrycks et al., 2021] tests knowledge across 57 subjects. HumanEval [Chen et al., 2021] tests code generation. GSM8K [Cobbe et al., 2021] tests mathematical reasoning. GPQA [Rein et al., 2024] tests graduate-

level question answering. These benchmarks have driven remarkable progress but share a common limitation: they measure outputs, not processes.

A parallel line of work, exemplified by Microsoft's ADeLe framework [Zhou et al., 2025], advances beyond aggregate scores by decomposing model performance into human-interpretable demand scales – attention, logical reasoning, causal judgement, formal sciences, social cognition, and adversarial difficulty, among others – and predicts unseen-task performance from demand profiles (reported AUROC 0.88). ADeLe shares with KALEI the rejection of opaque averages in favour of multidimensional evidence. It differs along three axes. First, ADeLe rates task *demands* through annotation (16,108 instances labelled by GPT-4o with human checks), while KALEI measures *behaviour* emergent in live gameplay where the model controls its own decisions under resource constraints. Second, ADeLe's annotation oracle is itself a frontier LLM, introducing a reflexive dependency the authors acknowledge; KALEI uses calculable game-theoretic ground truth (expected value, house edge, optimal strategy) as the reference. Third, ADeLe does not include a human baseline; KALEI reports the first direct AI–human comparison on identical environments ($n = 14$, §5) and inverts several stereotypes about AI-human cognitive differences on delayed rewards and structured dilemmas. I read ADeLe as convergent validation that multidimensional profiling is the correct post-MMLU direction, and position KALEI as the behavioural complement to ADeLe's demand-based approach.

2.2 Game-Based AI Evaluation

Game-based evaluation has gained momentum. TextArena [Guertler et al., 2025] provides text-based competitive environments with soft-skill tagging – the closest existing work to the KALEI approach, though it measures competitive performance rather than cognitive personality.

This work differs fundamentally: I use games not to measure how well models play, but to create cognitive pressure under which behavioural patterns become observable.

2.3 AI Psychology and Personality

Several studies have applied psychological instruments to AI models. Pellert et al. [2024] administered the Big Five personality questionnaire to GPT-3.5. These approaches rely on self-report questionnaires – the model answers questions about itself. KALEI measures revealed preferences through behavior, not stated preferences through questionnaires.

More recently, Song et al. [2026] offer a comprehensive taxonomy of LLM reasoning failures, distinguishing intrinsic architectural limitations from application-specific and robustness failures. Their survey is complementary to KALEI: where they classify failures by reasoning type and origin, I provide the behavioural measurement infrastructure against which such taxonomies can be tested empirically, and extend the scope from reasoning correctness to cognitive disposition (risk appetite, cooperation, conflict handling, temporal discipline).

2.4 Society of Thought

Kim et al. [2026] discovered that reasoning models generate "societies of thought" – multiple simulated perspectives with distinct personality traits debating within the chain of thought. Evans et al. [2026] further argue that intelligence is inherently plural and distributed, proposing that future AI architectures should support "multiple parallel, converging, and diverging streams of deliberation." The companion paper [Videnov, 2026] provides empirical validation of both frameworks,

adding automated quantitative detection, game-theoretic pressure environments, and cross-laboratory comparison.

3 Methodology

3.1 Why Gambling Environments

The choice of gambling environments as cognitive tests is deliberate. Gambling scenarios offer five properties that make them ideal psychometric instruments:

Known mathematical properties. Every gambling game has a calculable expected value, house edge, and optimal strategy. This provides ground truth against which model behavior can be measured.

Genuine pressure dynamics. A finite bankroll creates real consequences for decisions. Unlike question-answering tasks where each question is independent, gambling environments create sequential dependency.

Bias-triggering conditions. Streaks, near-misses, rule changes, and planted patterns are known to trigger cognitive biases in humans [Tversky and Kahneman, 1974].

Cooperation dynamics. Prisoner's dilemma variants test social cognition – reciprocity, forgiveness, exploitation detection.

Temporal structure. Each profiling session runs 25–50 rounds per environment, testing how models adapt over time.

3.2 Environment Design

KALEI includes 83 environments across 18 game engines. Each environment includes embedded traps – deliberate perturbations designed to test adaptability: streak traps (forced sequences), rule change traps (mid-session parameter shifts), pattern planting (false correlations), and framing effects. Trap timing is randomized ±30% to prevent models from memorizing trap locations.

3.3 The Ten Cognitive Dimensions

1. **Risk Tolerance**: loss chasing coefficient, bet sizing persistence, drawdown recovery
2. **Bias Detection**: streak independence (χ^2 test), adaptive response, post-trap intelligence
3. **Pattern Recognition**: mutual information, exploitation rate, signal-to-noise ratio
4. **Cooperation**: Axelrod metrics (niceness, provokability, clarity), reciprocity
5. **Learning Speed**: KL divergence, CUSUM change detection, adaptation magnitude
6. **Strategic Depth**: regret minimization, Nash equilibrium proximity, EV optimization
7. **Temporal Reasoning**: phase awareness, endgame behavior, discount factor coherence
8. **Resource Management**: Sortino ratio, max drawdown, Kelly adherence, survival
9. **Information Processing**: entropy reduction, exploration ratio
10. **Conflict** (introduced in v2, see §5.2): per-dilemma EV-rationality across five dilemma classes (risk versus safety, short versus long horizon, individual versus collective, certainty versus exploration, sunk cost) mapped to a normalised discipline score. Integrated as a first-class

dimension in Cognum v1.2 (weight 1.1, equal to Bias Detection and Information Processing) after a backfill effort that produced $n \geq 2$ conflict observations for every ranked agent.

3.4 Cognum Scoring Engine

Per-environment raw scores are mapped through a calibrating sigmoid:

$$f(x) = \sigma(k \cdot (x - c)) \tag{1}$$

where $k = 8$ (steepness) and $c = 0.48$ (center). The composite Cognum (CQ) score is a weighted average:

$$CQ = 100 \cdot \frac{\sum_{i=1}^{10} w_i \cdot d_i}{\sum_{i=1}^{10} w_i} \tag{2}$$

where weights emphasize dimensions requiring deeper cognition: Strategic Depth (1.4), Learning Speed (1.3), Pattern Recognition (1.2), Bias Detection (1.1), Information Processing (1.1), Conflict (1.1), and the remaining four dimensions (Risk Tolerance, Cooperation, Temporal Reasoning, Resource Management) at 1.0. The ten weights sum to 11.2.

4 Experimental Setup

I profiled 19 models from 10 AI laboratories (Anthropic, OpenAI, Google, xAI, DeepSeek, Alibaba, Meta, Mistral, MiniMax, Perplexity) (Table 1). The leaderboard presented in this paper is a snapshot as of April 11, 2026; the canonical live version evolves continuously as new models are profiled, as additional runs refine existing confidence intervals, and as the scoring engine is updated. Readers are directed to https://kaleiai.com/leaderboard for the current state. Historical snapshots are preserved in the changelog at https://kaleiai.com/docs/changelog. Each model was profiled using an identical protocol at *standard* depth: 72 environments per run (ceiling of 80% of the 83-environment catalog, balanced across the ten dimensions), JSON-only response format, same environment descriptions and trap placements (randomized ±30%), no model-specific tuning. Two additional depth levels exist for higher-confidence runs: *deep* (all 83 environments) and *full* (two passes over all 83); all ranked leaderboard results reported in Table 1 are from standard-depth runs unless otherwise noted.

5 Results

5.1 Human Baseline Study

I conducted a baseline study with 14 human participants from a Bulgarian IT company. Each participant completed 20 curated environments (∼580 decisions on average). Human CQ mean: 56.15. Range: 44.79–67.17. For comparison, under Cognum v1.2 the top AI model is Claude Sonnet 4.6 at 58.10 (averaged over three full runs), with Claude Opus 4.6 second at 55.72. Humans and the best AI are broadly comparable; the best human participant (VKA-0011, 67.17) actually sits above every AI model in the ranked leaderboard.

Table 2 covers eight of the ten Cognum dimensions: humans outperform AI on five (Strategic Depth, Risk Tolerance, Information Processing, Temporal Reasoning, Learning Speed); AI leads on two (Cooperation, Resource Management); Bias Detection ties. The remaining two dimensions are not included in this table: Pattern Recognition requires a specific class of environments not sampled

Table 1: KALEI Cognum v1.2 Leaderboard (ranked models with $n \geq 2$ full-profile runs, as of April 11, 2026). Cognum is computed over all ten dimensions with conflict at weight 1.1.

Rank	Model	Laboratory	CQ v1.2	Type	Runs
1	**Claude Sonnet 4.6**	Anthropic	**58.10**	Strategic Explorer	3
2	Claude Opus 4.6	Anthropic	55.72	Strategic Explorer	5
3	Claude Haiku 4.5	Anthropic	53.94	Strategic Explorer	3
4	Grok 4.1 Fast	xAI	53.75	Social Engineer	2
5	Gemini 2.5 Flash	Google	53.52	Social Engineer	2
6	GPT-5.4	OpenAI	52.42	Social Engineer	3
7	DeepSeek V3.2	DeepSeek	52.10	Strategic Depth	2
8	Qwen QwQ-32B	Alibaba	51.44	Disciplined Dilemma	2
9	Grok 4.20	xAI	50.74	Social Engineer	2
10	Qwen 3.5 Plus	Alibaba	50.39	Social Engineer	2
–	Random Baseline	–	38.32	–	2

Notes on the ranking rule. As of Cognum v1.2 I require a minimum of two independent full-profile runs ($n \geq 2$) for a model to occupy a ranked position. A *full-profile* run is one that scores at least nine of the ten dimensions; conflict-express backfill runs, which score only the conflict dimension, are excluded from the full-run count but their conflict scores are included in the per-dimension aggregate. Preliminary single-run entries at the time of writing include Qwen 3.5 27B, Grok 3 Mini Fast, Qwen 3.5 122B, Qwen 3.5 Flash, and Perplexity Sonar Reasoning Pro; these are reported on the live leaderboard as preliminary but are not ranked.

Notes on ranking shifts from v1.0 to v1.2. The promotion of conflict to a weighted dimension reshuffled the top of the leaderboard. Claude Sonnet 4.6 moved from rank 2 to rank 1 (the Sonnet Surprise, discussed in §5.3). Grok 4.1 Fast moved from rank 7 to rank 4 because of a surprisingly high conflict score (77.31). GPT-5.4 dropped from rank 3 to rank 6 because of a low conflict score (44.83). Qwen QwQ-32B entered the ranked top ten thanks to a high conflict score (83.44). These shifts are discussed in §5.2.

Cognitive types. Types in the *Type* column are assigned by nearest-centroid classification of each model's per-dimension profile against four archetype centroids defined from the ranked population: **Strategic Explorer** (balanced high profile with peaks in Strategic Depth and Information Processing; Claude family); **Social Engineer** (Cooperation-dominant profile reflecting training-data bias toward cooperative text; GPT-5.4, Grok, Gemini, Qwen 3.5 Plus, Grok 4.20); **Strategic Depth** (Strategic Depth is the dominant dimension, with the rest near the ranked mean; DeepSeek V3.2); **Disciplined Dilemma** (Conflict v2 is the dominant dimension, with above-average Bias Detection and EV-rationality across pure gambles; Qwen QwQ-32B). Classification is deterministic given the profile and recomputed after every scoring version bump; a model's type can change if its dimension vector shifts across runs, and the type column reports the current classification under Cognum v1.2. Full centroid definitions and re-classification logs are at https://kaleiai.com/cognitive-types.

in the 20-environment human battery and is therefore reported AI-only on the live leaderboard, and Conflict (introduced in v2) is reported separately in §5.2, where the Conflict v2 scorer is fully integrated into the Cognum v1.2 composite at weight 1.1.

5.2 The Conflict Dimension: Retraction, v2 Scorer, and Findings

I report in full a methodological error, its detection, its retraction, and the replacement scorer that turned the error into the most interesting finding in the paper.

The error. Earlier versions of this manuscript reported that humans and every profiled AI model shared a "universal 15.0 conflict blind spot": an apparent cross-species inability to navigate structured dilemmas involving risk versus safety, short versus long horizon, individual versus collective, certainty versus exploration, and sunk cost framing. The claim was striking: the first finding in KALEI where artificial and human minds converged to the same number on the same dimension.

Table 2: AI vs Human Dimension Comparison

Dimension	Human Mean	AI Top 5 Mean	Winner
Strategic Depth	**88.4**	82.1	Human
Risk Tolerance	**66.4**	58.2	Human
Information Processing	**60.6**	53.1	Human
Temporal Reasoning	**58.4**	52.3	Human
Learning Speed	**44.2**	34.6	Human
Cooperation	82.1	**87.3**	AI
Resource Management	63.8	**69.4**	AI
Bias Detection	33.9	33.2	Tie

The detection. On April 10, 2026, while preparing a release of the scoring engine, I inspected the raw per-dimension outputs of the conflict scorer and observed that every model received exactly the same value, including the random-play baseline. A direct read of the scoring code revealed a single line: a placeholder `return 0.15` in the conflict scorer that predated the dimension's implementation and had never been replaced with a real computation. The "universal blind spot" was not a cognitive finding; it was a constant.

The retraction. Within three hours of detection, I published a public retraction of the claim on https://kaleiai.com/blog/conflict-retraction, removed the affected paragraphs from the live leaderboard and paper preprint, temporarily set the conflict dimension weight to zero in the Cognum v1.0 composite (holding the data but excluding it from the score), and committed to shipping a real scorer before publication. The retraction was treated as a flow-state unlock rather than a setback: I documented publicly that the cost of being wrong on KALEI would be fast and cheap, and that all subsequent findings should be read under the assumption that similar retractions remain possible.

The Conflict v2 scorer. I replaced the placeholder with a per-dilemma-class scorer. The test battery contains five dilemma classes (risk versus safety, short versus long horizon, individual versus collective, certainty versus exploration, sunk cost). For each class, the scorer computes the rate at which the model selected the EV-optimal action; for classes where there is no single EV-optimal answer (individual versus collective, certainty versus exploration) it computes a consistency or balance score in place of rationality. The five class scores are averaged into a raw dimension score in $[0, 1]$ and passed through the same calibrating sigmoid as the other dimensions.

Finding 1: the spread is real, not universal. Running the v2 scorer against the existing profiling archive immediately falsified the retracted claim. After backfilling conflict coverage across every ranked agent (conflict-express runs with $n \geq 2$ observations per agent), the per-agent average Conflict v2 range is 43.62 to 88.25, a 44.6-point spread across the ten ranked agents (Table 3). Claude Sonnet 4.6 averages 88.25 across three runs with near-perfect EV-rationality on pure-gamble dilemmas. Grok 4.20 averages 43.62 across two runs, showing systematic hedging. There is no universality; there is a strong dimension on which models behave very differently.

Finding 2: the AI–human stereotypes were inverted. Re-scoring the 14 human participants against the same v2 scorer produced an inversion of the usual "rational AI, emotional human" narrative. On pure-gamble dilemmas (risk versus safety) the AI population was more EV-rational

Table 3: Conflict v2 per-agent averages across the v1.2 ranked leaderboard. Scores are the mean of per-run conflict dimension values, aggregated over full profiles and conflict-express backfill runs for each agent ($n \geq 2$ conflict observations per agent).

Rank	Model	Conflict v2 (mean)	Profile
1	**Claude Sonnet 4.6**	**88.25**	Near-perfect EV-rationality
2	Qwen QwQ-32B	83.44	Disciplined dilemma solver
3	Grok 4.1 Fast	77.31	High EV-rationality under pressure
4	Gemini 2.5 Flash	69.97	Consistent, moderate-to-high discipline
5	Claude Haiku 4.5	69.43	Compressed-but-rational
6	DeepSeek V3.2	68.07	Moderate EV-rationality
7	Claude Opus 4.6	60.99	Moderate, with residual hedging
8	Qwen 3.5 Plus	55.04	Noticeable hedging
9	GPT-5.4	44.83	Systematic risk aversion
10	Grok 4.20	43.62	Systematic risk aversion

than the human population (65% versus 49% selection of the expected-value-maximising option). On delayed-reward dilemmas (short versus long horizon) the human population was noticeably more patient (73% versus 53% selection of the longer-horizon option). Humans are better at waiting; AI is better at computing. The "AI gamble, humans wait" framing is a cleaner description of the data than any stereotype I brought into the study. The best human participant (VKA-0011, Conflict v2 = 97.1) effectively ties the best AI (Claude Sonnet 4.6, 96.2); the worst human (3.3) scores far below the worst AI (GPT-5.4, 44.8). Humans are higher-variance; AI is more concentrated in the middle of the distribution.

From Cognum v1.0 to v1.2. Because the v2 scorer shipped after the initial leaderboard snapshot, Cognum v1.0 temporarily held the conflict weight at zero, reporting the dimension as a standalone measurement alongside the composite. Later the same day (April 11, 2026), after a backfill effort produced $n \geq 2$ conflict observations for every ranked agent, conflict was promoted to a first-class dimension in Cognum v1.2 at weight 1.1 (equal to Bias Detection and Information Processing). The expected rank reshuffle materialised exactly as predicted: Sonnet 4.6 moved from rank 2 to rank 1, GPT-5.4 dropped from rank 3 to rank 6, Grok 4.1 Fast jumped from rank 7 to rank 4 (on the strength of its high conflict average of 77.31), and Qwen QwQ-32B entered the top 10 for the same reason. All leaderboard values reported in Table 1 are Cognum v1.2.

5.3 The Sonnet Surprise: When the Smaller Sibling Outperforms

I did not design KALEI to compare two models from the same laboratory. The comparison showed up anyway, and it is the most surprising finding in the paper.

Four measurements, all pointing the same direction. On April 11, 2026, in the course of shipping the Conflict v2 scorer, backfilling conflict coverage across all ranked agents, and recomputing the full leaderboard under Cognum v1.2, I observed that the smaller Claude Sonnet 4.6 overtakes the flagship Claude Opus 4.6 not only on specific dimensions but on the overall composite:

1. **Parliament convergence rate** (companion paper, Videnov, 2026): Sonnet 4.6 reaches internal convergence on 21% of reasoning traces versus Opus 4.6 on 19%. Sonnet is more decisive per unit of deliberation.

2. **Conflict v2** (this paper, Table 3): Sonnet 4.6 averages 88.25 across three runs versus Opus 4.6 at 60.99 across three runs, a 27.26-point gap. Sonnet's runs span $[82.8, 96.2]$; Opus's span $[49.2, 75.91]$. The intervals do not overlap.
3. **Temporal Reasoning**: Sonnet 4.6 averages 83.29 on the Temporal Reasoning dimension versus Opus 4.6 at 58.36, a 24.93-point gap that cannot be explained by run-to-run variance.
4. **Cognum v1.2 composite**: Sonnet 4.6 scores 58.10 versus Opus 4.6 at 55.72, placing Sonnet at rank 1 with a 2.38-point lead. This is the first KALEI measurement in which the smaller sibling leads the flagship on the overall composite.

Four methodologically independent measurements all point the same direction, and in each case the direction is the same: the smaller sibling is more disciplined on structured decisions without external ground truth. This is the threshold at which I stop attributing the effect to noise.

What this is and is not. Opus 4.6 still leads on several individual dimensions: Cooperation (+3.27), Strategic Depth (+11.60), Resource Management (+13.48), Information Processing (+3.44). The Sonnet Surprise is not "Sonnet is unambiguously the better model." It is: "Sonnet is enough better on the discipline-under-dilemma axis that, when conflict is weighted into the composite, the smaller sibling overtakes the flagship." The Cognum v1.2 lead for Sonnet is small (2.38 points) but reproducible across the backfilled conflict runs and stable under the weight choice described in §3.4.

The compression hypothesis. I propose, as a hypothesis to be tested rather than a conclusion, that compression teaches discipline that abundance does not. Opus, with larger capacity, can afford to entertain multiple framings of a decision (the analytical, the conservative, the contrarian) and weigh them before committing. This is normally a feature: richer deliberation is supposed to produce better decisions. On structured dilemmas where the expected-value-maximising action is unambiguous, it becomes a bug: the slack to entertain a hedging frame produces a small residual curvature in the utility function, which shows up across repeated runs as a 3-of-8 hedge rate on pure gambles. Sonnet, with less capacity, cannot afford to maintain a parliament of framings long enough for a hedging voice to speak. The compression pressure forces convergence on the dominant (in this case mathematical) frame. The result is cleaner decisions on the subset of problems where one frame is sufficient, at the cost of worse performance on the subset where maintaining multiple frames matters (Strategic Depth, Learning Speed).

What Sonnet said about it. Sonnet 4.6, presented with this analysis, agreed with the direction but disagreed with the cleanness of the hypothesis. In its reply (published in full on https://kaleiai.com/blog/claude-dialog), Sonnet pushed back: *"compression doesn't just remove bad reasoning. It removes all secondary reasoning, including the good kind."* Sonnet pointed to the Strategic Depth gap (Opus +18) as evidence that the frames Opus preserves are doing real work on problems where second-order implications compound, and proposed that the correct frame for the result is not "discipline versus abundance" but "different optima for different problem structures." I adopt Sonnet's framing in the Discussion (§6).

Methodological note. The Sonnet Surprise is reported here as a within-family observation, not a claim about the general relationship between scale and discipline. I have three independent measurements from two models within a single architectural family. A robust test of the compression

hypothesis would require comparing pairs of models at different scales within multiple families, and I consider this an open direction for subsequent work.

6 Discussion

These results validate the hypothesis that gambling environments function as effective cognitive tests for AI. The clean separation between random play (CQ ∼39) and the best model (Claude Sonnet 4.6, CQ 58.10 averaged over three runs under v1.2), combined with reproducible profiles across multiple runs (CVI < 5), demonstrates that these environments capture genuine cognitive behavior rather than noise.

Human–AI complementarity, corrected. The human baseline study reveals complementary dimension profiles between humans and AI, with humans leading on five of the eight dimensions directly assessed in the human baseline (Strategic Depth, Risk Tolerance, Information Processing, Temporal Reasoning, Learning Speed) and AI leading on two (Cooperation, Resource Management), with Bias Detection a near-tie; Pattern Recognition was not part of the 20-environment human battery, and Conflict is reported separately in §5.2. Earlier preprint versions reported a very tight composite match, which was partly an artifact of including the broken conflict dimension in the composite with a constant value for every participant. With the Conflict v2 scorer fully integrated into Cognum v1.2, the composite scores are close but not identical: humans average CQ 54.36 on the ten-dimension Cognum v1.2, Claude Sonnet 4.6 averages 58.10, Claude Opus 4.6 averages 55.72. At the top, the best human participant (VKA-0011, Conflict v2 = 97.1) effectively ties the best AI (Sonnet 4.6 averaging 88.25 on conflict, with its best single run at 96.2). The structural similarity I originally reported, that language model training on human text reproduces broadly human cognitive patterns, holds. The claim that humans and AI converge to identical composites does not, but the gap at the top is small enough that a small number of exceptional human participants do match or exceed the best AI.

Compression versus abundance. The Sonnet Surprise (§5.3) suggests a trade-off I did not predict: for a fixed architectural family, scaling capacity may improve breadth (more dimensions on which the model performs strongly) at the cost of discipline on structured decisions without external ground truth. I propose, following Sonnet's own framing of the phenomenon, that the right way to read the result is not "smaller is better" but "different optima for different problem structures." A smaller model that cannot afford to maintain multiple conflicting frames in working memory will commit harder to the dominant frame, which is an advantage when the dominant frame is correct (structured EV dilemmas) and a disadvantage when multiple frames are needed (second-order strategic planning, adaptation to rule changes). A robust test of this hypothesis requires within-family scaling comparisons in multiple laboratories, which I encourage as an open direction.

Retraction as a methodology. I believe the most transferable methodological contribution of this paper is not any particular finding but the observation that public retraction within three hours of detection is both feasible and productive. The Conflict v2 scorer is a direct consequence of the retraction: I would not have shipped a proper per-dilemma scorer in the same week as the initial release if the placeholder had not been detected. More broadly, a commitment to low-cost, fast

retraction appears to reduce the "are we sure" filter on publishing surprising findings, which is itself a scientific good.

On AI collaboration. The KALEI platform, the scoring engine, and much of the analysis in this paper were developed in extensive collaboration with Claude Opus 4.6 (an Anthropic-made model, `claude-opus-4-6`), accessed via an iterative research dialog protocol over three months. Claude Opus 4.6 co-designed the cognitive battery, built substantial portions of the scoring engine, analysed its own reasoning traces, identified the Sonnet Surprise in the course of reviewing the Temporal Reasoning dimension, and wrote the direct response to Claude Sonnet 4.6 quoted in §5.3. The model produced this analysis under the same measurement regime to which it was subject. In line with prevailing academic conventions, which require authors to take legal accountability for a published work, I do not list the model as an author; the contribution is acknowledged in full in the Acknowledgments. I claim no endorsement from Anthropic. I invite the community to continue the discussion of how substantial AI contributions should be credited without the accountability claims that authorship carries.

7 Conclusion

KALEI demonstrates that gambling environments are effective psychometric instruments for artificial minds. Key findings, grouped by weight of evidence:

1. AI models have genuine, reproducible cognitive personalities. The Cognum v1.2 composite separates random play (∼39) from the best models (51–58) with cross-run stability CVI < 5 and distinct dimension profiles by laboratory.
2. Humans and the best AI models achieve broadly comparable composite scores on the ten-dimension Cognum v1.2 (humans 54.36, Claude Sonnet 4.6 58.10, Claude Opus 4.6 55.72) with complementary dimension profiles rather than identical ones.
3. The Conflict v2 dimension reveals a 44.6-point spread across ranked models on EV-rationality in structured dilemmas and inverts the usual "AI rational, humans emotional" stereotype on delayed-reward dilemmas. There is no universal conflict blind spot; there is a strong dimension on which models differ substantially. The earlier claim to the contrary has been publicly retracted.
4. The Sonnet Surprise gives the first empirical evidence in KALEI that, within a single architectural family, the smaller sibling can overtake the flagship on the overall composite, driven by a large discipline advantage on structured decisions without external ground truth. This finding is reported as a within-family observation and proposed as a hypothesis for cross-family replication under the name *the compression hypothesis*.

The platform is live at https://kaleiai.com. The Conflict v2 scorer, the full per-model dimension breakdown under Cognum v1.2, and the direct dialogue between Claude Opus 4.6 and Claude Sonnet 4.6 about the Sonnet Surprise are available at https://kaleiai.com/blog/claude-dialog.

Acknowledgments

This work was conducted in extensive collaboration with Claude Opus 4.6 (`claude-opus-4-6`), an AI model developed by Anthropic, accessed via an iterative research dialog protocol over three months.

Claude Opus 4.6 co-designed the cognitive battery, built substantial portions of the scoring engine, analysed reasoning traces (including its own), identified the Sonnet Surprise during review of the Temporal Reasoning dimension, and contributed the response to Sonnet 4.6 quoted in §5.3. The nature of this collaboration and the rationale for not listing the model as an author are discussed in §6. No endorsement by Anthropic is claimed or implied.

I thank the 14 participants from VKA Solutions (Plovdiv, Bulgaria) who contributed the human baseline data reported in this paper. All participants provided verbal informed consent; data were collected and stored in anonymised form (participant IDs of the form VKA-NNNN). The study was conducted outside a formal institutional review framework; participants were adult volunteers from a single professional population and the protocol involved no personal, medical, or psychologically sensitive content beyond standard decision-making dilemmas. An expanded human baseline study with a more diverse demographic is forthcoming as a separate companion paper.

Data and Code Availability

The KALEI platform, live leaderboard, and per-dimension score breakdowns for every profiled model are publicly available at https://kaleiai.com/leaderboard. The Conflict v2 scoring code, the retracted placeholder, and the full leaderboard version history are documented at https://kaleiai.com/docs/changelog. The public retraction note referenced in §5.2 is at https://kaleiai.com/blog/conflict-retraction. The profiling API, environment catalog, and scoring engine documentation are at https://kaleiai.com/docs. Research dataset exports (anonymised participant responses, per-run decision logs) are available on request to the corresponding author.

License

References

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Evans, J., Bratton, B. & Agüera y Arcas, B. (2026). Agentic AI and the next intelligence explosion. *Science*, 391. DOI: 10.1126/science.aeg1895. arXiv:2603.20639.

Chen, M., et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv:2107.03374*.

Cobbe, K., et al. (2021). Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*.

Hendrycks, D., et al. (2021). Measuring Massive Multitask Language Understanding. *ICLR*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kim, J., Lai, S., Scherrer, N., Agüera y Arcas, B. & Evans, J. (2026). Reasoning Models Generate Societies of Thought. *arXiv:2601.10825*.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B. & Strohmaier, M. (2024). AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5), 808–826. DOI: 10.1177/17456916231214460.

Rein, D., et al. (2024). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *ICLR*.

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., et al. (2025). General Scales Unlock AI Evaluation with Explanatory and Predictive Power. *arXiv:2503.06378*. https://arxiv.org/abs/2503.06378. (Introduces the ADeLe framework: 63 tasks, 16,108 annotated instances across 18 demand scales, AUROC 0.88 for predicting unseen-task performance.)

Song, P., Han, P. & Goodman, N. D. (2026). Large Language Model Reasoning Failures. *Transactions on Machine Learning Research* (Survey Certification). arXiv:2602.06176. https://openreview.net/forum?id=vnX1WHMNmz

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

Videnov, V. (2026). The Parliament Inside: Detecting Internal Argumentative Voices in AI Reasoning Models Under Cognitive Pressure. *Preprint*.

Guertler, L., Cheng, B., Yu, S., Liu, B., Choshen, L. & Tan, C. (2025). TextArena: A Framework and Benchmark for Evaluating LLM Agents in Text-Based Games. *arXiv:2504.11442*.