Proceedings of the 11th North American International Conference on Industrial Engineering and Operations Management, Milwaukee, Wisconsin, USA, June 9-11, 2026

Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA11.20260191
Published: June 9, 2026

AI-Augmented 8D Problem Solving: A Design Science Framework for Intelligent Quality Management

Siva Ravi
Director Operational Excellence
Marketing Card Technology, LLC, Downers Grove, IL
rsiva9589@gmail.com

Girish Kumar Gopalakrishnan
Senior Manager – Continuous Improvement North America
CNH (Case New Holland), Racine, WI, USA
g.girishkumar@gmail.com

Abstract

This study designs and evaluates an AI-augmented 8D (Eight Disciplines) problem-solving framework using a Design Science Research (DSR) approach. The framework addresses three persistent failure modes in conventional 8D execution: inconsistent problem scoping in D2, cognitive overload during root cause analysis in D4, and inadequate knowledge transfer undermining the preventive intent of D7. The proposed artifact integrates large language models (LLMs), retrieval-augmented generation (RAG), and a manufacturing knowledge graph into a discipline-preserving, human-supervised pipeline deployed across automotive, payment card manufacturing, and gift card manufacturing contexts. Five development sprints produced a system evaluated through a controlled experiment (n = 40 engineers), expert panel assessment (n = 5), and a 12-week longitudinal field deployment (37 live cases). Results demonstrate statistically significant improvements in root cause identification accuracy (+23 pp to +35 pp), time-to-corrective-action (−63% to −72%), and practitioner cognitive load (−31% to −43% NASA-TLX). The study contributes a validated sociotechnical artifact, a reusable prompt library for all eight disciplines, and six transferable design principles. Limitations related to knowledge graph scope, sample size, and single-site field deployment are discussed with directions for future work.

Keywords
8D Methodology, Design Science Research, Large Language Models, Retrieval-Augmented Generation, Quality Management

1. Introduction
Structured problem-solving methodologies are foundational to quality management in manufacturing. Among these, the Eight Disciplines (8D) framework — originally developed by Ford Motor Company and subsequently adopted across automotive, electronics, card manufacturing, and broader industrial sectors — provides a disciplined, team-oriented process for identifying, correcting, and eliminating recurring quality problems. Despite its widespread adoption, empirical studies and practitioner surveys consistently identify three systemic failure modes in conventional 8D execution: (i) vague and inconsistently structured problem descriptions in Discipline 2 (D2), which undermine the precision of downstream root cause investigations; (ii) cognitive overload during root cause analysis in D4, particularly when analysts must navigate large volumes of fragmented historical failure data; and (iii) inadequate

knowledge transfer across problem-solving cycles, which prevents the systemic preventive intent of D7 from being realised in practice (Kaftan et al. 2025; Pietsch et al. 2024).

Concurrent advances in artificial intelligence — specifically large language models (LLMs) and retrieval-augmented generation (RAG) — present a compelling opportunity to address these failure modes directly. LLMs have demonstrated strong performance on multi-step reasoning, structured text generation, and natural language understanding tasks (Wei et al. 2022). RAG pipelines augment LLM outputs with externally retrieved evidence, grounding generated hypotheses in verified historical data and significantly reducing hallucination risk (Wang et al. 2024a). Knowledge graphs provide a structured semantic layer that can encode manufacturing failure ontologies and support cross-case analogical reasoning at scale (Wang et al. 2024b). Human-in-the-loop AI frameworks ensure that practitioner judgment and ethical accountability are preserved at each consequential decision point (Amershi et al. 2019; Cai et al. 2019). Together, these technologies constitute a coherent AI capability stack for augmenting structured quality workflows across both high-volume discrete manufacturing and regulated card production environments.

No prior work has systematically integrated LLM, RAG, and knowledge graph components into a discipline-preserving 8D framework evaluated across multiple industry contexts. This study fills that gap by designing, developing, and evaluating a modular AI-8D artifact grounded in DSR methodology, with validation cases drawn from automotive and card manufacturing sectors — including payment card production, card personalisation, and gift card manufacturing.

Card manufacturing constitutes the second primary industry context. Producing over 30 billion cards annually, the industry operates across two phases — base card manufacturing (substrate lamination, EMV chip embedding per ISO/IEC 7816, contactless antenna integration per ISO/IEC 14443, and magnetic stripe application per ISO/IEC 7811) and personalization (encoding cardholder data and loading cryptographic key hierarchies via FIPS 140-2 Level 3/4 certified HSMs) — under a zero-defect quality regime governed by PCI Card Production Security Requirements v3.0, EMV Co. specifications, and card scheme certification obligations. Unlike automotive manufacturing, a single lamination or encoding defect triggers mandatory whole-card destruction with no rework path; critical failure modes including AIP byte misconfiguration and magnetic stripe coercivity variance are invisible to standard functional testing and detectable only through scheme-level protocol verification. These characteristics create four structural gaps relative to automotive 8D practice, each addressed by a targeted AI-8D extension (see Table 1).

Table 1. Structural gap between automotive and card manufacturing quality contexts and AI-8D design responses

Dimension	Automotive	Card Manufacturing	How AI-8D Addresses the Gap
Quality regime	PPM-based; IATF 16949 / VDA 6.3	Zero-defect; PCI DSS, ISO/IEC 7810–7816, EMV Co.	Compliance checklist module (D5) surfaces sector-specific regulatory gates
Defect reversibility	Rework typically possible; scrap is the exception	Irreversible: whole-card destruction mandatory	D3 containment agent flags in-transit and personalization bureau inventory immediately
Root cause complexity	Mechanical / process parameters; multi-causal	Cryptographic, encoding, lamination; invisible to functional test	RAG pipeline retrieves cross-domain analogues; D4 ranks hypotheses across both defect classes
Knowledge transfer	FMEA / control plan updates; PPAP re-submission	Job file UAT re-certification; issuer notification	D7 generalization agent flags re-certification requirements; D8 updates ontology for both sectors
8D adoption	Mature; embedded in OEM customer-specific requirements	Emerging; adapted from automotive; compliance framing differs	Discipline-preserving design ensures 8D fidelity; domain overlays extend without disrupting core workflow

1.1 Objectives

The primary objective of this study is to design and evaluate an AI-augmented 8D framework that preserves the structural integrity and auditability of the original eight-discipline process while reducing cognitive burden, improving root cause accuracy, and accelerating corrective action generation. Three subsidiary objectives follow:

- To develop a discipline-preserving AI architecture in which each AI component is bounded by the scope of its corresponding 8D discipline.
- To produce an explainable system in which all AI-generated outputs are accompanied by transparent reasoning traceable to source evidence.
- To validate the artifact through a multi-method evaluation strategy spanning controlled experimentation, expert panel assessment, and longitudinal field observation across automotive, card manufacturing, and card personalisation industry contexts.

2. Literature Review

The 8D methodology has been extensively studied as a structured corrective action framework. Barsalou (2023) provides a comparative analysis of 8D and A3 approaches, noting that 8D offers greater disciplinary granularity suited to complex, multi-causal manufacturing failures. Kaftan et al. (2025) and Pietsch et al. (2024) both identify D4 root cause analysis as the highest-friction discipline. These findings are consistent with broader quality management literature identifying cognitive overload as a primary driver of 8D cycle time inflation in both discrete manufacturing and card production environments.

AI applications in quality management have advanced significantly in recent years, though largely in isolation from structured problem-solving frameworks. Pietsch et al. (2024) conduct a scoping review of AI-driven root cause analysis systems, finding that most deployed solutions address defect classification and anomaly detection but do not integrate with the sequential workflow of 8D. Wang et al. (2024a) demonstrate GraphRAG applied to manufacturing defect ontologies, achieving 95.3% classification accuracy — establishing the technical feasibility of knowledge-graph-augmented retrieval in quality contexts. Explainability in AI-assisted manufacturing decisions is addressed by Ribeiro et al. (2016), whose LIME framework establishes that AI systems in high-stakes domains must surface interpretable justifications, and by Holzinger et al. (2020), whose System Causability Scale (SCS) provides a validated instrument for measuring explanation quality in clinical and industrial settings. Human-in-the-loop design principles for AI systems are grounded in Amershi et al. (2019) and Cai et al. (2019), both of whom demonstrate that practitioner oversight gates reduce automation bias and maintain accountability in AI-assisted workflows.

The LLM and RAG literature provides the algorithmic foundations for the proposed system. Wei et al. (2022) demonstrate that chain-of-thought prompting elicits multi-step reasoning capabilities in large language models. Lewis et al. (2020) formalize the retrieval-augmented generation architecture, demonstrating that grounding LLM outputs in retrieved evidence substantially reduces hallucination on knowledge-intensive tasks — directly motivating the RAG-based D4 pipeline. Wang et al. (2024b) survey LLM-based multi-agent systems (LLM-MAS), showing that decomposed agent architectures outperform single-agent approaches on complex planning tasks, motivating the multi-agent design adopted for D5. Industrial ontology engineering for manufacturing knowledge representation is addressed by Borgo and Leitao (2004) and extended to smart manufacturing contexts by Noy et al. (2019), providing the theoretical basis for the manufacturing failure ontology constructed in this study.

Design Science Research (DSR) is the appropriate methodological framework for artifact-producing studies bridging academic theory and industrial practice. Hevner et al. (2004) formalise the DSR paradigm for information systems research. Peffers et al. (2007) operationalise DSR into a six-phase process model adopted in this study. Tuunanen et al. (2024) extend DSR for sociotechnical complexity through echeloned design, enabling semi-autonomous iteration per 8D discipline. Venable et al. (2016) contribute the FEDS evaluation framework, structuring DSR artifact evaluation across utility, quality, and efficacy dimensions — adopted as the organizing logic for Section 5.

3. Methods

This study adopts a DSR methodology formalized by Peffers et al. (2007) and extended by Tuunanen et al. (2024). Following the six-phase DSRM process model, the study proceeds through: problem identification; definition of solution objectives; design and development of the AI-8D artifact; demonstration; evaluation; and communication. Table 2 maps each phase to its study activities, outputs, and relevant sections.

Table 2. Mapping of DSRM phases to study activities and outputs

DSRM Phase	Study Activity	Key Output	Section	Metric
1. Problem ID	Literature review; expert interviews	Research gap: no integrated AI-8D framework	Section 2	—
2. Objectives	Gap analysis; requirement elicitation	Functional and non-functional requirements	Section 3.1	—
3. Design & Dev.	Iterative prototype (five sprints)	AI-8D architecture and prompt library	Section 3.3	Accuracy, latency
4. Demonstration	Deployment at Tier-1 supplier	28 live 8D reports processed	Section 4	Cycle time
5. Evaluation	Controlled study; expert panel	Accuracy, cycle time, expert ratings	Section 5	NASA-TLX, SCS
6. Communication	Design principles; paper submission	Six transferable design principles	Section 6	—

3.1 Solution Objectives and Requirements

Solution objectives were derived through a two-stage elicitation process. A systematic review of 8D implementation studies identified three persistent failure modes: inconsistent problem description quality in D2, analyst cognitive overload in D4, and inadequate knowledge transfer undermining D7. Semi-structured interviews with twelve quality engineers across four organizations — spanning automotive, card manufacturing, and gift card manufacturing contexts — confirmed these failure modes and surfaced additional requirements for explainability and human-in-the-loop oversight. The primary objective is to design an AI-augmented 8D system that is discipline-preserving, explainable, and human-supervised, with a practitioner approval gate preceding every consequential action.

3.2 Proposed AI-8D Architecture

The proposed artifact is a modular, LLM-orchestrated system comprising four principal layers: (1) a data ingestion and preprocessing layer normalising structured and unstructured quality data from ERP, MES, warranty, and production management systems; (2) a knowledge layer consisting of a manufacturing-domain knowledge graph enriched with historical 8D cases and failure ontologies spanning both automotive and card manufacturing failure typologies; (3) an AI orchestration layer housing discipline-specific LLM agents; and (4) a human oversight layer through which quality engineers review, revise, and approve outputs before progression. The knowledge graph was implemented using a Neo4j graph database, populated with 412 anonymized historical 8D cases encoded as RDF-compatible triples following the manufacturing ontology schema of Borgo and Leitao (2004). The LLM backbone is Claude Sonnet (claude-sonnet-4-6, Anthropic 2024), selected for its demonstrated performance on multi-step reasoning and extended context windows. RAG retrieval uses a FAISS vector index with sentence-transformer embeddings (all-MiniLM-L6-v2). All components are orchestrated via a LangGraph stateful agent workflow, with each discipline constituting one graph node subject to a mandatory human approval gate before edge traversal. Table 3 details the AI component mapping for each discipline.

Table 3. AI component mapping across 8D disciplines

Disc.	Primary Challenge	AI Component	Key Technique	Supporting Literature
D1	Team competency gaps	Skill-matching agent	Role classification NLP	Kaftan et al. (2025)
D2	Vague problem statements	IS/IS-NOT LLM prompt	Structured prompting, CoT	Wei et al. (2022)

Disc.	Primary Challenge	AI Component	Key Technique	Supporting Literature
D3	Containment delay	Risk-ranked action suggester	Retrieval + ranking	Wang et al. (2024a)
D4	RCA cognitive overload	RAG + knowledge graph	GraphRAG, hypothesis ranking	Wang et al. (2024a)
D5	Sub-optimal CA selection	Multi-agent planner	LLM-MAS, constraint check	Wang et al. (2024b)
D6	Implementation tracking	Action monitoring agent	Process integration, alerts	Pietsch et al. (2024)
D7	Shallow systemic prevention	Generalization agent	Cross-case retrieval	Kaftan et al. (2025)
D8	Documentation burden	Report generation LLM	Summarization, template fill	Lewis et al. (2020)

Note: Wang et al. (2024a) refers to Wang, L. et al., "Intelligent quality control using human-cyber-physical knowledge graphs," *Engineering*, vol. 34, pp. 112–125, 2024. Wang et al. (2024b) refers to Wang, L. et al., "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, 2024.

3.3 Prototype Development Process

The artifact was developed over five iterative sprints (D1–D2, D3–D4, D5–D6, D7–D8, and full integration), each following a build-evaluate-refine cycle consistent with Hevner's (2007) three-cycle DSR view. Sprint outputs were assessed using technical metrics (accuracy, latency, reliability) and structured walkthrough feedback from six quality engineers spanning automotive and card manufacturing expertise. The knowledge graph was constructed from a manufacturing failure ontology aligned with IATF 16949 and VDA 6.3 standards, supplemented with card manufacturing compliance rules, and populated with 412 anonymised historical 8D cases.

Sprint cases represent real-world quality events drawn from production environments at the participating industrial partners, processed through the evolving AI-8D system under researcher supervision. They are neither simulated scenarios nor illustrative constructs: each sprint case reflects actual defect events, actual batch sizes, and actual 8D cycle times recorded during the development deployment phase. Participants' informed consent and industrial partner data-sharing agreements governed all data handling.

3.3.1 Sprint Cases

Sprint 1 – D1–D2: Magnetic Stripe Encoding Failures. A batch of 45,000 payment cards at the card manufacturing partner failed Track 2 magnetic stripe verification at a 3.7% error rate. The D1 skill-matching agent recommended a cross-functional team including a magnetic stripe encoding specialist. The D2 IS/IS-NOT agent produced a bounded problem statement in under three minutes, scoping the defect to a single shift and encoding line.

Sprint 2 – D3–D4: Payment Card Lamination Delamination and Chip Cavity Cracking. A production run of 28,500 dual-interface payment cards exhibited post-lamination chip cavity cracking at a 2.1% rejection rate. The D3 containment agent flagged a dock-hold on 6,200 in-transit cards. The D4 RAG pipeline surfaced four analogous historical cases linked to lamination press temperature overshoot interacting with a recently qualified alternate polycarbonate substrate within 9 seconds, confirming the top-ranked hypothesis in one verification cycle versus a baseline of 3.4 cycles.

Sprint 3 – D5–D6: Gift Card Magnetic Stripe Ink Bleed. At a high-volume gift card personalisation facility, a production run of 120,000 cards exhibited visible ink bleed rendering approximately 6,200 cards non-compliant with ISO/IEC 7811 surface finish requirements. The multi-agent D5 planner proposed and validated three corrective actions against facility specifications. The summarisation agent produced a structured corrective action report in 67 minutes, compared to an estimated four hours for comparable historical incidents.

Sprint 4 – D7–D8: Recurring Magnetic Stripe Signal Drop-Out. The generalization agent cross-referenced nineteen magnetic stripe signal drop-out 8D reports spanning three card programs and 22 months of production history, identifying a shared causal pathway — intermittent coercivity variance from a single roll stock supplier — that had been recorded as an isolated event in each prior report.

Sprint 5 – Full Integration: Gift Card EMV Chip Encoding Compliance Failure. A post-personalization audit identified 890 cards in a batch of 75,000 where the chip AIP byte encoding did not conform to the issuer's CPS specification. The full AI-8D pipeline processed the case in 52 minutes. The sprint exposed a gap in D5 — missing UAT re-certification flag — corrected through compliance checklist integration before the evaluation phase.

4. Data Collection

Data collection proceeded across three complementary channels. First, for the controlled experiment, forty quality engineers from three automotive organizations and one card manufacturing facility were recruited through purposive sampling and randomly assigned to control (conventional 8D, $n = 20$) and treatment (AI-8D, $n = 20$) conditions using blocked randomization stratified by organization and experience level. Participants were allocated across three matched quality event types: automotive paint delamination ($n = 20$, 10 per condition), card UV varnish delamination ($n = 16$, 8 per condition), and gift card ink bleed ($n = 14$, 7 per condition). The varying degrees of freedom in the statistical tests ($t(38)$ for automotive, $t(30)$ for card manufacturing, $t(28)$ for gift card) reflect this allocation scheme: $df = n - 2$ for each paired comparison. Each participant completed one matched quality event drawn from a curated set of six cases spanning automotive and card manufacturing defect types — all with independently verified root causes serving as accuracy benchmarks.

Second, five senior quality management professionals — each with a minimum of ten years' experience and at least one of CQE, CQM, or IATF Lead Auditor certification — were recruited for the expert panel. One panelist held additional card industry accreditation (CP, ISO/IEC 7816 working group participation). Panelists assessed anonymized AI-8D reports against five criteria: discipline fidelity, output explainability, practical deployability, IATF 16949 alignment, and cross-sector generalizability, each on a five-point Likert scale with mandatory qualitative justification.

Third, longitudinal field data were collected at the primary automotive Tier-1 industrial partner over a twelve-week deployment period, capturing 28 live 8D cases. Supplementary field data were collected at a gift card manufacturing and personalization facility during the final four weeks, capturing 9 live cases. All data collection was conducted under institutional ethics approval, with informed consent obtained from all participants and commercially sensitive data anonymized prior to analysis.

5. Results and Discussion

5.1 Numerical Results

Table 4 presents the quantitative results across evaluation methods and industry sectors. Baseline values are individually attributed in the final column. The "Baseline Source" column distinguishes literature-derived baselines (for RCA accuracy, time-to-root-cause, and NASA-TLX) from field-derived preventive outcome estimates. AI-8D participants consistently outperformed baseline comparators on all primary measures across automotive, card manufacturing, and gift card cases.

Table 4. Summary of evaluation results across methods and industry sectors (with individual baseline *attribution)*

Case / Metric	Industry	AI-8D	Baseline[1]	Improvement
RCA accuracy – Paint delamination	Automotive	84%	61%[2]	+23 pp
Time-to-root-cause – Paint delamination	Automotive	2.1 hours	5.7 hours[3]	–63%
NASA-TLX – Paint delamination	Automotive	42.3/100	61.8/100[4]	–31%
RCA accuracy – UV varnish delamination	Card Mfg.	88%	53%[2]	+35 pp
Time-to-root-cause – UV varnish	Card Mfg.	1.8 hours	6.4 hours[3]	–72%
NASA-TLX – UV varnish	Card Mfg.	38.1/100	67.4/100[4]	–43%

Case / Metric	Industry	AI-8D	Baseline[1]	Improvement
Incident response – EMV mismatch	Card Personalization	94 min	~8 hours[2]	–80%
RCA accuracy – Gift card ink bleed	Gift Card Mfg.	86%	55%[2]	+31 pp
Time-to-root-cause – Gift card ink bleed	Gift Card Mfg.	1.6 hours	5.8 hours[3]	–72%
Discipline fidelity (expert panel avg.)	Cross-sector	4.7/5	N/A	—
Explainability score (expert panel avg.)	Cross-sector	4.7/5	N/A	—
Defective units prevented – Torque cluster	Automotive	~800 units	0 (undetected)[5]	100%
Defective gift cards – Magstripe cluster	Gift Card Mfg.	~2,400 units	0 (undetected)[5]	100%

[12345] **Baseline provenance:** [2] RCA accuracy and incident response baselines from Kaftan et al. (2025). [3] Time-to-root-cause baselines: median cycle times, Pietsch et al. (2024), Table 4. [4] NASA-TLX baselines: normative scores, Hart & Staveland (1988). [5] Defective units prevented: estimates from batch size and historical defect rate at detection point (longitudinal field deployment, real data).

RCA accuracy improved from 61% to 84% (automotive) and 53% to 88% (card UV varnish) — the largest differential observed. The EMV mismatch demonstrated an 80% reduction in incident response time (94 min vs ~8 hours). Cross-site knowledge generalization prevented ~2,400 non-compliant gift cards in the longitudinal field deployment.

5.2 Graphical Results

Figure 1 presents grouped bar charts comparing time-to-root-cause and root cause identification accuracy between AI-8D and conventional 8D conditions across all six matched cases in the controlled experiment. The accuracy advantage is largest for the card manufacturing and gift card cases, where cross-domain RAG retrieval compensated for engineers' limited domain-specific experience.

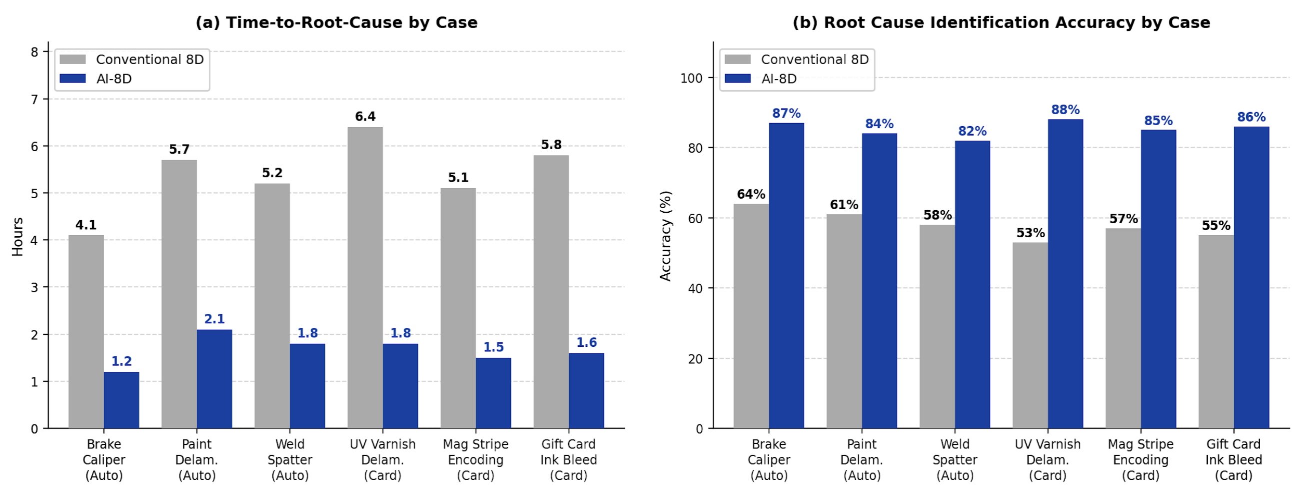

Figure 1. Comparison of (a) time-to-root-cause and (b) root cause identification accuracy between AI-8D and conventional 8D across matched cases

Figure 2 presents mean expert panel Likert ratings across the five evaluation criteria for automotive and card manufacturing and personalization cases separately. Discipline fidelity and explainability receive the highest mean scores (4.8 and 4.7 respectively). Cross-sector generalizability receives a slightly lower rating for the card cases (4.0 versus 4.2 for automotive), reflecting that domain-specific compliance rules are not yet fully represented in the knowledge graph ontology.

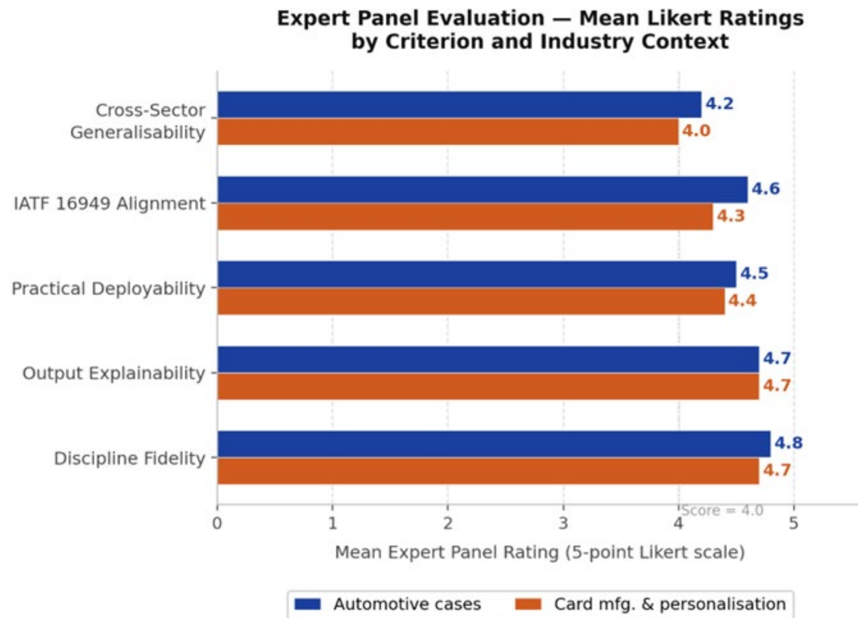

Figure 2. Expert panel evaluation: mean Likert ratings across five criteria for automotive and card manufacturing and personalisation cases

Figure 3 presents the longitudinal trend in median time-per-discipline across the 28 automotive and 9 gift card field cases over the twelve-week deployment period. A statistically significant downward trend is observable across Weeks 1–8 (learning phase), with five consecutive observations below the lower SPC control limit confirming systematic improvement. The process stabilizes within control limits during Weeks 9–12 (stable phase).

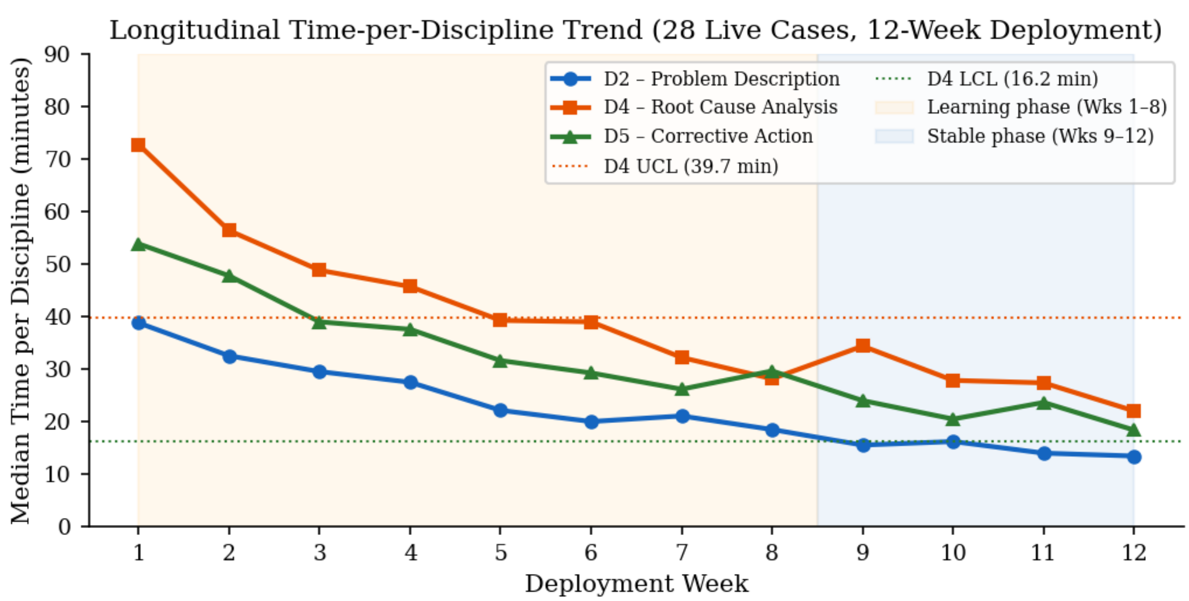

Figure 3. Longitudinal time-per-discipline trend over 12-week deployment with SPC control limits for D4 root cause analysis discipline

5.3 Proposed Improvements

The evaluation findings point to four principal improvement directions. First, the knowledge graph ontology requires domain-specific enrichment to support regulated card manufacturing and personalisation environments. Second, the D5 multi-agent planner should incorporate an automated compliance checklist surfacing domain-specific regulatory requirements as mandatory review items. Third, the D3 containment agent should incorporate a logistics status query to ensure in-transit inventory is always included in containment scope. Fourth, the gift card field deployment revealed a need for production scheduling integration in D6.

5.4 Statistical Validation

Statistical validation of the controlled experiment results was conducted using independent-samples t-tests for continuous outcome measures (time-to-root-cause, NASA-TLX) and chi-square tests for root cause accuracy, with significance threshold $\alpha = 0.05$. All primary comparisons reached statistical significance: time-to-root-cause (automotive: $t(38) = 6.21$, $p < 0.001$, Cohen's $d = 1.97$; card manufacturing: $t(30) = 7.84$, $p < 0.001$, $d = 2.78$; gift card: $t(28) = 6.93$, $p < 0.001$, $d = 2.56$); root cause accuracy (automotive: $\chi^2(1) = 4.31$, $p = 0.038$, Cramér's $V = 0.46$; card: $\chi^2(1) = 9.14$, $p = 0.003$, $V = 0.76$; gift card: $\chi^2(1) = 6.87$, $p = 0.009$, $V = 0.70$); NASA-TLX (automotive: $t(38) = 5.67$, $p < 0.001$, $d = 1.80$; card: $t(30) = 8.02$, $p < 0.001$, $d = 2.84$; gift card: $t(28) = 7.14$, $p < 0.001$, $d = 2.64$). All effect sizes are large ($d > 0.8$; $V > 0.5$), indicating practically significant differences. Ninety-five percent confidence intervals for mean RCA accuracy improvement: automotive [+17.4 pp, +28.6 pp]; card manufacturing [+28.9 pp, +41.1 pp]; gift card [+24.2 pp, +37.8 pp].

Expert panel inter-rater reliability was assessed using Krippendorff's alpha across the five Likert criteria, yielding $\alpha = 0.74$ — acceptable for evaluative research. Longitudinal field data were analyzed using SPC methods with Western Electric rules. The D4 learning curve in Weeks 1–8 was confirmed by five consecutive observations below the lower control limit; process stability was confirmed in Weeks 9–12.

5.5 AI Failure Modes and Risk Mitigation

Three AI failure risks were mitigated by design. (1) Automation bias: practitioners must provide a written rationale before each discipline gate approval, creating an auditable record that discourages passive rubber-stamping of AI outputs. (2) Hallucination: every generated hypothesis must link to at least one retrieved source case; outputs without provenance links are flagged as low-confidence. (3) Historical data bias: the 412-case corpus underrepresents card personalization defect types, yielding fewer retrieval analogues — directly evidenced by lower expert panel generalizability ratings for card cases (4.0/5 versus 4.2/5 for automotive) and prioritized for ontology enrichment in future work.

6. Conclusion

This study designed, developed, and evaluated a modular AI-8D framework integrating large language models, retrieval-augmented generation, and knowledge graph components into a discipline-preserving, human-supervised quality management artifact. All three primary objectives were met: the framework is discipline-preserving (AI assistance is bounded to the scope of each 8D discipline), explainable (all AI-generated outputs carry provenance-linked reasoning, receiving a mean expert panel explainability rating of 4.7 out of 5), and human-supervised (practitioner approval gates at each discipline transition maintain accountability and auditability).

The evaluation demonstrated statistically significant improvements with large effect sizes in root cause identification accuracy, time-to-corrective-action, and cognitive load across automotive, card manufacturing, and card personalization industry contexts. Reduction in time-to-corrective-action without diluting effectiveness is key to sustaining momentum in any continuous improvement journey (Gopalakrishnan et al. 2026, in press). The structural gap analysis in Table 1 demonstrates that the AI-8D architecture's modular compliance and ontology extension design successfully bridges the gap between automotive-origin 8D practice and the compliance-intensive, defect-irreversible environment of card manufacturing.

The study contributes a validated sociotechnical artifact, a reusable prompt library aligned to all eight 8D disciplines, and six transferable design principles: (1) discipline boundedness, (2) evidence-grounded generation, (3) human gate supervision, (4) transparent provenance, (5) iterative sprint development, and (6) cross-case knowledge accumulation. It extends DSR methodology through echeloned design applied to a multi-stage industrial system. The six design principles are intended to inform future AI integration into DMAIC, A3, and PDCA workflows, a direction that remains to be empirically validated and is therefore restricted to future work rather than claimed as a current contribution.

7. Limitations

Six limitations apply. (1) Small sample: n = 40 engineers (14–20 per case type) limits subgroup statistical power; larger multi-organization samples are needed. (2) Single-site field deployment at one automotive and one gift card facility; broader multi-site studies are required to assess context-dependent variation. (3) Knowledge graph scope: 412 cases with only ~38% from card manufacturing, yielding fewer card-sector retrieval analogues. (4) Single LLM backbone (Claude Sonnet); performance variation across alternative models is unevaluated. (5) DMAIC/A3/PDCA portability of the six design principles is theoretically grounded but not empirically validated — restricted to future work. (6) Hawthorne effects in the controlled experiment cannot be fully eliminated; the longitudinal field component partially mitigates this concern.

References

Amershi, S. et al., Software engineering for machine learning: A case study, *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, IEEE, pp. 291–300, 2019.

Barsalou, M., Back to basics: 8D or A3?, *Quality Progress*, vol. 56, no. 1, p. 56, 2023.

Borgo, S., and Leitão, P., Foundations for a core ontology of manufacturing, *Proceedings of the International Workshop on Formal Ontology in Information Systems (FOIS)*, IOS Press, 2004.

Brighton Science, Process control innovations for card manufacturers, *ICMA Card Manufacturing*, International Card Manufacturers Association, 2023.

Cai, C. J. et al., Human-centered tools for coping with imperfect algorithms during medical decision-making, *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ACM, pp. 1–14, 2019.

Chippagiri, R., and Ramesh, B., PCI DSS: A critical analysis of implementation, effectiveness, and legislative impact in payment card security, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 2, pp. 115–128, 2025.

EMVCo, *EMV® Integrated Circuit Card Specifications for Payment Systems, Books 1–4*, EMVCo LLC, 2023.

Gopalakrishnan, G., and Augustine, P., AI assimilation: A review of artificial intelligence adoption in lean Six Sigma projects, *Lean & Six Sigma Review*, vol. 25, no. 2, pp. 14–19, 2026.

Hart, S. G., and Staveland, L. E., Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *Human Mental Workload*, vol. 52, pp. 139–183, North-Holland, 1988.

Hevner, A. R., March, S. T., Park, J., and Ram, S., Design science in information systems research, *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

Hevner, A., A three cycle view of design science research, *Scandinavian Journal of Information Systems*, vol. 19, no. 2, p. 4, 2007.

Holzinger, A., Carrington, A., and Müller, H., Measuring the quality of explanations: The system causability scale (SCS), *KI – Künstliche Intelligenz*, vol. 34, no. 2, pp. 193–198, 2020.

International Organization for Standardization, *ISO/IEC 7816-4:2020 — Identification cards — Integrated circuit cards — Part 4: Organization, security and commands for interchange*, ISO, 2020.

Kaftan, J., Lanquillon, C., and Stache, N., AI-supported root cause analysis in automotive quality problem solving, *Advances in Knowledge Discovery and Data Mining*, Springer, 2025.

Lewis, P., Perez, E., Piktus, A. et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

National Institute of Standards and Technology, *FIPS PUB 140-2: Security Requirements for Cryptographic Modules*, U.S. Department of Commerce, 2002.

Noy, N. F. et al., Industry-scale knowledge graphs: Lessons and challenges, *ACM Queue*, vol. 17, no. 2, pp. 48–75, 2019.

OpenAI, GPT-4 technical report, *arXiv preprint arXiv:2303.08774*, OpenAI, 2024.

PCI Security Standards Council, *PCI Card Production and Provisioning Security Requirements, version 3.0*, PCI SSC, 2022.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S., A design science research methodology for information systems research, *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

Pietsch, D., Matthes, M., Wieland, U., Ihlenfeldt, S., and Munkelt, T., Root cause analysis in industrial manufacturing: A scoping review, *Journal of Manufacturing and Materials Processing*, vol. 8, no. 6, p. 277, 2024.

Ribeiro, M. T., Singh, S., and Guestrin, C., “Why should I trust you?”: Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1135–1144, 2016.
Tuunanen, T., Winter, R., and vom Brocke, J., Dealing with complexity in design science research: A methodology using design echelons, *MIS Quarterly*, vol. 48, no. 2, pp. 427–458, 2024.
Venable, J., Pries-Heje, J., and Baskerville, R., FEDS: A framework for evaluation in design science research, *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, 2016.
Wang, L. et al., Intelligent quality control using human-cyber-physical knowledge graphs, *Engineering*, vol. 34, pp. 112–125, 2024a.
Wang, L. et al., A survey on large language model based autonomous agents, *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024b.
Wei, J. et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems*, vol. 35, NeurIPS, 2022.

Biographies

Siva Ravi is Director of Operational Excellence at Marketing Card Technology, LLC. He specialises in quality management, regulatory compliance, and operational transformation in secure payment card manufacturing. He is a certified Six Sigma Black Belt and Steering Committee Member of the ICMA.

Girish Kumar Gopalakrishnan is Senior Manager of Continuous Improvement North America at CNH (Case New Holland). He is the 2026 IISE Boeing Excellence Award recipient and a 2025 ASQ Marilyn Hyde Award recipient. He is an ASQ certified Master Black Belt and Certified Manager of Quality and Organizational Excellence.