Kognitos™


# Trusted AI

## An AI Transformation Guide for Enterprise Centers of Excellence

Download Now

Kognitos™

# What is Required to Use AI Safely in the Enterprise?

1. AI Enabled Processes Should Use Logic Where Certainty is Required
2. AI Tools Should Provide an Auditable Record of Human and Model Actions
3. AI Should Pause and Obtain Human Guidance When Uncertain
4. AI Should Provide Controls to Prevent Hallucinations

Enterprise transformation, digital, strategy and IT teams face a perplexing challenge: How to meet the mandates set by C-Suite to develop and implement “AI” or more specifically “Generative AI” strategies, while adhering to the guidelines and uncertainty set forward by legal and infosec organizations.

Because of this challenge, a recent survey reported that only 10% of organizations have any sort of AI solution actually deployed, and most of those in non-critical functions like personal productivity or marketing.

This paper provides practitioners and executives guidance for what to consider in choosing and implementing AI safely in your organization, and actionable directions for how to get started.

# What is Required to Use **AI Safely** in the Enterprise?

First, it must be said that AI, due to its predictive and intuitive basis, fails to operate with 100% certainty or accuracy. True to our form, humans also acting on intuition also make mistakes. Conversely, fully logical and deterministic systems (computers), are able to achieve both certainty and accuracy.

In this respect, each of the following requirements to deploy a "Trusted AI" system in the enterprise, are based on a relative spectrum, not a hard set of specific criteria or thresholds. Yet, they should prove useful for Centers of Excellence stakeholders in need of direction.

# 1. AI Enabled Processes Should Use Logic Where Certainty is Required:

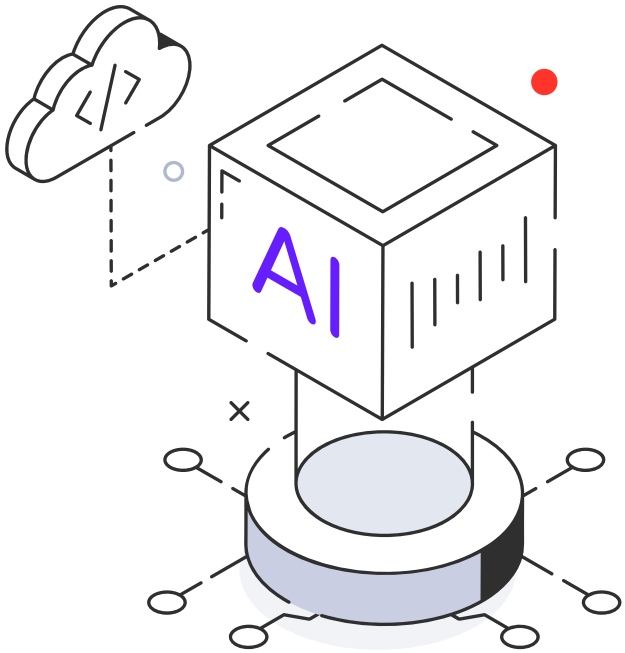
Today, many organizations are experimenting with Proofs of Concept (POCs) where they are leveraging large language models (LLMs) to automate business processes. However, very few of these processes ever make it past a POC stage and into production. One of the major reasons for this, is that most of the time, the uncertainty level of the LLM outputs are too high when compared with the risk level. LLMs are not deterministic or logical systems. Today they use incredible pattern recognition and prediction to generate outputs and take actions, but they are, as many put it, "a black box." It is not clear why the LLM took the steps it took. Similarly, these steps cannot be "tweaked' by a user, and may vary from prompt to prompt. Because of this, LLMs by themselves are not sufficient for the task of automating mission critical business processes. Instead, a system that is certain is required.

## Solution

The human brain is a masterful machine because of its ability to combine logic and intuition. When certainty is required, humans use logic to think through each step and execute an action based on the logical chain. This is hard work, but will result in the same action every time, or the closest to certainty we can enact. It also provides us with a record of how such an action or decision came to be.

For mission critical processes, especially in areas like finance and accounting, Enterprises must use Trusted AI systems that operate using deterministic computing. The steps of a process that require consistency and certainty, must use a logical system to execute each task, so that organizations can have confidence that at scale (millions of runs or more), the process will function the same way, every single time. Prediction and pattern recognition is not sufficient in these cases, and will instead lead to significant errors, and lack of control.

## 2. AI Tools Should Provide an Auditable Record of Human and Model Actions:




In the Enterprise records are important. Detailed past decisions and actions must be preserved for both regulatory compliance and for assessing areas for improvement. Today this often does not occur at a human level as people forget to document things properly. However in a world of LLMs acting independently, the problem could be far more problematic. AI running at a scale of millions or more and even communicating across multiple models, could have a drastic impact on an organization without incorrect actions being documented. Trying to remedy such a situation would truly be untenable.

### Solution

Trusted AI systems provide an inherent system of record, enabling authorized users to see exactly what actions are taken by both human and AI. This provides a critical governance control for the Enterprise. If an error or mistake is made, organizations can choose to re-run that particular process, and take any remediation steps required. Or if they simply need to conduct an investigation into the error, each output, each prompt, each data point logged in a simple, easy to understand record for review. This creates and provides an auditable record of all actions, giving the Enterprise a clear view into the risk levels and error rates that are tolerable for any given process.

## 3. AI Should Pause and Obtain Human Guidance When Uncertain:

One of the great fears with AI, highlighted in pop-culture films over several decades including Terminator, The Matrix and iRobot, is the cold, ruthless execution of decisions by the intelligent system. The fear is that AI could carry out dangerous directives without human review. In a less dramatic but still quite serious context, the Enterprise, LLMs or agents developed on LLMs could execute a decision or action based on its training, that is fundamentally wrong, but to the machine seems plausible and correct without pausing for human input. For example, imagine a reconciliation use case where bank statements are matched with payments plagued by a dizzying array of possibilities and rules. An AI system that does not doubt, could match different combinations of payments and transactions, marking the supplier as having paid, when in fact no such payment has occurred. In this scenario, it's plausible to imagine a system capable of preventing the collection of millions of dollars.

### Solution

Trusted AI systems, instead have the ability to pause and have a degree of "doubt" built into the system, that can be controlled by ranges of certainty. This type of mechanism is already in place in OCR systems, where confidence scores are frequently used to match the comfort level with the risk accompanying an incorrect action. If it is a critical process, a higher confidence score is used. In Trusted AI systems, users can set confidence levels on tasks such as extraction, and additionally, the system will pause when it encounters something it does not expect and await input from a human.

# 4. AI Should Provide Controls to Prevent Hallucinations

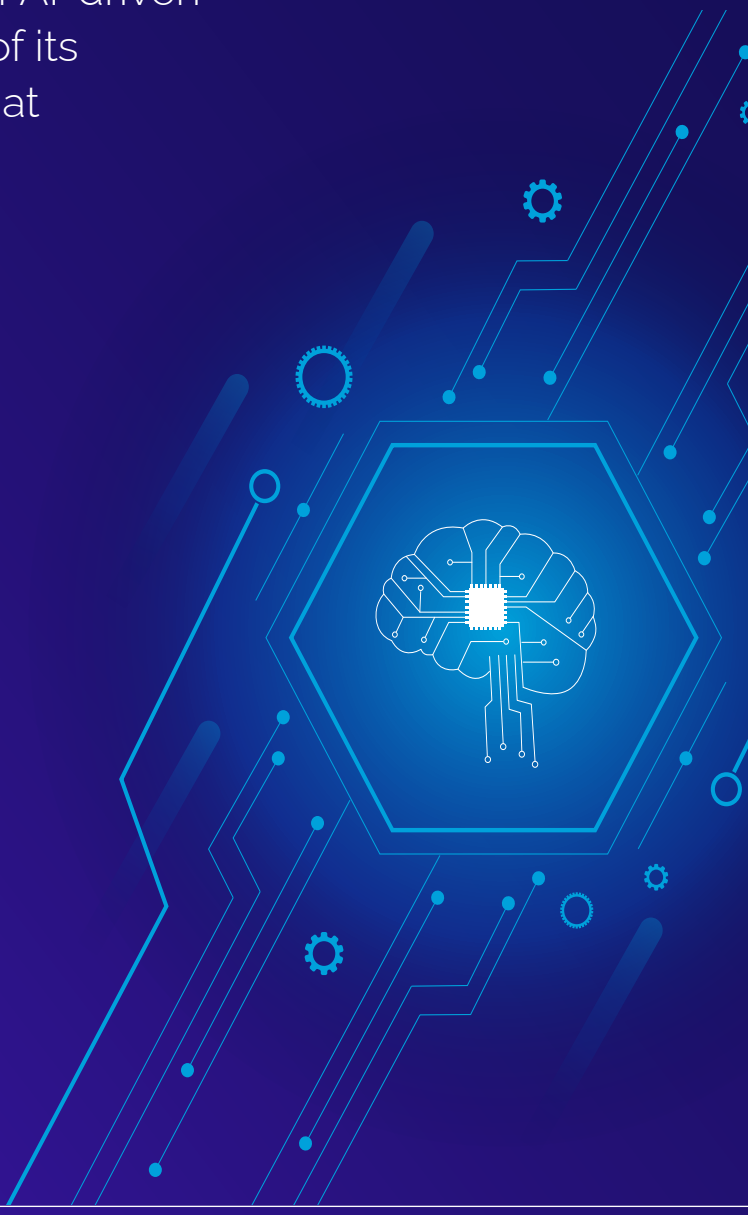
Hallucinations are likely one of the most well discussed challenges of putting AI projects into Enterprise production. The news has reported widely on "hallucination" issues with many of the more common large language models. To put simply, a hallucination is an output of a generative AI model that is factually incorrect, or nonsensical. These exist because the model attempts to intuit or provide a response beyond its training even if it doesn't understand the question or factually know the answer. It is easy to see why hallucinations would create major problems for mission critical processes. If random, nonsensical outputs are common, the system can't be trusted.

## Solution

To be clear, hallucinations are not a problem that are solved today. Instead, building guardrails around hallucinations is the best way to address this behavior today and still use LLM outputs in production. Kognitos has several guardrails in place for scenarios where LLMs are needed to extract or generate content in a workflow. First, Kognitos logs every LLM prompt and response in each run's workflow. This gives an organization an auditable record of all LLM activity at scale, in case any mistakes need to be addressed. Second, users can build steps into a Kognitos workflow to guide the system in understanding if an output is sufficient to process without engaging a human, or if human review or input is required. Expected results can be defined in the step after a prompt is generated and a matching function set to occur. For example, "the output should be in the form of xyz" If the LLM output is similar to the expected result, Kognitos will proceed. If not, due to the confidence functionality described above, Kognitos will prompt a user for help.

# CONCLUSION

The adoption of AI, particularly Generative AI, within enterprise environments necessitates a nuanced approach that balances innovation with managing risk. While the allure of AI-driven automation is undeniable, the reality of its implementation reveals challenges that demand careful consideration by organizational leaders. This sentiment drives the need to deploy “Trusted AI” systems that prioritize logical certainty, auditable transparency, human oversight, and safeguards against hallucinations. By adhering to these principles, organizations can navigate the complexities of AI transformation, minimizing potential pitfalls and maximizing the benefits of intelligent automation.



Kognitos™

## About Kognitos

Kognitos is the private & safe Generative AI to automate any business process in real-time, using plain human language. Robotic process automation and workflows rely on consultants, data scientists, software engineers and IT staff to model and mimic existing business processes. This traditional approach is resource, time and $ intensive, and does not fully address the major pain points in automation: conversational exception handling and document processing. Kognitos' Generative AI solution, Koncierge, self-learns and adapts to existing business processes and works as a force-multiplier within business units and centers of excellence, enabling the business users to focus on informed business decisions and supercharging their capabilities to stay ahead in the age of AI.

Contact Us
www.kognitos.com