Open Science
BRAZIL

DOI: https://doi.org/10.37497/opsbrazil.44

PREPRINT

DELEGATING INFORMATION SECURITY GOVERNANCE WORK TO AUTONOMOUS ARTIFICIAL INTELLIGENCE AGENTS: DELEGABILITY CRITERIA, AUTONOMY LEVELS, AND ACCOUNTABILITY

Delegação de Trabalho de Governança de Segurança da Informação a Agentes Autônomos de Inteligência Artificial: Critérios de Delegabilidade, Níveis de Autonomia e Responsabilização

MATHEWS HENRIQUE DA CRUZ
MBA em Segurança da Informação, Universidade Católica de Brasília, Akuris Pesquisa e Desenvolvimento
E-mail: mathews.cruz@nexure.com.br

ABSTRACT | The increasing adoption of autonomous artificial intelligence agents in corporate environments is redefining the boundaries between human work and automated execution in information security governance. Activities such as access reviews, request handling, compliance evidence collection, alert triage, and inventory updates involve significant operational components; however, their automation raises a governance question that is not adequately addressed by the binary distinction between manual and automated work: what degree of autonomy is admissible for a given task, and under what conditions can such delegation remain accountable, auditable, and defensible? This article proposes a task-based model for determining the admissible degree of delegation of information security governance activities to autonomous artificial intelligence agents. The model considers five attributes: criterion determinism, outcome verifiability, effect reversibility, regulatory consequence, and the cost of error at scale. These attributes are associated with five delegation levels, ranging from human execution to autonomous execution with exception-based control. Progression between levels is conditioned on measured and stable divergence between agent decisions and human decisions rather than on perceptions of technological maturity. The model is examined through a single case study conducted in a medium-sized Brazilian organization. During calibration, divergence between agent-proposed decisions and human decisions was 14%, decreasing to 3% after the decision criteria stabilized. Average resolution time also decreased from approximately 48 hours to approximately 10 minutes. The

Submetido: 02-09-2026
Postado: 02-09-2026
e-ISSN: 2764-9822

Endereço de correspondência:
Mathews Henrique da Cruz – E-mail: mathews.cruz@nexure.com.br

findings provide preliminary empirical evidence supporting the proposition that autonomy should be assigned according to task characteristics and demonstrated performance, while accountability remains linked to an identifiable human owner.

Keywords | Autonomous Artificial Intelligence Agents; Information Security Governance; Artificial Intelligence Governance; Automation; Identity And Access Management; Human Oversight; Accountability; Autonomy Levels.

1. INTRODUCTION

Information security governance has traditionally relied on a combination of formal policies, organizational responsibilities, technical controls, operational procedures, and human judgment. A significant portion of the work performed in this context, however, is repetitive and procedural. Access request analysis, routine ticket handling, compliance evidence collection, information correlation across corporate systems, inventory maintenance, and first-level technical triage can consume a considerable share of organizational operational capacity without necessarily requiring continuous specialized judgment in every case.

The distinction between defining a criterion and applying that criterion becomes particularly relevant with the development of autonomous artificial intelligence systems. In many governance processes, the most intellectually complex activity occurs when an organization establishes a policy, determines an acceptable condition, defines an authorization boundary, or establishes what constitutes an exception. Once these criteria have been defined, a substantial portion of the subsequent work consists of applying them consistently to individual cases.

The development of artificial intelligence systems capable of operating as agents has expanded the technical possibilities for delegating this operational layer. Unlike conventional automation based exclusively on deterministic scripts or predefined workflows, contemporary agents can interpret contextual information, consult multiple sources, interact with enterprise applications, use tools, and execute sequences of actions. Consequently, the system no longer functions solely as an information-processing mechanism but acquires operational capabilities that allow it to modify the state of corporate systems.

A governance problem therefore emerges that differs from the traditional question of whether a given process can be automated. The question becomes whether authority to execute a particular task should be delegated to an autonomous system, to what degree such authority should be granted, and which controls must remain under human responsibility.

This distinction is particularly relevant because both the literature and business practice often continue to approach automation from a binary perspective: a task is either performed manually or automated. This classification is insufficient for autonomous artificial intelligence agents. Between fully human execution and fully autonomous execution without intervention, there are multiple possibilities for distributing responsibilities, including informational support, decision preparation, human approval, autonomous execution, post-execution auditing, and exception-based intervention.

The idea that automation should be understood as a spectrum rather than a binary condition is not new. Studies in human factors have demonstrated that different levels of automation modify the distribution of functions between people and automated systems, with consequences for workload, performance, situational awareness, and control. The work of Endsley (1997) and Parasuraman et al. (2000), among others, established important foundations for understanding automation as a graduated distribution of functions between humans and technical systems.

Current artificial intelligence governance frameworks add important foundations to this discussion. The NIST Artificial Intelligence Risk Management Framework provides a voluntary and context-independent framework for managing risks associated with artificial intelligence, organized around the functions Govern, Map, Measure, and Manage. Governance operates as a cross-cutting element throughout the AI system life cycle.

Similarly, ISO/IEC 42001:2023 establishes requirements for an artificial intelligence management system, emphasizing organizational processes for the responsible management of AI systems. ISO/IEC 27001:2022, in turn, establishes a management foundation for information security. These frameworks provide relevant foundations for governance, accountability, risk management, and continual improvement, but they do not, by themselves, provide a sufficiently specific mechanism for addressing an operational question: what degree of autonomy should be assigned to a particular information security governance task?

This article starts from the proposition that this decision should not be determined primarily by the type of task, the sophistication of the artificial intelligence technology, or the organization's subjective perception of technological maturity. Rather, the admissible degree of autonomy should be determined primarily by the characteristics of the task to be delegated.

To operationalize this proposition, five attributes are assessed: criterion determinism, referring to the extent to which a decision can be expressed through explicit and auditable rules; outcome verifiability, referring to the possibility of objectively confirming whether the resulting outcome is correct; effect reversibility, referring to the possibility and cost of reversing an incorrect action; regulatory consequence, referring to the potential for an action to generate legal or regulatory effects or obligations toward third parties; and the cost of error at scale, referring to the potential for a systematic error to be repeatedly applied across a large volume of operations.

Based on these attributes, the article proposes a five-level delegation model. At Level 0, the agent provides informational support only and does not execute actions within the process. At Level 1, the agent collects, correlates, and prepares information while a human makes and executes the decision. At Level 2, the agent prepares the complete action, but execution requires explicit human approval. At Level 3, the agent executes autonomously while humans perform post-execution sampling audits and exception analysis. Finally, Level 4 allows autonomous execution with exception-based control and is reserved for high-volume processes with stable criteria, fully reversible effects, and appropriate controls.

Progression between these levels should not occur automatically. The model proposes that a task should only be promoted to a higher level after demonstrating measured and stable performance within predefined thresholds. In this way, autonomy ceases to be a decision based exclusively on perceptions of technological maturity and becomes a decision grounded in operational evidence.

A central indicator proposed in this study is the divergence rate between the decision produced by the agent and the decision that would have been made by the human reference process. During an initial period of assisted operation, agent decisions can be compared with human decisions. Once the divergence rate stabilizes within a predefined threshold, the task may be considered eligible for promotion to a higher level of autonomy.

The second fundamental proposition of this article concerns accountability. Delegating execution does not automatically transfer responsibility for the process.

An agent may execute an action, but accountability for the delegated process remains linked to the organization and to an identifiable human owner. For delegation to remain defensible in the context of an audit, incident investigation, or potential regulatory scrutiny, it must be possible to reconstruct not only the action performed but also its context, the criteria applied, the permissions used, and the associated human responsibility.

Four conditions are therefore considered fundamental: a distinct identity for the agent; least-privilege permissions and a defined operational scope; a formally identified human owner for each delegated process; and an audit trail capable of reconstructing the action and its context. An interruption and rollback mechanism that does not depend on the agent itself must also exist.

This concern is also related to the risk known as excessive agency in systems based on large language models. OWASP identifies excessive agency as a relevant risk when an AI system is provided with functionalities, permissions, or autonomy beyond what is necessary for the task it performs. Excessive authority can cause an incorrect or unexpected output to produce significant operational consequences.

The proposal presented in this article therefore treats autonomy as a controlled allocation of authority rather than as an absolute characteristic of an artificial intelligence system. The object of the governance decision is not simply the agent, but the task delegated to the agent.

The empirical basis of the study is a single case involving the design and implementation of autonomous agents in a medium-sized Brazilian organization whose identity is protected by a confidentiality agreement. The implementation includes a proprietary identity and access management solution with automated provisioning, role-based profiles, periodic access reviews, integration with the corporate directory, and integration with the corporate ticketing system.

At the time of analysis, autonomous agents performed all first-level technical support activities and the full set of access management processes covered by the implementation, under continuous sampling-based audit.

More than one thousand requests were processed autonomously between March 2026 and the period in which this study was written. During the calibration phase, divergence between agent-proposed decisions and human decisions was 14%. After stabilization of the criteria used by the agent, divergence decreased to 3%.

A significant reduction in average resolution time was also observed. Under the previous model, operated by the support team, average resolution time was approximately 48 hours. Following the adoption of autonomous execution, this time decreased to approximately 10 minutes from the moment the user opened the ticket.

These results are relevant not because they demonstrate that autonomous agents should replace human professionals, but because they illustrate how operational evidence can support governance decisions regarding the admissible level of autonomy. In the case analyzed, the transition to autonomous execution occurred only after the divergence rate had stabilized. Sampling-based auditing remained in place after this transition and therefore constituted a permanent control rather than merely a temporary implementation stage.

The case also demonstrates why the model should not be interpreted as a universal prescription for autonomy. The same organization may assign different autonomy levels to different tasks depending on their characteristics and consequences. A highly deterministic and reversible access management operation may justify autonomous execution, whereas an action involving irreversible consequences, direct regulatory obligations, or a high degree of contextual judgment may require human decision-making even when performed by the same technological agent.

The research question guiding this study is:

> **Which attributes of an information security governance task determine its admissible degree of delegation to autonomous artificial intelligence agents, and which control conditions must be preserved for such delegation to remain defensible under audit and regulatory scrutiny?**

The article presents three main contributions. First, it proposes a task-based model for assessing the admissible degree of delegation of information security governance activities. Second, it proposes a five-level delegation model linking task characteristics to different degrees of operational autonomy. Third, it proposes the use of measured divergence and its stabilization as an empirical condition for progression between autonomy levels, examined through a real-world case study.

The remainder of the article presents the theoretical foundation, describes the methodology, develops the five delegability criteria and five autonomy levels, presents the case study and its results, discusses the implications and limitations of the proposal, and concludes the study.

2. THEORETICAL FOUNDATION AND LITERATURE REVIEW

2.1 Automation as a Spectrum of Delegation Between Humans and Systems

Understanding automation as a spectrum of different levels of human participation provides a relevant conceptual foundation for analyzing the delegation of work to autonomous artificial intelligence agents. Although current systems offer capabilities significantly beyond traditional forms of automation, the fundamental problem remains the distribution of functions and authority between humans and technical systems.

Endsley (1997) argues that different levels of automation produce distinct consequences for human-system interaction, particularly with regard to situational awareness, decision-making, and control. Automation does not simply represent the replacement of a person with a machine; rather, it represents a redistribution of functions within a process.

Parasuraman, Sheridan, and Wickens (2000) proposed a framework for understanding levels of automation across different stages of information processing, including information acquisition, analysis, alternative selection, and execution. This perspective is particularly relevant to autonomous agents because it allows different components of the same task to be understood as having different degrees of automation.

For example, in an access management process, collecting the information required for a decision may be highly automatable, whereas deciding how to handle an exception may require human judgment. Likewise, executing a previously approved authorization may present a different risk profile from the one associated with initially determining the authorization.

The practical consequence of this perspective is that the question “Is the process automated?” has limited usefulness for autonomous agent governance. A more appropriate question is: “Which parts of the process are delegated, with what degree of authority, and under which controls?”

This shift in perspective is central to the model proposed in this article. The unit of analysis is neither the organization as a whole nor the agent as an abstract entity, but the specific task subject to delegation.

2.2 Autonomous Agents and Operational Authority

The evolution of systems based on language models has introduced an important difference from traditional automation. A conventional system generally operates according to a predefined sequence of rules or commands. An artificial intelligence agent may interpret information, select tools, construct a sequence of actions, and interact with different systems to achieve a given objective.

This capability significantly expands automation potential but also changes the nature of risk. When a system merely produces a recommendation, the result remains subject to

human decision-making. When the same system is authorized to execute an action, an interpretation error can immediately produce a change in the corporate environment.

The issue therefore extends beyond the accuracy of the output produced by the system to include the extent of authority granted to transform that output into action.

The concept of excessive agency developed by OWASP is particularly relevant in this context. The organization identifies the granting of functionalities, permissions, or autonomy beyond what is necessary to perform a given task as a risk. The problem is not necessarily the existence of autonomy but the absence of proportionality between the authority granted and the function performed.

From this perspective, an important principle emerges: the reliability of an agent should not, by itself, be used as justification for expanding its permissions.

Even an agent with a high accuracy rate can produce systematic errors. A flaw in its decision criterion may be repeatedly applied to hundreds or thousands of requests before being detected. Therefore, autonomy assessment must consider not only the probability of error but also the potential impact of its repetition at scale.

2.3 Artificial Intelligence Governance and Accountability

Artificial intelligence governance shifts a significant part of the technological discussion to the organizational level. The issue is not merely to develop capable systems but to establish structures that allow organizations to understand, measure, manage, and assign accountability for the outcomes produced by such systems.

The NIST Artificial Intelligence Risk Management Framework establishes four fundamental functions: Govern, Map, Measure, and Manage. The framework emphasizes that risk management should occur continuously throughout the AI system life cycle.

For autonomous agents, this principle implies that the decision to grant operational authority should not be considered permanent. System behavior, environmental conditions, and associated risks may change over time. Consequently, delegation must remain subject to monitoring and review.

ISO/IEC 42001:2023 complements this perspective by establishing requirements for artificial intelligence management systems. The standard introduces an organizational approach to managing risks and opportunities associated with AI use.

When autonomous agents are employed in information security processes, this governance operates at the intersection of two domains: AI governance and information security governance.

The agent must be governed as an artificial intelligence system but also as an operational entity receiving access to corporate resources. This characteristic reinforces the need to distinguish three concepts that are frequently treated as equivalent: technical capability, operational authorization, and accountability.

A system may be technically capable of performing a particular action without being authorized to perform it autonomously. Similarly, a system may be authorized to execute a task without accountability for the outcome being transferred to the system.

The model proposed in this article is based precisely on this distinction.

2.4 Non-Human Identity, Least Privilege, and Zero Trust

Delegating work to autonomous agents requires these systems to be treated as identifiable operational entities.

The use of human credentials by agents creates significant challenges for auditing and accountability. If an action is executed using an individual's identity, it becomes more difficult to determine whether the operation was performed directly by that individual, by an agent acting on their behalf, or by another automated mechanism.

The Zero Trust architecture provides an applicable principle for this problem. NIST SP 800-207 establishes that implicit trust should not exist simply because a user or asset is located within a particular network. Access should be explicitly authenticated and authorized based on the requested resource and relevant conditions.

Applied to autonomous agents, this principle supports the use of a dedicated identity for each agent or autonomous function, associated with least-privilege permissions and a predefined operational scope.

This separation also improves auditability. The organization can identify that a specific action was executed by a particular agent, at a particular time, using a particular permission and within a defined process.

The principle of least privilege assumes additional importance because it limits the potential impact of an error. If an agent possesses only the permissions required to perform its authorized task, a decision error is likely to have a more limited impact than it would if the same agent possessed broad administrative privileges.

Thus, identity, authorization, and autonomy should not be treated independently. The degree of autonomy granted to an agent must be accompanied by a proportional definition of its technical authority.

2.5 Identity and Access Management as an Application Domain

Identity and access management constitutes a particularly relevant domain for studying delegation because it combines high operational volume, relatively structured rules, and potentially significant consequences.

Processes such as user provisioning, profile changes, periodic access reviews, and permission termination can be based on objective information, including organizational role, employment status, contractual status, and segregation-of-duties rules.

At the same time, exceptions may require contextual analysis. A user may have a temporary access requirement, an unusual role, or a combination of permissions that is not adequately represented by a standard rule.

This characteristic makes identity and access management appropriate for demonstrating why different tasks within the same domain may justify different levels of autonomy.

A routine, deterministic, and reversible operation may be executed autonomously. An exception that significantly changes a user's privileges may require human approval. The agent does not need to be classified globally as autonomous or non-autonomous; autonomy can be determined according to the type of operation.

This approach allows organizations to preserve automation efficiency gains without transforming the entire process into an indiscriminate delegation of authority.

2.6 Verifiability, Reversibility, and Operational Risk

Two attributes are particularly important in determining the appropriate level of autonomy: the ability to verify the outcome and the ability to reverse its effects.

A task whose outcome can be immediately verified provides stronger conditions for post-execution control. If an operation is performed and an objective mechanism exists to determine whether the result corresponds to the expected criterion, the organization has greater capacity to detect deviations.

Reversibility operates as a complementary attribute. An incorrect action that can be quickly undone presents a different risk profile from an action whose effects are permanent or whose reversal is technically or legally complex.

The combination of these attributes makes it possible to distinguish tasks suitable for post-execution auditing from those requiring prior approval.

For example, a configuration change that can be automatically reversed and whose compliance can be immediately verified has characteristics different from an external communication issued on behalf of the organization, permanent deletion of data, or an action that creates an obligation toward a regulatory authority.

Regulatory consequence adds another layer of caution. Even when an agent demonstrates strong performance, certain decisions may remain subject to human responsibility because they produce legal or regulatory effects that cannot be treated simply as technical operations.

2.7 Systematic Errors and the Risk of Scale

An error made by a human worker and an error made by an autonomous agent have an important structural difference.

A human error may occur as an isolated event, although it may also result from systemic failures. An agent, however, can reproduce exactly the same error across a large number of operations.

This characteristic changes the nature of risk. The average error rate is insufficient to assess the safety of delegation. It is also necessary to determine whether errors are concentrated in particular classes of cases and whether an inappropriate criterion is being applied systematically.

For this reason, the proposed model incorporates the cost of error at scale as one of the five delegability criteria.

A task involving thousands of daily operations and a possibility of systematic error requires different controls from those applicable to a low-volume task.

Sampling-based auditing can be effective when errors are approximately randomly distributed. However, if a flawed criterion causes every decision in a particular category to be incorrect, a seemingly adequate sample may not immediately reveal the problem.

This consideration also justifies the need for continuous monitoring and independent interruption mechanisms.

2.8 Identified Gap and Proposed Contribution

The automation literature provides foundations for understanding different levels of human participation. AI governance frameworks provide structures for risk management, accountability, and organizational control. Security frameworks provide principles concerning identity, authorization, least privilege, and auditing. Literature and best practices concerning AI agents add the concern of excessive authority and agency.

However, an operational gap remains: how can these principles be translated into an objective mechanism for deciding how much autonomy a specific task can receive?

The model presented in this article seeks to address this gap through five observable task attributes: determinism, verifiability, reversibility, regulatory consequence, and cost of error at scale.

The proposal is not intended to replace existing frameworks. Rather, it is intended to function as a complementary operational layer, enabling organizations to translate general governance principles into concrete delegation decisions.

The principal conceptual contribution is therefore to shift the question from “Is the agent trustworthy?” to “Is this task appropriate for this level of autonomy, given its characteristics and available control mechanisms?”

This shift allows the same technology to be used with different levels of authority across different processes, preserving proportionality between technical capability, operational risk, and organizational accountability.

3. METHODOLOGY

3.1 Research Design

This study adopts a qualitative and applied approach structured as a single case study, supported by a prior theoretical foundation and empirical analysis of an implementation of autonomous agents in real information security governance and operational processes.

The choice of a single case study reflects the applied nature of the research problem. The objective is not to produce statistical generalization across all organizations using autonomous agents, but to examine a concrete implementation in depth and use the observed evidence to assess the applicability of the proposed delegation criteria and levels.

The study focuses on the relationship between task characteristics, the degree of autonomy granted to agents, and the control mechanisms preserved by the organization. Accordingly, the unit of analysis is not the artificial intelligence model in isolation but the sociotechnical structure formed by the agent, organizational process, corporate systems involved, and mechanisms for supervision and accountability.

The investigation was conducted using three principal sources of evidence: documentary analysis of execution records, comparison between agent-proposed decisions and human decisions during assisted operation, and direct observation of the design, deployment, and stabilization of the solution.

3.2 Case Study Context

The case analyzed concerns the implementation of autonomous agents in a medium-sized Brazilian organization. The organization's identity is withheld due to a confidentiality agreement.

The implementation covered processes related to information technology technical support and identity and access management. Within the identity and access domain, a proprietary solution was developed and used to automate activities related to user provisioning, role-based profile assignment, periodic access reviews, and integration with the corporate directory and ticketing system.

These processes were selected because of their operational characteristics. They involve a significant volume of requests, recurring use of structured information, and application of predefined organizational criteria.

At the same time, these processes have relevant security consequences because inappropriate access decisions may result in unauthorized resource exposure, segregation-of-duties violations, or retention of privileges inconsistent with the user's role.

The case therefore provides suitable conditions for examining the central tension investigated in this study: the possibility of achieving efficiency gains through autonomy without eliminating the control mechanisms required for information security governance.

3.3 Implementation Process

The implementation did not begin at the maximum autonomy level. The process started with assisted operation, in which the agent generated a decision or recommendation and the corresponding decision was subject to human evaluation.

This period enabled systematic comparison between agent behavior and the expected behavior according to the human reference criteria.

The comparison was used to identify divergences, classify their causes, and assess the stability of the criteria employed by the agent.

During this phase, divergence between the agent's proposed decision and the human decision was 14% of the analyzed cases. This initial divergence was treated as part of the calibration process rather than as sufficient evidence that the technology was unsuitable for the task.

As the criteria stabilized and results were monitored, the divergence rate decreased to 3%.

Only after this stabilization did the organization transition to execution without prior human approval for the processes covered by the study. Post-execution sampling-based auditing remained in place.

This implementation sequence is relevant to the proposed model because it provides empirical observation of the hypothesis that progression between autonomy levels should depend on performance evidence rather than solely on perceptions of technological maturity.

3.4 Evidence Sources and Collection Procedures

The first source of evidence consisted of execution records from the automated processes. These records enabled observation of operations performed by the agents, cases processed, and outcomes produced.

The second source consisted of comparisons between agent-generated decisions and human decisions during assisted operation. This comparison enabled calculation of the divergence rate and monitoring of its evolution throughout implementation.

The third source was direct observation of the implementation and operational processes. This observation made it possible to identify aspects that might not be captured by structured records, including dependence on source-data quality, the need to adjust decision criteria, changes in integrated systems, and challenges associated with maintaining human supervisory capacity.

The combination of these sources enabled an evidence-based analysis grounded both in system execution and in the organizational context in which the agents were deployed.

3.5 Variables and Analytical Criteria

The analysis was structured around the five delegability attributes proposed in this article.

The first attribute was criterion determinism, assessed according to the extent to which the expected decision could be expressed through clear, consistent, and auditable rules.

The second was outcome verifiability, referring to the possibility of objectively confirming whether the execution performed by the agent produced the expected result.

The third was effect reversibility, considering the possibility of undoing an incorrect action and the technical and operational costs associated with reversal.

The fourth attribute was regulatory consequence, considering whether an action could produce legal effects, regulatory obligations, or direct consequences for data subjects, third parties, or authorities.

The fifth was the cost of error at scale, considering the potential impact of systematically repeating an incorrect decision across a high volume of operations.

In addition to these attributes, the empirical analysis used the divergence rate between human decisions and agent-proposed decisions as an operational indicator of the evolution of autonomy.

Divergence was not interpreted simply as a system error rate. A divergence represents a difference between two decisions and may result from agent failure, inconsistency in the human reference process, insufficient information, policy ambiguity, or the need to revise the underlying criterion itself.

For this reason, interpreting divergence requires qualitative analysis of its causes rather than relying solely on its percentage value.

3.6 Progression Criterion Between Levels

The methodological proposal establishes that promotion of a task to a higher level of autonomy should occur only after a period of operation at the preceding level and observation of performance within predefined thresholds.

This principle was empirically applied in the case studied.

Assisted operation identified an initial divergence rate of 14%. Following stabilization of the criteria, the rate decreased to 3%. This reduction was used as evidence supporting the transition from execution requiring prior approval to autonomous execution with post-execution auditing.

The procedure does not assume that a particular divergence rate is universally acceptable for every task. The appropriate threshold depends on the characteristics, consequences, and risks of the operation under analysis.

Consequently, the 3% rate observed in this study should not be interpreted as a general safety threshold for autonomous agents. It constitutes case-specific evidence and demonstrates the value of establishing metrics before promoting a task to a higher level of autonomy.

3.7 Operational Indicators

Two indicators were used to characterize implementation outcomes.

The first was the divergence rate between agent decisions and human decisions during calibration and stabilization.

The second was average request resolution time.

Before implementation, the reported average resolution time for support activities performed by the support team was approximately 48 hours. After autonomous execution was implemented, observed average resolution time decreased to approximately 10 minutes from the moment the user opened the ticket.

An additional saving of more than one million Brazilian reais was recorded in the area's budget as a consequence of automating the analyzed processes and related activities.

Although this economic result is relevant for characterizing the organizational impact of the implementation, it is not used as an isolated indicator of security or autonomy suitability. Financial gains do not replace governance, security, auditability, or accountability criteria.

3.8 Validity Considerations and Methodological Limitations

The methodology presents limitations inherent to the single-case-study design.

The first limitation concerns the absence of statistical generalization. Results observed in a medium-sized Brazilian organization cannot automatically be extrapolated to organizations in different sectors, sizes, or maturity levels.

The second limitation concerns possible dependence on the organization's specific technological context. Infrastructure characteristics, data quality, integrated systems, and internal policies may significantly influence agent performance.

The third limitation derives from the nature of the divergence measure itself. The human decision used as the reference may also contain inconsistencies or limitations. Therefore, a reduction in divergence does not necessarily mean that all agent decisions are objectively correct.

The fourth limitation is temporal. Autonomous agents operate in environments subject to change. Changes in integrated systems, internal policies, user profiles, or regulatory requirements may alter the suitability of a criterion that was previously stable.

These limitations reinforce the need to interpret the model as a framework for governance and continuous assessment rather than as a definitive certification mechanism for autonomy.

4. DELEGABILITY CRITERIA

The principal proposition of this study is that the admissible degree of autonomy should be determined by the characteristics of the delegated task. To operationalize this proposition, five delegability criteria are defined.

The criteria should not be applied as a purely mechanical classification. A task may exhibit favorable characteristics under some criteria and unfavorable characteristics under others. The objective is to provide a structured assessment of the combination of these factors.

4.1 Criterion Determinism

The first criterion is the degree of determinism of the rule or criterion guiding the decision.

A task exhibits high determinism when its decision can be described through explicit, consistent, and auditable rules. The lower the dependence on contextual interpretation, the greater the potential for delegation.

For example, a rule specifying that a particular user profile should receive a defined set of permissions based on job role may present a high degree of determinism when the required data are correctly structured.

By contrast, a decision that depends on interpretation of exceptional circumstances, conflicting interests, or information that cannot be adequately represented through defined criteria presents lower determinism.

Determinism therefore does not mean that a process is simple. A task may involve large amounts of data and still have a highly deterministic criterion.

The relevant question is whether the decision criterion can be explicitly defined and audited.

The higher the determinism, the greater the potential admissible level of autonomy.

4.2 Outcome Verifiability

The second criterion concerns the ability to verify whether the execution produced by the agent was correct.

A task is more suitable for autonomy when an objective mechanism exists to confirm the outcome after execution.

Verifiability allows part of the control function to shift from prior approval to post-execution auditing.

For example, if an agent executes an access change and the system allows immediate verification that the user possesses exactly the permissions prescribed for the user's role, conditions are favorable for post-execution verification.

Conversely, when an outcome can only be evaluated much later, depends on subjective judgment, or cannot be objectively reconstructed, prior approval becomes more important.

Verifiability must also consider the quality of the mechanism used to confirm the result. A superficial verification does not provide the same level of assurance as an independent and technically reliable verification mechanism.

4.3 Effect Reversibility

The third criterion is the possibility of reversing an incorrect action.

Reversibility reduces the potential impact of errors because it allows the previous state to be restored.

An access-granting operation may, under certain architectures, be reversed by removing the granted permission. An action that permanently deletes information, generates external communication, or creates a legal obligation may have substantially different characteristics.

Reversibility also has an economic and operational dimension. It is not sufficient for an action to be theoretically reversible. The organization must assess the time, effort, and resources required for reversal and the consequences that may occur before the reversal is completed.

The more reversible the effect and the lower the cost of reversal, the greater the potential for post-execution control.

4.4 Regulatory Consequence

The fourth criterion considers whether an action may produce legal or regulatory consequences.

Certain actions should not be treated as simple technical operations because they generate effects involving data subjects, customers, third parties, authorities, or regulators.

The existence of a direct regulatory consequence reduces the admissible level of autonomy, particularly when the action involves a formal declaration, external communication, fulfillment of a legal obligation, or exercise of a decision requiring human accountability.

In this context, the agent's technical reliability does not eliminate the need for human decision-making.

The proposed principle is that legal and regulatory authority should not be implicitly transferred to an autonomous system merely because technical execution can be automated.

4.5 Cost of Error at Scale

The fifth criterion considers the impact of systematic errors.

An autonomous agent can execute a task hundreds or thousands of times. This capability is one of the principal sources of technological efficiency but also increases the potential impact of a flawed decision criterion.

If an agent incorrectly applies a rule to a single case, the impact may be limited. If the same flawed rule is automatically applied to thousands of users, the problem assumes an entirely different dimension.

The cost of error at scale must therefore consider not only the severity of an individual error but also the potential number of operations affected before the error is identified and corrected.

This criterion reinforces the importance of continuous monitoring, detection of divergence patterns, and independent interruption mechanisms.

A high-volume task may be an excellent candidate for automation, but precisely for that reason it requires controls capable of rapidly identifying systematic errors.

5. FIVE-LEVEL DELEGATION MODEL

Based on the criteria presented above, this study proposes a five-level model for classifying the delegation of information security governance work to autonomous agents.

The levels represent increasing degrees of operational authority.

5.1 Level 0 — Human Execution

At Level 0, the agent has no authority to execute actions within the process.

Its use is restricted to consultation, information organization, research, analytical support, or other functions that do not directly alter the state of the process.

Decision-making and execution remain entirely under human responsibility.

This level is appropriate for tasks involving high contextual complexity, significant regulatory consequences, low verifiability, or limited reversibility.

Level 0 may also represent the initial assessment stage for a new AI application before any operational authority is granted.

5.2 Level 1 — Assisted Preparation

At Level 1, the agent participates directly in decision preparation.

It may collect information, query systems, correlate data, identify patterns, and prepare a recommendation or set of actions.

The decision remains human, and execution is also performed by a person.

This level is particularly suitable for tasks in which information collection and organization are labor-intensive but final judgment depends on context.

The primary advantage of Level 1 is that it enables productivity gains without transferring operational authority to the agent.

5.3 Level 2 — Execution with Prior Approval

At Level 2, the agent can prepare the complete action and potentially perform the technical operation, but execution requires explicit human approval.

This level represents the recommended entry point for productive processes in which the organization intends to test agent capability before granting higher autonomy.

Human approval functions as a control barrier before the system state is changed.

During this stage, the organization can measure divergence between agent-proposed decisions and decisions that would have been made by humans.

Level 2 is particularly suitable for calibration and operational learning.

5.4 Level 3 — Execution with Post-Execution Audit

At Level 3, the agent executes the task autonomously without individual human approval for each operation.

Human control is shifted to the post-execution stage through sampling-based audits, exception analysis, and monitoring of process indicators.

This level is appropriate for tasks with relatively deterministic criteria, verifiable outcomes, and sufficiently reversible effects.

Transition from Level 2 to Level 3 should not occur merely because the agent has produced satisfactory results for a given period. The organization's defined indicators must demonstrate stable performance within predefined thresholds.

The case study analyzed in this article operates under this condition.

5.5 Level 4 — Autonomous Execution with Exception-Based Control

Level 4 represents the highest degree of autonomy defined by the model.

At this level, the agent normally executes operations and only routes cases that violate predefined conditions for human review.

Control shifts from predominantly post-execution sampling to exception-based mechanisms capable of identifying situations outside the expected pattern.

This level should be reserved for high-volume tasks with highly stable criteria, high verifiability, fully reversible effects, and low potential for direct regulatory consequences.

Even at this level, autonomy does not mean the absence of governance.

The organization must maintain a dedicated agent identity, least privilege, audit trails, performance monitoring, and an external interruption mechanism.

5.6 Progression Between Levels

Progression between levels constitutes a governance decision and must be evidence-based.

The proposed model establishes the following logic:

Level 0 → Level 1 → Level 2 → measurement → stabilization → Level 3 → continuous monitoring → potential Level 4.

Movement from one level to another should consider:

- divergence rate;
- divergence-rate stability;
- nature of identified divergences;
- severity of errors;
- outcome verifiability;
- action reversibility;
- changes in the technological environment;
- changes in internal policies;
- regulatory changes; and
- human intervention capability.

Autonomy should therefore not be treated as a permanent state achieved by the agent. It should remain conditional on the continued existence of the conditions that justified its authorization.

If those conditions cease to exist, the task must be capable of returning to a lower autonomy level.

This controlled degradation of autonomy is a direct consequence of the proposed approach: if promotion depends on evidence, continued autonomy must also depend on the persistence of that evidence.

6. ACCOUNTABILITY LAYER

Delegating work to autonomous agents does not eliminate the need for human accountability. On the contrary, the greater the degree of autonomy granted to a system, the greater the need to clearly establish who authorized the delegation, what limits were defined, and who is accountable for the process.

The central premise of this section is that execution may be delegated, but responsibility for the process is not automatically transferred to the agent.

The agent does not, by itself, constitute an accountable subject for the organizational process. Its actions occur within a governance structure defined by the organization and must remain linked to identifiable human owners.

For delegation to remain defensible during an audit, incident investigation, or potential regulatory scrutiny, four fundamental conditions must be preserved.

6.1 Distinct Agent Identity

Each agent capable of executing actions in corporate systems should have a distinct identity separate from identities used by human personnel.

This identity must make it possible to determine unequivocally that a particular action was executed by the agent rather than directly by a human user.

Using a personal credential to allow an agent to perform operations creates a break in the accountability chain. System records may indicate that a particular person performed an action when the operation was actually executed by an automated mechanism.

A distinct identity also enables specific access policies to be applied to the agent, its permissions to be restricted, and its activity to be independently monitored.

The agent should possess only the privileges necessary to perform the function for which it has been authorized.

6.2 Human Process Owner

Each delegated process should have a formally identified human owner.

The owner does not need to personally execute every operation performed by the agent but should be accountable for defining the criteria, authorizing delegation, monitoring indicators, and reviewing the conditions that justify maintaining autonomy.

This structure prevents automation from creating diffuse accountability in which no individual can explain why a task was delegated, what limits were established, or who should act in response to a failure.

Accountability should remain associated with the process rather than with the technical capabilities of the agent.

Thus, even when a process operates at Level 4, the organization must be able to identify the human owner responsible for governing that activity.

6.3 Audit Trail and Decision Reconstruction

Auditing an autonomous agent cannot be limited to recording the final action.

To adequately reconstruct a decision, the audit trail should make it possible, to the extent technically feasible, to identify the context that originated the operation, the information used, the criterion applied, the authorization available, the action executed, and the resulting outcome.

The objective is not necessarily to record every internal operation of the artificial intelligence model but to preserve sufficient evidence to reconstruct the operational decision-making process.

A subsequent investigation should be able to answer questions such as:

- Which agent executed the operation?
- When did the operation occur?
- Which object or user was affected?
- What information was available?

- Which rule or policy governed the operation?
- What permissions did the agent possess?
- What action was executed?
- What outcome was produced?
- Who was the human owner of the process?

The ability to answer these questions constitutes a fundamental element of auditability.

6.4 Independent Interruption Mechanism

Every delegated process should have an interruption mechanism that does not depend exclusively on the agent itself.

This condition is particularly important when the agent exhibits inappropriate behavior, systematic errors, or silent degradation.

If the same system responsible for execution is also the only mechanism capable of interrupting its own activity, the organization may face an operational dependency incompatible with an adequate control structure.

The interruption mechanism should therefore be operationally independent from the agent and available to process owners.

Whenever technically possible, a rollback mechanism should also exist to restore the previous state or limit the effects of an inappropriate operation.

6.5 Accountability and Autonomy

The four elements above establish a fundamental distinction between operational autonomy and organizational accountability.

An agent may operate autonomously without the organization relinquishing control over the process.

Autonomy means that execution of a given operation does not require individual human intervention.

Accountability means that a human and organizational structure remains capable of responding for the definition, authorization, monitoring, and review of that operation.

This distinction prevents increased autonomy from being interpreted as a transfer of responsibility to the system.

7. CASE STUDY AND EMPIRICAL RESULTS

7.1 Case Characterization

The case study concerns the implementation of autonomous agents in a medium-sized Brazilian organization whose identity is protected by a confidentiality agreement.

The solution involved proprietary agents and systems applied to information technology processes, particularly first-level technical support and identity and access management.

Within identity and access management, the solution included automated provisioning, role-based profile assignment, periodic access reviews, and integration with the corporate directory and ticketing system.

The architecture allowed agents to receive requests, consult relevant information, apply predefined criteria, and execute operations in corporate systems within their authorized permissions.

At the time of analysis, first-level technical support and all access management processes covered by the implementation were executed autonomously under continuous sampling-based auditing.

7.2 Operational Volume

Between March 2026 and the period in which the study was written, more than one thousand requests were processed autonomously.

This volume is relevant because it allows system behavior to be observed under real operational conditions rather than limiting evaluation to controlled tests or laboratory demonstrations.

Production use also allowed the identification of issues involving source-data quality, changes in integrated systems, and rule behavior across different request types.

However, processing volume alone should not be interpreted as evidence of safety or reliability. The number of operations demonstrates sufficient operational exposure for evaluation, but delegation quality depends on the combination of volume, task characteristics, performance, and control mechanisms.

7.3 Divergence During Calibration

The principal indicator used to monitor agent evolution was divergence between the system's proposed decision and the decision produced by the human reference process.

During calibration, observed divergence was 14%.

This result indicated that the operation had not yet achieved sufficient stability to justify removing individual human approval.

The agent therefore remained under assisted operation while the criteria were analyzed and adjusted.

After stabilization of the criteria, the divergence rate decreased to 3%.

This represents an 11-percentage-point reduction between the initial and stabilized periods.

More important than the absolute value is the trajectory observed.

The experience demonstrated that autonomy assessment can be conducted progressively: first the agent proposes, then its decisions are compared with human decisions, and only after stability has been observed can operational authority be expanded.

This result provides empirical support for the proposition that progression between autonomy levels should be conditioned on measured performance.

7.4 Reduction in Resolution Time

Another observed outcome was a reduction in average request resolution time.

Under the previous model, performed by the support team, the reported average resolution time was approximately 48 hours.

Following implementation of autonomous agents, average resolution time decreased to approximately 10 minutes from the moment the user opened the ticket.

This difference represents a significant operational improvement.

However, reduced resolution time does not, by itself, justify expanding autonomy.

Speed gains must be evaluated together with decision quality, action reversibility, audit mechanisms, and the other delegability criteria.

This distinction is important because automation capable of rapidly executing an inappropriate rule can increase the speed at which an error is produced.

7.5 Operational Savings

Automation of the analyzed processes and related activities resulted in savings of more than one million Brazilian reais in the area's budget.

This economic result demonstrates that delegation of work can produce a significant financial impact in addition to reducing resolution time.

However, the economic benefit is treated in this study as an organizational outcome of implementation rather than as a sufficient criterion for determining the level of autonomy.

A task should not receive greater authority merely because its automation generates savings. Delegation decisions must remain linked to process characteristics and required controls.

7.6 Case Positioning Within the Proposed Model

Following an extended period of assisted operation and stabilization of the divergence rate, first-level support and access management began operating without prior human approval for each request.

At the same time, post-execution sampling-based auditing was maintained.

This configuration corresponds to **Level 3 — Execution with Post-Execution Audit** in the proposed model.

This classification is relevant because it demonstrates that operational autonomy does not mean absence of control.

The agent possesses authority to execute the defined operations, but its activity remains subject to subsequent verification and possible intervention.

The case is not classified as Level 4 because execution still depends on sampling-based auditing rather than exclusively on exception-based control.

7.7 Relationship Between Divergence and Autonomy Progression

The reduction from 14% to 3% constitutes the principal empirical result associated with the proposed progression mechanism.

During the initial phase, divergence indicated that application of the criteria by the agent had not yet achieved sufficient stability.

After calibration, the reduction and stabilization of divergence provided evidence supporting the transition from Level 2 to Level 3.

The result does not demonstrate that 3% is a universally appropriate threshold.

What the case demonstrates is the usefulness of establishing a threshold before promotion, measuring performance against that threshold, and conditioning increased autonomy on observed stability.

This distinction is essential to prevent the model from being interpreted as a universal mathematical formula.

The acceptable percentage should be determined according to the nature, impact, and reversibility of the task.

8. DISCUSSION

8.1 The Task as the Unit of Governance Decision

The case study results reinforce the proposition that the task should constitute the primary unit for autonomy decisions.

The same agent may operate across different processes with different levels of authority.

This approach is more appropriate than classifying an entire organization or an entire agent as “autonomous” or “non-autonomous.”

A technology may be used to prepare decisions in a high-risk activity while simultaneously executing routine and reversible operations autonomously in another process.

This granularity allows organizations to capture productivity gains without transforming agent adoption into a generalized grant of authority.

8.2 Autonomy as a Function of Risk and Control

The results also indicate that autonomy should not be understood exclusively as a technical property.

An agent may demonstrate substantial processing capabilities, access multiple systems, and perform satisfactorily on a particular task. Nevertheless, this does not mean that it should be granted authority to execute every possible action.

Admissible autonomy results from the relationship between task characteristics and available controls.

A highly deterministic, verifiable, and reversible task presents favorable conditions for autonomous execution.

A task dependent on contextual judgment, difficult to verify, irreversible, or associated with direct regulatory consequences requires greater human participation.

The model therefore establishes a relationship between risk and authority.

The greater the potential consequence of a decision, the greater the requirement for prior control or human intervention.

8.3 The Importance of Measurement

The principal empirical evidence from the case concerns the reduction in divergence from 14% to 3%.

This evolution demonstrates the value of transforming autonomy progression into a measurable process.

Under approaches based exclusively on subjective judgment, an organization may increase autonomy because it considers a system “mature,” “reliable,” or “stable.”

Such assessments may be useful but are insufficient for a governance decision.

Metrics provide a verifiable condition.

The proposed model does not establish a universal divergence rate. Instead, it requires the organization to establish in advance the threshold considered acceptable for a given task and verify whether agent behavior remains within that threshold.

This approach brings autonomy governance closer to a continuous control process.

8.4 The Problem of Systematic Error

The analysis also reinforces the importance of distinguishing random errors from systematic errors.

Sampling-based auditing is useful for identifying deviations but may be insufficient when an agent consistently applies an inappropriate rule.

If a particular class of requests is incorrectly handled in every case, a small sample may not immediately reveal the problem.

For this reason, autonomy controls should combine sampling-based auditing with indicators capable of detecting behavioral changes, abnormal concentrations of divergence, and changes in the execution environment.

This need is particularly important in integrated systems, where changes in a source system may alter the meaning of data used by the agent without generating an obvious technical failure.

8.5 Silent Degradation

One of the limitations observed in the study is the possibility of silent degradation.

An agent may continue to execute its operations technically while the context in which its criteria were defined changes.

A change in an internal policy, organizational structure, integrated system, or data record may cause a previously appropriate rule to produce inappropriate outcomes.

This type of problem is particularly dangerous because it does not necessarily generate technical errors.

The system continues to operate, tickets continue to be processed, and actions continue to be recorded.

The problem appears in the content of the decisions rather than in system availability.

This characteristic reinforces the need for periodic review of criteria and monitoring of the environment in which autonomy operates.

8.6 Data Quality as a Condition for Delegability

The case also demonstrates that source-data quality is an important condition for delegation.

An agent does not automatically eliminate inconsistencies existing in corporate systems.

When a decision depends on incorrect, incomplete, or outdated information, automation may reproduce and amplify the problem.

This characteristic reinforces the idea that autonomy should be evaluated not only in relation to the agent but also in relation to the entire process.

A task that is suitable for autonomy in an organization with reliable data may not be suitable in another organization with poorly structured records.

Information quality should therefore be treated as a prerequisite for delegating certain activities.

8.7 Preservation of Human Competence

Another limitation identified concerns the potential loss of team competence.

When a process is fully delegated for extended periods, human professionals may no longer practice it regularly.

This may reduce the team's ability to manually perform the activity if the agent becomes unavailable and may also make it more difficult to assess the quality of automated decisions.

There is therefore an automation paradox: the more efficient a system becomes, the less human exposure to the process may occur and, consequently, the smaller the residual human capability available to supervise it.

Governance of autonomous agents should account for this risk.

Human oversight cannot depend on professionals who no longer understand the process they are supervising.

8.8 Accountability as a Permanent Condition

The experience analyzed also reinforces the need to maintain accountability regardless of the autonomy level.

The transition from Level 2 to Level 3 did not eliminate human accountability.

The agent began executing without individual approval, but remained subject to auditing and linked to an accountability structure.

This characteristic distinguishes autonomy from loss of control.

Operational autonomy should be accompanied by mechanisms for identification, authorization, logging, supervision, and intervention.

Accountability should therefore not be viewed as an additional stage of automation but as a permanent layer of the governance architecture.

8.9 Organizational Implications

The model presents several practical implications for organizations seeking to use autonomous agents in information security processes.

First, organizations should inventory the tasks they intend to delegate rather than evaluate autonomy solely at the system level.

Second, each task should be assessed according to the five proposed criteria.

Third, organizations should initiate new processes at an autonomy level compatible with their risk, preferably using assisted operation when significant uncertainty exists.

Fourth, progression should be metric-based.

Fifth, each agent should have a distinct identity, least-privilege permissions, and a human owner.

Sixth, organizations should maintain independent interruption and rollback mechanisms.

Finally, autonomy should be reviewed whenever significant changes occur in the process, integrated systems, internal policies, or regulatory requirements.

9. STUDY LIMITATIONS

This study presents limitations that must be considered when interpreting its findings.

The primary limitation is the single-case-study design. Analysis of one organization does not allow the observed results to be assumed to occur similarly in organizations with different sectors, sizes, technological architectures, or maturity levels.

The second limitation is dependence on the specific implementation context. Data quality, process structure, internal policies, and the systems used may have contributed to the observed results.

The third limitation concerns the divergence indicator. Comparing agent decisions with human decisions does not necessarily represent a comparison between an incorrect decision and a correct decision. The human reference process may also contain inconsistencies.

The fourth limitation concerns the observation period. Although more than one thousand requests were processed and the system was used in a real operational environment, longer observation periods are required to adequately assess phenomena such as silent degradation, policy changes, and shifts in request distributions.

The fifth limitation is the absence of systematic comparison across multiple organizations. Replication in other environments would be necessary to assess the robustness of the model and identify which criteria have greater explanatory power across different contexts.

These limitations do not invalidate the proposal but define the scope of its conclusions. The model should be regarded as an initial proposition supported by empirical evidence from an applied case, whose external validity requires further investigation.

10. CONCLUSION

The adoption of autonomous artificial intelligence agents in information security processes creates significant opportunities to reduce operational workload, increase speed, and improve efficiency. At the same time, the ability of these systems to interpret information and directly execute actions within corporate environments introduces a governance question that cannot be adequately addressed by simply distinguishing between manual and automated processes.

This article proposed that delegation decisions should be made at the task level rather than exclusively at the technology level.

To operationalize this proposal, five criteria were defined: criterion determinism, outcome verifiability, effect reversibility, regulatory consequence, and cost of error at scale.

These criteria enable organizations to assess whether a particular task presents characteristics compatible with greater or lesser autonomy.

Based on these criteria, a five-level delegation model was proposed, ranging from fully human execution to autonomous execution with exception-based control.

The model establishes that autonomy should increase progressively and remain conditioned on evidence of performance.

The principal empirical evidence obtained from the case study was the reduction in divergence between agent decisions and human decisions from 14% during calibration to 3% after stabilization of the criteria.

This reduction supported the transition from Level 2, characterized by execution subject to human approval, to Level 3, characterized by autonomous execution with post-execution auditing.

The case also demonstrated significant operational gains. Average resolution time decreased from approximately 48 hours to approximately 10 minutes, while automation of the analyzed and related processes resulted in savings of more than one million Brazilian reais in the area's budget.

These results demonstrate that work delegation can produce significant benefits but also reinforce that efficiency should not be used in isolation as a justification for expanding autonomy.

Execution speed does not replace security, auditability, and accountability criteria.

The analysis also identified important limitations. Source-data quality directly influences the quality of automated decisions. Systematic errors can be reproduced at scale. Changes in systems, policies, or data can produce silent degradation. Prolonged delegation may also reduce the human competence required to supervise or reassume the process.

These factors demonstrate that autonomy should not be treated as a permanent state.

Maintaining a given level of autonomy should depend on the continued existence of the conditions that justified its authorization. When those conditions cease to exist, the organization should be able to reduce the autonomy level or interrupt automated execution.

Accountability constitutes the second central element of the proposal.

Execution may be delegated to an agent, but responsibility for the process remains linked to the organization and to an identifiable human owner.

For delegation to remain defensible, the organization must preserve a distinct agent identity, least privilege, a human process owner, an audit trail, and an independent mechanism for interruption and rollback.

The central contribution of this study is therefore the proposal of a framework that enables autonomy to be treated as a measurable and risk-proportionate governance decision.

Rather than simply asking whether an organization should or should not use autonomous agents, the proposed framework allows organizations to ask which tasks can be delegated, at what level, based on which evidence, and under which control conditions.

This approach also helps overcome a potential false dichotomy between automation and human work.

The issue is not to replace people entirely with agents but to distribute functions proportionally to task characteristics.

Deterministic, verifiable, and reversible tasks may support greater autonomy. Tasks dependent on contextual judgment, involving irreversible effects or direct regulatory consequences, should retain greater human participation.

The case study provides initial evidence supporting this approach, particularly by demonstrating that autonomy promotion accompanied by objective measurement can be implemented progressively.

However, because the study is based on a single case, the results should not be considered generalizable without further investigation.

Future research may apply the model across organizations in different sectors, compare different types of agents, and longitudinally evaluate the stability of delegation criteria.

It will also be important to investigate quantitative methods for defining divergence thresholds, mechanisms for detecting systematic errors, and strategies for preserving human competence in highly automated environments.

The principal conclusion of this study is therefore that autonomy should not be granted to an agent based on its technological capability alone, but to a task based on its characteristics, associated risks, and the existence of controls capable of preserving accountability.

Defensible delegation is not delegation that eliminates human intervention. It is delegation that clearly establishes where human intervention is no longer necessary, what evidence justifies that decision, and which mechanisms remain available to regain control when conditions change.

REFERENCES

Endsley, M. R. (1997). Designing for situation awareness: An approach to user-centered design. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1*, 272–276.

International Organization for Standardization. (2022). *ISO/IEC 27001:2022 information security, cybersecurity and privacy protection — Information security management systems — Requirements*. ISO.

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 information technology — Artificial intelligence — Management system*. ISO.

National Institute of Standards and Technology. (2020). *Security and privacy controls for information systems and organizations* (NIST Special Publication 800-53, Rev. 5). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-53r5

National Institute of Standards and Technology. (2020). Zero trust architecture (NIST Special Publication 800-207). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-207

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

OWASP Foundation. (2025). OWASP Top 10 for large language model applications. OWASP Foundation.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

AUTHOR'S BIOGRAPHY

Mathews Henrique da Cruz é profissional da área de tecnologia e segurança da informação, com formação em MBA em Segurança da Informação pela Universidade Católica de Brasília. Atua no desenvolvimento e na aplicação de soluções voltadas à automação, governança de segurança da informação, gestão de identidade e acesso e utilização de agentes autônomos de inteligência artificial em ambientes corporativos. Seus interesses profissionais e acadêmicos concentram-se na integração entre inteligência artificial, segurança cibernética, automação de processos e modelos de governança, com ênfase em autonomia, controle, auditabilidade e responsabilização.

Mathews Henrique da Cruz *is a technology and information security professional with an MBA in Information Security from the Catholic University of Brasília. He works on the development and implementation of solutions focused on automation, information security governance, identity and access management, and the use of autonomous artificial intelligence agents in corporate environments. His professional and academic interests include the integration of artificial intelligence, cybersecurity, process automation, and governance frameworks, with particular emphasis on autonomy, control, auditability, and accountability.*