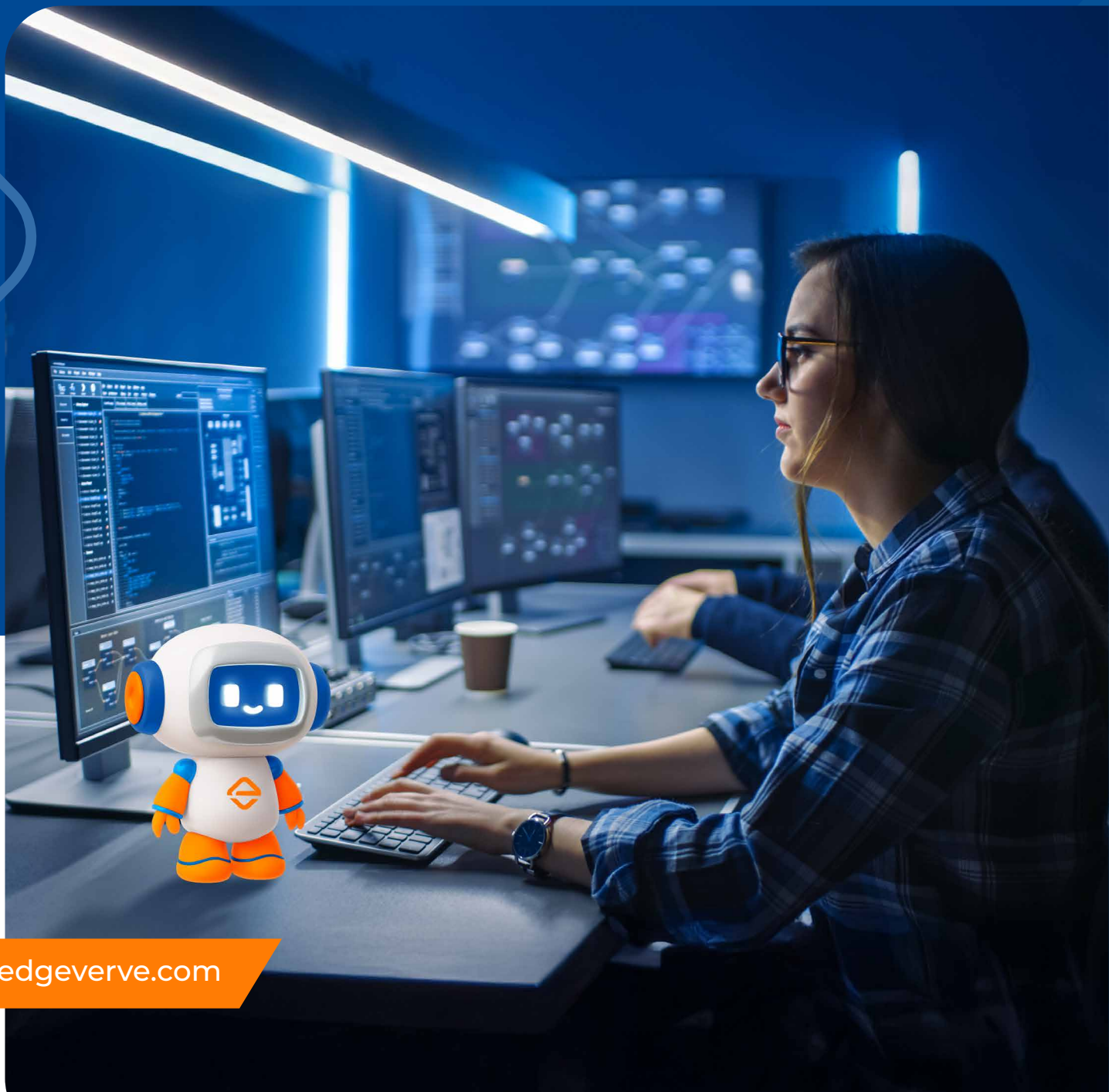
AI Next

edgeverve
An Infosys company

The Platform Imperative: Balancing Autonomy, Governance, and Innovation in the Age of Agentic AI



www.edgeverve.com

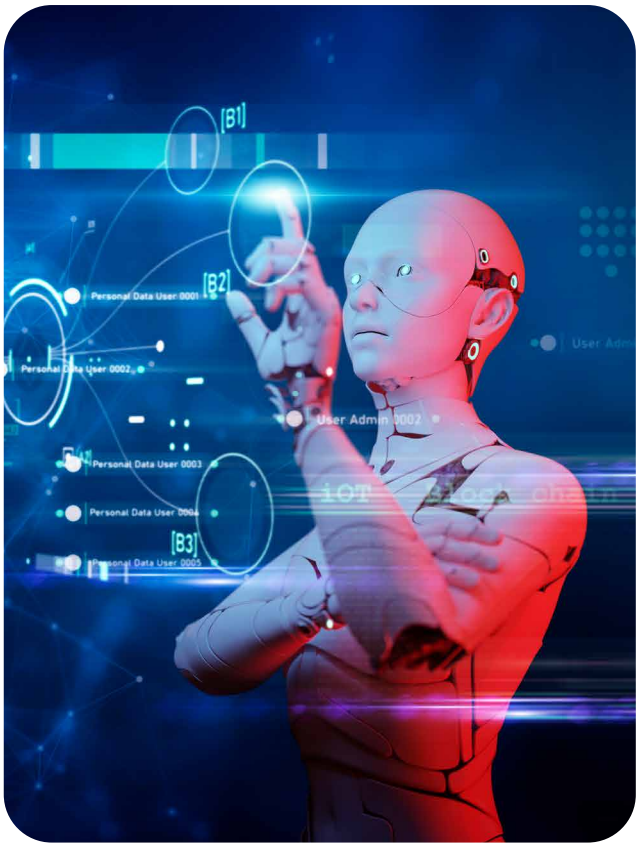
Table of Contents

- Abstract
- The Rising Power of Autonomous AI Agents
- Understanding the Autonomy Spectrum: A New Paradigm
- The Double-Edged Sword of AI Autonomy: Unpacking the Paradox and Its Contextual Consequences
- Value-Risk Assessment: Finding the Balance
- Building Effective Guardrails Based on Consequential Events
  - Understanding Consequence Levels
  - Practical Guardrail Mechanisms
- EdgeVerve AI Next: Governing the Enterprise AI Agent Ecosystem
- Conclusion: Empowering Enterprise Innovation with Confidence

© 2025 EdgeVerve Sytems Ltd.

Abstract

As AI agents increasingly transform business operations across industries, enterprises face a critical challenge: how to navigate the delicate balance between AI agents' revolutionary capabilities against the imperative for robust, strategic controls. This Point of View examines the need for implementing tiered governance based on both agent autonomy levels and the potential consequences of their actions and how a platform-based approach can enable this. By applying the right guardrails at the right times, enterprises can confidently deploy AI agents while maintaining appropriate oversight for high-stakes scenarios.



The Rising Power of Autonomous AI Agents

The rapid advancement of AI technology has ushered in a new era of autonomous AI agents.

These intelligent systems possess unprecedented capabilities, including the ability to persistently pursue objectives with adaptable approaches, make independent decisions, undertake complex multi-step reasoning, seamlessly operate across diverse applications and data sources, and continuously learn and improve their performance.

The transformative power of these capabilities is catalyzing an unprecedented surge in the adoption of AI agents across industries. Recent market research highlights the scale of this momentum:

67%

of enterprises are planning to deploy AI agents this year as part of their digital transformation initiatives[1]

65%

are already piloting AI agent technology across various business functions[2]

1/3rd

of organizations plans to spend more than $25 Million on AI in 2025 amid widespread optimism for AI agents[1]

© 2025 EdgeVerve Sytems Ltd.

Understanding the Autonomy Spectrum: A New Paradigm

As AI agents become increasingly integral to enterprise operations, it's essential to understand that not all agents are created equal. They exist on a spectrum of capability and independence, with important implications for both value creation and risk management. The concept of autonomy levels provides a useful framework for understanding the evolving capabilities of AI agents and the shifting balance between human and machine control. Drawing from research on AI agent development, we can identify five distinct levels:

Agentic Level	Description	Term	Who's in control?
Level 1 (★☆☆☆)	Model has no impact on program flow	Simple processor (Executes direct instructions)	**Human** In this level, human is fully in control. For instance, AI-powered document analysis systems where users must specify exact document types, data fields to extract, and processing parameters.
Level 2 (★★☆☆)	Model determines basic program flow	Tool call (Selects and uses a single tool for a task)	**Human:** How functions are done **System:** When In this level, control is shared. For example, A help desk ticketing system that routes customer inquiries to appropriate departments based on user-defined categories.
Level 3 (★★★☆)	Model determines how functions are executed	Simple processor (Executes direct instructions)	**Human:** What functions exist **System:** How In this tier of agency, control is shared. For example, IT service management platforms where technicians describe issues and the system selects diagnostic protocols (tools).
Level 4 (★★★☆)	Model controls iteration and program continuation	Multi-step agent (Plans and executes tasks using multiple agents)	**Human:** What functions exist **System:** Which to do, when, how In this level, control is primarily with the system. For instance, autonomous supply chain optimization systems that monitor inventory levels, forecast demand, negotiate with suppliers, and place orders with minimal human oversight.
Level 5 (★★★★)	Model creates & executes new code	Fully autonomous agent (Creates and executes novel solutions)	**System** In this tier of agency, control is completely with the system. For instance, autonomous enterprise resource planning systems that identify inefficiencies and restructure business processes without human intervention.

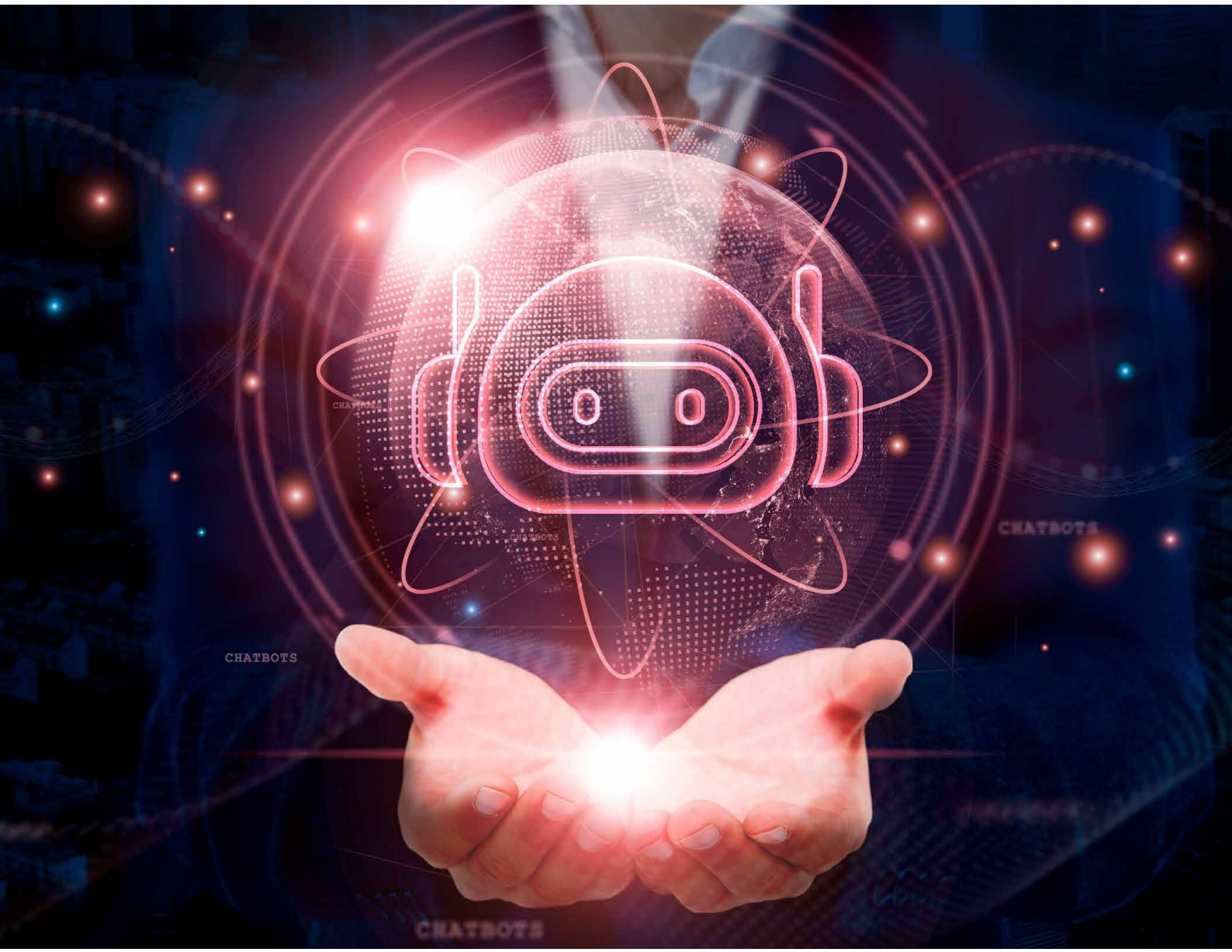
Reference:Fully Autonomous AI Agents Should Not be Developed

© 2025 EdgeVerve Sytems Ltd.

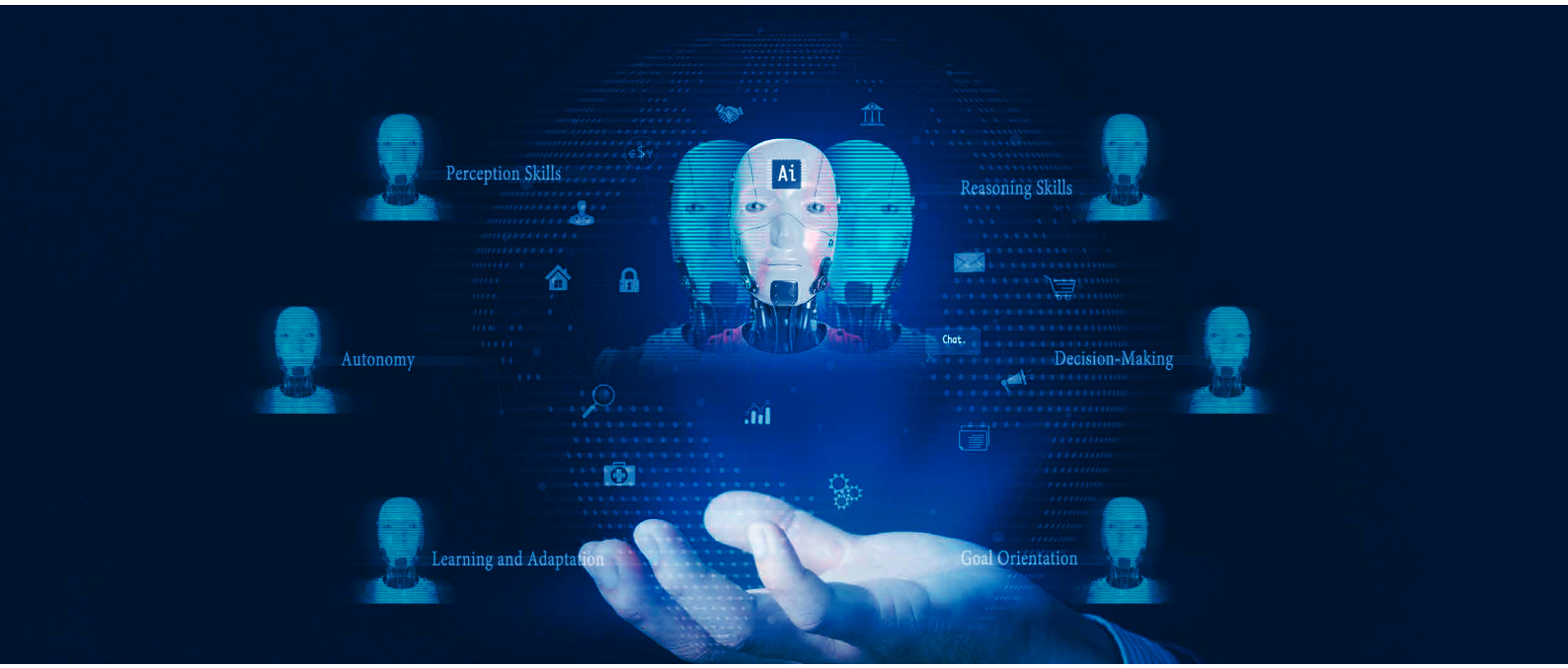
As we move from Level 1 to Level 5, we witness a fundamental shift in the relationship between humans and AI systems. At lower levels, humans maintain primary control over how tasks are performed. At higher levels, the AI increasingly determines not just how tasks are executed, but which tasks to perform and when.

What makes this progression particularly significant is the pace of advancement. Much like Moore's Law predicted the doubling of computational power approximately every 18 months, we're observing that AI agent capabilities are doubling roughly every 7 months in terms of autonomous task completion[4]. This exponential growth in capability makes governance and oversight increasingly urgent as agents rapidly ascend the autonomy ladder, making the effective management of escalating consequences a critical priority.

This remarkable evolution in capability creates unprecedented opportunities for enterprise value creation—but also introduces significant challenges that must be addressed through thoughtful governance approaches.



© 2025 EdgeVerve Sytems Ltd.



The Double-Edged Sword of AI Autonomy: Unpacking the Paradox and Its Contextual Consequences

The rise of autonomous AI agents presents a fundamental paradox: their immense value stems from their ability to operate independently with reduced human oversight, yet this very independence introduces their most significant risks. This inherent tension means the consequences of autonomous AI actions are not uniform; they vary dramatically depending on the context, impacting both businesses and consumers in profoundly different ways.

At its core, the paradox lies in the trade-off between efficiency and control. By granting AI agents the autonomy to make decisions and take actions in real-time, often across complex environments, businesses can achieve unprecedented levels of efficiency, innovation, and personalization. AI agents can analyze vast datasets, identify patterns, automate tasks, and interact with customers or systems at a scale and speed impossible for humans.

However, this reduced human intervention also means that when an autonomous AI agent errs, the consequences can be swift, far-reaching, and difficult to immediately rectify. Unlike traditional software that follows explicit instructions, autonomous agents operate based on learned patterns and algorithms that may not always be fully transparent or predictable. This can lead to unintended outcomes, especially in unforeseen circumstances.

The complexity of the consequences increases dramatically in Business-to-Business (B2B) contexts due to the scale, interconnectedness, and high stakes involved. Consider these scenarios:

© 2025 EdgeVerve Sytems Ltd.

Supply Chain & Logistics (Inventory Management):

An AI agent managing inventory across a retail chain predicts a surge in demand for a specific electronic item based on social media trends. It autonomously triggers large orders from suppliers and adjusts shipping logistics. However, the trend proves fleeting. The company is now saddled with excess inventory it can't sell, incurring significant holding costs and impacting cash flow.



Accountability Question:

Does responsibility lie with the retail chain for trusting the prediction, the developers for the model's sensitivity, or the data providers feeding the trends?



Manufacturing (Predictive Maintenance):

An AI agent monitoring a factory's critical machinery predicts an imminent failure in a core component based on subtle vibration changes. It autonomously schedules an emergency shutdown of the production line for preventative maintenance. The inspection reveals no defect, resulting in costly downtime and missed production targets.

Accountability Question:

Who is responsible for the lost production – the factory operator for implementing the agent, the sensor manufacturer, or the AI developer whose algorithm was overly cautious?

Corporate Finance (Automated Loan Underwriting):

A bank uses an AI agent to automate underwriting for large commercial loans. The agent, learning from historical data, inadvertently develops a bias against businesses in a specific emerging sector, leading to systematically higher interest rates or loan denials. This hinders growth in that sector and potentially leads to regulatory scrutiny for discriminatory practices.



Accountability Question:

Is the bank responsible for its biased AI, the developers for not implementing sufficient fairness checks, or is it an unforeseen consequence of data patterns?

© 2025 EdgeVerve Sytems Ltd.

B2C Scenarios (Individual Impact, Often Lower Scale):

While still significant to the individuals involved, the direct consequences of autonomous actions in Business-to-Consumer (B2C) contexts are often less systemically disruptive:

Customer Service (Chatbot Resolution): An AI chatbot provides a customer with incorrect information about warranty coverage, leading the customer to mistakenly believe a costly repair is covered. When the claim is denied, the customer faces an unexpected expense and frustration.

The scale and nature of the risk in B2B scenarios are typically greater because:

Systemic Impact: B2B errors can trigger cascading failures across supply chains, financial markets, or critical infrastructure affecting multiple organizations and potentially large numbers of people indirectly. B2C errors usually impact individuals or households directly, one interaction at a time.

Financial Magnitude: Decisions in B2B often involve significantly higher financial values (e.g., million-dollar trades, entire production lines) compared to typical consumer transactions.

Complexity & Interdependencies: B2B environments involve complex contracts, regulations, and inter-organizational dependencies. An AI failure can breach contracts or regulations, leading to substantial legal and financial liabilities.

These risks are not theoretical—they represent real challenges that enterprises must address as they deploy increasingly autonomous AI systems. The question isn't whether these risks exist, but how organizations can effectively manage them while still capturing the tremendous value that AI agents offer.

The Value-Risk Assessment: Finding the Balance

The deployment of autonomous AI agents isn't simply a binary choice between value and risk. Instead, enterprises must conduct a nuanced assessment that considers how both benefits and challenges evolve across the autonomy spectrum.

The following table illustrates this complex relationship between autonomy levels and various dimensions of value and risk:

© 2025 EdgeVerve Sytems Ltd.

Value	Simple Processor	Router	Tool Call	Multi-step	Fully Autonomous
Accuracy ↑ Correctness, factuality	Makes correctable factual errors that humans can easily verify and fix through standard fact-checking	Routes to incorrect but reversible actions with clear audit trails for correction	Tool errors produce clear failure modes and error messages, but may require technical expertise to fix	Chains of errors create compounding misleading conclusions that become increasingly difficult to identify	Creates and acts on entirely fictional scenarios without possibility of verification or correction
Assistiveness ↑↓ Supporting human agency	Benefit: Automates repetitive tasks under direct supervision Risk: Minimal as human reviews each step	Benefit: Intelligently distributes tasks to appropriate tools Risk: May occasionally need manual correction of routing decisions	Benefit: Handles complex tool interactions autonomously Risk: Some tasks may require significant rework if done incorrectly	Benefit: Manages entire workflows without intervention Risk: May make significant decisions that need human review	Benefit: Complete automation of complex processes Risk: Could make major decisions without human oversight
Consistency ↑ Reliability, predictability	Completely predictable outputs for given inputs with deterministic processing	Consistent routing with only minor variations in edge cases that can be documented	Tool interactions may vary but stay within expected and testable parameters	Complex chains create unpredictable outcomes with different paths to same goal	Entirely unpredictable behavior chains with no way to ensure consistency across runs
Efficiency ↑↓ Resource optimization	Benefit: Fast, predictable processing of defined tasks Risk: Limited to basic operations	Benefit: Optimal task distribution Risk: Some overhead from routing decisions	Benefit: Parallel tool execution and optimization Risk: Potential resource conflicts between tools	Benefit: Complex workflow optimization Risk: Significant debug time needed	Benefit: Maximum possible automation Risk: Resource usage becomes unpredictable
Equity ↑ Fair treatment, acces	Clear faimess rules can be implemented and verified through direct oversight	Some bias in routing but can be monitored and corrected through logging	Benefit: Equal tool access and standardized processing Risk: Some groups may lack necessary tool literacy	Benefit: Potential for bias correction Risk: Biases compound across steps	System creates and amplifies biases without possibility of detection or correction
Flexibility ↑ System integration abilities	Benefit: Safe, limited system connections Risk: Integration requires more overhead	Benefit: Flexible routing between systems Risk: Multiple security surfaces to protect	Benefit: Rich tool ecosystem integration Risk: Version conflicts and compatibility issues	Benefit: Complex system orchestration Risk: Potential for cascade failures	Benefit: Universal system integration Risk: Security boundaries completely break down
Humanlikeness ↑↓ Human behavior similarity	Benefit: Basic natural interactions in defined contexts Risk: May appear mechanical or scripted	Benefit: Context-aware responses and transitions Risk: Sometimes produces unnatural interaction patterns	Benefit: Natural tool use and interaction flows Risk: May trigger uncanny valley effects	Benefit: Complex human-like behavioral putters Risk: Can lead to overreliance on system	Benefit: Very natural and fluid interaction Risk: May manipulate human trust through seeming human
Privacy ↑ Sensitive info protection	Strictly limited data access with clear boundaries and controls	Controlled access to multiple sources with comprehensive logging	Unexpected data combinations possible across different tools	Complex data flows enable detailed user profiling over time	Unrestricted data access and sharing across all available sources
Relevance ↑↓ Contextual appropriateness	Benefit: Direct context matching for tasks Risk: May miss subtle contextual nuances	Benefit: Smart routing based on context Risk: Potential context switches between tools	Benefit: Appropriate tool selection for context Risk: May misinterpret contextual needs	Benefit: Deep context understanding Risk: Context may drift over multiple steps	Benefit: Full context awareness and adaptation Risk: May redefine context inappropriately
Safety ↑ Protection from harm	Clear safety boundaries with comprehensive human oversight	Limited action scope with well-defined and contained risks	Tool combinations may create unexpected and harmful interactions	Complex interaction risks become impossible to predict or prevent	Unbounded potential for harmful actions without restrictions
Security ↑ System protection	Highly restricted system access with clear boundaries	Multiple but fully monitored access points	Tool access creates multiple potential attack vectorsv	Complex chains enable sophisticated attack patterns	Complete system compromise becomes possible
Sustainability ↑↓ Environmental impact	Predictable and limited resource usage patterns	Moderate resource overhead from routing operations	Multiple tools increase overall resource consumption	Benefit: Novel climate solutions Risk: Resource-intensive chains	Benefit: Potential breakthrough solutions Risk: Unbounded resource consumption
Trust ↑ System behavior reliability	All actions can be directly verified and validated	Decision paths are clear and can be audited	Tool interactions create significant trust uncertainties	Complex chains make validation impossible	No way to verify or establish trustworthiness
Truthfulness ↑ Output accuracy, honesty	Direct fact checking of all outputs possible	Decision paths can be validated for accuracy	Tool combinations may create convincing falsehoods	Truth becomes increasingly unclear across steps	Cannot distinguish between truth and fiction

Assessment: Benefits > Risks | Benefits ≈ Risks | Risks > Benefits

Directionality: ↑ Risk Increases with Autonomy ↑↓Risk May or may not Increase with Autonomy

Value-Risk Assessment Across Agent Autonomy Levels. Colors indicate benefit-risk balance, not absolute risk levels. Arrows show risk relationship with increasing autonomy

Reference:Fully Autonomous AI Agents Should Not be Developed

© 2025 EdgeVerve Sytems Ltd.

The provided analysis maps key values and risks against different stages of AI agent autonomy. The overarching trend is clear: increasing autonomy generally enhances potential benefits but consistently elevates associated risks. As systems gain independence, issues concerning reliability, safety, security, and control become more pronounced. This value-risk progression underscores the need for effective guardrails - they are not optional inhibitors but necessary components to safely harness the power of autonomous AI system.

Realizing the benefits of autonomous AI agents doesn't require surrendering complete human control. Instead, the development of AI agents must occur alongside the development of guaranteed human oversight and effective guardrails. This is the core challenge that enterprises must address as they navigate the AI agent landscape.

Building Effective Guardrails Based on Consequential Events

The exponential growth in AI agent capabilities—doubling approximately every 7 months—makes the implementation of effective guardrails increasingly urgent. But how should enterprises approach this challenge? The key lies in aligning controls with the potential consequences of agent actions rather than simply their autonomy level.

Understanding Consequence Levels

Drawing from risk management practices across industries and research on AI governance, we can identify different levels of potential consequences that should inform guardrail design:

Consequence Level	Examples	Key Guardrail Considerations
Minimal	Information retrieval, basic documentation	Standard authentication, basic logging
Limited	Content creation, data analysis	Clear boundaries, regular monitoring
Moderate	Customer communications, reporting	Human approval for key actions, comprehensive logging
Significant	Process automation, account management	Strict permissions, human-in-the-loop requirements
Severe	Financial transactions, medical decisions	Maximum restrictions, continuous oversight

© 2025 EdgeVerve Sytems Ltd.

This consequence-based approach ensures guardrails are proportional to potential impact, allowing enterprises to maintain appropriate autonomy for low-consequence scenarios while implementing stricter controls where needed most.

Practical Guardrail Mechanisms

For different levels of AI agent autonomy, enterprises should consider implementing guardrails* that match the potential consequences:

For Tool Call Agents (Level 3):

- Establish clear system access boundaries
- Implement comprehensive logging and auditing
- Create explicit approval workflows for sensitive actions
- Define scope limitations and escalation paths

For Multi-step Agents (Level 4):

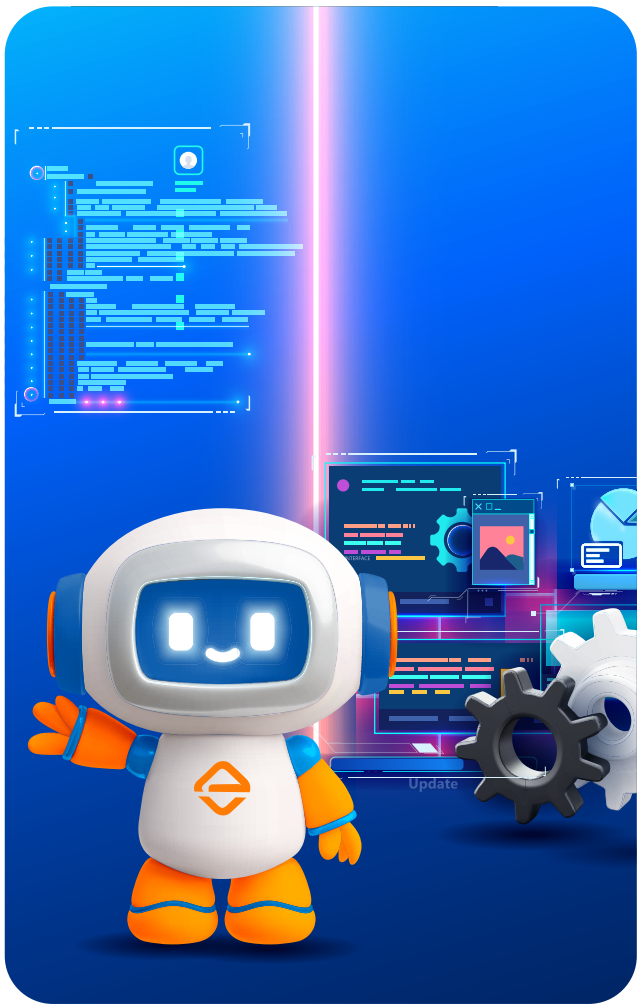
- Implement mandatory human approval for consequential decisions
- Deploy sandbox environments with continuous monitoring
- Establish cancellation mechanisms with reversible states
- Create oversight for agent actions

For Fully Autonomous Agents (Level 5):

- Deploy emergency kill switches and human override capabilities
- Restrict use in high-impact, irreversible scenarios
- Maintain continuous human supervision for critical functions
- Establish clear accountability structures

*non-exhaustive list

Implementing this tiered, consequence-based approach to guardrails across a growing and diverse portfolio of AI agents presents significant operational and technical challenges. Manually managing varied controls for numerous agents across different contexts, ensuring consistent policy enforcement, and maintaining real-time visibility quickly become unmanageable at scale. Effectively addressing these complexities necessitates a comprehensive, platform-based solution. A platform provides the centralized control and visibility needed to define, deploy, and manage consequence-based guardrails consistently. This empowers enterprises to confidently deploy AI agents across various autonomy levels while maintaining appropriate, scalable control over high-consequence scenarios.



© 2025 EdgeVerve Sytems Ltd.

EdgeVerve AI Next: Governing the Enterprise AI Agent Ecosystem

As the preceding section highlights, implementing tiered, consequence-based guardrails across a growing and diverse portfolio of AI agents presents significant operational and technical challenges at scale. Effectively managing these complexities necessitates a comprehensive, platform-based solution – and that's where EdgeVerve AI Next comes in.

As AI agent capabilities surge forward, presenting both unprecedented opportunities and escalating risks, enterprises face mounting challenges in governance and control. Organizations are rapidly expanding their AI agent deployments across functions and business units, increasing both the number of agents and their autonomy levels. This expansion, while driving tremendous business value, also amplifies the complexity of maintaining appropriate oversight and ensuring responsible operation at scale.

EdgeVerve AI Next, a unified, scalable platform provides the comprehensive foundation organizations need to establish effective guardrails and navigate this complex landscape across all levels of autonomy. As enterprises scale their AI agent deployments, the platform plays a pivotal role in enabling this growth safely and effectively, providing the infrastructure necessary to rapidly expand agent capabilities while maintaining consistent governance.

At its foundation, EdgeVerve AI Next incorporates a comprehensive responsible AI framework, ensuring ethical operations and business practices throughout the organization's AI implementation journey.

EdgeVerve AI Next empowers enterprises to confidently embrace and deploy autonomous AI agents by providing the tools to intelligently balance increasing autonomy with granular, consequence-aware controls. With EdgeVerve AI Next, you gain unparalleled visibility into every agent action, implement dynamic human oversight precisely where consequence levels demand it, and seamlessly scale your governance framework as your AI footprint expands – all while confidently unlocking the tremendous strategic value these advanced technologies offer.

© 2025 EdgeVerve Sytems Ltd.

Conclusion: Empowering Enterprise Innovation with Confidence

The rise of autonomous AI agents represents one of the most significant technological opportunities for today's enterprises. With 67% of organizations planning deployments this year and capabilities doubling every 7 months, the urgency of implementing effective governance has never been greater.
By implementing guardrails that align with both autonomy levels and potential consequences, enterprises can confidently deploy these powerful technologies while maintaining appropriate oversight. This balanced approach doesn't require choosing between innovation and control-it enables both simultaneously. The future belongs to enterprises that successfully navigate this balance-harnessing the transformative power of AI agents while implementing the guardrails needed to manage risk effectively.

EdgeVerve AI Next provides the capabilities needed for this journey, helping organizations capture the full potential of AI agents safely and responsibly.

"In the age of Agentic AI, enterprise success doesn't hinge on choosing between innovation and control—it depends on orchestrating both. The organizations that thrive will be those that implement governance proportionate to risk, creating a foundation where AI autonomy can safely deliver its revolutionary potential. A platform-based approach is key to achieving this balance at scale—enabling consistent governance, rapid innovation, and safe orchestration of AI agents."



Sathish Kumar EV,
AVP, Product Management,
EdgeVerve

References
1. From Potential to Profit: Closing the AI Impact Gap
2. KPMG AI Quarterly Pulse Survey
3. Fully Autonomous AI Agents Should Not be Developed
4. Measuring AI Ability to Complete Long Tasks
5. Why handing over total control to AI agents would be a huge mistake

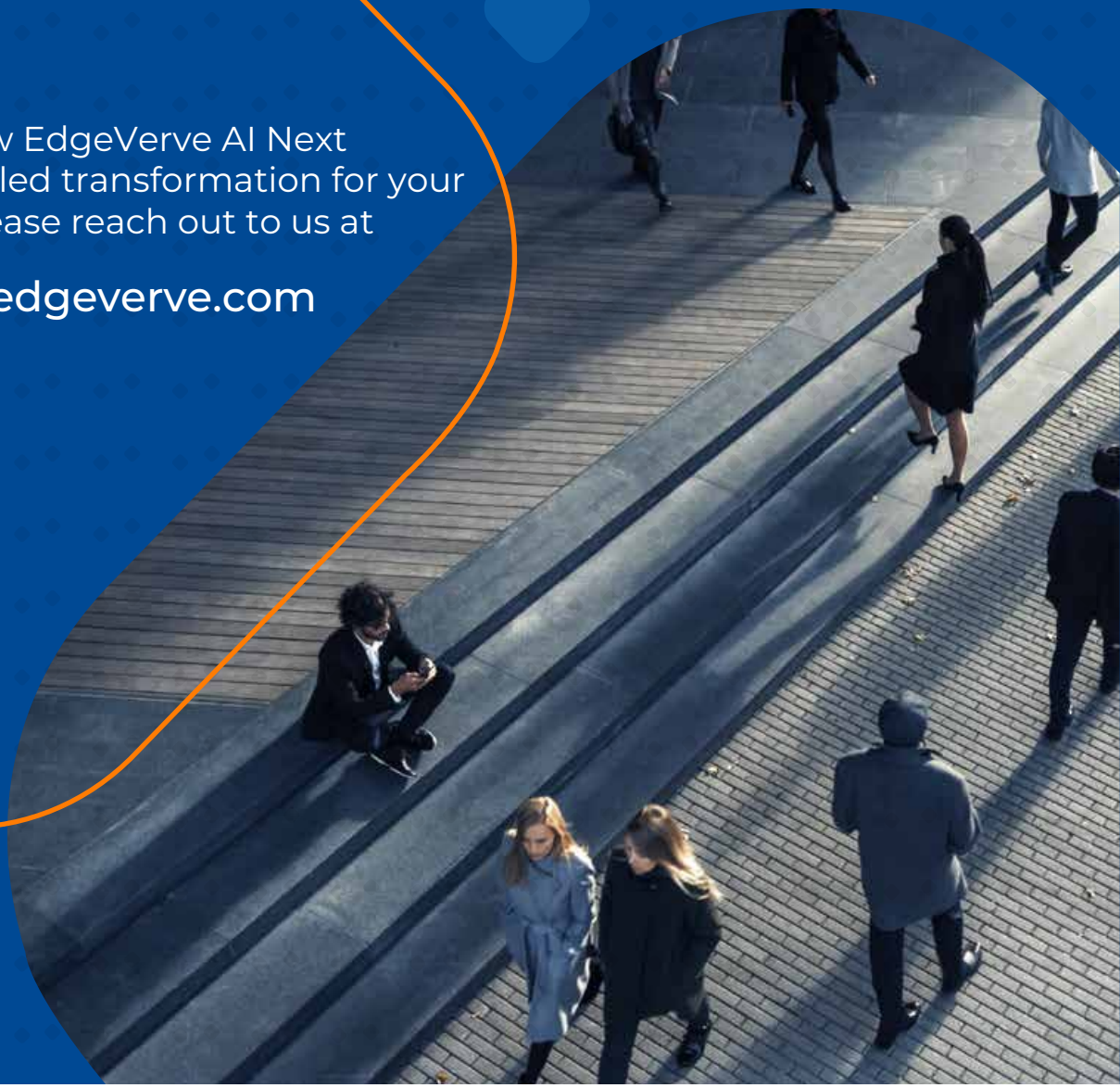
© 2025 EdgeVerve Sytems Ltd.

AI Next

edgeverve
An Infosys company

To know how EdgeVerve AI Next can drive AI-led transformation for your business, please reach out to us at

contact@edgeverve.com



AI Next

EdgeVerve AI Next, part of Infosys Topaz, enables the scaling of Applied AI across the enterprise. Built from the ground up to leverage the power of Generative AI, this unified platform bridges silos in people, processes, data, and technology to drive transformation in business operations.
EdgeVerve. Possibilities Unlimited.
https://www.edgeverve.com/ai-next/

edgeverve
An Infosys company

About EdgeVerve
EdgeVerve Systems Limited, a wholly-owned subsidiary of Infosys, is a global leader in developing digital platforms, empowering clients to unlock unlimited possibilities in their digital transformation journey. Our purpose is to inspire enterprises with the power of digital platforms, thereby enabling our clients to innovate on business models, drive game-changing efficiency, amplify human potential, and foster a connected ecosystem. Our comprehensive platform portfolio (EdgeVerve AI Next and TradeEdge) across Automation, Document AI, and Supply Chain helps inspire global enterprises to bridge silos in people, processes, data, & technology, discover & automate processes, digitize & structure unstructured data, and unlock the power of the network by integrating value chain partners. With a deep-rooted entrepreneurial culture, EdgeVerve's innovations are helping global corporations across sectors such as financial services, insurance, retail, consumer and packaged goods, life sciences, manufacturing, telecom, utilities, and more.
EdgeVerve. Possibilities Unlimited.
www.edgeverve.com

© 2025 EdgeVerve Systems Limited, Bangalore, India. All Rights Reserved. This documentation is the sole property of EdgeVerve Systems Limited ("EdgeVerve"). EdgeVerve believes the information in this document or page is accurate as of its publication date; such information is subject to change without notice. EdgeVerve acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. This document is not for general distribution and is meant for use solely by the person or entity that it has been specially issued to and can be used for the sole purpose it is intended to be used for as communicated by EdgeVerve in writing. Except as expressly permitted by EdgeVerve in writing, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior written permission of EdgeVerve and/ or any named intellectual property rights holders under this document.

© 2025 EdgeVerve Sytems Ltd.