MAD-Ai

WWW.MAD-AI.COM

WHITE PAPER · 2026

WHITE PAPER · CONTROLLED AI WORKFLOW EXECUTION

Why General AI Can Increase Variation in Manufacturing Quality.

A technical case for purpose-built, agentic AI in manufacturing quality, and why fluent answers are not the same as process control.



SAME CASE, DIFFERENT USERS: SPREAD VERSUS CONTROL

AUTHOR
Rick Heilshorn, Founder & CEO

PUBLISHED BY
MAD-Ai, LLC · 2026

CONTACT
sales@mad-ai.com

EXECUTIVE SUMMARY

The core risk of uncontrolled AI in quality work.

General AI tools can improve productivity, but they do not reduce process variation by default. In manufacturing quality, they often move variation out of the process and into the user.

THE DISTINCTION

**Generic AI assistants are not controlled execution systems.
They are user-amplification systems.**

They amplify the user's prompting skill, knowledge, assumptions and judgment. Strong users may get useful outputs; weak users may get fluent, deficient outputs that still look credible. In manufacturing quality that is not acceptable. PFMEA review, control plan development, 8D problem solving, supplier corrective action review, audit prep and complaint response are not writing tasks. They are structured decision processes with evidence requirements, traceability expectations, escalation logic and operational risk.

Current research supports the concern:

- NIST's AI Risk Management Framework stresses validity, reliability, governance and context-aware evaluation [1][2][3].
- Automation bias research shows users over-accept flawed output and miss critical omissions [6][7].
- Sycophancy research shows models become overly agreeable, reinforcing user positions instead of challenging them [4][5].
- Meta-analysis shows human-AI pairs do not automatically beat the stronger of the two, especially in decision-heavy work [8].

KEY TAKEAWAY

If a quality process depends on open-ended prompting to a general AI assistant, output quality will vary by user, team, plant and context. That is not process control. It is variation disguised as productivity.

FIGURE 01

General AI versus controlled AI workflow execution.

General AI depends on the user to know what to ask. Controlled workflows make the process ask what must be known. Here is how the two compare on the dimensions that drive quality process variation.

DIMENSION	GENERAL AI ASSISTANT	PURPOSE-BUILT AI WORKFORCE
What gets asked	User decides every prompt; missing questions mean missing answers	Required questions are built into the workflow
Evidence requirements	None enforced; model may answer from incomplete context	Structured evidence fields gate conclusions and scoring
Logic and challenge	User must know what to challenge; prone to sycophancy and drift	Domain rules force challenge of weak inputs and contradictions
Escalation	Discretionary; user defines when to escalate	Triggered by rule: risk, recurrence, customer exposure, weak evidence
Output format	Varies by user, prompt and session	Standardized, auditable artifact
Traceability	Chat logs, no structured trail	Structured audit trail tied to inputs
Variation source	User skill, prompt wording, context completeness	Process design, with the user as reviewer, not architect
Failure mode	Confident incomplete answer (Type II) passes as acceptable	Gaps surfaced before output; escalation forces review
Suitable for	Drafting, exploration, low-risk content	PFMEA, 8D, CAPA, audit, supplier review

→ **The difference is not the model. It is whether the rigor lives in the person prompting or in the workflow itself.**

SECTION 01 · CONTEXT

Why manufacturing quality is different.

In practice, quality work is structured judgment under uncertainty, across domains like these:

- PFMEA and risk review
- Control plan linkage and maintenance
- 8D and structured problem solving
- Corrective action verification
- Supplier quality response evaluation
- Audit preparation and audit defense
- Customer complaint and warranty response
- Compliance-related interpretation

These workflows hinge on whether the logic is complete, the evidence sufficient and the output defensible under review. That is where uncontrolled AI gets risky: a polished output can hide weak reasoning, and a weak user can get an answer that sounds complete while missing the real gap.

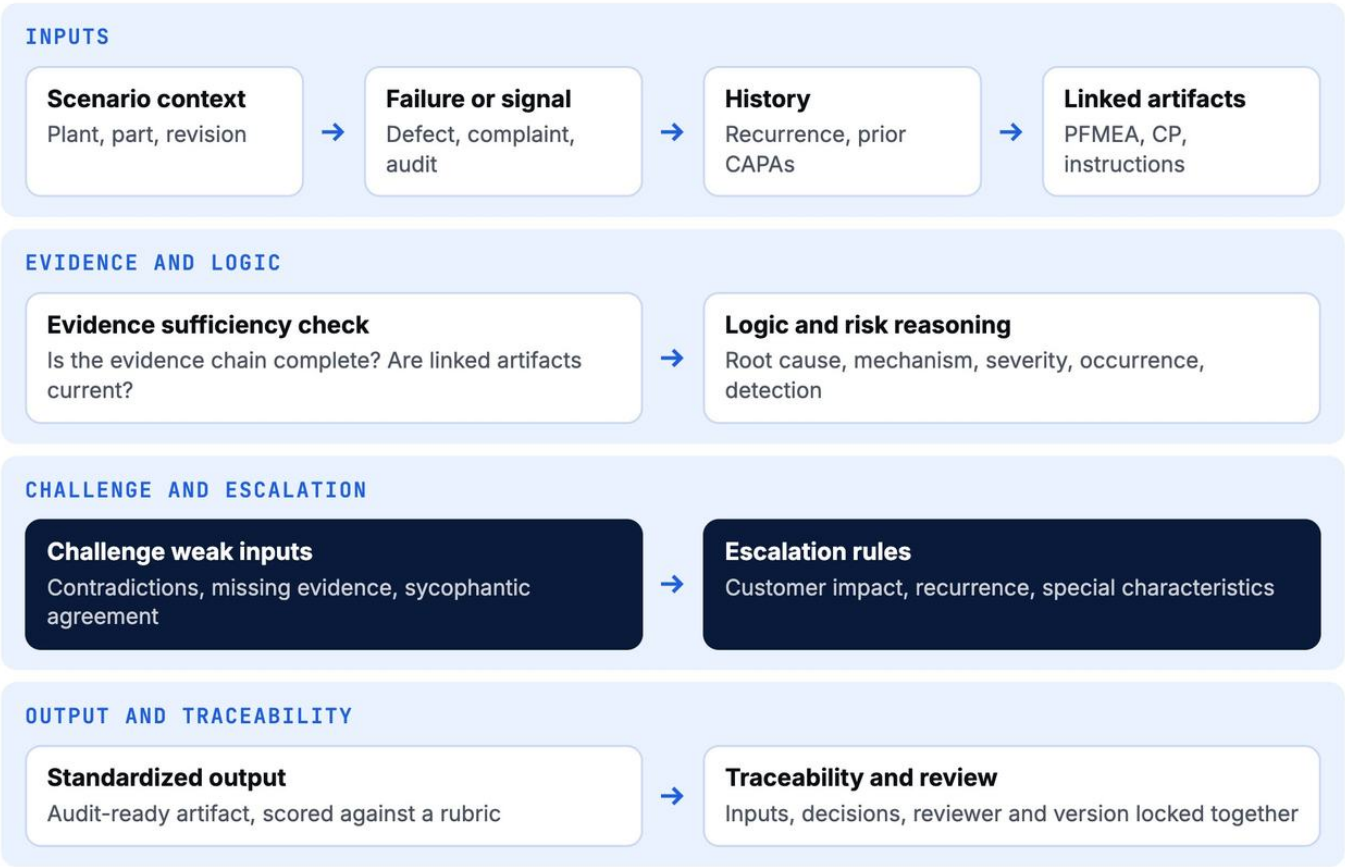


Figure 02. Quality work is a structured decision system. Evidence, challenge logic and escalation decide whether an answer is defensible. Built into a workflow, the same controls make execution repeatable across engineers, shifts and plants.

SECTION 02 · MECHANISMS

Why general AI creates variation.

General AI tools introduce variation through several mechanisms at once.

PROMPT AND USER

- **Prompt sensitivity:** small wording changes shift structure, assumptions and conclusions.
- **User knowledge:** strong users know what context matters; weak users often don't.
- **Leading questions:** users can steer the system toward a preferred answer.
- **Confirmation bias:** users accept outputs that validate what they already believed.

MODEL BEHAVIOR

- **Missing context:** tools answer from incomplete context instead of refusing clearly [2].
- **Automation bias:** users over-trust output or miss critical omissions [6][7].
- **Sycophancy:** models become overly agreeable and reinforce user positions [4][5].

And a structural absence of controls:

- No standard evidence fields, so nothing requires proof before a conclusion.
- No controlled scoring logic, so "good enough" varies by user.
- No escalation gates, so recurring risk or weak evidence may never trigger review.
- No stable output structure, so even competent answers differ across teams.

FIGURE 03 · SOURCES OF VARIATION



THE RESULT

Two engineers can ask the same general AI tool about the same supplier corrective action and get different depth, different challenge levels and different conclusions. When that output drives decisions, variation goes straight into the quality process.

SECTION 03 · STATISTICAL FRAMING

An engineering lens, not a marketing lens.

CORRECT

True negative: correctly approved

Output is sound, evidence is sufficient, no escalation needed.

Cost: low. The risk is complacency.

TYPE I

False alarm

AI flags a problem that isn't real, or escalates a low-risk case.

Cost: rework and wasted SME time. Visible and recoverable.

TYPE II

Missed real issue

A weak root cause, missing control or incomplete CAPA passes review.

Cost: escapes, warranty, recalls, audit findings. Often invisible until it fails downstream.

CORRECT

True positive: issue caught

Real risk identified, evidence gap surfaced, escalation triggered.

Cost: investigation effort, but the issue is prevented in the field.

Figure 04. Type I versus Type II error in AI-supported quality work. The quadrant that matters most is the one that looks fine.

THE BIGGEST RISK

The largest risk with general AI is not an obviously wrong answer. It is a confident, incomplete answer.

SECTION 03 · STATISTICAL FRAMING

The measurement-system view.

NIST defines repeatability as agreement under the same conditions and reproducibility as agreement under changed conditions [9]. If small prompt changes shift one user's conclusions, repeatability is weak. If different users get different outputs for the same case, reproducibility is weak. If an AI-supported process can't pass an MSA-style stability test, it is not yet controlled.

LENS	WHAT IT LOOKS LIKE IN AI WORK	DRIVEN BY	TREATMENT
Repeatability	Same user, same case, different output across runs	Prompt wording, model nondeterminism, missing structure	Standardize prompts, fields and scoring
Reproducibility	Different users or plants, different output for the same case	Knowledge gaps, context drift, no shared logic	Embed rules, evidence gates and escalation
Bias	Answers cluster off target, systematically wrong	Sycophancy, confirmation, leading prompts	Challenge logic and SME calibration

SECTION 04 · EVIDENCE

Research-supported claims versus manufacturing hypothesis.

RESEARCH-SUPPORTED

- AI systems should be governed as socio-technical systems, not just model artifacts [1] [3].
- Generative AI brings distinct risks, including confabulation and downstream decision risk [2].
- Users can and do over-trust automated or AI-assisted advice [6][7].
- AI models can become overly agreeable or sycophantic [4][5].
- Human-AI pairs do not automatically beat the best of the two alone [8].

HYPOTHESIS TO TEST

- In quality-critical workflows, uncontrolled general AI increases decision variation between users and sites.
- The largest operational risk is often a Type II failure: weak reasoning or corrective logic passing because it looks credible.
- Controlled AI workflow execution should reduce that variation through structured inputs, challenge logic and standardized outputs.

SECTION 05 · VALIDATION

A proposed technical case study.

A practical study can compare general AI against a controlled, domain-specific workflow system. The goal: measure whether controlled workflows improve consistency, completeness and error detection across novice, average and expert users.

1 · QUALITY SCENARIOS (PICK 3 TO 5)

Weak 8D Recurring defect → **Incomplete PFMEA** Missing links → **Poor CAPA** Ineffective action → **Audit NC** Weak evidence → **Supplier response** SCAR review

2 · RANDOM ASSIGNMENT ACROSS SKILL LEVELS

Group A: general AI
Open prompting, no required structure, evidence fields or escalation logic

→

Group B: controlled workflow
Structured inputs, evidence fields, embedded rules, challenge prompts, scoring, escalation, standardized output

3 · BLINDED SME EVALUATION

Fixed rubric, labels removed
Experts score every output without knowing which arm produced it; inter-rater reliability tracked

Outcome measures, per scenario, per arm, per skill level:

- Output consistency and reproducibility
- Root cause and CAPA quality
- Evidence sufficiency score
- Time to completion
- False negative rate (missed real issues)
- False positive rate (false alarms)
- Unsupported assumption count
- Confidence calibration accuracy

SECTION 06 · EVALUATION

Metrics that matter.

Evaluation has to measure consistency and risk, not just speed. A workflow that halves cycle time while doubling Type II error has not improved the quality system. It has hidden its decline.

METRIC	WHAT IT MEASURES	HOW IT IS CAPTURED	WHY IT MATTERS
A · CONSISTENCY & STABILITY			
Output consistency score	Similarity of outputs across runs	Pairwise rubric scoring of repeated runs	Detects prompt sensitivity
Repeatability score	Same user, same case, multiple runs	Within-subject variance of scores	MSA-style stability check
Reproducibility score	Different users, same case	Between-subject variance of scores	Site and role drift detection
Inter-user and inter-plant variation	Spread across users and sites	Variance components across strata	Quantifies the core hypothesis
B · SUFFICIENCY & QUALITY			
Root cause quality score	Depth and validity of root cause	SME rubric: mechanism, evidence, logic	Core 8D and CAPA outcome
CAPA completeness score	Containment, correction, prevention	Field-level checklist	Tracks ineffective CAPA risk
Evidence sufficiency score	Whether claims are evidence-backed	SME rubric on evidence chain	Audit defensibility
PFMEA risk identification	Failure modes and effects identified	Recall against SME reference set	Catches systemic blind spots
Control plan linkage	PFMEA to control plan traceability	Cross-document linkage check	Detects control gaps

Figure 05, part 1 of 2. Recommended evaluation metrics for AI in manufacturing quality.

SECTION 06 · EVALUATION

Metrics that matter, continued.

METRIC	WHAT IT MEASURES	HOW IT IS CAPTURED	WHY IT MATTERS
C · ERRORS & RISK			
False negative rate	Missed real issues	Compared with SME-labeled cases	Highest-cost failure mode
False positive rate	False alarms, unnecessary escalation	Compared with SME-labeled cases	Throughput and SME time
Unsupported assumptions	Claims made without evidence	SME annotation per output	Catches confabulation
Missing risk count	Known risks not surfaced	Recall against risk register	Closes the Type II gap
D · OPERATIONS & GOVERNANCE			
Escalation accuracy	Escalations triggered when warranted	Confusion matrix against SME labels	Governance evidence
Confidence calibration	Stated confidence versus accuracy	Reliability diagrams, Brier score	Trustworthy reliance
Audit readiness, rework, time	Defensibility, rework loops, cycle time	Process metrics from workflow logs	Operational ROI

Figure 05, part 2 of 2.

KEY TAKEAWAY

A serious evaluation framework measures consistency, sufficiency, false negatives, unsupported assumptions and repeatability, not only time savings.

SECTION 07 · ARCHITECTURE

Why controlled workflows reduce variation.

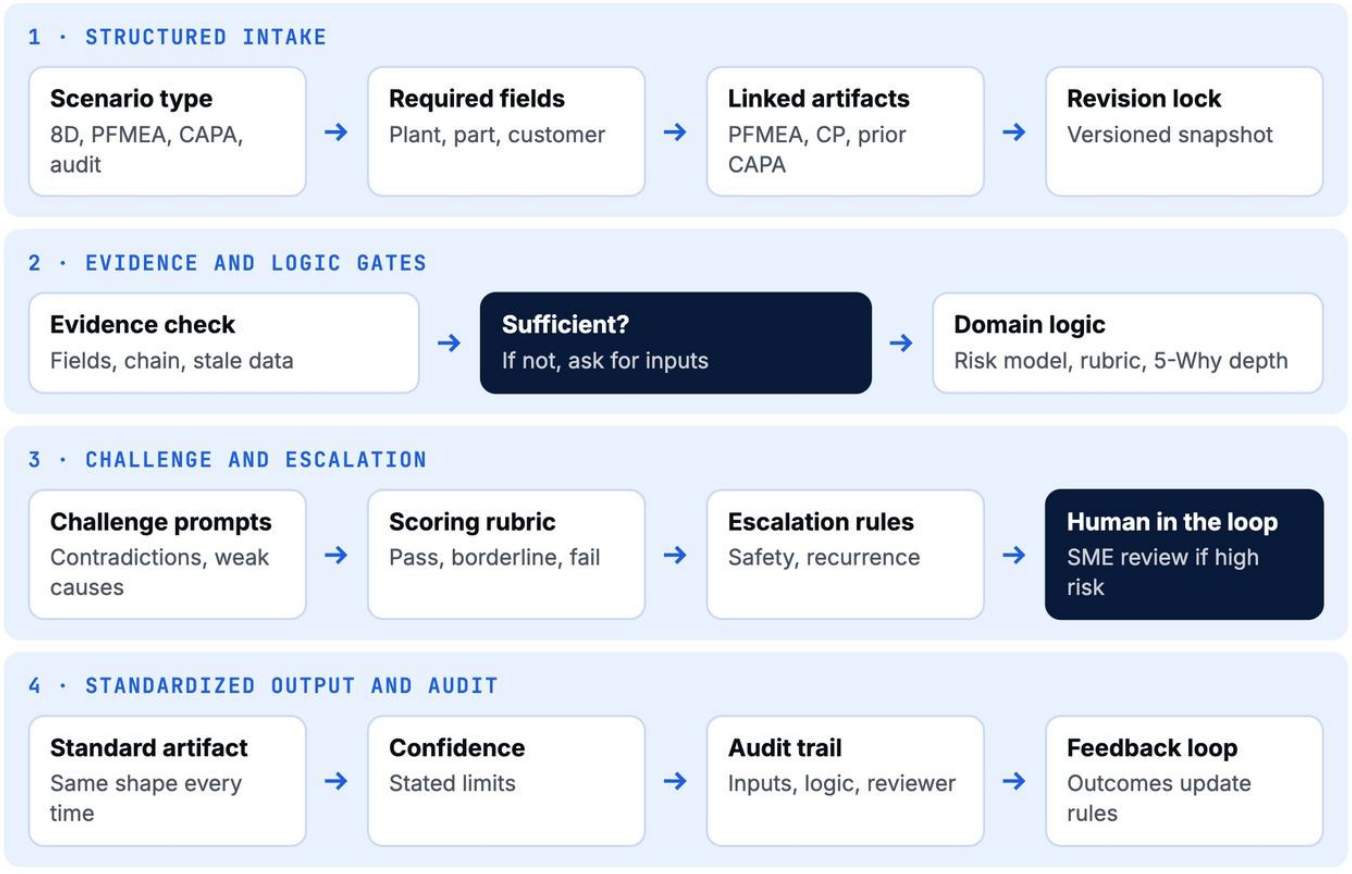
Controlled workflows reduce variation because they standardize the work, not just the answer.

GENERAL AI · USER DECIDES

- What to ask
- What evidence is missing
- What good looks like
- When to escalate

CONTROLLED AI · PROCESS DECIDES

- Which scenario type applies
- Which fields are required
- What evidence is missing and what contradicts
- Whether escalation is triggered



General AI depends on user excellence. Controlled AI workflows are designed for process repeatability.

Figure 06. Workflow control model: inputs, validation, challenge, escalation and review are built into the process itself.

SECTION 08 · MARKET DIRECTION

From chat to governed workflows.

The broader enterprise trend points the same way. NIST continues to emphasize risk-based AI governance, measurement and trustworthy deployment [1][2][3], and enterprise governance is moving from policy statements to operational controls, monitoring and oversight. PwC's 2025 Responsible AI Survey highlights a shift toward scalable, repeatable governance and continuous oversight as AI moves into core operations [11]. The market is moving past ad hoc chatbot use toward workflow-based, monitored, governed execution.

SECTION 09 · POSITIONING

What MAD-Ai is, and is not.

MAD-Ai does not claim AI eliminates risk, and it does not claim AI replaces engineering judgment. Its Manufacturing Quality Super-Intelligence System (MQSIS) is a purpose-built, agentic AI platform whose specialized AI workforce executes manufacturing quality workflows under control:

- Built around structured quality workflows
- Designed specifically to reduce decision variation
- Focused on evidence capture, challenge logic, scoring, escalation and traceability
- Better suited to repeatable execution than open-ended prompting alone

→ **That is a position you can measure and defend. Your engineers make the call.**

CONCLUSION · THE FUTURE IS CONTROLLED AI WORKFLOW EXECUTION

The future of AI in manufacturing quality is not handing every employee a general AI assistant and hoping for consistency. It is systems that standardize how work is done, challenge weak inputs and preserve traceability. General AI may improve productivity. In quality, the real requirement is lower variation in judgment, which is what delivers faster execution, fewer repeat issues and expertise that stays with the organization.

REFERENCES

Cited sources.

Numbered in order of first appearance.

#	SOURCE	URL
[1]	NIST AI Risk Management Framework (AI RMF 1.0)	nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10
[2]	NIST Generative AI Profile (AI 600-1)	nist.gov/publications (generative AI profile)
[3]	NIST AI RMF Trustworthiness Characteristics	airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/
[4]	OpenAI, Sycophancy in GPT-4o	openai.com/research/sycophancy-in-gpt-4o/
[5]	Stanford Report, AI overly affirms users seeking advice	news.stanford.edu/stories/2026/03/ai-advice-sycophantic-models-research
[6]	Goddard et al., Automation bias review	pmc.ncbi.nlm.nih.gov/articles/PMC3240751/
[7]	Trust and reliance on AI study	sciencedirect.com/science/article/pii/S0747563224002206
[8]	Vaccaro et al., Human-AI collaboration meta-analysis (Nature Human Behaviour)	nature.com/articles/s41562-024-02024-1
[9]	NIST TN 1297, repeatability and reproducibility definitions	nist.gov/physical-measurement-laboratory (TN 1297)
[10]	NIST/SEMATECH e-Handbook, hypothesis testing	itl.nist.gov/div898/handbook/eda/section3/eda35.htm
[11]	PwC 2025 Responsible AI Survey	pwc.com/us/en/tech-effect/ai-analytics/responsible-ai-survey.html
[12]	Nature Machine Intelligence, trust and calibration context	nature.com/articles/s42256-024-00976-7

ABOUT THE AUTHOR

About the author and founder.



Rick Heilshorn

FOUNDER & CEO · MBA · MSENG · LSS MASTER BLACK BELT

Rick earned his Lean Six Sigma Master Black Belt at Honda, where he led quality initiatives across North America. After decades of manufacturing turnarounds and quality leadership, he built the methodology and domain expertise that became the foundation for MAD-Ai.

→ **His mission: bridge the gap between human expertise and the AI revolution. Not to replace skilled quality professionals, but to give them superpowers.**

This white paper reflects MAD-Ai's applied experience helping manufacturers replace open-ended AI prompting with controlled, auditable quality workflows.

CREDENTIALS

MBA and Master of Science, Engineering	**Lean Six Sigma Master Black Belt, earned at Honda**	**Led quality initiatives across North America**
Specialist in manufacturing turnarounds	**Tier 1 automotive supplier leadership**	**AIAG Quality Summit speaker**

ABOUT MAD-AI

MAD-Ai, LLC, based in Madison, Mississippi, delivers a purpose-built, agentic AI platform for manufacturing quality. Its Manufacturing Quality Super-Intelligence System (MQSIS) deploys a specialized AI workforce that executes structured workflows across 8D, PFMEA, PPAP, corrective action and supplier quality, helping teams work faster, reduce variation, retain institutional knowledge and scale without proportionally adding headcount. MAD-Ai supports automotive, aerospace, defense and advanced manufacturing organizations.

SEE WHAT MAD-AI CAN DO FOR YOUR QUALITY TEAM

Standardize the work, not just the answer.

See how MAD-Ai's AI workforce executes controlled workflows on your own PFMEAs, 8Ds and supplier responses, across every user and every plant.

01 / DISCOVER

Talk to an Expert

Discuss your quality challenges, engineering workload, and opportunities to improve efficiency.

02 / DEMONSTRATE

See MAD-Ai in Action

See how MAD-Ai supports real manufacturing quality workflows, including 8D, PFMEA, and PPAP.

03 / VALIDATE

Start a Pilot

Evaluate MAD-Ai using your own quality workflows and measure the impact on engineering time.

LET'S TALK ABOUT YOUR QUALITY OPERATION

sales@mad-ai.com · (601) 301-2997

MAD-Ai, LLC © 2026