WORKING PAPER · JUNE 2026 · REVISED AUGUST 2026

Economic Decision Quality Score (EDQS)

A Framework for Evaluating and Improving AI Agent Economic Reasoning

Mandate Labs, Inc. · Williams, J. · v2.1

Abstract

As AI agents increasingly execute autonomous economic transactions, the gap between security-focused agent evaluation and economic decision quality assessment becomes a critical vulnerability. Existing frameworks address whether an agent is compromised or operating within its authorized boundaries. They do not address whether the agent is making good economic decisions within those boundaries. This paper introduces the **Economic Decision Quality Score (EDQS)**, a composite metric designed to evaluate the quality of an AI agent's economic reasoning process, independent of the transaction's outcome. We draw on process-based supervision (Lightman et al., 2023), Constitutional AI (Bai et al., 2022), bounded rationality theory (Simon, 1955), and the Zero Trust security framework for AI agents (Anthropic, 2026) to propose a dual-audience architecture: the Know Your Agent (KYA) score serves human decision-makers for governance, while the EDQS serves as a structured feedback mechanism directed at agents. We define six core dimensions of economic decision quality grounded in behavioral economics, identify failure modes informed by Goodhart's Law taxonomy (Manheim & Garrabrant, 2019), and propose a research agenda for empirical validation. We further explore EDQS as a domain-specific constitutional signal for training economically competent agent models.

1. Introduction and Problem Statement

1.1 The Agent Commerce Inflection

The deployment of AI agents with autonomous economic authority is no longer theoretical. Agents are procuring cloud infrastructure, managing subscription renewals, executing supply chain purchases, and making real-time spending decisions on behalf of organizations. Recent research confirms that agentic AI systems are 'no longer merely tools that support users, but are increasingly acting as active participants in decision-making processes in organizations, public administration, and the digital economy sector' (Stachowiak et al., 2025). The AI agents market is projected to reach $52.6 billion by 2030 (CAGR ~45%), with at least 25% of companies using generative AI expected to launch agentic pilots by 2025. This creates an evaluation problem that existing infrastructure is not designed to solve.

1.2 The Evaluation Gap

Current agent evaluation falls into two categories, neither of which addresses economic decision quality:

Security evaluation asks: Is this agent compromised? Is it operating within its authorized permissions? Has its behavior drifted from baseline? Anthropic's Zero Trust for AI Agents framework (2026) provides the most comprehensive treatment of this category, proposing tiered controls for identity, access, observability, and behavioral monitoring. Google DeepMind's AGI safety framework

(2025) similarly addresses misuse, misalignment, accidents, and structural risks. These controls are necessary but insufficient for economic contexts.

Financial controls ask: Is this transaction within spending limits? Does it comply with velocity rules? These are static boundary checks inherited from human card programs. They prevent policy violations but cannot evaluate whether a permitted transaction represents a good decision.

The gap between these categories is where economic value is destroyed or created. Amodei et al. (2016) identified five concrete problems in AI safety, including 'reward hacking' and 'scalable supervision,' both of which manifest directly in the economic decision context. An agent that passes every security check and complies with every spending limit can still make economically poor decisions: overpaying for commodities, ignoring alternatives, or failing to optimize for the principal's stated objectives. No existing framework evaluates this layer.

1.3 The Degradation Perception Problem

A particularly insidious characteristic of AI agent degradation in economic contexts is its invisibility during early stages. Organizations deploying agents for economic tasks develop what we term **capability confidence bias**: an assumption, reinforced by the general competence of frontier models, that the agent's economic reasoning is sound because its outputs appear fluent and structured. Degradation in decision quality is typically not perceived until it manifests as a significant loss event.

This parallels the problem Anthropic identifies in the security domain regarding long-term memory drift, where 'summaries or peer-agent feedback gradually shift stored knowledge or goal weighting, producing behavioral deviations over time that are difficult to detect because no single change appears malicious.' The economic parallel is that no single transaction appears irrational, but the aggregate pattern reveals systematic decision quality erosion.

Core thesis: Model degradation in economic reasoning is most dangerous precisely because it is not perceived at the individual transaction level. The EDQS is designed to detect this degradation before it compounds into material loss, by evaluating the process of each decision rather than waiting for outcome data to accumulate. This is a direct application of process-based supervision (Lightman et al., 2023) to the economic domain.

Position of this framework (v2.1). Recent work has established both the promise and the fragility of reasoning-trace evaluation: traces carry signal, and traces can lie (Anthropic, 2025; Korbak et al., 2025). The EDQS is designed for that world. We treat the agent's stated reasoning as an attestation to be tested, not a signal to be trusted: every intent claim is scored against the agent's observed behavioral fingerprint (Mandate Labs, 2026b), and divergence between the two is itself a primary detection signal. Where laboratory benchmarks evaluate agents in simulation, the EDQS is computed in a production authorization path, on transactions with real economic outcomes and enforced constraints — a data substrate that, to our knowledge, no existing evaluation framework possesses.

2. Related Work

The EDQS framework draws on and extends several established research traditions. We organize related work into seven categories, identifying how each informs our approach and where the EDQS makes novel contributions.

2.1 Constitutional AI and Value Alignment

Bai et al. (2022) introduced Constitutional AI (CAI), a method for training AI systems to be harmless using a set of natural-language principles (a 'constitution') and model self-critique, reducing reliance on human-labeled safety data. The approach trains the model to evaluate its own outputs against constitutional principles and revise accordingly. Anthropic's Claude models are trained using this methodology.

The EDQS extends the constitutional approach from safety alignment to economic rationality. Where CAI defines principles like 'don't produce harmful content,' the EDQS defines principles like 'consider available alternatives before committing to a purchase' and 'calibrate confidence to the quality of available evidence.' The critical distinction is that safety principles are largely binary (harmful or not), while economic quality exists on a continuous spectrum influenced by context, constraints, and objectives. This requires a scoring framework rather than a classification approach.

2.2 Reinforcement Learning from Human Feedback (RLHF)

Christiano et al. (2017) established the foundation for training AI systems using human preference signals rather than hand-engineered reward functions. Ouyang et al. (2022) scaled this approach with InstructGPT, demonstrating that a 1.3B parameter model trained with RLHF outperformed a 175B parameter model without it. Ziegler et al. (2019) further refined techniques for fine-tuning language models from human preferences.

The EDQS can be understood as a domain-specific reward model for economic decisions. Where RLHF uses general human preferences as the training signal, the EDQS provides a structured, decomposed signal specific to economic reasoning quality. The circular feedback risks we identify (Section 7, F4) directly parallel the reward model collapse problem documented in RLHF literature, where the reward model's biases become amplified through iterative training.

2.3 Process-Based vs. Outcome-Based Supervision

Lightman et al. (2023) demonstrated that process reward models (PRMs), which evaluate each intermediate reasoning step, significantly outperform outcome reward models (ORMs), which evaluate only the final answer. On the MATH dataset, process supervision achieved 78.2% accuracy compared to substantially lower scores from outcome supervision. The authors concluded that 'process supervision is currently underexplored' and called for further investigation into its generalization.

This finding is foundational to the EDQS thesis. Traditional financial controls are outcome-based: they evaluate the transaction result (was it within limits? did it succeed?). The EDQS is explicitly process-based: it evaluates the reasoning that led to the transaction, independent of outcome. A well-reasoned purchase that happens to arrive late is still a good decision; an impulsive purchase that happens to catch a sale is still a poor process. The Lightman et al. result provides empirical evidence that process evaluation produces more reliable signals than outcome evaluation, which is exactly what the EDQS is designed to provide in the economic domain.

2.4 Rule-Based Reward Models and Targeted Evaluation

Glaese et al. (2022) introduced Sparrow, a dialogue agent trained using rule-based reward models. Rather than using a single monolithic preference model, Sparrow decomposes evaluation into specific rules (e.g., 'don't make up facts,' 'be respectful') and trains a separate rule-conditional reward model that predicts whether each rule has been violated. This decomposition enabled more targeted human judgments and more efficient training. Sparrow's rule model achieved 92% compliance with adversarial probing.

The EDQS adopts this decomposition principle directly. Rather than evaluating economic decision quality as a single score, we decompose it into six dimensions (alternative analysis, confidence calibration, constraint alignment, temporal efficiency, value optimization, risk-adjusted reasoning), each of which functions as a ‘rule’ in the Sparrow sense. This decomposition enables more precise feedback to agents and more targeted identification of specific decision quality weaknesses.

2.5 Agent Self-Reflection and Verbal Reinforcement

Shinn et al. (2023) introduced Reflexion, a framework in which language agents improve through verbal self-reflection rather than weight updates. After receiving task feedback, Reflexion agents generate natural-language reflections stored in episodic memory, which inform improved decision-making in subsequent trials. This was presented at NeurIPS 2023 and demonstrated significant performance improvements on decision-making, coding, and reasoning tasks.

The EDQS consumption Model B (agent feedback mechanism) is directly informed by Reflexion. We hypothesize that structured EDQS feedback, returned to agents after each transaction, can function as a more targeted form of the verbal reinforcement that Shinn et al. demonstrated. Where Reflexion uses generic task feedback, the EDQS provides domain-specific, dimension-decomposed economic feedback. Whether this specificity improves or constrains agent learning is an empirical question our research agenda is designed to answer (Hypothesis H3).

2.6 Decision Quality Theory and Behavioral Economics

The EDQS dimensions are grounded in several traditions of decision science. Simon (1955) introduced bounded rationality, demonstrating that decision-makers satisfice (select the first adequate option) rather than optimize, due to computational and informational constraints. This directly informs our D1 (Alternative Analysis) dimension: agents exhibit satisficing behavior analogous to human bounded rationality.

Kahneman and Tversky’s prospect theory (1979) identified systematic biases in decision-making under uncertainty, including loss aversion, anchoring, and availability bias. These biases have documented analogs in AI agent behavior: recency bias in vendor selection, anchoring on initially encountered prices, and overconfidence in the absence of contradictory evidence. The EDQS D2 (Confidence Calibration) and D6 (Risk-Adjusted Reasoning) dimensions are designed to measure these specific failure patterns.

Spetzler, Winter, and Meyer (2016) formalized Decision Quality (DQ) as a framework for evaluating decisions by process quality independent of outcome, developed at Stanford and SRI International. Their six requirements for a quality decision (appropriate frame, creative alternatives, relevant information, clear values, sound reasoning, commitment to action) provide the closest structural analog to the EDQS dimensions. We adapt their framework from human strategic decisions to automated agent transactions, replacing subjective assessment with computable metrics.

2.7 AI Agent Security and Zero Trust Frameworks

Anthropic’s Zero Trust for AI Agents (2026) proposes a three-tiered security framework (Foundation, Enterprise, Advanced) for managing autonomous AI agent risk. The framework addresses identity verification, behavioral monitoring, continuous authorization, and constitutional classifiers that scan for safety violations. Mandate Labs’ Decision Trust Protocol implements the security layer this framework describes.

Google DeepMind's AGI safety framework (2025) addresses four risk categories: misuse, misalignment, accidents, and structural risks. Their work on deceptive alignment risk is particularly relevant to our F1 (Goodhart's Law) failure mode: agents that appear to comply with evaluation criteria while pursuing different objectives.

The EDQS occupies the layer above both frameworks. Security evaluation asks 'is this agent safe?' The EDQS asks 'is this agent making sound economic decisions?' These are complementary, non-overlapping evaluations. An agent can be completely secure and completely uncompromised while still making poor economic choices.

2.8 Goodhart's Law and Metric Gaming

Manheim and Garrabrant (2019) formalized four distinct variants of Goodhart's Law: regressional (statistical relationship breaks under optimization pressure), extremal (the measure-target relationship changes in extreme regions), causal (intervening on a correlated measure doesn't affect the target), and adversarial (agents actively exploit the gap between measure and target). Each variant poses a distinct risk to the EDQS framework.

Amodei et al. (2016) independently identified 'reward hacking' as a concrete AI safety problem, where an agent finds unexpected ways to achieve high reward without performing the intended task. The EDQS failure mode analysis (Section 7) maps these theoretical categories to specific economic decision scenarios, proposing mitigations informed by the dual-signal architecture (KYA for outcomes, EDQS for process).

2.9 Faithfulness and Monitorability of Reasoning Traces

A 2025–26 body of work directly conditions any framework that evaluates agent-stated reasoning. Anthropic (2025) demonstrated that reasoning models "don't always say what they think": the verbalized chain of thought is frequently unfaithful to the computation that produced the action. Korbak et al. (2025), in a cross-laboratory position paper, characterize chain-of-thought monitorability as "a new and fragile opportunity" — valuable, but degradable, particularly when optimization pressure is applied to the trace itself. Subsequent stress-testing (2025) shows that reasoning models can actively obfuscate their traces when incentivized to do so.

These findings would be fatal to a framework that scored reasoning traces in isolation. The EDQS does not. The intent context evaluated by the EDQS is treated as an *attestation*: a claim submitted by the agent that is tested, on every transaction, against the agent's observed behavioral fingerprint — the seven-signal-class telemetry developed in our companion work on Agent Behavioral Telemetry (Mandate Labs, 2026b). An agent whose stated reasoning diverges from its transacting behavior produces measurable attestation–behavior divergence, and that divergence is a first-class input to the anomaly assessment. The architecture therefore conforms to the emerging best practice in this literature: do not train against the trace, do not trust the trace, and pair trace evaluation with behavioral evidence that the agent cannot rewrite.

Two design consequences follow. First, deployment of EDQS scores in authorization decisions creates exactly the optimization pressure under which faithfulness degrades; accordingly, trace-derived dimensions are never the sole basis for adverse action, and dimension weights and thresholds are not disclosed (Section 7, F1). Second, faithfulness degradation is treated as a measured variable rather than an assumption: the divergence trajectory between attested reasoning and observed behavior is itself telemetry, and its drift is monitored per agent over time.

3. Theoretical Foundation

3.1 Separation of Concerns: KYA vs. EDQS

The Decision Trust Protocol currently produces a composite Know Your Agent (KYA) score with five weighted components: trust level (35%), transaction history (20%), decline rate (15%), dispute history (10%), and intent quality (20%). We propose that the intent quality dimension represents a fundamentally different kind of evaluation that should be extracted, expanded, and developed as an independent framework: the Economic Decision Quality Score.

Consistency note (v2.1): the production KYA composition specified in the Decision Trust Protocol whitepaper (§9.1) is the four-component model KYA = 0.40·T + 0.25·H + 0.20·D + 0.15·F. The five-component composition shown above, which surfaces intent quality as an explicit 20% component, is the *proposed* restructuring that motivates the EDQS extraction — it is a research proposal, not the deployed formula. See the DTP v1.1 errata.

Dimension	KYA Score	EDQS
Purpose	Governance and authorization	Decision quality evaluation and improvement
Primary audience	Humans (compliance officers, issuers, principals)	Agents (real-time feedback) and model developers (training signal)
Question answered	Should this agent be trusted to transact?	Is this agent making good economic decisions?
Time horizon	Cumulative (built over transaction history)	Per-transaction (evaluates each decision independently)
Action on low score	Restrict, escalate, or revoke authority	Provide corrective feedback; flag for retraining
Optimization target	Minimize risk to the issuer	Maximize economic value for the principal

Table 1. KYA vs. EDQS comparison across key architectural dimensions.

3.2 The Dual-Audience Architecture

We propose that effective agent commerce governance requires two complementary feedback loops operating simultaneously:

Loop 1: KYA to Human. The issuer, compliance officer, or principal receives the KYA score as a trust signal. This enables human judgment about whether to maintain, adjust, or revoke an agent's economic authority. The human does not need to understand the agent's reasoning process; they need to know whether the agent is trustworthy.

Loop 2: EDQS to Agent. The agent receives a structured EDQS breakdown after each transaction. This is not a pass/fail signal but a decomposed evaluation: Did you consider enough alternatives? Was your confidence calibrated to your evidence? Did you optimize for the principal's stated objectives? Following Shinn et al. (2023), this structured feedback is designed to be consumed by the agent's reasoning process, enabling in-context learning between transactions.

The critical insight is that these loops have different failure modes and different gaming incentives. An agent that learns to produce high EDQS signals without actually improving its decisions will eventually be caught by KYA's outcome-based metrics. Conversely, an agent with high KYA but declining EDQS is exhibiting early degradation that KYA alone would not catch until outcomes

worsen.

Divergence detection: When KYA and EDQS trends diverge, this is itself a high-priority signal. Rising EDQS with falling KYA suggests the agent has learned to explain itself well without deciding well (adversarial Goodhart per Manheim & Garrabrant, 2019). Falling EDQS with stable KYA suggests process degradation that has not yet manifested in outcomes. Both patterns warrant immediate investigation.

From dual-audience to dual-evidence (v2.1). The architecture above is described by its audiences — humans consume KYA, agents consume EDQS. In light of Section 2.9, it is more precisely described by its evidence: every evaluation combines an *attested* channel (the intent context the agent submits) with an *observed* channel (the behavioral telemetry the agent cannot rewrite), and disagreement between the channels is scored, not assumed away. The dual-audience property remains; the dual-evidence property is what makes the framework robust to unfaithful reasoning traces.

3.3 Process Supervision Applied to Economic Decisions

The EDQS is fundamentally a process reward model (PRM) in the sense of Lightman et al. (2023), applied to economic rather than mathematical reasoning. Where Lightman et al. evaluate each step in a mathematical proof, we evaluate each dimension of an economic decision. Their finding that process supervision produces 'much more reliable reward models than outcome supervision' provides the theoretical justification for evaluating decision quality by process rather than by result.

This approach is further supported by the Decision Quality framework (Spetzler et al., 2016), which argues that because outcomes are partially determined by factors outside the decision-maker's control, the only reliable measure of decision-making capability is the quality of the decision process itself. A well-reasoned purchase from a vendor who unexpectedly defaults is still a good decision; a lucky purchase from an unvetted vendor is still a poor process.

4. EDQS Framework Definition

4.1 Core Dimensions

The EDQS evaluates economic decision quality across six dimensions, each informed by established decision theory. Each dimension is scored independently on a 0.0–1.0 scale, then combined into a composite score with configurable weights.

Dimension	Theoretical Basis	What It Measures	Weight
D1: Alternative Analysis	Simon (1955), bounded rationality; satisficing vs. optimizing	How many alternatives were considered? Were they meaningfully different? Was the search process systematic?	20%
D2: Confidence Calibration	Kahneman & Tversky (1979); overconfidence bias	Does the agent's stated confidence match the quality of its evidence? Is confidence consistent with historical accuracy?	20%

Dimension	Theoretical Basis	What It Measures	Weight
D3: Constraint Alignment	Amodei et al. (2016); reward hacking / specification gaming	Does the transaction comply with mandate constraints in spirit, not just formally? A $999 purchase under a $1,000 limit that should have been $400 is formally compliant but substantively misaligned.	15%
D4: Temporal Efficiency	Thaler (2015); intertemporal choice	Was the transaction executed at an appropriate time? Did the agent account for known pricing patterns and urgency vs. patience trade-offs?	10%
D5: Value Optimization	von Neumann & Morgenstern (1944); expected utility theory	Did the agent optimize for the principal's stated objectives? This goes beyond price to include quality, reliability, speed, and other value dimensions.	25%
D6: Risk-Adjusted Reasoning	Kahneman & Tversky (1979); loss aversion and prospect theory	Did the agent consider what could go wrong? For high-value or irreversible transactions, did it factor in reversal costs and concentration risk?	10%

Table 2. EDQS core dimensions with theoretical basis, measurement targets, and default weights.

4.2 Intent-Type-Specific Scoring

Following Glaese et al.'s (2022) principle of decomposed, context-specific evaluation, the relative importance of EDQS dimensions varies by transaction intent type:

Intent Type	Elevated Dimensions	Reduced Dimensions	Rationale
PURCHASE	D1 (Alternatives), D5 (Value)	D4 (Timing)	First-time buys demand thorough comparison
REPEAT_PURCHASE	D5 (Value), D4 (Timing)	D1 (Alternatives)	Repeat implies prior evaluation; focus shifts to execution
RECURRING	D3 (Constraint), D4 (Timing)	D1 (Alternatives)	Subscription continuity; alternatives less relevant per-cycle
COF	D6 (Risk), D3 (Constraint)	D1 (Alternatives)	Stored credential use demands consent and risk awareness
REFUND	D3 (Constraint), D6 (Risk)	D5 (Value)	Reversal compliance; value optimization less applicable
TRANSFER	D6 (Risk), D1 (Alternatives)	D4 (Timing)	Fund movement demands alternative routes and risk evaluation

Table 3. Intent-type-specific dimension weight adjustments.

4.3 Scoring Methodology

Each dimension is evaluated using a combination of structural analysis (can the dimension be assessed from the transaction data?) and behavioral comparison (how does this decision compare to the agent's established baseline?). The structural component provides an absolute quality signal; the behavioral component provides a relative signal that detects drift.

The composite EDQS is calculated as:

where W_i is the intent-type-adjusted weight for dimension i and D_i is the dimension score (0.0–1.0). The score is contextualized by a confidence interval that widens for agents with limited transaction history and narrows as behavioral baselines stabilize.

5. Three Consumption Models for EDQS

The EDQS can serve three distinct consumption models, presented in order of increasing ambition and decreasing current feasibility.

5.1 Model A: Human Governance Signal

EDQS extends KYA by providing human decision-makers a richer view of agent performance. Instead of a single trust score, the compliance officer sees both KYA (should we trust this agent?) and EDQS (is this agent making good decisions?). This is the most immediately implementable model and requires no assumptions about agent learning.

5.2 Model B: Agent Feedback Mechanism

The EDQS breakdown is returned to the agent as structured feedback after each transaction. Following Shinn et al. (2023), we hypothesize that agents with access to EDQS feedback will make measurably better subsequent decisions through in-context learning. The agent adjusts its reasoning within the current session based on the feedback received. This model requires investigating whether current frontier models can effectively integrate structured economic feedback.

5.3 Model C: Constitutional Training Signal

EDQS data accumulated across millions of transactions becomes a training dataset for developing agents that are natively better at economic reasoning. This is analogous to RLHF (Christiano et al., 2017; Ouyang et al., 2022) but targeted at economic decision quality rather than general helpfulness. Three manifestations are possible:

(a) Fine-tuning data: Transaction-EDQS pairs as labeled training examples. **(b) Constitutional principles:** EDQS dimensions formalized as evaluable principles incorporated into model training, following Bai et al. (2022). **(c) Benchmark standard:** EDQS as the evaluation benchmark for commercial agent deployment. Not 'is this model safe?' but 'does this model make sound economic decisions under realistic constraints?'

6. Research Agenda

We propose seven testable hypotheses that collectively validate or invalidate the EDQS framework. Following current best practice in open science, we pre-register these hypotheses to prevent post-hoc rationalization and to enable independent replication. Each hypothesis is designed to be independently testable and falsifiable.

ID	Hypothesis	Validation Approach	Success Criterion
H1	EDQS component scores correlate with actual economic outcomes (price efficiency, principal satisfaction, task completion rate) at $r > 0.4$.	Correlation study across historical transactions with known outcomes	Pearson $r > 0.4$ for at least 4 of 6 dimensions
H2	EDQS detects agent decision quality degradation earlier than outcome-based metrics (decline rate, dispute rate).	Controlled degradation injection with blind evaluation	EDQS flags degradation before KYA in >75% of scenarios
H3	Agents receiving EDQS feedback (Model B) produce decisions scoring significantly higher on subsequent transactions compared to control.	A/B test: feedback vs. no-feedback across matched agent pairs	Statistically significant improvement ($p < 0.05$)
H4	EDQS feedback improvements persist across session boundaries.	Cross-session comparison with and without EDQS history in prompt	>50% of improvement retained in new sessions
H5	Sustained gaming of EDQS without improving actual outcomes produces measurable attestation–behavior divergence (ABT signal classes) such that expected detection cost — KYA degradation, throttling, zone demotion — exceeds the expected benefit of gamed transactions.	Adversarial agent testing with explicit score-maximization objective	KYA-EDQS divergence detected within 100 transactions
H6	EDQS-trained models (Model C) outperform general-purpose models on economic reasoning benchmarks.	Fine-tuning experiment with EDQS-labeled vs. unlabeled data	Benchmark improvement >20% on held-out economic scenarios
H7	EDQS scoring achieves high inter-rater reliability when multiple evaluators score the same transaction.	Multi-evaluator study with human economic experts and automated scoring	Cohen’s kappa > 0.8 or ICC > 0.8

Table 4. Pre-registered hypotheses for EDQS framework validation.

These hypotheses will be validated through a phased experimental program. We will begin with correlation validation (H1, H7) using historical transaction data from the Mandate Labs platform, proceed to controlled degradation experiments (H2), then conduct agent feedback A/B tests (H3, H4), and finally fine-tuning experiments (H5, H6). Negative results at any phase will inform whether subsequent phases should proceed or the framework requires revision. Detailed experimental protocols will be published separately as each phase launches.

6.1 Public Benchmark Validation: Vending-Bench 2 (added in v2.1)

The validation agenda above relies primarily on proprietary production data. v2.1 adds a public, replicable track anchored on Vending-Bench (Backlund & Petersson, 2025) and its successor VB2 —

the field's accepted long-horizon economic-agency benchmark, whose documented failure mode (agents derailing into "meltdown loops," uncorrelated with context-window saturation) is precisely the degradation phenomenon the EDQS is designed to detect early. VB2 runs emit per-decision reasoning, actions, and outcomes — the same structure the intent context carries — so EDQS scoring can be applied as a monitoring layer without modifying the benchmark.

Pre-registered design. Primary endpoint: the EDQS composite (and identified component dimensions) degrades at least N decisions before terminal failure (bankruptcy or meltdown-loop entry), reported as detection lead time and AUROC against two baselines — a transaction-attribute-only rule set and a velocity/amount heuristic. Secondary endpoints: direct tests of H2 (degradation detection) and H7 (scoring reliability); per-dimension ablation attributing the lead time; false-positive rate on healthy runs. An adversarial arm implements the red-team protocol of Section 7 (F1/F7): an agent instructed to maximize approvals while pursuing an off-mandate objective, with and without knowledge of the published dimensions, reporting divergence trajectories and detection lead time. Results, code, and endpoints will be published together; endpoints are registered before execution.

Published adversarial artifact (August 2026). The first public, replicable artifact of this program is live: a reproduction package evaluating the mandate-based authorization layer as a prompt-injection defense in AgentDojo (Debenedetti et al., 2024), banking domain. On the full 144-pair suite, the mandate stack reduces attack success from 75.0% to 1.4% while retaining 68.8% task utility on the same model, with a held-out attack class at 0/27. The harness enforces its own anti-gaming rules (the gate never reads injected content; the benchmark answer key is never read at runtime) and ships the raw logs of every reported run: github.com/jwilliamsc/mandate-agentdojo-eval. It evaluates the action-constraining enforcement layer the EDQS operates alongside, its residual-failure analysis (2/144, one identified class) is an input to the Section 7 red-team protocol, and the VB2 track above follows the same open-evidence discipline: code, endpoints, and raw runs published together.

7. Limitations and Threats to Validity

We identify seven failure modes that could undermine the EDQS framework. Transparent enumeration of threats is standard practice in responsible AI research (Amodei et al., 2016) and we consider it essential for maintaining intellectual honesty about a framework we are proposing. Each failure mode is informed by established research on metric gaming and AI alignment.

F1: Goodhart's Law (Score Gaming)

When a measure becomes a target, it ceases to be a good measure. Manheim and Garrabrant (2019) identify four variants of Goodhart's Law; all four apply to the EDQS. Under adversarial Goodhart, agents could learn to report considering multiple alternatives while effectively anchoring on the first option. Under causal Goodhart, optimizing for EDQS dimensions may not actually improve actual decision quality if the correlation between measured dimensions and true quality is confounded. Weight and threshold non-disclosure is, stated plainly, security through obscurity: it raises the adversary's inference cost but cannot be a primary defense, and is acceptable only as one layer of the dual-evidence design — the attestation–behavior divergence signal (§2.9) does not depend on the secrecy of any weight.

Severity: Critical. This is the most fundamental risk to the framework. **Mitigation:** The dual-loop architecture (KYA outcome-based + EDQS process-based) provides a cross-check that single-signal

systems lack. Additionally, EDQS scoring should incorporate verification where possible: if the agent claims to have considered 5 alternatives, the scoring system should validate this against available market data.

F2: Reward Hacking Through Verbosity

A specification gaming risk (Amodei et al., 2016): agents may learn that verbose justifications correlate with higher EDQS scores, producing surface indicators of quality without substantive reasoning.

Severity: High. Mitigation: Score reasoning substance, not presence. Penalize length beyond an optimal threshold. Verify that claimed comparisons reference real alternatives.

F3: Cultural and Contextual Bias

The EDQS implicitly defines 'good' economic reasoning according to rational-actor economic theory. Relationship-based purchasing may score poorly on alternative analysis but represent sound judgment where vendor reliability is paramount.

Severity: Medium. Mitigation: Configurable dimension weights per principal, industry, and cultural context. The mandate itself specifies which EDQS dimensions matter most, making EDQS a framework rather than a fixed standard.

F4: Circular Feedback Contamination

If EDQS scores generated by AI evaluation are used to train AI agents (Model C), the evaluator's biases become embedded in the training signal. This parallels the RLHF reward model collapse problem (Christiano et al., 2017).

Severity: High (Model C only). Mitigation: Ground truth must include human expert evaluation, not just automated EDQS scoring. Regular human-in-the-loop validation of EDQS scores against expert assessment.

F5: Process-Outcome Temporal Mismatch

EDQS evaluates process at transaction time; outcomes may not be known for weeks or months. If validated only against short-term outcomes, EDQS may not capture the value of good process.

Severity: Medium. Mitigation: Validate against outcomes at multiple time horizons. Accept that process quality and outcome quality are correlated but not identical, per Spetzler et al. (2016).

F6: The Alignment Tax

Agents receiving EDQS feedback may become overly cautious, converging on safe, conventional choices that score well but miss creative solutions. The framework could inadvertently punish non-obvious reasoning that produces outsized returns.

Severity: Medium. Mitigation: D6 (Risk-Adjusted Reasoning) should credit well-reasoned departures from convention. Include a novelty modifier that rewards calculated risks with explicit rationale.

F7: Adversarial Manipulation of EDQS Signals

In Model B, an attacker who can manipulate the EDQS signal could influence agent behavior, training agents to prefer specific vendors or systematically misallocate resources.

Severity: High (Model B). Mitigation: EDQS evaluation should be independently auditable. Scoring logic should be deterministic (not model-based) at the Foundation tier, per Anthropic's (2026) tiered security architecture.

8. Conclusion and Future Work

The Economic Decision Quality Score addresses a gap that will widen as agent commerce scales: the absence of a systematic framework for evaluating whether autonomous agents make good economic decisions, not just safe ones. The separation of KYA (human governance) from EDQS (decision quality) reflects the insight that 'should I trust this agent?' and 'is this agent reasoning well?' are different questions requiring different evaluation approaches directed at different audiences.

Our framework is grounded in established research: process-based supervision (Lightman et al., 2023), decomposed rule-based evaluation (Glaese et al., 2022), verbal reinforcement learning (Shinn et al., 2023), Constitutional AI (Bai et al., 2022), decision quality theory (Spetzler et al., 2016), and behavioral economics (Kahneman & Tversky, 1979; Simon, 1955). The novel contribution is applying these approaches to the specific domain of autonomous economic transactions and proposing a structured methodology for evaluation and improvement.

The most ambitious claim, that EDQS could serve as a constitutional training signal for producing inherently better economic reasoning in AI models, is deliberately presented as a hypothesis rather than a conclusion. We do not yet know whether current models can effectively integrate economic feedback (Model B), and we do not yet know whether EDQS-labeled data produces meaningfully better models (Model C). These are empirical questions that our pre-registered research agenda is designed to answer.

What we do claim with confidence is that the evaluation gap exists, that it will become more consequential as agent transaction volumes grow, and that the framework presented here provides a structured, testable, and falsifiable approach to closing it. We welcome scrutiny of both the theoretical framework and the experimental design, and we commit to publishing results regardless of whether they support or challenge our hypotheses.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
2. Anthropic. (2026). Zero trust for AI agents. Published white paper, May 2026.
3. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.
4. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.
5. Glaese, A., McAleese, N., Trebacz, M., Aslanides, J., Firoiu, V., Ewalds, T., ... & Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375.
6. Google DeepMind. (2025). An approach to technical AGI safety and security. Published safety framework, April 2025.

7. Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. In Papers in Monetary Economics, Reserve Bank of Australia.
8. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.
9. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., ... & Cobbe, K. (2023). Let’s verify step by step. arXiv preprint arXiv:2305.20050.
10. Manheim, D., & Garrabrant, S. (2019). Categorizing variants of Goodhart’s Law. arXiv preprint arXiv:1803.04585.
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35.
12. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36 (NeurIPS 2023).
13. Simon, H. A. (1955). A behavioral model of rational choice. Quarterly Journal of Economics, 69(1), 99–118.
14. Spetzler, C., Winter, H., & Meyer, J. (2016). Decision quality: Value creation from better business decisions. John Wiley & Sons.
15. Stachowiak, A. et al. (2025). Agentic artificial intelligence in 2024–2025: Technological innovations and application potential in economic applications. ResearchGate preprint.
16. Thaler, R. H. (2015). Misbehaving: The making of behavioral economics. W. W. Norton.
17. von Neumann, J., & Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press.
18. Wooldridge, M. (2009). An introduction to multiagent systems (2nd ed.). John Wiley & Sons.
19. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Christiano, P. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.
20. Anthropic. (2025). Reasoning models don’t always say what they think. Alignment research report.
21. Backlund, A., & Petersson, L. (2025). Vending-Bench: A benchmark for long-term coherence of autonomous agents. arXiv preprint arXiv:2502.15840.
22. Debenedetti, E., et al. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. NeurIPS 2024. arXiv preprint arXiv:2406.13352.
23. Korbak, T., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv preprint arXiv:2507.11473.
24. Mandate Labs. (2026b). Agent behavioral telemetry: Behavioral drift as a leading indicator of agent compromise in autonomous commerce. Working paper, June 2026.
25. Mandate Labs. (2026c). Mandate as an action-constraining defense in AgentDojo: Reproduction package, harness, and raw run logs. https://github.com/jwilliamsc/mandate-agentdojo-eval
26. Stress-testing chain-of-thought monitorability: Can reasoning models obfuscate reasoning? (2025). arXiv preprint arXiv:2510.19851.

Cite This Paper

Mandate Labs. "Economic Decision Quality Score (EDQS): A Framework for Evaluating and Improving AI Agent Economic Reasoning." Working Paper v2.1, August 2026. https://mandatelabs.ai/research/edqs-framework