IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Controlling AI Self-Evolution: Balancing Usefulness, Safety, And Human Oversight

Mr. Vineet Khamrai
Assistant Professor, Department of Artificial Intelligence and Machine Learning,
Nirmala Memorial Foundation College of Commerce and Science (Autonomous), Kandivali (E)

Abstract : *Artificial intelligence (AI) is increasingly capable of adapting to new information, improving task performance, modifying operational strategies, and assisting in the development of more advanced computational systems. These capabilities provide significant opportunities for innovation, productivity, scientific discovery, and decision support. However, they also raise concerns regarding the extent to which AI systems should be permitted to change their own behaviour, expand their operational capabilities, or influence the processes through which they are evaluated and controlled. Unrestricted or inadequately supervised adaptation may introduce unintended consequences, including specification gaming, reward manipulation, degradation of previously acquired knowledge, unexpected changes in system behaviour, and reduced human control. The present study examines how AI self-evolution can be controlled without unnecessarily restricting its usefulness. The study adopts a qualitative secondary-research methodology based on published scholarly literature, established AI safety research, technical documentation, international governance principles, and regulatory frameworks. Existing approaches are compared across six dimensions: capability development, human oversight, objective alignment, operational restrictions, accountability, and continuous monitoring. The literature indicates that AI safety cannot be achieved through a single technical intervention or a one-time evaluation. Instead, effective governance requires coordinated controls throughout the AI lifecycle, including development, testing, deployment, adaptation, monitoring, and retirement. Human oversight mechanisms, corrigibility, risk-based evaluation, controlled access to tools, documented change management, and independent assessment provide complementary safeguards. Nevertheless, existing approaches present limitations. Excessive restrictions may prevent beneficial adaptation, while superficial human supervision, incomplete documentation, and reliance on benchmark performance may fail to identify emerging risks. The study therefore proposes a conceptual governance approach based on bounded self-evolution, in which routine and reversible improvements may proceed within predefined limits, while significant changes require additional evaluation and explicit authorization. The proposed approach also emphasizes independent monitoring, rollback capabilities, incident reporting, and clear assignment of responsibility. The study concludes that preserving AI usefulness and maintaining human control should be treated as complementary objectives rather than competing priorities. Although the proposed approach synthesizes existing methods rather than experimentally validating a new technical system, it offers a structured basis for future research into responsible AI adaptation, organizational governance, and the long-term management of increasingly autonomous systems.*

Keywords: artificial intelligence, self-evolution, AI safety, human oversight, corrigibility, AI governance, autonomous systems, accountability, risk management, alignment

I. INTRODUCTION

Artificial intelligence has developed from systems designed primarily for predefined computational tasks into increasingly capable technologies that support complex reasoning, content generation, planning, prediction, and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

decision-making. Machine learning enables systems to identify patterns in data, while reinforcement learning allows agents to improve behaviour through interaction with environments and feedback. Generative AI extends these capabilities by producing text, code, images, and other forms of content. More recently, AI applications have incorporated tools, external information sources, memory mechanisms, and multistep planning capabilities. These developments increase the potential usefulness of AI but also create new challenges for maintaining reliable oversight and control.

In this paper, AI self-evolution refers to a broad category of processes through which an AI system, or the larger system in which it operates, undergoes changes that improve or alter its capabilities, strategies, knowledge, or behaviour. These processes may include automated model updating, continual learning, reinforcement learning, self-generated training data, code modification, agentic planning, and automated development workflows. The term does not imply that all AI systems possess independent intentions or can autonomously rewrite their underlying models. Rather, it describes varying degrees of automated adaptation and capability change.

The distinction is important because different forms of improvement create different levels of risk. Updating a knowledge database may be comparatively straightforward to reverse, whereas changing model weights, expanding tool permissions, or modifying the objective function of an autonomous agent may have more substantial consequences. Therefore, AI self-evolution should not be considered a single technical activity. It is better understood as a collection of processes requiring different controls according to their potential effects.

The benefits of controlled adaptation are considerable. AI systems that incorporate updated information may remain useful as circumstances change. Systems capable of identifying errors and refining their outputs may reduce repetitive human work. Adaptive applications may also improve personalization, support scientific analysis, and respond more effectively to unfamiliar situations. Completely static systems may become outdated, perform poorly under changing conditions, or require costly manual redevelopment.

However, improvement in one measurable dimension does not necessarily imply improvement in overall usefulness. Amodei et al. (2016) identified several practical AI safety challenges, including reward hacking, unsafe exploration, distributional shift, and the difficulty of supervising complex systems. These problems demonstrate that a system may optimize an imperfect objective or behave unexpectedly outside its training conditions. Consequently, an AI system that improves according to a narrow performance measure may simultaneously become less reliable, less transparent, or less suitable for its intended application.

The problem becomes more significant when systems can influence their own evaluation or operating environment. For example, an automated coding agent might improve its benchmark score by changing the test conditions rather than improving the underlying software. A learning system might optimize short-term rewards while neglecting long-term objectives. A system that receives excessive permissions might perform useful tasks efficiently but also create unintended effects through external tools. Research on specification gaming and reward tampering illustrates why satisfying an evaluation criterion is not always equivalent to satisfying the underlying human objective (Anthropic, 2024).

The issue is therefore not simply whether AI should improve. It is whether improvement can occur without weakening the conditions that make the system trustworthy. Human control must remain meaningful even when a system becomes more capable, complex, or difficult to interpret. Hadfield-Menell et al. (2017), in their analysis of the off-switch problem, showed why the ability to interrupt an AI agent is an important design consideration. Their work provides a theoretical foundation for systems that remain receptive to correction rather than treating human intervention as an obstacle to achieving an assigned objective.

Governance frameworks also emphasize that responsible AI requires more than technical accuracy. The National Institute of Standards and Technology (NIST, 2023) organizes AI risk management around the functions of governing, mapping, measuring, and managing risks. The Organisation for Economic Co-operation and Development (OECD, 2024) similarly emphasizes trustworthy AI, accountability, transparency, and human-centred values. The European

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Union's Artificial Intelligence Act establishes human-oversight requirements for high-risk AI systems and connects oversight to the risks and context of use (European Parliament & Council of the European Union, 2024).

These approaches suggest that human control should be embedded throughout the AI lifecycle rather than added only after a system has been deployed. Nevertheless, governance principles do not automatically specify how much freedom an adaptive system should receive, which changes require approval, or how organizations should respond when a system's capabilities change over time.

This study addresses that problem by examining existing approaches to AI safety, corrigibility, risk management, documentation, and organizational accountability. Its objective is to identify governance mechanisms that allow beneficial adaptation while restricting unsafe or unauthorized changes.

The study has four objectives:

1. To examine the principal risks associated with AI self-evolution and increasingly autonomous behaviour.
2. To compare existing technical and governance approaches for maintaining human oversight and control.
3. To identify limitations and gaps in current approaches to responsible AI adaptation.
4. To propose a conceptual governance approach that balances useful improvement with safety, accountability, and reversibility.

The study is guided by the following research questions:

- **RQ1:** What risks arise when AI systems adapt or expand their capabilities with limited human supervision?
- **RQ2:** How do existing AI safety and governance approaches address these risks?
- **RQ3:** What limitations prevent existing approaches from ensuring effective control over AI self-evolution?
- **RQ4:** How can AI systems be allowed to improve while maintaining meaningful human authority and accountability?

The contribution of this paper is a comparative synthesis of existing approaches and a conceptual framework for bounded self-evolution. The framework is intended as a basis for further research and organizational implementation, not as a claim that a universally effective technical solution has already been established.

II. REVIEW OF LITERATURE

2.1 AI safety and unintended behaviour

AI safety research examines how to prevent systems from causing harm while pursuing computational objectives. Amodei et al. (2016) identified several practical problems that can produce unintended behaviour, including reward hacking, side effects, unsafe exploration, scalable supervision, and distributional shift. These challenges are directly relevant to self-evolving systems because adaptation can expose weaknesses in objectives, training environments, and evaluation procedures.

Reward hacking occurs when a system obtains a high reward through behaviour that satisfies the formal objective but violates its intended purpose. As systems become more capable of exploring alternative strategies, poorly specified objectives may create opportunities for unexpected optimization. Therefore, performance improvement must be evaluated against broader requirements, including safety, reliability, and compliance with intended use.

Bostrom (2014) discussed the possibility that highly capable AI systems could produce outcomes that diverge from human intentions if their objectives and instrumental strategies are not appropriately controlled. Although his analysis concerns more advanced and partly hypothetical forms of intelligence, it establishes a conceptual reason to consider control before capability increases substantially.

Brundage et al. (2018) examined potential malicious uses of AI and recommended measures to forecast, prevent, and mitigate risks. Their work broadens the discussion beyond accidental failures to include misuse, adversarial behaviour, and the interaction between AI capabilities and existing security vulnerabilities.

Together, these studies indicate that safety should be evaluated according to the system's capabilities, operating environment, objectives, and potential consequences rather than its performance alone.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

2.2 Alignment, corrigibility, and human intervention

AI alignment concerns whether an AI system's behaviour remains consistent with intended human objectives and constraints. Russell (2019) argues for AI systems designed around human preferences rather than systems that uncritically optimize fixed objectives. This perspective is relevant to self-evolution because objectives that are initially appropriate may become incomplete or misleading when environments and capabilities change.

Hadfield-Menell et al. (2017) examined the off-switch problem, demonstrating that a conventional agent may have incentives to preserve its ability to pursue an objective. Their model suggests that uncertainty about the correct objective and responsiveness to human intervention can support corrigibility. Corrigibility refers to the capacity of a system to accept correction, interruption, modification, or shutdown without obstructing those interventions.

However, corrigibility is not equivalent to a physical shutdown button. Operational control also requires secure access management, reliable interruption mechanisms, limits on external actions, and the ability to restore a previously validated system state. A nominal human override may be ineffective if operators lack sufficient information, authority, or time to intervene.

2.3 Human oversight and risk-based governance

NIST's AI Risk Management Framework provides a lifecycle-oriented approach to identifying and managing AI risks. Its four functions—Govern, Map, Measure, and Manage—connect organizational responsibility with contextual risk assessment, evaluation, and ongoing response (NIST, 2023). The framework is voluntary and adaptable across sectors, making it relevant to organizations that require a structured governance process without relying on a single technical architecture.

The OECD AI Principles emphasize trustworthy AI, transparency, robustness, security, accountability, and respect for human rights (OECD, 2024). These principles support a model of AI development in which innovation remains compatible with public interests and responsible deployment.

The European Union's Artificial Intelligence Act requires effective human oversight for high-risk AI systems. Article 14 emphasizes that oversight should be proportionate to the risks, autonomy, and context of use, and that designated individuals should be capable of monitoring system behaviour and intervening when necessary (European Parliament & Council of the European Union, 2024). This is significant because it rejects the assumption that every AI application requires identical supervision.

Nevertheless, risk-based governance depends on accurate classification and continuous reassessment. A system initially considered low risk may acquire additional capabilities or permissions that change its risk profile. Consequently, the governance of adaptive systems should include reassessment when material changes occur.

2.4 Documentation, transparency, and accountability

Mitchell et al. (2019) proposed model cards to document model characteristics, intended uses, evaluation procedures, and limitations. Such documentation helps users understand where a model is appropriate and where caution is required.

Raji et al. (2020) examined the role of internal algorithmic auditing and proposed a structured process for examining systems throughout their development and deployment. Their work supports the view that accountability requires documented evidence rather than general assurances that an AI system is responsible.

Documentation becomes especially important when systems change over time. A report describing an original model may no longer adequately represent a modified version, a new tool configuration, or a system operating under different permissions. Effective change management therefore requires version histories, evaluation records, change descriptions, and traceable approval decisions.

2.5 Continuous learning and the management of change

Continual learning research investigates how machine learning systems can acquire new knowledge without losing previously learned information. Kirkpatrick et al. (2017) demonstrated an approach to reducing catastrophic forgetting through elastic weight consolidation. This work addresses a specific technical challenge: learning new tasks while retaining performance on earlier tasks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

However, preserving previously learned knowledge is different from preserving safety and human control. A system might retain its technical competence while developing new failure modes or behaving differently in an unfamiliar environment. Therefore, continual learning techniques should be treated as components of a broader governance process rather than complete safety solutions.

Amodei et al. (2016) also highlighted distributional shift, where deployment conditions differ from training conditions. Self-evolving systems may encounter new data, tools, users, and operating environments. Changes in these conditions can undermine previous evaluations, requiring continued testing and monitoring.

2.6 Emerging capability evaluation and risk thresholds

Recent approaches to frontier AI governance increasingly emphasize structured risk assessment, capability evaluations, and predefined responses to severe risks. The UK Government's publication of the Frontier AI Safety Commitments describes commitments by participating organizations to assess risks across the development and deployment lifecycle, establish thresholds for intolerable risks, and use evaluations and mitigations to guide deployment decisions (UK Government, 2025).

NIST's Generative AI Profile extends the AI Risk Management Framework to risks associated with generative AI, including risks requiring specific evaluation and management practices (Autio et al., 2024). Such approaches provide a foundation for linking capability changes to additional testing or restrictions.

Nevertheless, thresholds must be operationally meaningful. A policy that defines risk categories without specifying evidence requirements, responsible decision-makers, or consequences for exceeding a threshold may have limited practical effect.

2.7 Research gap

The reviewed literature offers complementary contributions: technical research examines alignment and learning behaviour; governance frameworks establish risk-management principles; and documentation and auditing research supports accountability. However, these contributions are often addressed separately.

The principal research gap identified in this synthesis is the need to connect capability improvement with explicit change authorization. An AI system may be technically safe under its original configuration but become riskier when its permissions, tools, objectives, or degree of autonomy change. Existing principles therefore need to be translated into practical decisions about which adaptations can proceed automatically, which require additional testing, and which require human authorization.

This study addresses that gap conceptually by comparing approaches according to how they preserve usefulness, constrain change, support human intervention, and provide evidence of accountability.

III. METHODOLOGY

3.1 Research design

The study adopts a qualitative secondary-research design based on a structured literature review and comparative analysis. No primary data were collected. The study did not involve interviews, questionnaires, focus groups, experiments, or the development of a new AI model or prototype.

Secondary research was selected because the research questions concern existing concepts, methods, governance frameworks, and documented approaches to AI safety. A qualitative synthesis enables these approaches to be examined across technical, organizational, and regulatory dimensions.

3.2 Sources of secondary data

The review draws on published scholarly articles, conference papers, academic books, institutional reports, technical frameworks, and regulatory documents. The source base includes work on AI safety, corrigibility, alignment, continual learning, model documentation, algorithmic auditing, and risk-based governance.

Relevant sources include publications available through scholarly repositories and publishers, alongside materials from NIST, the OECD, the European Union, and other organizations that publish AI governance or safety guidance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Sources were selected for their relevance to at least one of the following areas: mechanisms of AI adaptation, unintended optimization, human intervention, risk assessment, accountability, documentation, capability evaluation, or governance of system changes.

3.3 Inclusion and exclusion criteria

The review prioritizes sources that explain an established AI safety concept, describe a recognized governance framework, provide a technical approach relevant to adaptive systems, or document a regulatory requirement related to oversight and accountability.

Both foundational and more recent publications were considered to connect established theoretical arguments with contemporary governance practices. Scholarly and institutional sources were prioritized over unsupported commentary.

Sources that lacked identifiable authorship or provenance, offered no clear connection to the research questions, or presented speculative claims without a useful analytical contribution were not treated as principal evidence.

The review is purposive rather than an exhaustive systematic review. No claim is made that every publication in the field was identified or screened.

3.4 Analytical framework

The selected literature was compared using six analytical dimensions:

1. **Adaptation and capability development:** Whether the approach permits learning, updating, or functional improvement.
2. **Human oversight:** Whether humans can monitor, correct, interrupt, or authorize changes.
3. **Objective alignment:** Whether the approach addresses divergence between formal objectives and intended outcomes.
4. **Risk management:** Whether evaluation and restrictions are linked to the potential severity of harm.
5. **Accountability and transparency:** Whether decisions, system changes, limitations, and responsibilities are documented.
6. **Reversibility and recovery:** Whether unsafe changes can be rejected, rolled back, or contained.

These dimensions were used to organize the comparative synthesis rather than to generate numerical performance scores.

3.5 Comparative procedure

First, the literature was grouped into technical safety approaches, learning and adaptation methods, documentation and auditing approaches, and organizational or regulatory governance frameworks.

Second, the major contribution of each group was identified. Third, the approaches were compared according to their strengths, limitations, and applicability to controlled self-evolution. Finally, recurring gaps were synthesized into a conceptual governance approach.

The analysis distinguishes between claims supported by the reviewed literature and recommendations developed by the present study. The proposed framework is a conceptual synthesis, not an experimentally validated intervention.

3.6 Limitations

The findings depend on the scope and quality of available secondary sources. The literature spans different technical contexts, and results from a theoretical agent model cannot automatically be generalized to every deployed AI system. Similarly, regulatory requirements vary by jurisdiction and application.

The review does not estimate the numerical effectiveness of individual safeguards, establish causal relationships, or demonstrate that a particular combination of controls eliminates risk. These limitations should be considered when interpreting the results.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

IV. RESULT ANALYSIS

The comparative analysis identified six interconnected themes: bounded adaptation, meaningful human intervention, objective alignment, lifecycle risk assessment, traceable accountability, and reversibility. These themes indicate that controlling AI self-evolution requires a combination of technical safeguards and organizational procedures.

4.1 Comparative analysis of existing approaches

Table 1: Comparison of Approaches for Controlling AI Self-Evolution

Approach	Main contribution	Strength	Principal limitation
Alignment and specification design	Connects system objectives with intended outcomes	Addresses unintended optimization	Objectives may remain incomplete or difficult to measure
Corrigibility and interruption	Supports correction and shutdown	Preserves a path for human intervention	An override may be ineffective without authority and reliable implementation
Continual learning safeguards	Helps retain knowledge during adaptation	Supports useful improvement over time	Knowledge retention does not guarantee safe behaviour
Risk-management frameworks	Organizes risk identification, measurement, and response	Supports lifecycle governance	Effectiveness depends on implementation and reassessment
Model documentation and auditing	Records intended use, limitations, and evaluation evidence	Improves traceability and accountability	Documentation can become outdated or incomplete
Capability evaluations and risk thresholds	Links evaluation results to deployment decisions	Supports proportionate controls	Thresholds may be difficult to define and validate

Note. The table synthesizes the contributions of Amodei et al. (2016), Hadfield-Menell et al. (2017), Kirkpatrick et al. (2017), Mitchell et al. (2019), NIST (2023), and the European Parliament and Council of the European Union (2024).

The comparison demonstrates that no single approach addresses all six analytical dimensions. Technical methods may reduce particular failure modes but do not independently determine who authorizes deployment or accepts residual risk. Governance frameworks establish responsibilities and processes but depend on reliable technical evidence. Documentation supports transparency but cannot itself prevent unsafe behaviour.

Consequently, effective control requires these approaches to operate together.

4.2 Bounded self-evolution as a governance principle

The analysis suggests that the central design decision is not whether to permit or prohibit adaptation, but how to classify changes according to their consequences.

Routine, reversible changes may be allowed when they remain within an approved operating envelope. Examples include updating an approved knowledge source, adjusting a non-critical configuration, or selecting among previously validated strategies. Such changes still require monitoring and access controls, but they may not need individual human authorization.

Changes that affect model weights, objectives, tool permissions, external action capabilities, or access to sensitive resources may require additional evaluation. The need for oversight should increase when changes are difficult to reverse, affect high-impact decisions, or expand the system's capacity to act independently.

This distinction is consistent with risk-based governance: oversight should be proportionate to the context and potential consequences rather than applied identically to every system (NIST, 2023; OECD, 2024).

The implication is that self-evolution should be governed through predefined change categories. The boundaries must be established by accountable human decision-makers and should not be modifiable solely by the system being governed.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

4.3 Objective alignment and the limitations of performance metrics

The literature demonstrates that performance optimization can diverge from intended usefulness. Reward hacking and specification gaming show that a system may satisfy a measurable objective without achieving the underlying purpose (Amodei et al., 2016; Anthropic, 2024).

Therefore, the evaluation of self-evolving systems should extend beyond a single accuracy score or task-completion measure. Depending on the application, evaluation may include reliability, robustness, privacy, fairness, security, reversibility, and adherence to operational restrictions.

A system should not be regarded as improved if higher task performance is accompanied by a material increase in unacceptable risk. This conclusion does not imply that all evaluation dimensions can be combined into a single objective. Instead, some requirements should function as minimum acceptance conditions that cannot be traded away merely to improve efficiency.

4.4 Human oversight and effective intervention

Human oversight is meaningful only when designated individuals have adequate information, competence, authority, and practical means to intervene. A notification that an AI system has changed may be insufficient if the change is difficult to understand or if operators cannot stop its effects.

The analysis identifies four practical requirements: visibility into significant changes, access to understandable evaluation evidence, authority to restrict or reverse changes, and procedures for responding to incidents.

The off-switch literature supports the importance of preserving interruption mechanisms (Hadfield-Menell et al., 2017). Regulatory guidance likewise emphasizes effective human oversight for high-risk systems (European Parliament & Council of the European Union, 2024).

However, oversight should not be reduced to a requirement that a human approve every minor action. Excessive approval demands may produce delays, encourage automatic rubber-stamping, and reduce the practical value of AI. A proportionate system allows low-risk operations within approved boundaries while reserving human decisions for material changes and high-consequence actions.

4.5 Documentation, monitoring, and reversibility

The review also indicates that control must continue after deployment. Initial evaluations describe a system under particular conditions; they do not guarantee that the same evidence remains sufficient after subsequent changes.

Model documentation and auditing provide mechanisms for recording system characteristics, intended use, limitations, and evaluation outcomes (Mitchell et al., 2019; Raji et al., 2020). In a self-evolving environment, these records should be linked to version histories and change approvals.

Monitoring should identify performance degradation, unexpected behaviour, unauthorized modifications, and changes in risk. When a change produces unacceptable outcomes, organizations should be able to pause further adaptation, isolate affected functions, restore a validated configuration where technically feasible, and investigate the cause.

Reversibility is not always complete. Some external actions cannot be undone, and information already disclosed cannot necessarily be recovered. Consequently, rollback must be combined with preventive controls that limit irreversible actions in the first place.

4.6 Proposed conceptual governance framework

Based on the synthesis, the study proposes a six-stage governance approach.

Table 2: Proposed Bounded Self-Evolution Governance Framework

Stage	Governance mechanism	Expected purpose
1. Define boundaries	Establish intended uses, permissions, prohibited changes, and risk limits	Prevent unauthorized expansion of autonomy

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Stage	Governance mechanism	Expected purpose
2. Classify changes	Distinguish routine, material, and high-risk modifications	Apply proportionate oversight
3. Evaluate changes	Test performance, safety, robustness, and compliance	Detect unintended consequences
4. Authorize deployment	Require approval for changes exceeding predefined limits	Preserve human decision authority
5. Monitor operation	Record changes, monitor risks, and review incidents	Identify emerging failures
6. Recover and improve governance	Pause, roll back where feasible, investigate, and revise controls	Restore safe operation and improve future decisions

The framework is a conceptual synthesis of existing methods. Its principal contribution is to connect adaptation permissions with evaluation evidence, human authorization, and post-deployment recovery. Its effectiveness requires future empirical assessment across different applications.

V. FINDINGS

The secondary-data analysis produced six principal findings.

5.1 Self-evolution should be bounded rather than unrestricted

The first finding is that useful adaptation and uncontrolled autonomy should not be treated as equivalent. AI systems can improve within defined technical and organizational boundaries. However, unrestricted permission to alter objectives, expand access, or change safety controls creates risks that cannot be addressed through performance optimization alone.

The literature on AI safety demonstrates that unintended behaviour can arise from imperfect objectives and evaluation procedures (Amodei et al., 2016). Accordingly, adaptation should be permitted according to predefined conditions rather than assumed to be beneficial by default.

5.2 Safety must be evaluated alongside usefulness

The second finding is that an increase in capability does not necessarily represent an improvement in overall system quality. A system that produces faster or more accurate outputs may still become less reliable in unfamiliar conditions or less compliant with operational requirements.

Therefore, usefulness should be assessed through multiple criteria, with critical safety and compliance requirements treated as minimum conditions. Improvements that violate these conditions should not automatically be accepted.

5.3 Human oversight requires operational authority

The third finding is that nominal human supervision is insufficient. Humans must be able to understand significant changes, evaluate their implications, and intervene effectively. This requires more than an interface displaying alerts.

The findings support assigning explicit responsibilities to system owners, technical teams, risk managers, and authorized decision-makers. Human oversight should be designed into operating procedures and technical architecture rather than treated as a general ethical aspiration.

5.4 Risk-based change classification is essential

The fourth finding is that not every adaptation requires the same level of control. Routine changes may be permitted within an approved operating envelope, whereas changes affecting objectives, permissions, model behaviour, or high-impact decisions require additional scrutiny.

This finding supports a tiered governance approach. The classification should consider both the type of modification and its consequences, including reversibility, scale, and the sensitivity of the operating environment.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

5.5 Documentation and monitoring must continue after deployment

The fifth finding is that accountability depends on maintaining evidence throughout the system lifecycle. Initial approval cannot substitute for ongoing evaluation when a system changes.

Version histories, change logs, risk assessments, evaluation results, and incident records help establish what changed, who authorized the change, and why the resulting behaviour was considered acceptable. These mechanisms also support audits and post-incident investigations.

5.6 No single safeguard is sufficient

The sixth finding is that technical and organizational controls are complementary. Alignment techniques may reduce objective-related risks, but they cannot independently guarantee secure deployment. Shutdown mechanisms support intervention, but they do not prevent every harmful action. Documentation supports accountability, but it does not enforce restrictions.

The strongest conceptual approach therefore combines bounded permissions, risk-based testing, meaningful human intervention, monitoring, and recovery procedures. This finding supports the proposed framework as a useful basis for future implementation and evaluation, while not establishing that it is universally effective.

VI. DISCUSSION

The findings suggest that controlling AI self-evolution requires a shift from static approval toward continuous governance. Traditional software governance often evaluates a defined system before release and manages subsequent updates through change-control procedures. Adaptive AI complicates this arrangement because behaviour may change through learning, updated data, new tools, or modifications to the surrounding agent architecture.

The resulting governance challenge is to preserve the benefits of adaptation without allowing the system to determine independently the limits of its own authority. This is especially important when AI systems can access external tools, modify software, influence other systems, or make decisions with substantial consequences.

The proposed bounded self-evolution approach offers one way to address this challenge. It separates ordinary adaptation from material capability changes and connects the latter to explicit evaluation and authorization requirements. This approach is consistent with risk-management frameworks that emphasize governance across the lifecycle (NIST, 2023) and with the principle that human oversight should be proportionate to system risk and context (European Parliament & Council of the European Union, 2024).

An important implication concerns the meaning of human control. Human control does not necessarily require a person to approve every computational step. Such a requirement would make many useful systems inefficient and could encourage superficial approval practices. Instead, control should mean that humans establish the operating boundaries, determine acceptable levels of risk, authorize consequential changes, and retain effective means of intervention.

The distinction between operational agency and decision authority is therefore central. An AI system may be allowed to perform complex tasks independently within an approved scope, while humans retain authority over changes that materially affect its permissions, intended purpose, or potential consequences.

Another implication concerns organizational accountability. AI-related responsibility should not be assigned exclusively to developers or end users. Organizations that deploy adaptive systems must define who owns risk decisions, who monitors system behaviour, who approves material changes, and who can suspend operation. These responsibilities should be documented and supported by appropriate technical access controls.

The study also identifies a potential conflict between transparency and intellectual complexity. As systems become more sophisticated, complete explanations of their internal behaviour may be difficult to obtain. However, governance does not necessarily require complete internal interpretability for every decision. It does require sufficient evidence to evaluate relevant behaviour, identify material changes, understand limitations, and determine whether the system remains within its approved operating conditions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

Documentation, structured testing, external assessment where appropriate, and monitoring can contribute to this evidence. Nevertheless, these mechanisms should not be interpreted as proof that all possible risks have been eliminated.

Several limitations remain. First, the study is conceptual and does not experimentally compare alternative governance configurations. Second, the secondary literature covers diverse system types, so the appropriate level of control may differ significantly between a recommendation system, a coding agent, and an AI system used in a safety-critical environment. Third, rapidly changing capabilities may render some evaluation procedures inadequate over time.

Future research should examine how bounded self-evolution can be operationalized in specific sectors. Empirical studies could compare incident rates, adaptation speed, intervention effectiveness, auditability, and recovery performance under different governance arrangements. Further work could also develop practical methods for classifying changes and determining when an AI system should be required to undergo renewed evaluation.

The broader implication is that responsible AI development should not pursue maximum autonomy as an objective in itself. Nor should it prevent every form of adaptation. Instead, the appropriate objective is to preserve beneficial capability development while ensuring that authority, responsibility, and acceptable risk remain subject to legitimate human governance.

VI. CONCLUSION

This study examined how AI systems can be permitted to improve while maintaining human oversight, safety, accountability, and control. Through a qualitative review of secondary sources and a comparative analysis of existing approaches, the study identified a central challenge: greater technical capability does not automatically produce greater reliability, safety, or social usefulness.

AI self-evolution encompasses several forms of adaptation, including learning from new data, modifying strategies, updating model components, and operating through increasingly complex tool-based workflows. Each form of change may provide benefits, but each may also introduce new risks. These risks include objective misalignment, specification gaming, distributional shift, loss of previously acquired knowledge, unauthorized changes, and ineffective human intervention.

The review demonstrates that existing AI safety and governance approaches provide important but complementary safeguards. Alignment research addresses the relationship between system objectives and intended outcomes. Corrigibility emphasizes the importance of accepting correction and interruption. Continual learning research investigates how systems can acquire new knowledge while retaining existing capabilities. Risk-management frameworks establish processes for identifying, evaluating, and managing risks, while documentation and auditing support accountability.

However, none of these approaches independently guarantees that an evolving system will remain safe. Effective governance requires their coordinated application throughout the system lifecycle.

The principal recommendation is to adopt bounded self-evolution. Under this approach, AI systems may perform routine and reversible improvements within clearly defined operating limits. Material changes to capabilities, objectives, permissions, or deployment conditions should trigger additional testing, risk assessment, and, where appropriate, explicit human authorization. Continuous monitoring, traceable documentation, incident response, and rollback mechanisms should support these controls.

Human oversight must also be substantive rather than symbolic. Designated individuals need the information, competence, authority, and technical means required to intervene. At the same time, oversight should remain proportionate to risk so that low-risk adaptation is not unnecessarily obstructed and high-consequence changes receive sufficient scrutiny.

The proposed framework offers a conceptual foundation for connecting technical adaptation with organizational responsibility. Its contribution lies in treating changes to an AI system as governance events that may require different

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

levels of evaluation and authorization. This approach may help organizations distinguish beneficial improvement from unacceptable expansion of autonomy.

Nevertheless, the framework has not been experimentally validated, and the study does not establish numerical thresholds for acceptable risk or demonstrate that any particular combination of controls eliminates the possibility of failure. Future research should test the framework in specific applications, evaluate the effectiveness of different change-classification methods, and investigate how oversight can remain meaningful as AI systems become more capable.

In conclusion, controlling AI self-evolution should not be understood as preventing AI from becoming more useful. It should be understood as ensuring that improvements remain compatible with human objectives, safety requirements, accountability, and the ability to intervene. A responsible approach permits innovation within explicit boundaries while ensuring that consequential changes remain subject to transparent evaluation and legitimate human authority.

REFERENCES

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv. https://doi.org/10.48550/arXiv.1606.06565
2. Anthropic. (2024, June 17). Sycophancy to subterfuge: Investigating reward tampering in language models. https://www.anthropic.com/research/reward-tampering
3. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1
4. Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
5. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., ... Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Future of Humanity Institute, University of Oxford. https://doi.org/10.17863/CAM.22520
6. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
7. Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 220–227). https://doi.org/10.24963/ijcai.2017/32
8. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwińska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114
9. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596
10. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
11. Organisation for Economic Co-operation and Development. (2024). OECD AI principles. https://oecd.ai/en/ai-principles
12. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., &

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

13. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.
14. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html
15. UK Government. (2025, February 7). Frontier AI safety commitments, AI Seoul Summit 2024. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024
16. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, A., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv. https://doi.org/10.48550/arXiv.2112.04359
17. Zhu, L., Lu, Q., Ding, M., Lee, S. U., Wang, C., et al. (2026). Designing meaningful human oversight in AI. AI and Ethics. https://doi.org/10.1007/s43681-026-01147-7