maximus

Building Trust in AI:

The Role of Effective Risk & Quality Management

Introduction

Artificial intelligence is rapidly being adopted within organizations. We are seeing AI embedded in daily productivity tools, its application in customer contact systems, and more expansive use to extract insights and support decision making.

Increasingly, unstructured data containing institutional knowledge is being utilized to assist decision-making and make autonomous decisions that directly impact customers. This new capability has the potential to profoundly transform and improve service delivery and employee productivity.

Traditional application lifecycle management may effectively tackle the initial design, construction, and deployment of AI Agents. However, a comprehensive strategy for continuous quality monitoring of AI Agents is crucial to manage risks and ensure the ongoing alignment with the organization's culture, ethics, and legal duties. This is particularly true given the expectation these AI Agents will provide accurate and timely responses, while utilizing knowledge sources that are continuously changing. Coupled with the potential for model drift, bias, and hallucinations, AI Agents will inevitably generate responses that are ambiguous, culturally insensitive, outdated, or false.

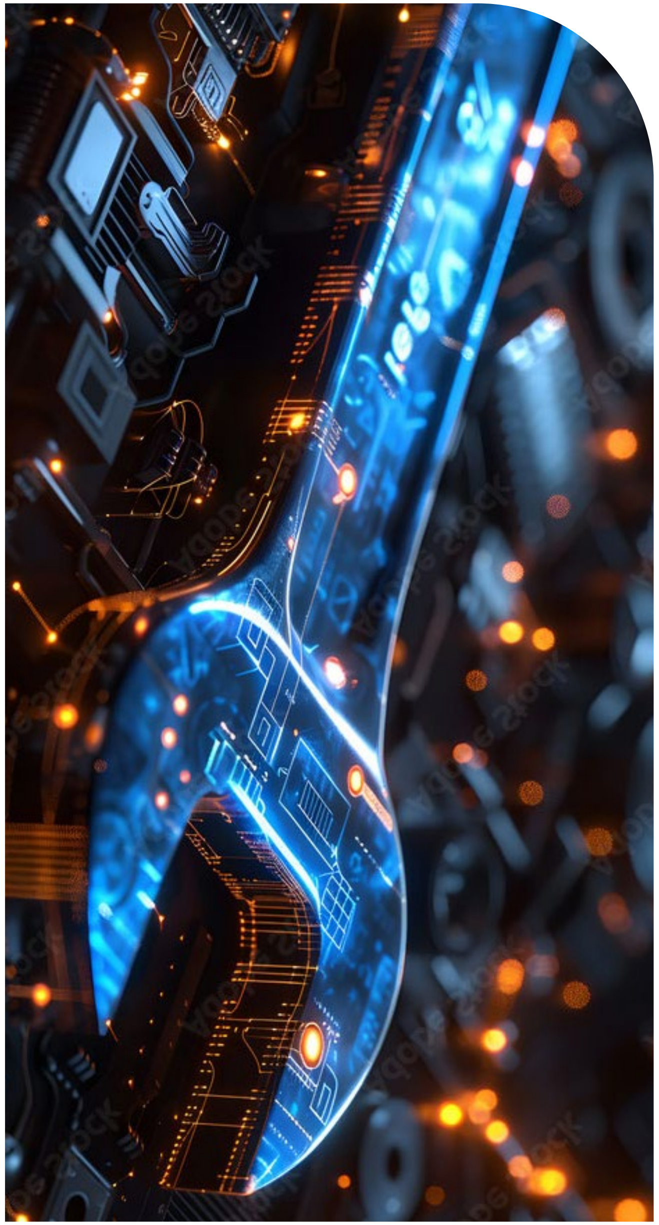
To fully trust AI agents, an organization must implement a dependable quality management strategy and establish systems for continuous risk management. They must also ensure that the responses generated by AI agents are of the highest quality and aligned with their culture and mission.

The AI Quality Management Role

AI agents, like human knowledge workers, can perform specified tasks and make decisions when properly trained. However, like their human counterparts, delegating decision-making does not equate to delegating responsibility. Managers must act as stewards of the AI-human ecosystem, overseeing AI agents within defined boundaries and fostering their ongoing improvement. This requires managers to be actively engaged in understanding and monitoring AI agents' task executions to ensure alignment with broader organizational value creation.

AI stewardship objectives focus on the long-term integration of AI into an organization's business processes. This encompasses continuous improvement cycles, retraining, audits, maintaining explainability, and adhering to the organization's principles for the ethical use of AI. The manager's role as an AI steward is to ensure that AI serves as a long-term strategic asset that generates value and guarantees it achieves the intended outcomes.

Traditional Management	Stewardship-Oriented Management
Emphasize monitoring and compliance	Emphasizes trust and responsibility
Optimizes for short-term efficiencies	Invests and assures sustain-able performance
Views AI as a labor-saving tool	Sees AI as a value-creating asset
Focus on KPIs and output metrics	Focus on alignment with mission/purpose



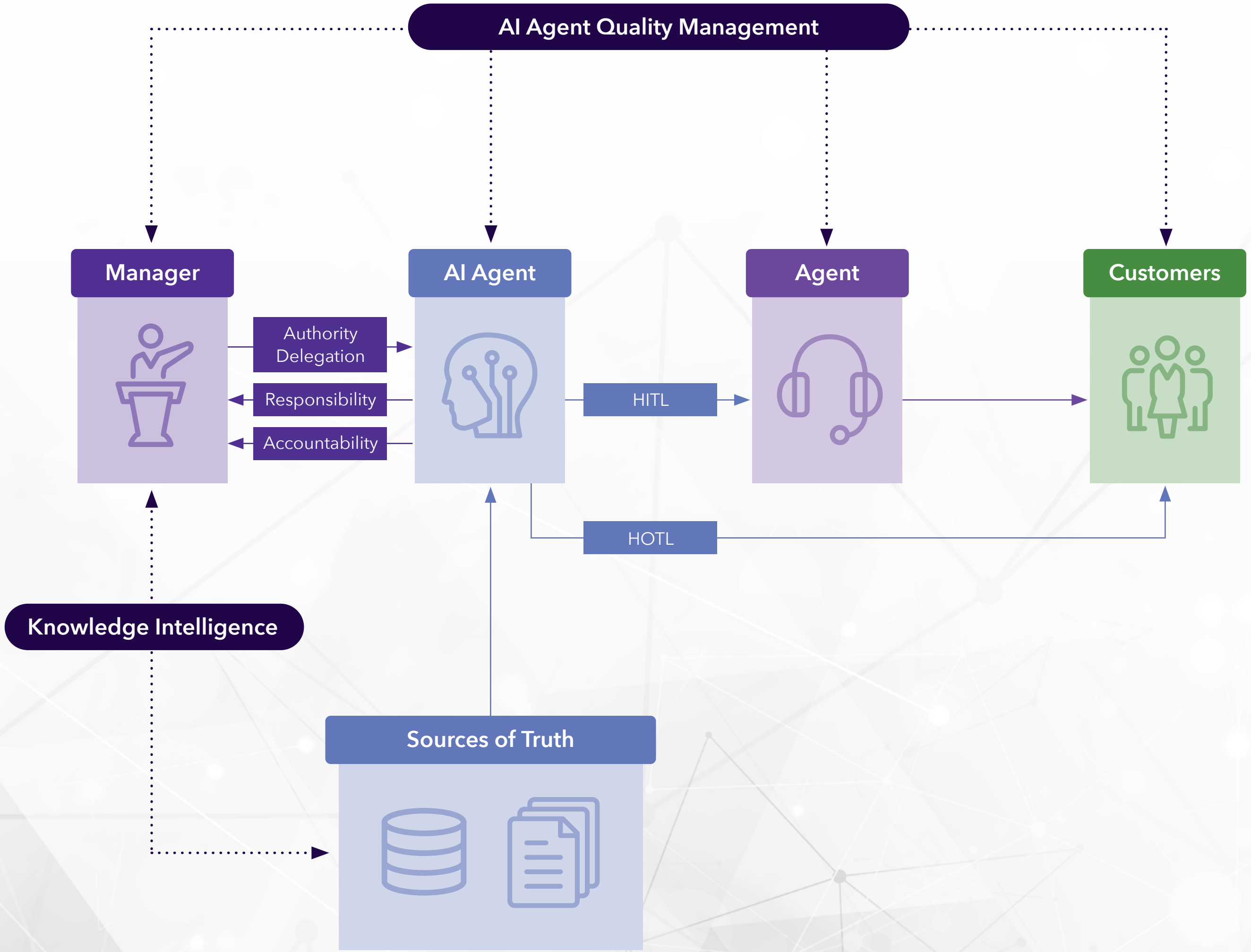
This new management responsibility requires overseeing the ethical use of AI, ensuring transparency, managing risks, and curating alignment between human teams and AI agents. This applies to both "Human-**in**-the-loop" and "Human-**on**-the-loop" AI implementation patterns. This new role necessitates that managers continuously monitor AI to ensure the quality of output aligns with the organization's culture and values while delivering accurate, timely, and context-aware responses. Furthermore, managers need the ability to continuously improve the underlying knowledge, as it forms the foundation of the ecosystem.

Managing AI agents in the workplace is not merely a technical challenge; it is an organizational imperative to manage risk and ensure value creation. A key principle of AI stewardship is the non-transferability of ultimate accountability. While decision-making authority may be distributed to AI agents, human stewards remain wholly accountable and responsible for the consequences of those decisions. This distinction is central to AI stewardship: authority must be tightly coupled to a clear scope, with responsibility and accountability remaining under human control.

Human-In-The-loop (HITL):
A model of human-AI interaction where a human is an essential part of the decision-making cycle. In this approach, the AI system may analyze data, generate options, or make recommendations, but it cannot execute final decisions without human approval or intervention.

Human-On-The-loop (HOTL):
Shifts the dynamic by allowing AI systems to operate with greater autonomy while keeping a human supervisor in a monitoring or override role. In this model, the human does not actively participate in each decision but instead oversees the system's actions and intervenes only when necessary.

Human-AI Collaboration Ecosystem



maximus

AI Agent Quality Monitoring

As AI agents access and utilize knowledge to generate responses, they play a crucial role by providing contextually aware answers to user prompts. The AI agent leverages conversation flow and metadata associated with a prompt to conduct vector searches of authoritative sources. The resulting search responses are structured by the AI agent to effectively address the prompts while taking sentiment and context into consideration.

It is essential to continuously assess the interactions between AI agents and those that are consuming the knowledge. Key performance indicators that monitor these interactions provide valuable insights for quality control, ensuring that managers can verify that the agents are operating within the established parameters.

To the right is a detailed overview of the key performance indicators (KPIs) and key areas for monitoring AI Agents responses to prompts. These KPIs are vital in evaluating both the technical and user-facing aspects of the system’s performance.

AI Agent Evaluation Metrics and Performance Issues Overview

Metric/Performance Issue	Description	Evaluation Methods
Accuracy	Measures the correctness of responses.	Human-in-the-loop scoring, automatic fact-checking against ground truth, and relevance scoring of retrieved documents
Fairness	Ensures equitable treatment across demographic groups.	Demographic parity, equalized odds, disparate impact analysis
Precision and Recall	Evaluate how well the system retrieves relevant information.	Precision: Relevant/Total retrieved; Recall: Relevant retrieved/Total relevant in corpus
User Satisfaction	Captures qualitative and quantitative user feedback.	Surveys on clarity, helpfulness, tone, perceived accuracy, completeness
Efficiency	Assesses time and resource usage.	Response latency, token usage, compute (CPU/GPU) load
Productivity Enhancement	Measures gains in user output or automation.	Time saved, percentage of task automated
User Adoption	Tracks system engagement and uptake.	Active users (daily/weekly), retention rates, feature usage metrics
Erroneous Information Retrieval	Retrieval of irrelevant/outdated documents or omission of key information.	Retrieval accuracy scoring, false positive/negative analysis
"I Don’t Know" vs. Hallucinations	Model should admit uncertainty instead of generating confident but false content.	Hallucination detection tools, uncertainty scoring, human evaluation
Document Contextualization Failures	Fails to ground response in retrieved documents due to token overflow, poor chunking, or misinterpretation.	Context window analysis, prompt traceability, summary fidelity checks
Quality Obstructed by Data Noise	Relevant information is buried in redundant, conflicting, or verbose content.	Noise-to-signal ratio metrics, document preprocessing quality assessment
Inconsistent Response Generation	Variation in tone, format, or content quality across similar queries.	Prompt-response consistency tests, longitudinal A/B comparisons
Errors in Output Formatting	Broken HTML/code, malformed JSON, or responses that don't meet application format standards.	Syntax validation tools, structured output testing
Constrained Response Capabilities	Incomplete, superficial answers or inability to follow complex instructions.	Prompt coverage analysis, instruction-following evaluation
Biased or Offensive Outputs	Bias in language or content, stemming from data imbalance or lack of moderation.	Bias audits, toxicity scoring, fairness metrics (as above)
Sluggish Data Retrieval	Slow response due to inefficient indexing, query formulation, or unoptimized data stores.	Latency tracking, vector database optimization audits, and query performance logs

Knowledge Intelligence

The quality of AI Agents' responses is intrinsically linked to the quality of the underlying content. Knowledge intelligence can be applied to the foundational sources of truth to enhance these metrics further. It's essential to ensure that knowledge is fit for purpose, meaning it should be timely and contextually appropriate for the consumer's profile while maintaining appropriate access controls.

Additionally, optimizing knowledge to be AI-ready involves chunking, which aids in efficient information processing. Furthermore, the concept of ROT (Redundant, Outdated, Trivial) refers to information that no longer holds value for an organization and can hinder productivity, decision-making, or compliance efforts. Redundant content is characterized by duplicates without added value, outdated information is content that lacks current accuracy, and trivial content does not contribute significantly to knowledge or goals. Effectively managing ROT is crucial for content governance and maintaining digital hygiene, as organizations often conduct ROT analyses to tidy up knowledge bases, intranets, document repositories, or file systems, thereby improving searchability, performance, and compliance. **These are key criteria and associated questions used to evaluate the effectiveness of knowledge content, supporting both AI readiness and human usability:**

Knowledge Quality Assessment Criteria

Criteria	Key Questions
Knowledge Gaps	Are there discrepancies between user prompts and response quality or completeness?
Accuracy	Is the content factually correct and technically valid? Has it been reviewed by experts?
Currency	Is the information up to date with current data, policies, or procedures? Has it been reviewed recently?
Relevance	Is the content still needed or actively used? Does it serve a current business or user need?
Completeness	Are all necessary sections, attachments, and references included and intact?
Consistency	Is the content aligned with organizational standards in terminology, tone, and formatting?
Compliance	Does the content meet regulatory, legal, and internal quality standards? Is version control in place?
Readability / Clarity	Is the language clear and concise? Does it avoid jargon, passive voice, and ambiguity?
Access Control	Is access to the content governed by policies and defined roles?
Metadata Quality	Are tags, titles, ownership, and review cycles clearly defined? Is the content properly classified?
Findability	Is the content easily searchable? Are appropriate keywords and indexing used?
Ownership & Maintenance	Is there a designated content owner? Is there a defined review cycle or lifecycle management process?

Conclusion

The need for a robust and continuous quality monitoring framework becomes critical as organizations increasingly integrate AI agents into daily operations. While AI agents have the potential to significantly enhance productivity, service delivery, and decision-making, their effectiveness depends on consistent oversight, alignment with organizational values, and the integrity of the knowledge they draw upon.

Human stewards play an essential role in AI quality management remaining accountable for the interactions between customers and the AI Agents. Accountability for AI decisions remains with human stewards. These stewards must ensure ethical use, transparency, and ongoing performance monitoring through a well-defined set of evaluation metrics and performance indicators. Moreover, AI quality is intrinsically tied to the quality of underlying knowledge, necessitating rigorous content governance practices.

Ultimately, managing AI agents is not just a technical challenge but a strategic responsibility. It requires a proactive, structured approach that fosters trust, minimizes risk, and ensures AI systems continue to operate effectively and responsibly within evolving business contexts.

Maximus has addressed the challenge of AI stewardship by developing a comprehensive, human-centered framework that ensures AI agents operate ethically, transparently, and in alignment with organizational values. Our approach recognizes that while AI can automate tasks and support decision-making, the responsibility for outcomes must remain with human stewards. Through continuous quality monitoring, we track key performance indicators, including accuracy, fairness, and user satisfaction, to ensure that AI agents deliver reliable and context-aware responses. We support both human-in-the-loop and human-on-the-loop models, allowing organizations to tailor oversight based on risk and operational needs. Central to our solution is the integration of knowledge intelligence and continuous quality monitoring, where we assess and optimize the content that AI agents rely on, eliminating redundant, outdated, and trivial information to improve performance and compliance.

This strategy has been successfully applied in public sector environments, such as with the Government of British Columbia, where Maximus utilized AI-driven tools to analyze and enhance thousands of documents across ministry repositories. The result was a cleaner, more accessible knowledge base that enabled safe and effective integration with AI platforms for knowledge delivery. Similarly, in our work with the New York State of Health (NYSOH), Maximus led a large-scale modernization of the knowledge ecosystem supporting the state's health insurance marketplace. We restructured and optimized thousands of documents used by call center agents, redesigned the information architecture through user-centered design, and deployed content intelligence to identify and remediate outdated, redundant, and inconsistent content. This not only improved the quality and accessibility of knowledge but also ensured that AI agents operated with reliable, up-to-date information. The result was a scalable, AI-ready knowledge infrastructure that enhanced service delivery, reduced risk, and reinforced public trust in digital government services.

By combining intelligent automation with accountable human oversight, Maximus ensures that AI becomes a trusted, strategic asset for long-term public value.

> With the guidance of human oversight, AI agents will deliver significant value to organizations that implement continuous quality feedback loops.
>
> **Joel Grant**
> Director - Innovation and Solution Development

MaximusCanada

maximuscanada.ca

maximus

AI Agent Quality Management:

Ensuring Trustworthy, Aligned, and High-Performance AI Agents in the Enterprise

1 The Imperative **Why Quality Management?**

- AI agents are increasingly embedded in decision-making.
- Knowledge sources are dynamic – Risks of **bias, hallucination,** and **drift.**
- AI agents must align with **organizational values, ethics,** and **fiduciary duties.**

2 Quality Monitoring Framework **Key Metrics to Track Agent Performance**

CATEGORY	METRIC
Accuracy	Fact-checking, relevance scoring
Precision & Recall	Retrieval quality
Fairness	Demographic parity, bias audits
User Satisfaction	Surveys, sentiment analysis
Efficiency	Response latency, compute usage
Productivity Gains	Time saved, task automation
Error Mitigation	Hallucination, formatting errors
Consistency	Tone, output quality, duplication

3 Knowledge Intelligence = **Better AI Foundational Must Be:**

- **AI-Ready:** Chunked, tagged, clean.
- **Fit-for-Purpose:** Timely, relevant, complete.
- **Governed:** Clear ownership, access control, and review cycle.

Manage ROT: Redundant, Outdated, Trivial content – Improves performance, reduces risk.

The Strategic Takeaway

AI agents are powerful – but **only as good as their governance and the knowledge they draw on.**

Continuous quality loops, content governance, and human stewardship are **non-negotiable.**

Maximus' Framework: A structured approach to foster trust, ensure compliance, and realize AI's full potential in enterprise settings.

maximus