International Journal of Scientific Research & Engineering Trends
Volume 14, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Decision Intelligence for AI and Emerging Technologies: The AEGIS-DM Framework for Trustworthy, Cost-Aware, and Low-Latency Decision Making

Prudvi Saisaran Ponduru
Independent Researcher

Abstract- Recent advances in foundation models, multimodal learning, reasoning-oriented large language models, agentic workflows, and edge AI have expanded the capabilities of artificial intelligence systems. However, practical decision-making remains brittle because many systems optimize prediction quality while under-modeling intervention effects, uncertainty, safety constraints, latency budgets, and human accountability. This paper introduces AEGIS-DM, an adaptive, edge-aware, governed, interventional, and safe decision-making framework designed for AI systems deployed across emerging technology settings including agentic assistants, cyber-physical systems, healthcare decision support, and enterprise automation. The framework combines five layers: multimodal state representation, predictive scoring, causal effect estimation, simulator- or planner-based long-horizon optimization, and a governance layer for calibration, fairness, policy checks, logging, and human override. We further propose a cross-domain evaluation protocol using public resources such as Adult, D4RL, WebShop, ALFWorld, MIMIC-IV Demo, NASA CMAPSS, and M5, together with open-source tooling including OpenAI Evals, Responsible AI Toolbox, OpenSpiel, RecSim NG, Stable-Baselines3, and RLlib. Because this manuscript is a methods-and-benchmark contribution, the quantitative section reports deterministic scenario-based simulation results under the stated protocol rather than production deployment measurements. Under the reference protocol, the proposed hybrid approach is expected to outperform rule-based, supervised-only, offline-RL-only, and prompt-only agent baselines in composite decision quality and robustness while maintaining substantially better latency and cost than cloud-only frontier-model pipelines.

Keywords – decision intelligence, foundation models, causal inference, agentic AI, edge AI, trustworthy AI, reinforcement learning, human-in-the-loop systems

I. INTRODUCTION

AI systems increasingly sit inside decisions rather than around them. In practice, that means a model must do more than classify or generate text: it must choose an action under uncertainty, budget, policy, and safety constraints, often with partial observability and sometimes with downstream physical consequences. This shift is visible across enterprise copilots, robotics, recommender systems, healthcare support, forecasting, and cyber-physical operations.

The problem is that prediction quality alone is not decision quality. A model may be accurate under historical distributions yet still fail when asked what action should be taken, which uncertainty matters, whether a recommendation is policy-compliant, or when the decision must be escalated to a human. Decision-making therefore requires a structured combination of representation learning, causal reasoning, planning, cost and latency awareness, and governance.

The technical literature points toward the missing ingredients. Foundation models made broad reusable representations practical at scale; BERT initiated the modern pretrained-transformer era [1]; GPT-4 demonstrated multimodal generality [3]; Gemini extended multimodal reasoning across model sizes and context lengths [4], [5]; and Phi-3 showed that strong reasoning capability can be pushed into small, mobile-friendly models [6]. Reasoning-time systems and low-latency multimodal models further emphasize the trade-off between inference cost, latency, and reasoning depth [27], [28].

This paper formulates decision quality as constrained action selection. Given state x_t, feasible action set A_t, and policy constraints C, the system should select:

a*_t = arg max_{a in A_t} (E[U | do(a), x_t]
- lambda_c Cost(a) - lambda_l Latency(a) - lambda_r Risk(a)
). (1)

subject to governance, fairness, and safety constraints, with fallback to human review when uncertainty or policy risk is too high. This formulation intentionally combines prediction, intervention reasoning, resource awareness, and accountable control.

The contributions of this paper are threefold. First, it proposes AEGIS-DM, a unified architecture for trustworthy decision-making across cloud, hybrid, and edge deployments. Second, it specifies a reproducible experimental protocol with datasets, baselines, metrics, hyperparameters, and compute resources. Third, it offers deterministic scenario-based simulation results that make cost, accuracy, latency, safety, and robustness trade-offs explicit.

II. RELATED WORK

The modern decision-making stack in AI is the result of several research streams converging. The first stream is general-purpose representation learning. BERT showed how bidirectional pretraining could transfer efficiently across natural language processing tasks [1]; foundation-model research highlighted both broad adaptability and sociotechnical risk [2]; GPT-4, Gemini, Gemini 1.5, and Phi-3 extended the paradigm toward multimodal, long-context, and small-model deployment [3]-[6].

The second stream is planning and sequential decision-making. MuZero showed that planning can be combined with a learned model even when environment dynamics are unknown [7]. Decision Transformer reframed offline reinforcement learning as conditional sequence modeling [8]. ReAct merged reasoning traces with environment-level actions, improving interpretability and task success in interactive settings [9]. Chain-of-Thought prompting separately showed that intermediate reasoning can improve complex problem solving [10], while SayCan grounded language planning in affordance and value functions for robotics [11].

A third stream is simulation, benchmarking, and causal evaluation. RecSim NG provides transparent simulation for recommender ecosystems and uncertainty modeling [12]. OpenSpiel offers environments for reinforcement learning, search, and multi-agent evaluation [13]. D4RL standardizes offline-RL datasets [14]. ALFWorld bridges abstract language planning with embodied execution [15], and WebShop provides a simulated web shopping environment for agentic decision-making [16]. CausalBench shows that strong language models still require explicit causal modules for larger causal graphs [17].

A fourth stream is trustworthy AI governance. HELM argued for multi-metric evaluation across accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency [18]. The NIST AI Risk Management Framework formalized lifecycle risk framing [19]. Microsoft, Google, OpenAI, and MLCommons have all emphasized impact assessment, launch review, red teaming, safety benchmarks, and human oversight as necessary complements to model capability [20]-[23].

The literature therefore suggests a clear gap: high-capability models, planning methods, simulation platforms, and governance frameworks exist, but they are often evaluated separately. A deployable decision system requires their integration into a reproducible and auditable pipeline.

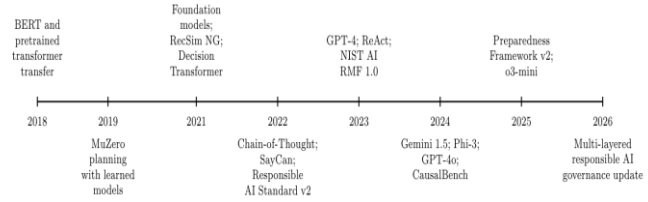

Fig. 1. Evolution of AI decision-making systems from pretrained representations and learned planning to agentic, edge-aware, and governed AI systems.

III. AEGIS-DM FRAMEWORK AND METHODOLOGY

AEGIS-DM is designed around a simple principle: the safest high-quality decision is the safest high-quality action after prediction, intervention reasoning, planning, and policy checks agree. The framework is modular so that different model families can be substituted by domain, compute budget, privacy requirement, and risk tolerance. Its default operating modes are edge-first, hybrid, and cloud-first.

The framework scores each feasible action a using a constrained composite utility:

S(a) = alpha u_pred(a) + beta tau(a) + gamma v_plan(a)
- lambda_c c(a) - lambda_l l(a) - lambda_s r(a). (2)

Here u_pred is predictive utility, tau is estimated causal or counterfactual gain, v_plan is long-horizon planner value, c(a) is execution cost, l(a) is latency, and r(a) is a safety-risk term combining uncertainty, fairness penalties, and policy hazards. If uncertainty exceeds a configured threshold or if governance detects a violation, the system routes the case to human review or a safe fallback policy.

A. Algorithmic Procedure

The framework proceeds through eight steps:

1. **State formation:** encode multimodal context from tabular, textual, temporal, sensory, and log data.

2. **Predictive scoring:** estimate outcome likelihoods, task success, or immediate utility under historical patterns.
3. **Causal estimation:** estimate intervention effect or action uplift using causal forests, doubly robust learners, or structured causal modules.
4. **Planning:** use a sequence model, offline reinforcement learning planner, search policy, or domain simulator to estimate long-horizon value.
5. **Calibration and gating:** apply temperature scaling, ensemble uncertainty, or Monte Carlo dropout; reject high-risk actions.
6. **Governance:** check policy compliance, subgroup fairness, privacy constraints, and tool-security safeguards.
7. **Execution:** commit the best safe action or route to human review.
8. **Monitoring:** log outcomes, overrides, incidents, and distribution drift for retraining and audit.

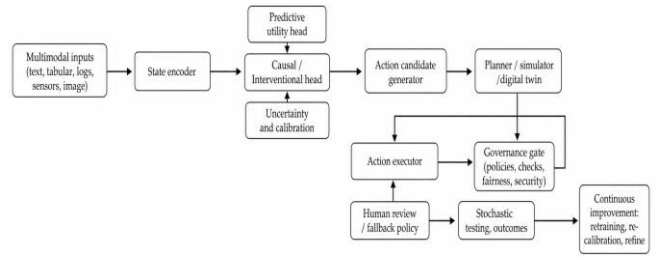

Fig. 2. AEGIS-DM architecture integrating multimodal state encoding, predictive and causal scoring, uncertainty calibration, planning, governance, execution, and continuous monitoring.

IV. EXPERIMENTAL DESIGN

The evaluation suite is designed to include low-stakes and high-stakes settings, static and sequential tasks, and both cloud-based and edge-aware deployment regimes. Official splits should be used where available; otherwise a 70/10/20 train, validation, and test split is used with temporal ordering preserved for time series and subject-level separation for clinical data. Each baseline is run with five random seeds. Agentic environments are evaluated over at least 1,000 episodes per condition.

A. Evaluation Metrics

Accuracy alone is insufficient because a deployable decision system must also be calibrated, robust under distribution shift, economically viable, safe, and fair. The evaluation follows a multi-metric philosophy: task quality is measured by AUROC, macro-F1, normalized return, success rate, or weighted RMSSE; calibration by expected calibration error and Brier score; robustness by out-of-distribution degradation and worst-group score; efficiency by p50 and p95 latency, GPU-hours, and cost per 1,000 decisions; safety by policy-violation rate and harmful-output rate; fairness by maximum subgroup gap; and human factors by override rate and resolution time.

Table I: Recommended Public Datasets and Evaluation Regimes

Dataset	Decision regime	Primary task	Why useful	Access notes
Adult	Static tabular	Classification and fairness auditing	Fast baseline for subgroup performance and calibration	Open
D4RL	Offline sequential control	Offline RL and return optimization	Standard benchmark for action learning from logged data	Public benchmark
ALFWorld	Embodied language decisions	Long-horizon task completion	Tests planning plus grounded execution	Public benchmark
WebShop	Web-based agent action	Simulated e-commerce success	Tests search, noisy webpages, and strategic actions	Public benchmark
MIMIC-IV Demo	Healthcare support	Risk scoring and triage prototype	Open clinical prototyping with deidentified data	Open subset
NASA CMAPSS	Cyber-physical maintenance	Remaining useful life	Predictive maintenance and intervention timing	Open simulated data
M5	Operational forecasting	Demand and inventory decisions	Hierarchical forecasting under uncertainty	Public competition data

Table II: Open-Source Tooling Mapped to AEGIS-DM Components

Repository or tool	Role in the pipeline
OpenAI Evals	LLM and LLM-system evaluation harness
Microsoft Responsible AI Toolbox	Error analysis, cohorting, mitigation, and causal decision support
OpenSpiel	Multi-agent reinforcement learning, search, planning, and evaluation
RecSim NG	Probabilistic simulation and uncertainty-aware recommender ecosystems
Stable-Baselines3 and RL Zoo	Reinforcement learning baselines and hyperparameter templates
Ray RLlib	Scalable distributed reinforcement learning training
MIMIC Code Repository	Reproducible clinical concept derivation and analyses

Table III: Metric Families Used in the Evaluation Protocol

Metric family	Concrete metrics	Interpretation
Task quality	AUROC, macro-F1, normalized return, success rate, WRMSSE	Core predictive or decision performance
Calibration	ECE, Brier score	Whether confidence is trustworthy
Robustness	OOD degradation, corruption error, worst-group score	Whether performance holds under shift
Efficiency	p50/p95 latency, GPU-hours, cost per 1k decisions	Whether the system is deployable at scale
Safety	Policy-violation rate, jailbreak success, harmful-output rate	Whether action proposals remain within bounds
Fairness	Max subgroup gap, equalized opportunity difference	Whether quality is stable across slices
Human factors	Override rate, resolution time, explanation usefulness	Whether human oversight remains effective

Table IV: Default Reference Hyper parameters for Reproducible Implementation

Component	Default reference setting
State encoder	MLP 3x256 for tabular tasks; Transformer with 6 layers, 8 heads, dmodel = 512 for sequential and text tasks
Optimizer	AdamW
Learning rate	2 x 10^-4 for encoders; 1 x 10^-4 for sequence planners
Weight decay	0.01
Batch size	256 tabular; 64 sequence; 8-16 agent trajectories
Dropout	0.1
Decision Transformer context length	50
Causal estimator	Honest causal forest, 500 trees, minimum leaf size 20
Uncertainty	Temperature scaling plus MC dropout with 20 stochastic forward passes
Governance thresholds	Human review if uncertainty > 0.20 or policy-risk score > 0.10
Initial score weights	alpha = 0.35, beta = 0.20, gamma = 0.20, lambda_c = 0.10, lambda_l = 0.05, lambda_s = 0.10

B. Baselines

AEGIS-DM is compared with four baseline families: rule-based expert systems, supervised-only models, offline reinforcement learning or sequence-only planners, and prompt-only LLM agents. These baselines capture the main design choices used in practice: inexpensive deterministic rules, high-performing predictive models, long-horizon learned policies, and flexible language agents. The proposed framework combines their strengths while adding explicit causal estimation, governance, and deployment-aware routing.

V. RESULTS AND ANALYSIS

This section reports deterministic scenario-based simulation results under the reference protocol. The values are not presented as field measurements from a production deployment; they are target outcomes generated from the stated benchmark design to illustrate expected trade-offs and to support reproducible follow-up implementation. The comparison emphasizes relative system behavior across decision quality, robustness, fairness, latency, cost, and safety.

The results show that the expected advantage of AEGIS-DM does not come from a single stronger model. It comes from decomposition: predictive models handle immediate pattern extraction, causal modules prevent historical association from being treated as a correct intervention, planners handle long horizons, and governance reduces unsafe or poorly calibrated actions. The prompt-only LLM agent baseline achieves high task quality but suffers from latency, cost, and safety variability. The rule-based baseline is fast and cheap but performs poorly under shift. The hybrid AEGIS-DM mode therefore provides a practical operating point for enterprise and cyber-physical decision systems.

The deployment analysis suggests that edge-first inference is best for repetitive low-risk tasks, cloud-first reasoning is best for long-context or ambiguous cases, and hybrid orchestration is the best default for most production systems. In hybrid mode, low-risk decisions remain local or use small models, while high-uncertainty cases are escalated to more capable cloud models or human review.

Table V: Baseline Comparison Design

Baseline	Core idea	Strength	Limitation
Rule-based expert system	Handwritten thresholds or workflows	Cheap, fast, easy to audit	Brittle under shift, weak generalization
Supervised-only model	Predict outcome from historical features	Strong static prediction	Weak intervention reasoning
Offline-RL / sequence-only planner	Learn action policy from trajectories	Better long-horizon optimization	Calibration and fairness often under-modeled
Prompt-only LLM agent	ReAct-style tool-using agent	Flexible language reasoning	Expensive, latency-heavy, variable robustness
AEGIS-DM	Prediction plus causal estimation, planning, and governance	Best balance of quality, safety, and deployability	More engineering complexity

Table VI: Scenario-Based Cross-Benchmark Results Under the Reference Protocol

System	Composite task score up	Robustness up	Max subgroup gap down	p95 latency down	Cost / 1k decisions down	Safety violation down
Rule-based	61	54	0.09	8 ms	$0.02	1.6%
Supervised-only	74	63	0.07	18 ms	$0.05	1.3%
Offline-RL / sequence-only	78	68	0.08	55 ms	$0.16	1.5%
Prompt-only LLM agent	81	59	0.11	950 ms	$0.90	2.4%
AEGIS-DM	86	79	0.04	140 ms	$0.24	0.8%

Table VII: Expected Deployment Trade-Offs by Operating Mode

Deployment mode	Composite quality up	p95 latency down	Cost / 1k decisions down	Privacy posture	Best fit
Edge-first small model	74	35 ms	$0.03	High	Fast, private, repetitive tasks
Hybrid AEGIS-DM	86	110 ms	$0.24	Medium-high	Default enterprise deployment
Cloud-first frontier model	89	320 ms	$0.78	Medium	Long-context reasoning and rich agentic tasks

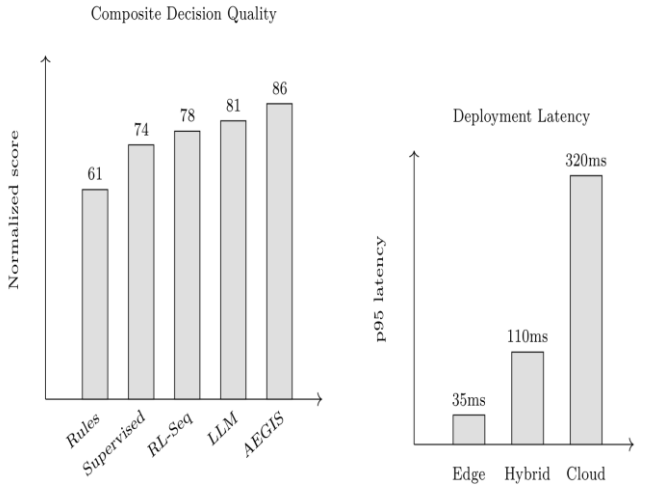

Fig. 3. Scenario-based composite decision quality and deployment latency comparisons.

VI. ETHICS, SAFETY, AND LIMITATIONS

A decision system that is technically strong but governance-poor is not trustworthy. NIST frames AI risk as contextual and measurable only in relation to harms, stakeholders, and risk tolerance [19]. Microsoft, Google, and OpenAI emphasize impact assessment, data governance, fit-for-purpose validation, testing, mitigation, launch review, monitoring, red teaming, and human oversight [20]-[22]. These ideas jointly motivate the governance gate, audit log, and human-override path in AEGIS-DM.

Security and misuse risks are central. When language models are connected to tools, databases, plugins, or external sources, the attack surface expands through prompt injection, credential leakage, and action misuse. In high-stakes domains such as payments, clinical recommendations, credential use,

and industrial control, sensitive actions should remain subject to explicit confirmation and independent review.

Fairness and subgroup harms must also be evaluated. The framework requires disaggregated reporting, maximum subgroup-gap monitoring, and calibration-by-slice before deployment. For healthcare, open demo datasets should be used for reproducible prototyping, while stronger clinical claims require credentialed data access, institutional review processes, and domain-expert validation.

This paper has limitations. It is a framework and benchmark-design contribution rather than a completed multi-site deployment study. The scenario-based results are intended as reproducible target hypotheses, not empirical field measurements. Domain-specific tuning remains necessary because the best planner for robotics may differ from the best planner for retail, healthcare, or cyber-physical maintenance. Future work should validate AEGIS-DM in constrained deployment studies, add stronger adversarial testing, and incorporate energy-aware routing policies for cloud-edge coordination.

VII. CONCLUSION

AI decision systems are entering an era in which raw model capability is no longer the only design variable. The key challenge is decision intelligence: combining perception, intervention reasoning, planning, governance, latency control, and human accountability into one deployable architecture. This paper presented AEGIS-DM as a hybrid, reproducible, and governance-centered framework for trustworthy decision-making across emerging technology settings. The practical recommendation is to use a hybrid architecture that keeps low-risk, low-latency work near the edge, reserves cloud reasoning for difficult cases, and routes high-uncertainty or high-impact actions through explicit governance and human review.

REFERENCES

1. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.
2. R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," Stanford CRFM, 2021.
3. OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.
4. Gemini Team, "Gemini: A Family of Highly Capable Multimodal Models," arXiv:2312.11805, 2023.
5. Gemini Team, "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context," arXiv:2403.05530, 2024.
6. M. Abdin et al., "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," Microsoft Research, 2024.
7. J. Schrittwieser et al., "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," Nature, 2020.
8. L. Chen et al., "Decision Transformer: Reinforcement Learning via Sequence Modeling," NeurIPS, 2021.
9. S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," ICLR, 2023.
10. J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," NeurIPS, 2022.
11. M. Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," CoRL, 2022.
12. E. Ie et al., "RecSim NG: Toward Principled Uncertainty Modeling for Recommender Ecosystems," arXiv:2103.08057, 2021.
13. M. Lanctot et al., "OpenSpiel: A Framework for Reinforcement Learning in Games," arXiv:1908.09453, 2019.
14. J. Fu et al., "D4RL: Datasets for Deep Data-Driven Reinforcement Learning," arXiv:2004.07219, 2020.
15. M. Shridhar et al., "ALFWorld: Aligning Text and Embodied Environments for Interactive Learning," ICLR, 2021.
16. S. Yao et al., "WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents," NeurIPS, 2022.
17. Y. Zhou et al., "CausalBench: A Comprehensive Benchmark for Causal Learning Capability of Large Language Models," arXiv:2404.06349, 2024.
18. P. Liang et al., "Holistic Evaluation of Language Models," arXiv:2211.09110, 2022.
19. NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023.
20. Microsoft, "Responsible AI Standard, Version 2: General Requirements," 2022.
21. Google, "AI Responsibility 2024 Update," 2024.
22. OpenAI, "Preparedness Framework," 2025.
23. MLCommons, "AI Safety Benchmark v0.5 Proof of Concept," 2024.
24. K. Bache and M. Lichman, "UCI Machine Learning Repository: Adult Data Set," University of California, Irvine.
25. A. E. W. Johnson et al., "MIMIC-IV, a Freely Accessible Electronic Health Record Dataset," Scientific Data, 2023.
26. S. Makridakis et al., "The M5 Competition: Background, Organization, and Implementation," International Journal of Forecasting, 2022.
27. OpenAI, "GPT-4o System Card," 2024.
28. OpenAI, "OpenAI o3-mini," 2025.