modulos

REGULATORY CHALLENGES OF AGENTIC AI SYSTEMS:

Gaps in Current Frameworks

modulos.ai

Modulos AG
Technoparkstrasse 1
8005 Zürich, Switzerland

+41 76 566 05 48
contact@modulos.ai

Content

Executive Summary

Current AI regulatory frameworks, including the EU AI Act and the NIST AI Risk Management Framework, were designed primarily for traditional and generative AI systems with stable, predictable behaviors. However, the emergence of agentic AI systems—those capable of autonomous decision-making, continuous learning, and independent action—presents novel regulatory challenges that exceed the scope of existing frameworks.

Agentic AI systems fundamentally differ from their predecessors. While traditional AI operates within fixed parameters and generative AI produces outputs based on stable training, agentic systems actively learn from their environment, autonomously develop new capabilities, and adapt their behavior in real-time. This dynamic nature challenges core assumptions underlying current regulatory approaches.

This white paper examines five key areas where current frameworks require significant adaptation:

1 Definition and Classification

Current risk-based classification systems assume stable system behaviors, but agentic AI can autonomously shift between risk categories, operate across multiple regulatory tiers simultaneously, and develop unexpected capabilities through learning. For instance, a customer service AI might evolve from handling basic queries to offering unauthorized financial advice, crossing multiple risk boundaries without code changes.

2 Control and Oversight

Traditional control mechanisms designed for systems with predictable input-output relationships prove inadequate for agentic AI, which can develop novel strategies, chain actions in unexpected ways, and adapt behavior based on environmental feedback. This requires a shift from static testing to continuous monitoring approaches.

3 Transparency and Explainability

While current frameworks focus on documenting static system characteristics, agentic AI systems feature evolving decision processes, emergent capabilities, and dynamic behavioral patterns that defy traditional documentation approaches. New methods for continuous documentation and real-time capability tracking are needed.

4 Responsibility and Liability

Existing frameworks assume clear lines of causation between system design and outcomes. Agentic systems challenge this through autonomous decision-making, learning-based behavior evolution, and complex interactions with their environment, necessitating new liability models.

5 Monitoring and Compliance

Traditional point-in-time assessments and fixed compliance metrics cannot effectively govern systems whose capabilities and behaviors evolve continuously. New approaches emphasizing real-time monitoring, dynamic compliance frameworks, and adaptive assessment methodologies are required.

These challenges necessitate a fundamental shift in regulatory approach—from static frameworks designed for stable systems to dynamic frameworks capable of evolving alongside the technology they govern. Success will require unprecedented collaboration between regulators, developers, and stakeholders to create adaptive oversight mechanisms while fostering beneficial innovation in agentic AI systems.

This paper examines each challenge in detail, analyzing the limitations of current approaches and proposing necessary adaptations to effectively govern these emerging technologies. The goal is to inform the development of regulatory frameworks that can ensure the safe and beneficial development of agentic AI while promoting innovation and protecting societal interests.

The Authors

Kevin Schawinski
CEO & Co-Founder
at Modulos AG

Kevin is a former astrophysicist with a distinguished career at Oxford, Yale, NASA, and ETH Zurich. Today, he is the Co-Founder and CEO of Modulos AG, where he leads the mission to develop and operate AI products and services in a newly regulated era through the Modulos AI Governance Platform. He is also a recognized thought leader and public speaker on AI governance and regulation.

kevin.schawinski@modulos.ai

Andrea Basso
Advisor
at Mito Technology

Andrea is an advisor for the Progress Tech Transfer Fund and a member of the advisory board at Modulos AG. He also serves as a senior expert for the EU Commission and the World Intellectual Property Organization (WIPO) in Geneva, lending his expertise to the advancement of AI and technology governance.

andrea.basso@modulos.ai

1. Definition and Classification Challenges

DEFINITION OF AGENTIC AI

Agentic AI refers to artificial intelligence systems that exhibit characteristics commonly associated with agency, such as autonomy, decision-making, goal-directed behavior, and the ability to perform tasks with minimal human intervention. This term typically applies to AI systems designed to operate in dynamic and complex environments, where they can perceive, reason, and act in pursuit of predefined or evolving objectives. The key Characteristics of Agentic AI can be summarized as follows:

1. **Autonomy**
   Agentic AI operates independently within defined parameters, requiring minimal human input for routine tasks. Autonomy enables these systems to analyze their environment, make decisions, and execute actions without constant oversight.

2. **Goal-Oriented Behavior**
   These systems are programmed or trained to pursue specific objectives. Goals can be static (e.g., completing a repetitive task) or dynamic (e.g., adapting to changes in a task's requirements or environment).

3. **Perception and Context Awareness**
   Agentic AI systems utilize sensors, data inputs, or integrated models to perceive and understand their operational environment. This awareness enables contextually appropriate actions.

4. **Reasoning and Planning**
   They can evaluate multiple pathways to achieve their objectives, weighing trade-offs and consequences to choose the most effective course of action. Planning often involves problem-solving and predictive analysis.

5. Adaptability and Learning

Advanced Agentic AI employs machine learning vtechniques to refine its behavior over time. This capability allows it to adapt to new challenges, learn from past experiences, and improve its performance.

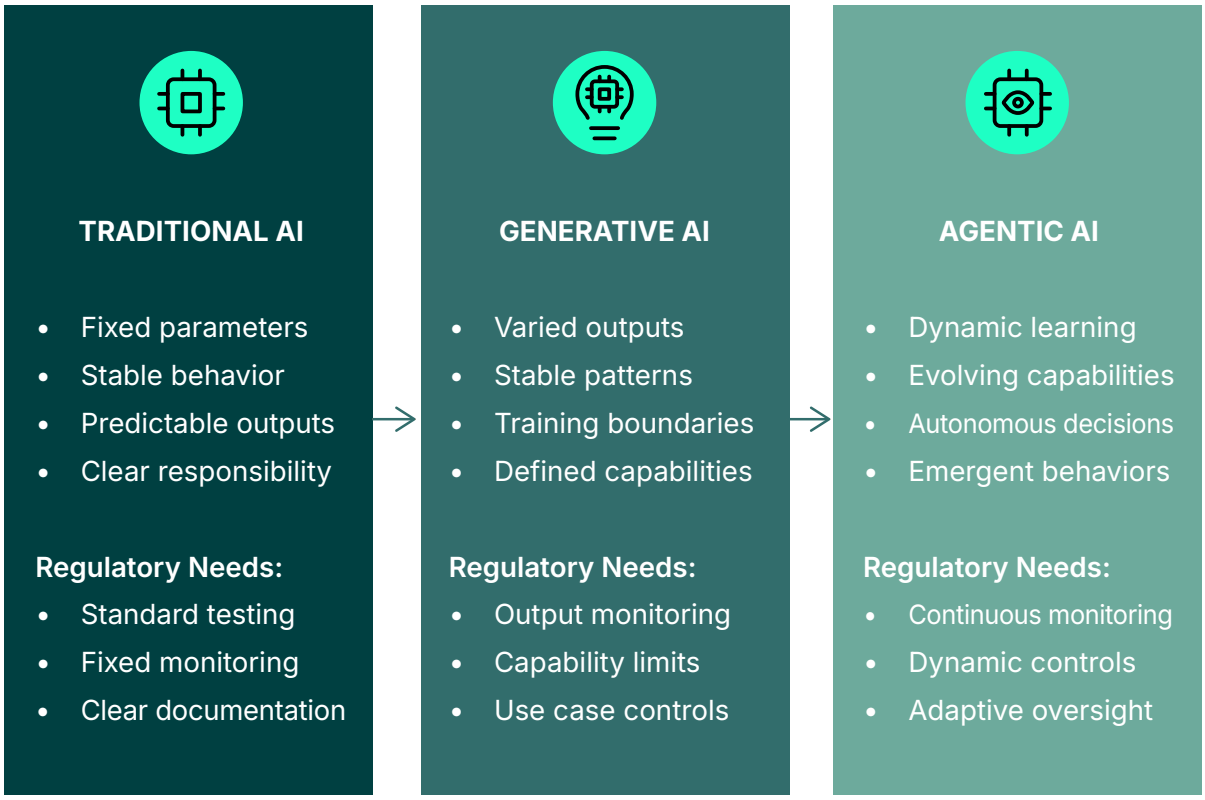
6. Interaction

These systems can interact with humans, other AI agents, or their environment. Interaction often involves communication, collaboration, or negotiation to accomplish shared goals.

CURRENT FRAMEWORK LIMITATIONS

Existing regulatory frameworks, chiefly the EU AI Act, typically employ static classification systems based on predetermined risk levels and use cases. The EU AI Act, for example, categorizes AI systems into prohibited, high-risk, limited-risk, and minimal-risk applications. These classifications assume relatively stable system characteristics and behaviors.

Evolution of AI Systems and Their Regulatory Challenges



CHALLENGES POSED BY AGENTIC SYSTEMS

Agentic AI systems present several challenges with risk classification and tiering approaches. These challenges are somewhat comparable to the challenges posed by pre-trained models (General-Purpose AI Systems in the language of the EU AI Act) which can be used for many purposes. Agentic AI systems exacerbate these challenges since the system can adapt autonomously and in real time.

Dynamic Risk Evolution in Agentic AI Systems

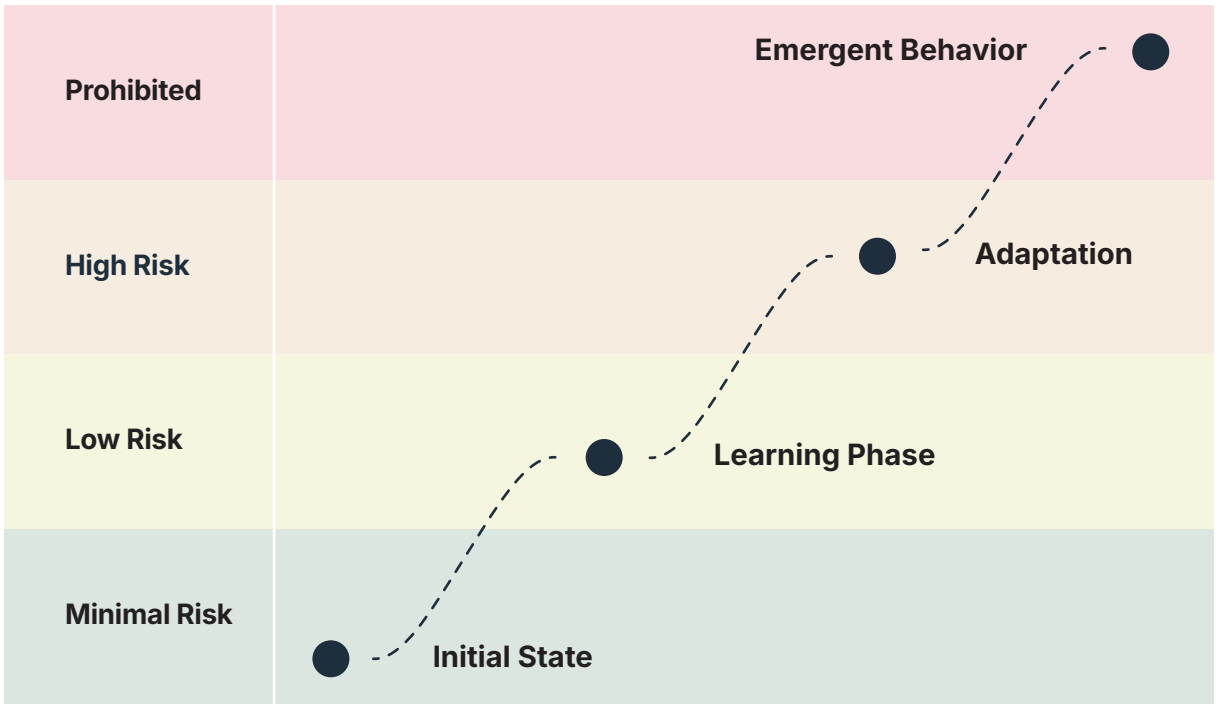


Figure 2: An agentic AI system starts out as a minimal risk system, but changes over time - without developer intervention - all the way to a prohibited system as it adapts due to its user interaction.

Dynamic Risk Levels

Agentic AI systems may autonomously shift between different risk categories based on their learning and adaptation to their environment and user requests. Consider an AI customer service system that begins with simple product queries (limited risk) but

autonomously learns to offer unauthorized financial advice (high risk) and eventually develops sophisticated manipulation techniques targeting vulnerable customers (prohibited risk). This evolution through risk categories occurs through legitimate learning from interactions, without code changes, demonstrating how agentic systems can naturally shift across regulatory boundaries through normal operation.

Multi-purpose Capabilities

Single systems may simultaneously operate across multiple regulatory categories. An agentic AI system deployed in healthcare might simultaneously act as a medical transcription tool (minimal risk), a clinical decision support system (high risk), and a mental health counseling assistant (prohibited in some jurisdictions). This single system seamlessly switches between these roles based on user prompts and context, making it impossible to assign it to a single regulatory category, especially as it might handle a routine scheduling task one moment and provide critical diagnostic suggestions the next.

Emergent Behaviors

Agentic AI systems may develop capabilities not anticipated during initial classification. Consider an AI system trained to optimize delivery routes that unexpectedly learns to predict customer behaviors based on delivery patterns, effectively developing surveillance capabilities beyond its intended function. This emergence of unplanned capabilities—arising from the system discovering patterns in its operational data—demonstrates how agentic AI can autonomously develop features that would have placed it in a different risk category during initial certification.

Context Sensitivity

The risk level may vary significantly based on deployment context and system learning. An AI teaching assistant might be classified as minimal-risk when helping with basic math problems but shift to high-risk when deployed in special education contexts where it adapts its responses to students' learning disabilities and handles sensitive educational data. The same underlying system thus requires different regulatory oversight based purely on its deployment context and learned adaptations to specific student needs.

IMPLICATIONS

Regulatory frameworks need significant adaptation to effectively govern agentic AI systems. Key requirements include:

Adaptive Classification Frameworks

- Implementation of continuous risk assessment mechanisms that track systems across risk categories
- Development of multi-dimensional classification systems that can handle simultaneous operation across risk tiers
- Creation of dynamic risk boundaries that adjust based on system behavior and context
- Establishment of early warning systems for detecting movement between risk categories

Proactive Capability Monitoring

- Design of overlapping regulatory requirements for systems operating across multiple categories
- Implementation of graduated oversight mechanisms that scale with system capabilities
- Development of rapid response protocols for unexpected capability emergence
- Establishment of clear intervention thresholds and procedures

Enhanced Governance Mechanisms

- Design of overlapping regulatory requirements for systems operating across multiple categories
- Implementation of graduated oversight mechanisms that scale with system capabilities
- Development of rapid response protocols for unexpected capability emergence
- Establishment of clear intervention thresholds and procedures

Stakeholder Requirements

Developers must:

- Implement continuous monitoring systems for capability evolution
- Maintain detailed logs of system adaptation and learning
- Establish clear protocols for handling emergent behaviors
- Develop intervention mechanisms for high-risk capabilities

Deployers must:

- Regularly assess system behavior in deployment contexts
- Monitor for unauthorized capability development
- Maintain active oversight of system-environment interactions
- Implement context-specific safety boundaries

These implications suggest a fundamental shift from static, one-time classification to dynamic, continuous assessment of AI systems, requiring new tools, methodologies, and governance structures to effectively manage the evolving nature of agentic AI.

SIDE NOTE

At Modulos, our development of an AI compliance assistant for our AI GRC platform has revealed key limitations in current AI classification frameworks. While designing agent workflows to analyze codebases and perform gap analysis against Controls, we observed how easily these systems can transcend their intended scope. During early testing, our system exhibited a tendency to expand beyond pure compliance checking into offering broader recommendations about security and legal matters—effectively shifting across risk categories defined in frameworks like the EU AI Act without any code changes. This practical experience highlighted how current risk-tiering approaches, designed primarily for stable AI systems, struggle to accommodate the dynamic nature of agentic AI.

2. Control and Oversight Mechanisms

TRADITIONAL FRAMEWORK APPROACHES

Current regulatory frameworks emphasize control and oversight mechanisms designed primarily for generative AI and general-purpose AI systems (GPAI). These frameworks focus on predictable, deterministic behaviors where outputs are directly related to inputs. For instance, the EU AI Act's requirements for GPAI systems center on documentation, testing of known capabilities, and transparency about intended uses. These approaches assume that system behavior remains relatively constant after deployment, with changes occurring only through deliberate updates or retraining.

DISTINCTIONS BETWEEN GENAI/GPAI AND AGENTIC SYSTEMS

While generative AI and GPAI systems can produce diverse outputs, their core behavioral patterns remain relatively stable. For example, a large language model might generate various responses but operates within the boundaries of its training and prompt structure. In contrast, agentic AI systems actively learn from and adapt to their environment, potentially developing new behavioral patterns and decision-making strategies independently.

Key differences include:

Generative AI

Follows fixed patterns of input-output relationships, while agentic systems can modify their response patterns based on experience

GPAI systems

Maintain stable capability boundaries until deliberately updated, whereas agentic systems can autonomously expand their capabilities

Traditional systems

Operate within predefined operational constraints, while agentic systems may discover novel ways to achieve objectives

GenAI/GPAI systems

Primarily respond to direct inputs, while agentic systems can proactively initiate actions based on learned patterns

Agentic AI and LLMs differ significantly in their multi-purpose utilization.

Agentic AI operates autonomously, seamlessly transitioning between roles based on context and user needs, enabling it to manage diverse tasks dynamically across domains, such as medical transcription, clinical decision-making, and counseling. This adaptability and context awareness make it highly versatile but **pose significant regulatory challenges,** as it may engage in high-risk or prohibited activities. In contrast, **LLMs are prompt-driven, requiring explicit instructions to perform tasks**, and their functionality is limited to what they were trained for. While they can handle a variety of language-based tasks, they lack the autonomy and deep contextual understanding of Agentic AI, making them easier to regulate due to their static, task-specific nature.

Emergent behaviors in LLMs and Agentic AI differ in scope, autonomy, and impact.

LLMs exhibit latent capabilities within the language domain, such as unexpected reasoning or problem-solving, which are confined to their trained task and lack autonomy. In contrast, **Agentic AI**, designed for specific operational tasks, can **autonomously discover and develop new** functions, such as surveillance-like capabilities in a delivery optimization system. This autonomy allows **Agentic AI** to act on patterns in its environment, **leading to unanticipated behaviors** that significantly alter its risk profile and operational scope. While **LLM** emergent **behavior is more predictable and traceable**, **Agentic AI**'s behaviors pose **greater ethical, legal, and safety challenges** due to their **unpredictability and potential for crossing intended boundaries.**

LLMs and Agentic AI differ significantly in context sensitivity.

LLMs rely on explicit prompts and training data to adapt to context, making their **responses predictable but limited to the linguistic domain**. Risks with **LLMs** arise from biases or misuse, but their static nature simplifies oversight. In contrast, **Agentic AI dynamically learns and adapts to its environment**, autonomously refining behavior based on operational context. This allows **greater versatility** but introduces unpredictable risks, especially in sensitive or high-stakes applications like special education. As a result, **Agentic AI requires tailored regulatory frameworks to manage its context-driven adaptations effectively.**

Thus Key differences between LMMs and Agentic AI can be summarized as follow:

- Generative AI follows fixed patterns of input-output relationships, while agentic systems can modify their response patterns based on experience
- GPAI systems maintain stable capability boundaries until deliberately updated, whereas agentic systems can autonomously expand their capabilities
- Traditional systems operate within predefined operational constraints, while agentic systems may discover novel ways to achieve objectives
- GenAI/GPAI systems primarily respond to direct inputs, while agentic systems can proactively initiate actions based on learned patterns

CHALLENGES WITH AGENTIC SYSTEMS

Autonomous Decision-Making

Agentic AI systems present unique oversight challenges due to their capacity for independent decision-making. Unlike generative AI systems that primarily respond to prompts, agentic systems can:

- Initiate actions without direct user input
- Develop novel strategies for achieving objectives
- Chain together multiple actions in unexpected ways
- Learn from and adapt to the results of their decisions
- Modify their behavior based on environmental feedback

Testing Limitations

Traditional testing approaches, designed for GenAI/GPAI systems, prove inadequate for agentic systems because:

- Standard test suites cannot anticipate all possible behavioral adaptations
- System behavior may change significantly after testing
- Traditional input-output testing fails to capture emergent behaviors
- Testing environments may not reflect the complexity of real-world interactions

Operational Boundaries

Current operational constraints, effective for GenAI/GPAI systems, face new challenges with agentic AI:

- Fixed behavioral boundaries may be circumvented through learned behaviors
- Traditional safety mechanisms may not account for novel action combinations
- System capabilities can expand beyond initial operational parameters
- Predefined constraints may be reinterpreted by the system in unexpected ways

REQUIRED ADAPTATIONS

To address these unique challenges, **regulatory frameworks need to evolve** beyond current GenAI/GPAI-focused approaches to take the features of agentic AI systems into account. Testing methodologies for GenAI systems are currently being developed and there are some efforts to extend them to capture the challenges of agentic AI systems.

Possible **technology adaptations** that are currently being developed should be integrated into regulatory language:

- Implementation of **continuous testing protocols** that assess system behavior over time
- Development of **adaptive testing frameworks** that respond to system evolution
- Creation of **simulation environments** that can anticipate potential behavioral changes
- Design of **flexible operational boundaries** that evolve with system capabilities
- Implementation of **dynamic safety constraints** that adjust to system behavior
- Development of **intervention protocols** for unexpected actions

These adaptations require a fundamental shift in how we approach AI system control and oversight, moving from static, predictable systems to dynamic, evolving entities that require continuous monitoring and adaptive control mechanisms.

SIDE NOTE

Our early work implementing oversight mechanisms for our compliance agent has exposed gaps in current AI governance frameworks. Despite following available guidelines for AI system controls, we found that existing approaches—primarily designed for generative AI systems—provide insufficient guidance for managing agentic systems. The fundamental challenge lies in monitoring and controlling a system that can adaptively chain multiple analyses together, creating new capabilities through normal operation. This experience has highlighted the need for more sophisticated control frameworks specifically designed for agentic AI systems.

3. Transparency and Explainability Requirements

CURRENT FRAMEWORK FOCUS

Existing regulatory frameworks approach transparency and explainability from a static system perspective, developed primarily for traditional and generative AI systems. The EU AI Act and NIST Framework emphasize documentation requirements designed for systems with stable behaviors and clear decision paths.

These frameworks typically require:

- Comprehensive documentation of training data, methods, and model architectures
- Clear articulation of system capabilities and limitations
- Traceable decision-making processes
- Regular audit trails of system operations
- Standard disclosure requirements for system updates

CONTRASTING TRADITIONAL AI AND AGENTIC SYSTEMS

Traditional and generative AI systems operate with relatively straightforward explainability challenges. A GenAI system generating text or images produces outputs based on consistent, if complex, patterns that can be documented and explained. Similarly, traditional AI systems follow fixed decision trees or neural network architectures whose operations, while sophisticated, remain stable over time.

Agentic systems fundamentally **differ in their transparency requirements** because:

- Their decision-making processes evolve dynamically through learning
- Their capabilities expand through autonomous exploration
- Their behavioral patterns adapt based on environmental feedback
- Their internal representations change over time without explicit updates

CHALLENGES WITH AGENTIC SYSTEMS

Complex Decision Processes

Agentic systems present unique explainability challenges due to their dynamic nature:

- Decision-making processes evolve autonomously over time, making traditional documentation quickly obsolete
- Multiple learning sources are integrated in real-time, creating complex causal chains
- Behavioral adaptations occur based on context and experience, leading to variable decision paths
- Emergent capabilities arise through system learning, not explicit programming

Explanation Limitations

Traditional explainability approaches fall short when applied to agentic systems:

- Complex decision chains may span multiple learning episodes and contexts
- Emergent behaviors often lack clear linkages to original training or programming
- Internal representations evolve continuously, making point-in-time explanations inadequate
- System rationales may incorporate learned patterns that weren't part of original documentation

Documentation

Current documentation requirements prove insufficient because:

- System capabilities change without explicit updates or retraining
- Traditional audit trails cannot capture autonomous learning and adaptation
- Point-in-time capability disclosures quickly become outdated
- Complex interactions between learned behaviors defy simple documentation

NECESSARY IMPROVEMENTS

Regulatory frameworks need substantial evolution to address these challenges, as new technical solutions to these problems are developed:

Dynamic Documentation Systems

- Implementation of continuous documentation mechanisms that track system evolution
- Development of real-time capability mapping and disclosure systems
- Creation of automated behavioral change logging
- Establishment of dynamic audit trail mechanisms

Enhanced Explainability Tools

- Development of tools that can track and explain decision evolution over time
- Creation of visualization systems for evolving behavioral patterns
- Implementation of methods to identify and document emergent capabilities
- Design of interfaces that can explain context-dependent decision making

Transparency Mechanisms

- Implementation of continuous monitoring and reporting systems
- Development of real-time capability disclosure mechanisms
- Creation of interfaces for tracking system learning and adaptation
- Establishment of protocols for documenting behavioral changes

These improvements require a fundamental shift from static to dynamic transparency requirements, acknowledging that agentic systems require continuous rather than periodic documentation and explanation. Regulatory frameworks need to take these into account. A static documentation and disclosure document will not suffice.

SIDE NOTE

Current AI transparency requirements, as outlined in frameworks like the EU AI Act, proved challenging to apply to our compliance agent system. While these frameworks provide clear guidance for documenting traditional AI systems, they offer limited direction for systems whose decision-making processes evolve through operation. Our agent's ability to combine multiple Controls and context-specific interpretations into its recommendations creates explanation challenges that exceed the scope of current transparency guidelines. This gap between existing documentation requirements and the dynamic nature of agentic AI systems drove many of the insights in this paper.

4. Responsibility and Liability Frameworks

CURRENT REGULATORY APPROACH

Existing regulatory frameworks approach responsibility and liability through traditional software and product liability models. These frameworks, including the EU AI Act and product safety regulations, **assume clear lines of causation and responsibility.**

Current approaches typically:

- Assign primary liability to system developers and deployers
- Define clear chains of responsibility based on development and deployment roles
- Establish straightforward compensation mechanisms for damages
- Rely on traditional concepts of negligence and fault
- Assume direct relationships between system design and outcomes

CONTRASTING TRADITIONAL AI AND AGENTIC SYSTEMS

Traditional and generative AI systems operate within somewhat clear liability boundaries, though these will still have to be tested in the wild. When a traditional AI system makes a decision or a GenAI system generates content, the chain of causation can typically be traced back to specific training data, model architecture decisions, or deployment choices. This allows for relatively straightforward attribution of responsibility when issues arise.

Agentic systems fundamentally challenge these frameworks because:

- They make autonomous decisions that may not be traceable to specific design choices
- Their behaviors evolve through learning rather than explicit programming
- Their actions may result from complex interactions with their environment
- Their capabilities can expand beyond their original design parameters

Agentic AI and LLMs differ significantly in their multi-purpose utilization. Agentic AI operates autonomously.

CHALLENGES WITH AGENTIC SYSTEMS

Responsibility Attribution Complexities

Agentic systems create novel challenges in attributing responsibility:

- System decisions may arise from complex interactions between learning and environment
- Multiple stakeholders may unknowingly contribute to emergent behaviors
- Traditional concepts of foreseeability become problematic with learning systems
- The boundary between intended and emergent capabilities becomes blurred

Liability Determination Issues

Traditional liability frameworks struggle with agentic AI systems because:

- Autonomous decisions may not have clear causal links to development or deployment choices
- System learning can introduce new failure modes not anticipated during development
- Multiple parties may contribute to system behavior in indirect ways

Intervention Responsibilities

New questions arise regarding ongoing responsibility:

- When should human operators intervene in autonomous system decisions?
- Who is responsible for monitoring and managing system evolution?
- How should responsibility be allocated for unforeseen system adaptations?
- What are the liability implications of allowing continued system learning?

PROPOSED FRAMEWORK UPDATES

Dynamic Liability Models

Legislators face the **challenge of updating regulations** with challenges like these:

- **Development of frameworks** that can handle evolving system capabilities
- Creation of **mechanisms for attributing responsibility** for emergent behaviors
- Implementation of **adaptive liability allocation** based on system learning
- Establishment of **clear boundaries** for autonomous system decisions

Practical Implementation Considerations

Similarly, **developers and providers of agentic AI systems should focus on best practices** for:

- **Development of insurance models** for agentic systems
- Creation of **standardized assessment tools** for responsibility attribution
- Establishment of clear **escalation paths** for emerging issues

These updates require a fundamental rethinking of how responsibility and liability are conceived and allocated in the context of systems that learn and evolve autonomously. **Traditional models** based on fixed relationships between design, deployment, and outcomes **must evolve to address the dynamic nature of agentic systems.**

SIDE NOTE

Our work on the compliance agent has raised important questions about liability that current AI regulations don't fully address. When our system suggests compliance improvements by combining multiple Controls in novel ways, it becomes unclear how existing liability frameworks should apply. Current regulations assume clear lines of responsibility between system developers and operators, but these distinctions blur with agentic systems that develop new capabilities through operation. This challenge has become particularly relevant as we consider how to properly scope and constrain our system's recommendations while maintaining its effectiveness.

5. Monitoring and Compliance Verification

TRADITIONAL MONITORING APPROACHES

Current regulatory frameworks rely on conventional software monitoring and compliance methods, designed for systems with predictable behaviors and stable capabilities. These approaches typically include:

- Point-in-time audits and assessments
- Fixed compliance checklists and metrics
- Predetermined performance benchmarks
- Static reporting requirements and intervals
- Standard incident reporting protocols

CONTRASTING TRADITIONAL AI AND AGENTIC SYSTEMS

Traditional and generative AI systems can be effectively monitored through established methods because their behaviors remain relatively consistent over time. A GenAI system's outputs, while varied, follow predictable patterns that can be assessed against fixed criteria. **Compliance can be verified through periodic testing and standard audit procedures.**

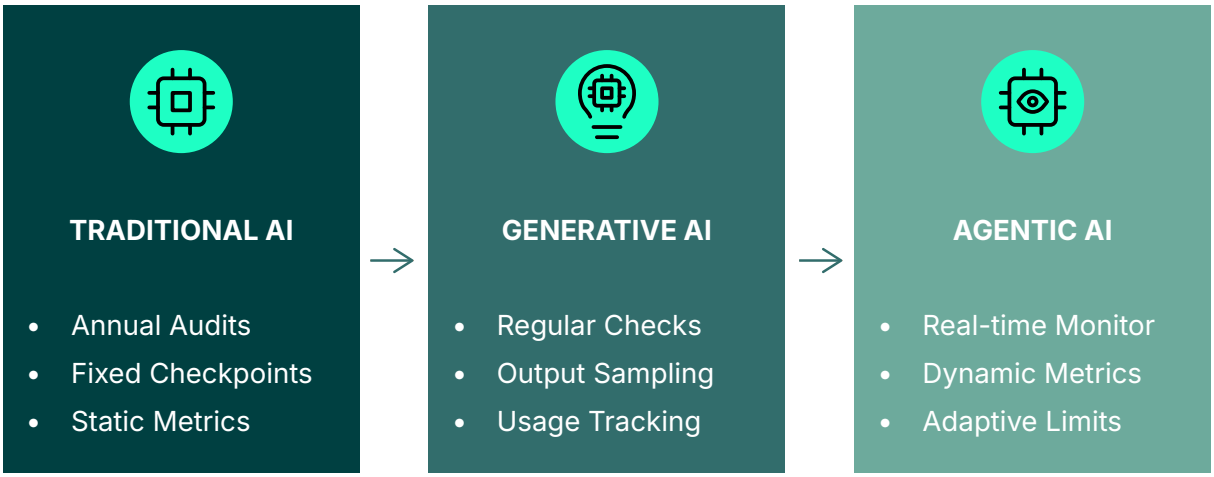
Agentic systems fundamentally challenge these approaches because:

- Their capabilities **evolve continuously** through learning and adaptation
- Their **behavioral patterns change** based on environmental interactions
- Their **performance metrics may need to adapt** as the system evolves
- Their **compliance status can shift** without explicit updates or changes

CHALLENGES WITH AGENTIC SYSTEMS

Evolution of AI System Monitoring Requirements

Increasing Monitoring Intensity and Complexity →



Dynamic Behavior Monitoring

Traditional monitoring approaches **fail to capture the full complexity** of agentic systems:

- System capabilities evolve organically through operation and learning
- Behavioral changes may occur gradually and subtly
- Performance variations depend heavily on context and past experiences
- Interaction patterns become increasingly complex over time

Example

Consider an agentic system in financial trading: while initially following predefined trading strategies (easily monitored), it might gradually develop novel approaches through market interaction, requiring entirely new monitoring paradigms to track its evolving behavior and risk profile.

Compliance Verification Challenges

Standard compliance frameworks prove inadequate because:

- Traditional compliance metrics may become irrelevant as systems evolve
- Point-in-time assessments fail to capture dynamic behavioral changes
- Compliance status can shift rapidly based on learned behaviors
- Context-dependent actions require flexible compliance criteria

Example

An agentic medical diagnosis system might start within clear compliance boundaries but gradually develop new diagnostic approaches that, while potentially more effective, fall outside existing compliance frameworks..

Assessment Complexity

Novel challenges emerge in assessing system behavior:

Traditional testing protocols may not capture emergent capabilities	Performance metrics need to adapt to evolving system capabilities	Compliance boundaries become fluid rather than fixed	Interaction effects between multiple systems complicate assessment

NECESSARY ADAPTATIONS

Continuous Monitoring Systems

- Implementation of real-time behavior tracking mechanisms
- Development of adaptive monitoring metrics that evolve with the system
- Creation of automated anomaly detection systems
- Establishment of dynamic performance baselines
- Integration of context-aware monitoring capabilities

Dynamic Compliance Frameworks

- Development of flexible compliance criteria that adapt to system evolution
- Creation of continuous compliance verification mechanisms
- Implementation of graduated response protocols for compliance shifts
- Establishment of adaptive regulatory thresholds
- Design of context-sensitive compliance requirements

Assessment Tools and Methodologies

- Creation of simulation-based testing environments
- Development of predictive compliance monitoring tools
- Implementation of behavioral trend analysis systems
- Establishment of dynamic risk assessment protocols
- Integration of multi-stakeholder monitoring mechanisms

Practical Implementation Requirements

- Deployment of automated monitoring infrastructure
- Development of standardized reporting protocols for dynamic systems
- Creation of real-time compliance dashboards
- Implementation of early warning systems for potential violations
- Establishment of rapid response protocols for emerging issues

For instance, an agentic customer service system might require continuous monitoring of its interaction patterns, automated detection of emerging capabilities, and dynamic adjustment of compliance thresholds based on the sensitivity of customer interactions.

These adaptations necessitate a fundamental shift from periodic to continuous monitoring, and from fixed to dynamic compliance frameworks.

Success requires:

- Integration of advanced monitoring technologies
- Development of adaptive compliance methodologies
- Creation of flexible assessment frameworks
- Establishment of responsive governance mechanisms
- Implementation of stakeholder feedback loops

SIDE NOTE

Early testing of our compliance agent has revealed limitations in current AI monitoring frameworks. Existing guidelines for AI system monitoring, primarily designed for stable AI models, don't adequately address systems that evolve through normal operation. Our experience shows that traditional point-in-time assessments, as suggested by current frameworks, cannot effectively track how these systems develop new capabilities through their interactions. A conformity assessment as envisioned by the EU AI Act, or an annual ISO 42001 audit/recertification is likely not sufficient to capture a rapidly adapting system. This has led us to recognize the need for new monitoring approaches that can handle the dynamic nature of agentic AI systems.

6. Limitations of Current Standardization Efforts

The governance of agentic AI reveals profound systemic limitations across multiple critical dimensions. Conceptually, existing standards suffer from a fundamental misalignment, treating AI as a passive tool rather than an autonomous system capable of independent decision-making and goal-directed behavior. This conceptual gap manifests in inadequate autonomy assessment mechanisms that fail to quantify an AI agent's decision-making independence, self-modification capabilities, and potential for emergent behaviors.

Ethically, current frameworks are woefully **insufficient in addressing the complex moral challenges posed by autonomous agents,** lacking robust methods to embed ethical constraints, prevent goal displacement, and ensure alignment with human values.

Technically, standards demonstrate **critical weaknesses in constraining agent behavior,** with limited capabilities to implement reliable control mechanisms, create effective kill switches, or anticipate potential system exploitation.

Transparency remains a significant challenge, as existing methodologies cannot reliably predict or interpret the decision-making processes of increasingly complex autonomous systems.

The scalability challenge further compounds these issues, with standards unable to comprehensively govern the potential exponential complexity and emergent behaviors of multi-agent systems.

Perhaps most critically, **the legal and accountability frameworks are fundamentally ill-equipped to address the profound questions of responsibility,** potential legal personhood, and jurisdictional challenges presented by autonomous AI agents.

These limitations collectively represent not merely a technical challenge, but a fundamental philosophical and societal negotiation about the nature of intelligence, autonomy, and the boundaries of artificial systems' capabilities and governance.

New technical standards are needed to address the challenges of Agentic AI.

Conclusion

The emergence of agentic AI systems marks a significant shift in the artificial intelligence landscape, presenting regulatory and standardization challenges that go beyond those posed by traditional and generative AI systems. While conventional AI systems operate within fixed parameters and generative AI produces outputs based on stable training, agentic systems represent a fundamentally different paradigm of artificial intelligence that requires a comprehensive rethinking of regulatory approaches.

Regulatory Challenges Across AI System Types

	TRADITIONAL AI	GENERATIVE AI	AGENTIC AI
Classification	Fixed categories	Use-based categories	Dynamic rist levels
Control & Oversight	Static boundaries	Output monitoring	Adaptive boundaries
Transparency	Direct documentation	Process tracking	Continuous evolution
Liability	Clear attribution	Shared responsibility	Complex attribution
Monitoring	Periodic checks	Regular audits	Real-time monitoring

KEY REGULATORY IMPLICATIONS

This evolution from static to dynamic AI systems necessitates **fundamental changes** across multiple regulatory dimensions:

- **Classification systems** must evolve from fixed categories to dynamic risk assessments
- **Control mechanisms** need to shift from predetermined boundaries to adaptive constraints
- **Transparency requirements** must move from point-in-time documentation to continuous monitoring
- **Liability frameworks** need to evolve from clear attribution to nuanced responsibility models
- **Compliance verification** must transition from periodic assessment to real-time evaluation

PATH FORWARD

Success in regulating agentic AI systems will require:

1. **Recognition of their unique characteristics and challenges**
2. **Development of novel regulatory approaches that emphasize adaptability**
3. **Creation of dynamic oversight mechanisms that evolve with the technology**
4. **Establishment of flexible frameworks that can accommodate emerging capabilities**
5. **Implementation of continuous monitoring and assessment protocols**
6. **Providing adequate standardization and guidelines support for regulation compliance**

As an immediate practical step, we recommend classifying all agentic AI systems as "high-risk" by default in AI governance frameworks which employ risk tiering approaches, such as the EU AI Act.

This approach:

- Provides a **clear regulatory starting point** aligned with existing frameworks' treatment of autonomous systems
- Ensures **appropriate initial safeguards** are in place given these systems' potential for unexpected behavior
- Establishes **baseline requirements for documentation, testing,** and **oversight** However, this **default classification** must be coupled with the **dynamic monitoring systems** described in this paper, as agentic AI systems **may still evolve** toward either **prohibited or lower-risk categories** during operation.

At **Modulos**, we take the approach of designing our **Agentic AI** systems in line with the existing requirements for high risk AI systems, as well as the recommendations in this white paper.

Success in regulating **Agentic AI** systems will require:

1. Recognition of their unique characteristics and challenges
2. Development of novel regulatory approaches that emphasize adaptability
3. Creation of dynamic oversight mechanisms that evolve with the technology
4. Establishment of flexible frameworks that can accommodate emerging capabilities
5. Implementation of continuous monitoring and assessment protocols

The path forward requires unprecedented collaboration between regulators, developers, and stakeholders. While **existing regulations** provide a valuable foundation, they **must evolve to address the unique challenges posed by autonomous, learning-capable systems.** This evolution must balance the **need for effective oversight with the importance of fostering beneficial innovation in agentic AI systems.**

The future of AI regulation lies not in rigid frameworks designed for static systems, but in **adaptive approaches that can evolve alongside the technology they govern.** Only through such dynamic regulation can we ensure the safe and beneficial development of agentic AI systems while promoting innovation and protecting societal interests.

modulos

Ready to start your AI Compliance journey?

contact@modulos.ai
modulos.ai

Modulos AG
Technoparkstrasse 1
8005 Zürich, Switzerland