

# Mars<sup>®</sup> Operating System: Structured Knowledge for High-Stakes Decision Intelligence

Causum, Inc.

## Abstract

Current AI decision-support systems face a fundamental tradeoff: expert systems provided structure but proved too brittle to deploy; large language models provide flexibility but lack the determinism required for high-stakes domains. This gap has persisted for three decades. Mars<sup>®</sup> Operating System bridges this divide through ontologically-structured knowledge graphs that enable runtime instantiation of composable worldviews. By reducing decision entropy through explicit symbolic representations while maintaining adaptability through human-in-the-loop knowledge capture, Mars<sup>®</sup> OS achieves what neither paradigm could alone.

We validate this approach on ValuePrism, the only publicly available proxy for Key Decision-Maker Attributes evaluation. Mars<sup>®</sup> OS demonstrates 82% mean accuracy across 900 independent evaluations—4.2 $\times$  better than zero-shot language models and 2.2 $\times$  better than retrieval-augmented generation. This performance gap reflects information-theoretic advantages: structured lookup within bounded exploration spaces versus high-entropy pattern matching across unbounded search. For organizations requiring explainable, consistent, and verifiable decision support in domains where expertise disagrees and consequences matter, Mars<sup>®</sup> OS provides the missing infrastructure.

## 1 The Thirty-Year Gap

### 1.1 Why Expert Systems Failed

The 1980s vision of expert systems promised encoded human expertise deployed at scale. MYCIN diagnosed bacterial infections; XCON configured VAX computers; DENDRAL identified molecular structures. These systems worked within their intended boundaries. The problem was never technical correctness, it was operational brittleness.

The knowledge change bottleneck proved to be insurmountable. Domain experts think in patterns, heuristics, and contextual judgments. Translating this into rigid rules required engineering as an intermediary. Each new case, each edge condition, each evolving domain standard demanded manual rule updates that propagated throughout the entire system. By the time a change took effect, it became obsolete by new standards. Maintenance costs exceeded development costs. Organizations were never able to catch up and abandoned the paradigm.

But the failure was not about symbolic AI itself. It was about *how* knowledge was structured as a monolithic, hand-coded, and inflexible strategy. When the world changed, the systems could not adapt.

## 1.2 What Large Language Models Changed

LLMs solved the flexibility problem by abandoning structure entirely. Training on massive corpora captures implicit patterns without explicit encoding. Ask an LLM just about anything and it responds fluently. No knowledge engineers required. No brittle rule bases. The system generalizes because it learned statistical regularities, not programmed logic.

This works brilliantly for many applications. But high-stakes decision domains expose the fundamental tradeoff. Statistical pattern matching cannot guarantee consistency. It cannot explain its reasoning in terms humans can verify. It cannot bound its outputs to validated knowledge. When experts disagree on the correct approach—the defining characteristic of difficult decisions—LLMs average across training distributions, producing plausible-sounding outputs that may align with no actual expert’s reasoning.

Recent research confirms these limitations persist despite sophistication. Stavarache (2025) showed that even 160,000 fine-tuning examples in the Banking BIAN domain yielded only 50-65% accuracy on domain-specific correctness [1]. Lampi et al. (2024) achieved just 76% accuracy in medical triage decision prediction despite rule-based KDMA models [2]. The problem is not insufficient training data—it is the wrong paradigm for deterministic reasoning.

## 1.3 The State of the Art

Until now, there was no commercially viable solution that bridged structured knowledge and adaptive flexibility. Organizations needing high-stakes decision support face an impossible choice:

- **Build custom expert systems:** Get structure and determinism, but face multi-year development cycles and maintenance nightmares
- **Deploy LLM solutions:** Get flexibility and rapid deployment, but sacrifice explainability and consistency guarantees
- **Maintain manual processes:** Keep human expertise authoritative, but lose scalability and decision velocity

This gap is not an engineering detail. It represents a fundamental architectural absence in the AI stack. Current approaches optimize for either flexibility or structure, never both.

# 2 Information Theory of Decision Entropy

## 2.1 The Lookup vs. Guessing Problem

Consider the information-theoretic difference between these two approaches:

**Retrieval-Augmented Generation:** Given a query, compute semantic embeddings, retrieve top- $k$  similar passages from a vector database, concatenate retrieved context with the query, and prompt an LLM to synthesize an answer. The entropy of this process is high: embeddings collapse high-dimensional meaning into fixed vectors, losing nuance; retrieval similarity is computed in latent space with no guarantees about conceptual relevance; the LLM’s generation process samples from a probability distribution over token sequences.

At each step, the system *guesses*. It guesses which passages are relevant based on embedding similarity. It guesses how to combine retrieved information based on statistical patterns from training. Small perturbations in query phrasing can yield different retrievals, different contexts, different outputs.

**Structured Knowledge Graphs:** Given a query, traverse explicitly defined relationships in an ontology to identify relevant knowledge nodes. Each edge represents a verified connection. Each node contains human-validated information. The system follows deterministic paths through a bounded space of possibilities.

This is *lookup*, not guessing. The system retrieves exactly what human experts encoded. If the knowledge structure lacks a connection, the system reports uncertainty rather than hallucinating plausibility. Consistency is guaranteed by graph structure: the same query always traverses the same paths.

## 2.2 Bounded Exploration

The key insight is that high-stakes decisions do not require unbounded exploration. Domain experts already know the relevant factors, the key tradeoffs, the critical edge cases. The challenge is capturing this knowledge in a form that is both precise and adaptable.

Mars<sup>®</sup> OS implements bounded exploration through composable ontologies. Rather than encoding every possible scenario as explicit rules (the expert system failure mode), we structure the *decision space itself*: what factors matter, how they relate, what boundaries govern their interaction. Runtime instantiation then explores this space guided by the specific context.

This combines the best of both paradigms:

- Structure that enables deterministic traversal and explainability
- Flexibility through compositional knowledge modules that adapt to context
- Bounded search spaces that prevent hallucination while allowing novel combinations

The expert systems lesson: structure is essential. The LLM lesson: flexibility is essential. Our synthesis: structure the *space of possibilities*, then explore within it.

## 3 Architecture: Ontological Knowledge Composition

### 3.1 Knowledge as Runtime Substrate

Mars<sup>®</sup> OS inverts the typical AI architecture. Rather than treating knowledge as training data or retrieval context, we treat it as the *runtime substrate*. The knowledge graph *is* the execution environment.

Decision intelligence emerges from traversal through structured representations. A decision query does not retrieve passages or prompt a model. It instantiates a worldview, activates relevant ontological regions, and explores the constrained space of valid reasoning paths. Together, decision intelligence enables autonomy and explainability in AI agents.

This requires three architectural components:

**1. Ontological Knowledge Graphs:** Domain knowledge represented as typed entities and relationships (internally called Edicts). Unlike generic knowledge graphs, these are purpose-built for decision domains: entities represent decision-relevant factors; edges encode validated relationships; properties capture boundary conditions and applicability constraints.

**2. Composable Worldviews:** Decision-makers operate from frameworks: utilitarian vs. deontological ethics, risk-averse vs. risk-seeking preferences, short-term vs. long-term optimization. Mars<sup>®</sup> OS represents these as composable modules (internally called Pragmas) that structure how the knowledge graph is traversed. A worldview specifies which factors take precedence, how conflicts are resolved, and what constitutes acceptable bounds.

**3. Runtime Instantiation:** Given a decision context and a worldview, the system instantiates a specific reasoning instance. This combines the general knowledge structure with context-specific constraints to produce a bounded exploration space. The system then deterministically traverses this space, identifying relevant factors and valid reasoning paths.

### 3.2 Human-in-the-Loop Knowledge Capture

The bottleneck of the expert system was knowledge acquisition. Mars<sup>®</sup> OS addresses this through interfaces designed for domain experts, not knowledge engineers.

Rather than requiring experts to formalize if-then rules, we provide tools for:

- **Relationship definition:** "When X applies, consider Y" captured as graph edges
- **Boundary specification:** "Factor Z matters only when conditions A and B hold" encoded as constraints
- **Worldview templates:** "In this domain, prioritize factors P, Q, R" defined as traversal heuristics

Critically, this knowledge capture is incremental and compositional. Experts build modules, not monoliths. New knowledge extends existing structures without requiring a comprehensive redesign. This solves the maintenance problem that killed expert systems while retaining their deterministic reasoning.

### 3.3 Information-Theoretic Advantage

The reduction in entropy is quantifiable. In RAG systems, the answer space includes all possible text generations conditioned on retrieved passages—an astronomically large space. In Mars<sup>®</sup> OS, the answer space includes only paths through the knowledge graph reachable from the query context under worldview constraints, a space that we explicitly bound.

This yields three operational advantages:

- **Consistency:** Same query + same worldview = same reasoning path
- **Explainability:** Outputs are graph traversals that humans can inspect
- **Verifiability:** Every conclusion traces to validated knowledge nodes

For high-stakes decisions, these are not optional. They are requirements.

## 4 Benchmark Evaluation

### 4.1 The KDMA Measurement Problem

DARPA’s In the Moment (ITM) program and related research identify Key Decision-Maker Attributes as critical to AI alignment in difficult domains [3]. KDMAs are individual-level characteristics that predict how decision-makers navigate scenarios where experts disagree, no single answer is correct, and time pressure plus conflicting values create genuine dilemmas.

There are no public benchmarks for the evaluation of KDMA. This absence poses a challenge: systems designed for KDMA-relevant domains have no evaluation standard.

### 4.2 ValuePrism as Proxy

We identified ValuePrism [4] as the best available proxy. The dataset contains 31,000 human-written situations paired with 218,000 values, rights, and duties. Each situation presents ethical complexity, which requires reasoning about conflicting principles.

ValuePrism does not measure KDMA directly. It assesses whether systems can identify which values are relevant to decision contexts—a necessary but insufficient condition for aligned decision-making.

However, the value system is non-deterministic, which poses a challenge in its use. Although many answers align, there are superlative, unknown, rules that introduce randomness indicative of real-world conditions with respect to value systems. Therefore, the purpose of this benchmark is never perfect alignment, and rather a unique representation intended to represent real-world challenges.

We evaluated Mars<sup>®</sup> OS on a random sample of 300 ValuePrism situations against an ontological representation of values-based decision guidance data as the knowledge base. The ontology parsing is native to Mars<sup>®</sup> OS while it was chunked and stored for retrieval in RAG with the LangChain library [5]. This was repeated on three independent runs (900 total evaluations).

**Baseline comparisons:**

- Zero-shot LLM: Direct prompting of frontier language model
- LangChain Basic RAG: Document retrieval + LLM generation
- LangChain Hardened RAG: Document retrieval with re-ranking and compression

**Evaluation metric:** Accuracy in identifying values matching ValuePrism’s expert annotations.

### 4.3 Results

Mars<sup>®</sup> OS achieved a greater than two-fold increase in accuracy relative to advanced RAG implementations under non-deterministic conditions.

Table 1: Performance Comparison on ValuePrism Benchmark (300 samples  $\times$  3 runs)

| System               | Run 1 | Run 2 | Run 3 | Mean        | SD (emp.) | SD (theory) |
|----------------------|-------|-------|-------|-------------|-----------|-------------|
| Mars <sup>®</sup> OS | 83.3  | 79.8  | 82.9  | <b>82.0</b> | 1.96      | 2.2         |
| LangChain Hardened   | 37.5  | 34.5  | 38.1  | 36.7        | 1.92      | 2.8         |
| LangChain Basic      | 33.3  | 35.1  | 28.3  | 32.2        | 3.50      | 2.7         |
| Zero-shot LLM        | 18.6  | 17.5  | 22.3  | 19.5        | 2.51      | 2.3         |

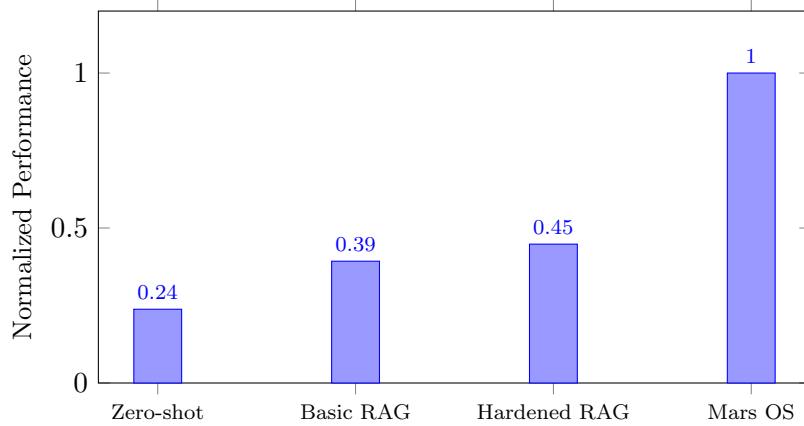


Figure 1: Performance normalized to Mars<sup>®</sup> OS baseline (1.0), with standard deviation across three runs. Mars<sup>®</sup> OS demonstrates 2.2-4.2 $\times$  advantage over alternative approaches.

## 5 Path Forward

### 5.1 The Benchmark Gap

The research community needs public KDMA evaluation frameworks. Current proxies like ValuePrism demonstrate architectural capabilities, but do not measure actual decision alignment under realistic conditions.

ValuePrism has a particular randomness quality that does not exist in more rigid procedural and regulatory guided operations. Yet, even under scenarios of standardized operations, an element of randomness exists based on the individual expertise or worldview of an operator. This specificity becomes the target for formalization:

- Decision scenarios with expert disagreement documented
- Individual decision-maker profiles capturing stable attributes
- Time pressure and resource constraint conditions
- Measurement of trust calibration and delegation behavior

Organizations deploying these systems in high-stakes domains deserve better evaluation standards. We invite collaboration to work with us to establish anonymized public or private datasets for benchmarking.

## 6 Data as Enforceability

The Mars<sup>®</sup> OS results reveals something more fundamental than architectural choices or performance metrics. It demonstrates that *enforceability can be an intrinsic property of data itself*.

Traditional approaches treat enforcement as external to knowledge: rules applied *after* information is retrieved, constraints checked *during* inference, validation performed *on top of* outputs. This externalization creates the fundamental brittleness that killed expert systems. When enforcement is separated from knowledge, every change requires updating both the knowledge base and the enforcement layer. They drift. Systems break.

Mars<sup>®</sup> OS inverts this relationship. Enforcement is embedded in the knowledge structure itself—in the edges that define valid relationships, in the constraints that bound node properties, and in the ontological types that determine composability. When you traverse the graph, you cannot reach invalid states because invalid connections do not exist. The structure *is* the enforcement.

This may be the first operational system to demonstrate this principle at scale. Not because the idea is novel—philosophers and logicians have understood that knowledge structure implies logical constraints for centuries—but because the technology stack finally exists to make it practical. Knowledge graphs provide the substrate. Modern compute enables real-time traversal. HITL interfaces make maintenance tractable. The pieces converged.

What this means for decision intelligence: we can finally escape the performance-correctness tradeoff. Fast systems guess and are often wrong. Correct systems are slow and brittle. Mars<sup>®</sup> OS is fast *because* it is structured correctly—bounded search spaces enable efficient traversal, and enforceability means every path yields valid results.

The deeper implication: high-stakes AI does not require bigger models or more training data. It requires *better data architecture*. When knowledge itself encodes the constraints that govern its use, decision-making systems automatically inherit those guaranties. This is not a new paradigm, it is the original paradigm of symbolic AI, finally equipped with the flexibility mechanisms it always needed.

What we learned building expert systems was correct: explicit knowledge representation matters. What we learned building LLMs was also correct: rigid encoding fails. The synthesis is not a compromise between the two—it is recognizing that they were solving different problems. Expert systems tried to encode *decisions*. LLMs learned to generate *language*. Mars<sup>®</sup> OS encodes *decision spaces* as ontological structures and traversal graphs within which valid decisions exist.

This opens research directions the field has not seriously pursued in decades. If enforceability is intrinsic to data structure, then:

- Knowledge acquisition becomes knowledge architecture—experts do not write rules, they structure possibility spaces
- Explainability is not a post-hoc technique but a natural consequence of traversal histories
- Alignment is not calibrated statistically but composed structurally through worldview modules

- Trust is not established through testing but guaranteed through graph properties

## 6.1 Research Collaboration

We invite collaboration with organizations working on:

- KDMA characterization and measurement frameworks
- Domain-specific decision ontology development
- Human-AI teaming in high-stakes environments
- Explainable AI for mission-critical applications

## 7 Conclusion

We demonstrated that enforceability is intrinsic to data. This shifts the entire AI safety problem. We have spent the last decade trying to make models behave correctly through training, fine-tuning, RLHF, constitutional AI—all attempting to impose constraints on systems that fundamentally operate in unbounded possibility spaces. That is the hard way.

The easier way: structure the data so only valid operations are possible. Build knowledge graphs where edges encode verified relationships. Implement worldviews that bound exploration to acceptable regions. Create ontologies that make harmful outputs unreachable not because they are filtered but because they do not exist in the traversal space.

This does not solve all AI problems. LLMs remain essential for language understanding and generation. But for decision intelligence of systems that must be reliable, explainable, and verifiable in domains where consequences matter, the path forward looks less like building smarter models and more like building smarter knowledge infrastructure.

Mars<sup>®</sup> OS is one solution in this direction. We expect others will follow. The convergence of symbolic structure and operational flexibility represents more than incremental progress. It suggests that the foundational problems of AI decision consistency, explainability, verifiability, and alignment were never about model architecture. They were about data architecture.

Mars<sup>®</sup> Operating System demonstrates what becomes possible when enforcement is intrinsic to knowledge structure: deterministic reasoning without brittleness, bounded exploration without limitation, symbolic precision with adaptive flexibility. These were never competing requirements. They were always complementary properties waiting for the right substrate.

For organizations operating in high-stakes domains, this changes the procurement question. The choice is not between AI that is powerful but uncontrollable versus systems that are safe but inflexible. The choice is between treating knowledge as training data versus treating it as operational infrastructure.

### Contact:

Causum, Inc.

[www.causum.com](http://www.causum.com)

[contact@causum.com](mailto:contact@causum.com)

## References

- [1] Stavarache, L. L. (2025). Enhancing Fine-Tuning LLM Evaluation: A Study on Calibration and Metrics for Industry-Specific AI Alignment. IBM, Columbus, USA.
- [2] Lampi, J., et al. (2024). Predictive Models of Decision Making in Medical Triage. Soar Technology, LLC.
- [3] Defense Advanced Research Projects Agency. In the Moment (ITM) Program. <https://www.darpa.mil/program/in-the-moment>, 2024.
- [4] Sorensen, T., et al. (2024). Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. AAAI Conference on Artificial Intelligence, 38(18), 19937-19947.
- [5] LangChain. (n.d.). LangChain: An open-source framework for developing applications with large language models. <https://www.langchain.com>
- [6] Sen, A., et al. (2025). Decision-Maker Alignment: Benchmark Datasets. Drexel University & Knexus Research.