Open Science
BRAZIL

DOI: https://doi.org/10.37497/opsbrazil.42

PREPRINT

DELEGAÇÃO DE TRABALHO DE GOVERNANÇA DE SEGURANÇA DA INFORMAÇÃO A AGENTES AUTÔNOMOS DE INTELIGÊNCIA ARTIFICIAL: CRITÉRIOS DE DELEGABILIDADE, NÍVEIS DE AUTONOMIA E RESPONSABILIZAÇÃO

Delegating Information Security Governance Work to Autonomous Artificial Intelligence Agents: Delegability Criteria, Autonomy Levels, and Accountability

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

RESUMO | A crescente adoção de agentes autônomos de inteligência artificial em ambientes corporativos está modificando os limites entre trabalho humano e execução automatizada na governança de segurança da informação. Atividades como revisão de acessos, tratamento de solicitações, coleta de evidências de conformidade, triagem de alertas e atualização de inventários possuem componentes operacionais significativos, mas sua automação produz uma questão de governança que não é adequadamente respondida pela distinção binária entre trabalho manual e automatizado: qual grau de autonomia é admissível para determinada tarefa e sob quais condições essa delegação permanece responsável, auditável e defensável? Este artigo propõe um modelo baseado na tarefa para determinar o grau admissível de delegação de atividades de governança de segurança da informação a agentes autônomos de inteligência artificial. O modelo considera cinco atributos: determinismo do critério, verificabilidade do resultado, reversibilidade do efeito, consequência regulatória e custo do erro em escala. Esses atributos são relacionados a cinco níveis de delegação, que variam da execução humana à execução autônoma com controle por exceção. A progressão entre os níveis é condicionada à divergência mensurada e estável entre as decisões do agente e as decisões humanas, e não à percepção de maturidade tecnológica. O modelo é analisado por meio de um estudo de caso único realizado em uma organização brasileira de médio porte. Durante a calibração, a divergência entre as decisões propostas

Submetido: 18-08-2026
Postado: 18-08-2026
e-ISSN: 2764-9822

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

pelo agente e as decisões humanas foi de 14%, reduzindo-se para 3% após a estabilização dos critérios. O tempo médio de resolução também caiu de aproximadamente 48 horas para cerca de 10 minutos. Os resultados fornecem evidências empíricas preliminares em favor da proposição de que a autonomia deve ser atribuída de acordo com as características da tarefa e com o desempenho demonstrado, enquanto a responsabilização permanece vinculada a um responsável humano identificável.

Palavras-chave | Agentes Autônomos De Inteligência Artificial; Governança De Segurança Da Informação; Governança De Inteligência Artificial; Automação; Gestão De Identidade E Acesso; Supervisão Humana; Responsabilização; Níveis De Autonomia.

ABSTRACT | The adoption of autonomous artificial intelligence agents in information security processes expands automation capabilities while creating new governance and accountability challenges. This article proposes a model for determining the appropriate degree of task delegation to autonomous agents based on five criteria: determinism, verifiability, reversibility, regulatory consequences, and the cost of errors at scale. The model establishes five levels of autonomy, ranging from human execution to autonomous execution with exception-based control. The proposal is examined through a case study conducted in a medium-sized Brazilian organization. During calibration, the divergence between agent decisions and human decisions decreased from 14% to 3%, while average resolution time decreased from approximately 48 hours to 10 minutes. The results indicate that autonomy should be assigned according to task characteristics and demonstrated performance, while human accountability remains preserved.

Keywords |: Autonomous artificial intelligence agents; information security governance; automation; identity and access management; accountability.

1. INTRODUÇÃO

A governança de segurança da informação tradicionalmente depende de uma combinação de políticas formais, responsabilidades organizacionais, controles técnicos, procedimentos operacionais e julgamento humano. Uma parcela significativa do trabalho realizado nesse contexto, entretanto, é repetitiva e procedural. A análise de solicitações de acesso, o tratamento de chamados rotineiros, a coleta de evidências para atividades de conformidade, a correlação de informações provenientes de sistemas corporativos, a manutenção de inventários e a triagem técnica de primeiro nível podem consumir parcela considerável da capacidade operacional das organizações sem necessariamente exigir, em todos os casos, o exercício contínuo de julgamento especializado.

A distinção entre definir um critério e aplicar esse critério torna-se particularmente relevante diante do desenvolvimento de sistemas autônomos de inteligência artificial. Em diversos processos de governança, a atividade de maior complexidade intelectual ocorre quando a organização estabelece uma política, determina uma condição aceitável, define uma fronteira de autorização ou estabelece o que constitui uma exceção. Uma vez definidos esses critérios, parte significativa do trabalho subsequente consiste em aplicá-los de maneira consistente a casos individuais.

O desenvolvimento de sistemas de inteligência artificial capazes de operar como agentes ampliou as possibilidades técnicas de delegação dessa camada operacional. Diferentemente da automação convencional baseada exclusivamente em scripts determinísticos ou

fluxos de trabalho previamente definidos, agentes contemporâneos podem interpretar informações contextuais, consultar diferentes fontes, interagir com aplicações corporativas, utilizar ferramentas e executar sequências de ações. Dessa forma, o sistema deixa de atuar exclusivamente como mecanismo de processamento de informação e passa a possuir capacidade operacional para modificar o estado de sistemas corporativos.

Surge, portanto, um problema de governança diferente da pergunta tradicional sobre se determinado processo pode ser automatizado. A questão passa a ser se a autoridade para executar determinada tarefa deve ser delegada a um sistema autônomo, em qual grau essa autoridade deve ser concedida e quais controles precisam permanecer sob responsabilidade humana.

Essa distinção é especialmente relevante porque a literatura e a prática empresarial ainda frequentemente tratam automação a partir de uma perspectiva binária: uma tarefa é executada manualmente ou é automatizada. Essa classificação é insuficiente para agentes autônomos de inteligência artificial. Entre a execução integralmente humana e a execução autônoma sem intervenção existem diferentes possibilidades de distribuição de responsabilidades, incluindo apoio informacional, preparação de decisões, aprovação humana, execução autônoma, auditoria posterior e intervenção baseada em exceções.

A ideia de que a automação deve ser compreendida como um espectro, e não como uma condição binária, não é recente. Estudos na área de fatores humanos demonstraram que diferentes níveis de automação modificam a distribuição das funções entre pessoas e sistemas automatizados, produzindo efeitos sobre carga de trabalho, desempenho, consciência situacional e controle. Trabalhos de Endsley (1997) e Parasuraman et al. (2000), entre outros, estabeleceram fundamentos importantes para compreender a automação como uma distribuição graduada de funções entre seres humanos e sistemas.

Os atuais frameworks de governança de inteligência artificial acrescentam fundamentos importantes a essa discussão. O NIST Artificial Intelligence Risk Management Framework apresenta uma estrutura voluntária e aplicável a diferentes contextos para gerenciamento de riscos relacionados à inteligência artificial, organizada nas funções Governar, Mapear, Medir e Gerenciar. A governança funciona como elemento transversal ao longo do ciclo de vida dos sistemas de IA.

Da mesma forma, a ISO/IEC 42001:2023 estabelece requisitos para um sistema de gestão de inteligência artificial, enfatizando processos organizacionais destinados ao gerenciamento responsável de sistemas de IA. A ISO/IEC 27001:2022, por sua vez, estabelece a base de gestão para a segurança da informação. Esses referenciais fornecem fundamentos relevantes para governança, responsabilização, gestão de riscos e melhoria contínua, mas não oferecem, por si só, um mecanismo suficientemente específico para responder a uma questão operacional: qual grau de autonomia deve ser atribuído a determinada tarefa de governança de segurança?

Este artigo parte da proposição de que essa decisão não deve ser determinada principalmente pelo tipo da tarefa, pela sofisticação da tecnologia de inteligência artificial ou pela percepção subjetiva de maturidade tecnológica da organização. O grau admissível

de autonomia deve ser determinado, principalmente, pelas características da própria tarefa que será delegada.

Propõe-se, para isso, a avaliação de cinco atributos: determinismo do critério, correspondente ao grau em que a decisão pode ser expressa por regras explícitas e auditáveis; verificabilidade do resultado, correspondente à possibilidade de confirmar objetivamente se o resultado produzido está correto; reversibilidade do efeito, correspondente à possibilidade e ao custo de desfazer uma ação incorreta; consequência regulatória, correspondente ao potencial de a ação produzir efeitos jurídicos, regulatórios ou obrigações perante terceiros; e custo do erro em escala, correspondente ao potencial de um erro sistemático ser repetido sobre grande volume de operações.

A partir desses atributos, o artigo propõe um modelo de cinco níveis de delegação. No Nível 0, o agente oferece apenas apoio informacional e não executa ações no processo. No Nível 1, o agente coleta, correlaciona e prepara informações, enquanto o ser humano toma e executa a decisão. No Nível 2, o agente prepara a ação completa, mas sua execução depende de aprovação humana explícita. No Nível 3, o agente executa autonomamente e o ser humano realiza auditoria posterior por amostragem e análise de exceções. Finalmente, o Nível 4 permite execução autônoma com controle baseado em exceções, sendo reservado a processos de alto volume, com critérios estáveis, efeitos plenamente reversíveis e controles adequados.

A progressão entre esses níveis não deve ocorrer automaticamente. O modelo propõe que uma tarefa somente seja promovida para um nível superior após apresentar desempenho mensurado e estável dentro de limites previamente estabelecidos. Dessa maneira, a autonomia deixa de ser uma decisão baseada exclusivamente na percepção de maturidade da tecnologia e passa a ser uma decisão fundamentada em evidências operacionais.

Um indicador central proposto neste estudo é a taxa de divergência entre a decisão produzida pelo agente e a decisão que seria tomada pelo processo humano de referência. Durante um período inicial de operação assistida, as decisões propostas pelo agente podem ser comparadas às decisões humanas. Uma vez que a taxa de divergência se estabilize dentro de um limite previamente estabelecido, a tarefa pode ser considerada apta à promoção para um nível superior de autonomia.

A segunda proposição fundamental do artigo diz respeito à responsabilização. A delegação da execução não transfere automaticamente a responsabilidade pelo processo.

O agente pode executar uma ação, mas a responsabilidade pelo processo delegado permanece vinculada à organização e a um responsável humano identificável. Para que a delegação seja defensável diante de auditoria, investigação de incidente ou eventual escrutínio regulatório, deve ser possível reconstruir não apenas a ação realizada, mas também seu contexto, os critérios aplicados, as permissões utilizadas e a responsabilidade humana associada.

Para isso, quatro condições são consideradas fundamentais: uma identidade própria e distinta para o agente; privilégios mínimos e escopo operacional definido; um responsável humano formalmente identificado para cada processo delegado; e uma trilha de auditoria

capaz de reconstruir a ação e seu contexto. Deve existir ainda um mecanismo de interrupção e reversão que não dependa do próprio agente para funcionar.

Essa preocupação também se relaciona ao risco conhecido como agência excessiva em sistemas baseados em grandes modelos de linguagem. A OWASP identifica a agência excessiva como um risco relevante quando um sistema de IA recebe funcionalidades, permissões ou autonomia superiores ao necessário para a tarefa que desempenha. A concessão excessiva de autoridade pode fazer com que uma saída incorreta ou inesperada produza consequências operacionais significativas.

A proposta deste artigo, portanto, trata autonomia como uma atribuição controlada de autoridade, e não como uma característica absoluta do sistema de inteligência artificial. O objeto da decisão de governança não é simplesmente o agente, mas a tarefa que será delegada ao agente.

A base empírica do estudo é um caso único envolvendo a concepção e implementação de agentes autônomos em uma organização brasileira de médio porte, cuja identificação é preservada em razão de acordo de confidencialidade. A implementação inclui uma solução proprietária de gestão de identidade e acesso, com provisionamento automatizado, perfis baseados em função, revisão periódica de acessos, integração com o diretório corporativo e integração com o sistema corporativo de chamados.

No momento da análise, agentes autônomos executam todo o atendimento de primeiro nível do suporte técnico de tecnologia da informação e a totalidade dos processos de gestão de acessos da organização, sob regime de auditoria contínua por amostragem.

Mais de mil solicitações foram processadas autonomamente entre março de 2026 e o período de redação do estudo. Durante a fase de calibração, a divergência entre as decisões propostas pelo agente e as decisões humanas foi de 14%. Após a estabilização dos critérios utilizados pelo agente, essa divergência caiu para 3%.

Também foi observada uma redução expressiva no tempo médio de atendimento. No modelo anterior, executado pela equipe de suporte, o tempo médio de resolução era de aproximadamente 48 horas. Após a adoção da execução autônoma, esse tempo passou para aproximadamente 10 minutos a partir da abertura do chamado pelo usuário.

Esses resultados são relevantes não porque demonstrem que agentes autônomos devam substituir profissionais humanos, mas porque ilustram como evidências operacionais podem ser utilizadas para fundamentar uma decisão de governança sobre o nível de autonomia admissível. No caso analisado, a transição para execução autônoma ocorreu somente depois da estabilização da taxa de divergência. A auditoria por amostragem permaneceu após essa transição e, portanto, constituiu um controle permanente, e não apenas uma etapa transitória de implantação.

O caso também demonstra por que o modelo não deve ser interpretado como uma prescrição universal de autonomia. Uma mesma organização pode atribuir níveis diferentes de autonomia a diferentes tarefas, dependendo das características e consequências de cada uma. Uma operação de gestão de acesso altamente determinística e reversível pode justificar execução autônoma, enquanto uma ação que envolva consequência irreversível,

obrigação regulatória direta ou elevado grau de julgamento contextual pode exigir decisão humana mesmo quando realizada pelo mesmo agente tecnológico.

A pergunta de pesquisa que orienta o estudo é:

> **Quais atributos de uma tarefa de governança de segurança da informação determinam seu grau admissível de delegação a agentes autônomos de inteligência artificial e quais condições de controle precisam ser preservadas para que essa delegação permaneça defensável perante auditoria e escrutínio regulatório?**

O artigo apresenta três contribuições principais. A primeira é a proposição de um modelo baseado na tarefa para avaliar o grau admissível de delegação de atividades de governança de segurança da informação. A segunda é a proposição de um modelo de cinco níveis de delegação, relacionando características da tarefa a diferentes graus de autonomia operacional. A terceira é a proposição da utilização da divergência mensurada e de sua estabilização como condição empírica para a progressão entre níveis de autonomia, examinada por meio de um estudo de caso real.

O restante do artigo apresenta a fundamentação teórica, descreve a metodologia empregada, desenvolve os cinco critérios de delegabilidade e os cinco níveis de autonomia, apresenta o estudo de caso e seus resultados, discute as implicações e limitações da proposta e, por fim, apresenta as conclusões do estudo.

2. FUNDAMENTAÇÃO TEÓRICA E REVISÃO DE LITERATURA

2.1 Automação como espectro de delegação entre humanos e sistemas

A compreensão da automação como um espectro de diferentes níveis de participação humana constitui uma base conceitual relevante para a análise da delegação de trabalho a agentes autônomos de inteligência artificial. Embora os sistemas atuais apresentem capacidades significativamente superiores às formas tradicionais de automação, o problema fundamental permanece relacionado à distribuição de funções e autoridade entre seres humanos e sistemas técnicos.

Endsley (1997) argumenta que diferentes níveis de automação produzem consequências distintas para a interação entre seres humanos e sistemas, especialmente em relação à consciência situacional, à tomada de decisão e ao controle. A automação não representa simplesmente a substituição de uma pessoa por uma máquina, mas uma redistribuição de funções dentro de um processo.

Parasuraman, Sheridan e Wickens (2000) propuseram uma estrutura para compreender níveis de automação a partir de diferentes estágios do processamento de informação, incluindo aquisição de informação, análise, seleção de alternativas e execução. Essa perspectiva é especialmente relevante para agentes autônomos porque permite compreender

que diferentes componentes de uma mesma tarefa podem apresentar diferentes graus de automatização.

Em um processo de gestão de acessos, por exemplo, a coleta das informações necessárias para uma decisão pode ser altamente automatizável, enquanto a decisão sobre uma exceção pode exigir julgamento humano. Da mesma forma, a execução de uma autorização previamente aprovada pode apresentar risco diferente daquele associado à definição inicial da autorização.

A consequência prática dessa perspectiva é que a pergunta "o processo é automatizado?" possui pouca utilidade para a governança de agentes autônomos. A pergunta mais adequada é "quais partes do processo são delegadas, com qual grau de autoridade e sob quais controles?".

Essa mudança de perspectiva é central para o modelo proposto neste artigo. A unidade de análise não é a organização como um todo, nem o agente como uma entidade abstrata, mas a **tarefa concreta submetida à delegação**.

2.2 Agentes autônomos e autoridade operacional

A evolução dos sistemas baseados em modelos de linguagem introduziu uma diferença importante em relação à automação tradicional. Um sistema convencional geralmente opera segundo uma sequência de regras ou comandos previamente estabelecidos. Um agente baseado em inteligência artificial pode interpretar informações, selecionar ferramentas, construir uma sequência de ações e interagir com diferentes sistemas para atingir determinado objetivo.

Essa capacidade amplia significativamente o potencial de automação, mas também modifica a natureza do risco. Quando um sistema apenas produz uma recomendação, o resultado permanece sujeito à decisão humana. Quando o mesmo sistema possui autorização para executar uma ação, um erro de interpretação pode produzir imediatamente uma alteração no ambiente corporativo.

A questão, portanto, deixa de ser exclusivamente a precisão da resposta produzida pelo sistema e passa a incluir a extensão da autoridade concedida para transformar essa resposta em uma ação.

O conceito de agência excessiva desenvolvido pela OWASP é particularmente relevante nesse contexto. A organização identifica como risco a concessão de funcionalidades, permissões ou autonomia superiores ao necessário para a execução de uma determinada tarefa. O problema não está necessariamente na existência de autonomia, mas na ausência de proporcionalidade entre a autoridade concedida e a função desempenhada.

A partir dessa perspectiva, um princípio importante emerge: **a confiabilidade do agente não deve ser utilizada como justificativa isolada para ampliar suas permissões**.

Mesmo um agente que apresente elevada taxa de acerto pode produzir erros sistemáticos. Uma falha em seu critério de decisão pode ser aplicada repetidamente sobre centenas ou milhares de solicitações antes de ser identificada. Assim, a avaliação da autonomia precisa

considerar não apenas a probabilidade de erro, mas também o impacto potencial de sua repetição em escala.

2.3 Governança de inteligência artificial e responsabilização

A governança de inteligência artificial desloca parte significativa da discussão tecnológica para o nível organizacional. A questão não é apenas desenvolver sistemas capazes, mas estabelecer estruturas que permitam compreender, medir, gerenciar e responsabilizar os resultados produzidos por esses sistemas.

O NIST AI Risk Management Framework apresenta quatro funções fundamentais: Governar, Mapear, Medir e Gerenciar. A estrutura enfatiza que o gerenciamento de riscos deve ocorrer continuamente e ao longo do ciclo de vida do sistema de IA.

Para agentes autônomos, esse princípio implica que a decisão de conceder autoridade operacional não deve ser considerada definitiva. O comportamento do sistema, as condições do ambiente e os riscos associados podem mudar ao longo do tempo. Consequentemente, a delegação precisa estar sujeita a monitoramento e revisão.

A ISO/IEC 42001:2023 complementa essa perspectiva ao estabelecer requisitos para sistemas de gestão de inteligência artificial. A norma introduz uma abordagem organizacional para o gerenciamento de riscos e oportunidades associados ao uso de IA.

Quando agentes autônomos são utilizados em processos de segurança da informação, essa governança passa a existir na interseção entre dois domínios: governança de IA e governança de segurança da informação.

O agente precisa ser governado como sistema de inteligência artificial, mas também como uma entidade operacional que recebe acesso a recursos corporativos. Essa característica reforça a necessidade de separar três conceitos que frequentemente são tratados como equivalentes: capacidade técnica, autorização operacional e responsabilização.

Um sistema pode ser tecnicamente capaz de realizar determinada ação sem estar autorizado a realizá-la autonomamente. Da mesma forma, um sistema pode estar autorizado a executar uma tarefa sem que a responsabilidade pelo resultado tenha sido transferida para ele.

O modelo proposto neste artigo parte justamente dessa separação.

2.4 Identidade não humana, privilégio mínimo e Zero Trust

A delegação de trabalho a agentes autônomos exige que esses sistemas sejam tratados como entidades operacionais identificáveis.

A utilização de credenciais humanas por agentes cria dificuldades importantes para auditoria e responsabilização. Se uma ação é executada utilizando a identidade de uma pessoa, torna-se mais difícil determinar se a operação foi realizada diretamente pelo indivíduo, pelo agente em seu nome ou por outro mecanismo automatizado.

A arquitetura Zero Trust fornece um princípio aplicável a esse problema. O NIST SP 800-207 estabelece que não deve existir confiança implícita baseada simplesmente na localização do usuário ou do ativo na rede. O acesso deve ser autenticado e autorizado de maneira explícita, considerando o recurso e as condições da solicitação.

Aplicado a agentes autônomos, esse princípio favorece a utilização de uma identidade própria para cada agente ou função autônoma, associada a privilégios mínimos e a um escopo operacional previamente definido.

Essa separação também melhora a capacidade de auditoria. A organização pode identificar que determinada ação foi executada por determinado agente, em determinado momento, com determinada permissão e dentro de determinado processo.

O princípio do menor privilégio assume importância adicional porque limita a extensão de um eventual erro. Se um agente possui somente os privilégios necessários para executar determinada tarefa, um erro de decisão tende a produzir impacto mais limitado do que ocorreria caso o mesmo agente tivesse privilégios administrativos amplos.

Assim, identidade, autorização e autonomia não devem ser tratados separadamente. O grau de autonomia concedido a um agente precisa ser acompanhado de uma definição proporcional de sua autoridade técnica.

2.5 Gestão de identidade e acesso como domínio de aplicação

A gestão de identidade e acesso constitui um domínio particularmente relevante para o estudo da delegação porque combina alto volume operacional, regras relativamente estruturadas e consequências potencialmente significativas.

Processos como provisionamento de usuários, alteração de perfis, revisão periódica de acessos e encerramento de permissões podem ser baseados em informações objetivas, como função, vínculo organizacional, situação contratual e regras de segregação de funções.

Ao mesmo tempo, exceções podem exigir análise contextual. Um usuário pode possuir uma necessidade temporária de acesso, uma função incomum ou uma combinação de permissões que não esteja adequadamente representada em uma regra padrão.

Essa característica torna a gestão de identidade e acesso adequada para demonstrar por que diferentes tarefas dentro de um mesmo domínio podem justificar diferentes níveis de autonomia.

Uma operação rotineira, determinística e reversível pode ser executada autonomamente. Uma exceção que altere significativamente os privilégios de um usuário pode exigir aprovação humana. O agente não precisa ser classificado globalmente como autônomo ou não autônomo; a autonomia pode ser determinada conforme o tipo de operação.

Essa abordagem permite preservar o ganho de eficiência da automação sem transformar todo o processo em uma delegação indiscriminada de autoridade.

2.6 Verificabilidade, reversibilidade e risco operacional

Dois atributos assumem importância particular na determinação do nível de autonomia: a possibilidade de verificar o resultado e a possibilidade de reverter seus efeitos.

Uma tarefa cujo resultado pode ser verificado imediatamente oferece melhores condições para controle posterior. Se uma operação é realizada e existe um mecanismo objetivo para verificar se o resultado corresponde ao critério esperado, a organização possui maior capacidade de detectar desvios.

A reversibilidade atua de maneira complementar. Uma ação equivocada que possa ser desfeita rapidamente apresenta risco diferente de uma ação cujo efeito seja permanente ou cuja reversão seja tecnicamente ou juridicamente complexa.

A combinação desses atributos permite diferenciar tarefas que podem ser submetidas a auditoria posterior daquelas que exigem aprovação prévia.

Por exemplo, uma alteração de configuração que possa ser revertida automaticamente e cuja conformidade possa ser verificada imediatamente apresenta características distintas de uma comunicação externa enviada em nome da organização, da eliminação definitiva de dados ou de uma ação que gere obrigação perante uma autoridade reguladora.

A consequência regulatória adiciona uma camada adicional de cautela. Mesmo que um agente apresente desempenho elevado, determinadas decisões podem permanecer sob responsabilidade humana porque produzem efeitos jurídicos ou regulatórios que não podem ser tratados simplesmente como operações técnicas.

2.7 Erros sistemáticos e o risco da escala

O erro cometido por um trabalhador humano e o erro cometido por um agente autônomo possuem uma diferença estrutural relevante.

Um erro humano pode ocorrer de maneira isolada, embora também possa resultar de falhas sistêmicas. Um agente, entretanto, pode reproduzir exatamente o mesmo erro sobre um grande número de operações.

Essa característica altera a natureza do risco. A taxa média de erro não é suficiente para avaliar a segurança de uma delegação. Também é necessário determinar se os erros estão concentrados em determinadas classes de casos e se existe algum critério inadequado sendo aplicado sistematicamente.

Por esse motivo, o modelo proposto incorpora o **custo do erro em escala** como um dos cinco critérios de delegabilidade.

Uma tarefa que envolva milhares de operações diárias e uma possibilidade de erro sistemático exige controles diferentes daqueles aplicáveis a uma tarefa de baixo volume.

A auditoria por amostragem pode ser eficiente quando os erros são distribuídos de maneira aproximadamente aleatória. Entretanto, se uma falha de critério fizer com que todas as decisões de determinada categoria sejam incorretas, uma amostra aparentemente pequena pode não revelar imediatamente o problema.

Essa consideração também justifica a necessidade de monitoramento contínuo e de mecanismos independentes de interrupção.

2.8 Lacuna identificada e contribuição proposta

A literatura sobre automação fornece fundamentos para compreender diferentes níveis de participação humana. Os frameworks de governança de IA fornecem estruturas para gerenciamento de riscos, responsabilização e controle organizacional. Os frameworks de segurança fornecem princípios relacionados à identidade, autorização, menor privilégio e auditoria. A literatura e as boas práticas relacionadas a agentes de IA acrescentam a preocupação com excesso de autoridade e agência.

Entretanto, permanece uma lacuna operacional: como transformar esses princípios em um mecanismo objetivo para decidir **quanto de autonomia uma tarefa específica pode receber**?

O modelo apresentado neste artigo busca preencher essa lacuna por meio de cinco atributos observáveis da tarefa: determinismo, verificabilidade, reversibilidade, consequência regulatória e custo do erro em escala.

A proposta não pretende substituir frameworks existentes. Ao contrário, pretende funcionar como uma camada operacional complementar, permitindo que organizações traduzam princípios gerais de governança em decisões concretas sobre delegação.

A principal contribuição conceitual consiste, portanto, em deslocar a pergunta de "o agente é confiável?" para "esta tarefa é adequada a este nível de autonomia, dadas suas características e seus mecanismos de controle?".

Essa mudança permite que a mesma tecnologia seja utilizada com diferentes níveis de autoridade em diferentes processos, preservando a proporcionalidade entre capacidade técnica, risco operacional e responsabilidade organizacional.

3. METODOLOGIA

3.1 Delineamento da pesquisa

Este estudo adota uma abordagem qualitativa e aplicada, estruturada como estudo de caso único, com fundamentação teórica prévia e análise empírica de uma implementação de agentes autônomos em processos reais de governança e operação de segurança da informação.

A escolha do estudo de caso único decorre da natureza aplicada do problema investigado. O objetivo não é produzir uma generalização estatística sobre todas as organizações que utilizam agentes autônomos, mas analisar em profundidade uma implementação concreta e utilizar as evidências observadas para examinar a aplicabilidade dos critérios e níveis de delegação propostos.

O estudo concentra-se na relação entre características das tarefas, grau de autonomia concedido aos agentes e mecanismos de controle preservados pela organização. Dessa forma, o objeto de análise não é o modelo de inteligência artificial isoladamente, mas a estrutura sociotécnica formada pelo agente, pelo processo organizacional, pelos sistemas corporativos envolvidos e pelos mecanismos de supervisão e responsabilização.

A investigação foi conduzida a partir de três fontes principais de evidência: análise documental dos registros de execução, comparação entre as decisões propostas pelos agentes e as decisões tomadas por humanos durante o período de operação assistida, e observação direta do processo de concepção, implantação e estabilização da solução.

3.2 Contexto do estudo de caso

O caso analisado corresponde à implementação de agentes autônomos em uma empresa brasileira de médio porte. A identificação da organização é preservada em razão de acordo de confidencialidade.

A implementação compreendeu processos relacionados ao suporte técnico de tecnologia da informação e à gestão de identidades e acessos. No domínio de identidade e acesso, foi desenvolvida e utilizada uma solução proprietária para automatização de atividades relacionadas ao provisionamento de usuários, aplicação de perfis baseados em função, revisão periódica de acessos e integração com o diretório corporativo e o sistema de chamados.

A escolha desses processos para a implementação está relacionada às características operacionais apresentadas por eles. Trata-se de atividades que envolvem volume significativo de solicitações, utilização recorrente de informações estruturadas e aplicação de critérios organizacionais previamente definidos.

Ao mesmo tempo, esses processos possuem consequências de segurança relevantes, uma vez que decisões inadequadas de acesso podem resultar em exposição indevida de recursos, violação de segregação de funções ou manutenção de privilégios incompatíveis com a função desempenhada pelo usuário.

O caso, portanto, oferece condições adequadas para examinar a tensão central investigada neste estudo: a possibilidade de obter ganhos de eficiência por meio da autonomia sem eliminar os mecanismos de controle necessários à governança da segurança da informação.

3.3 Processo de implementação

A implementação não foi conduzida diretamente no nível de autonomia máxima. O processo iniciou-se com a operação assistida, na qual o agente produzia uma decisão ou recomendação e a decisão correspondente era submetida à avaliação humana.

Esse período permitiu comparar sistematicamente o comportamento do agente com o comportamento esperado segundo os critérios humanos de referência.

A comparação foi utilizada para identificar divergências, classificar suas causas e avaliar a estabilidade dos critérios empregados pelo agente.

Durante essa fase, a divergência observada entre a decisão proposta pelo agente e a decisão humana foi de 14% dos casos analisados. A existência dessa divergência inicial foi tratada como parte do processo de calibração e não como evidência suficiente para concluir que a tecnologia era inadequada para a tarefa.

Com a estabilização dos critérios e o acompanhamento dos resultados, a taxa de divergência caiu para 3%.

Somente após essa estabilização ocorreu a transição para execução sem aprovação humana prévia nos processos abrangidos pelo estudo. A alteração de nível de autonomia foi acompanhada pela manutenção da auditoria posterior por amostragem.

Essa sequência de implantação é relevante para o modelo proposto porque permite observar empiricamente a hipótese de que a progressão entre níveis de autonomia deve depender de evidências de desempenho e não apenas da percepção de maturidade tecnológica.

3.4 Fontes e procedimentos de coleta de evidências

A primeira fonte de evidência foi constituída pelos registros de execução dos processos automatizados. Esses registros permitiram observar as operações realizadas pelos agentes, os casos processados e os resultados produzidos.

A segunda fonte consistiu na comparação entre as decisões produzidas pelo agente e as decisões humanas durante o período de operação assistida. Essa comparação permitiu calcular a taxa de divergência e acompanhar sua evolução ao longo da implementação.

A terceira fonte foi a observação direta do processo de implantação e operação. Essa observação permitiu identificar aspectos que não seriam necessariamente capturados pelos registros estruturados, incluindo dependência da qualidade dos dados de origem, necessidade de ajustes nos critérios, alterações nos sistemas integrados e dificuldades relacionadas à manutenção da capacidade humana de supervisão.

A combinação dessas fontes permitiu realizar uma análise baseada em evidências provenientes tanto da execução do sistema quanto do contexto organizacional no qual ele foi implantado.

3.5 Variáveis e critérios de análise

A análise foi estruturada a partir dos cinco atributos de delegabilidade propostos neste artigo.

O primeiro atributo foi o **determinismo do critério**, avaliado a partir da possibilidade de expressar a decisão esperada por meio de regras claras, consistentes e auditáveis.

O segundo foi a **verificabilidade do resultado**, relacionada à possibilidade de confirmar objetivamente se a execução realizada pelo agente produziu o resultado esperado.

O terceiro foi a **reversibilidade do efeito**, considerando a possibilidade de desfazer uma ação incorreta e os custos técnicos e operacionais associados à reversão.

O quarto atributo foi a **consequência regulatória**, considerando se a ação poderia produzir efeitos jurídicos, obrigações regulatórias ou consequências diretas para titulares, terceiros ou autoridades.

O quinto foi o **custo do erro em escala**, considerando o impacto potencial da repetição sistemática de uma decisão incorreta sobre um volume elevado de operações.

Além desses atributos, a análise empírica utilizou a taxa de divergência entre decisões humanas e decisões propostas pelo agente como indicador operacional para avaliar a evolução da autonomia.

A divergência não foi interpretada simplesmente como uma taxa de erro do sistema. Uma divergência representa uma diferença entre duas decisões e pode decorrer de falha do agente, inconsistência no processo humano de referência, informação insuficiente, ambiguidade de política ou necessidade de revisão do próprio critério.

Por essa razão, a interpretação da divergência exige análise qualitativa de suas causas, e não apenas a observação de seu valor percentual.

3.6 Critério de progressão entre níveis

A proposta metodológica estabelece que a promoção de uma tarefa para um nível superior de autonomia deve ocorrer somente após um período de operação no nível anterior e após a observação de desempenho dentro de limites previamente definidos.

Esse princípio foi aplicado empiricamente no caso estudado.

A operação assistida permitiu identificar uma taxa inicial de divergência de 14%. Após a estabilização dos critérios, a taxa caiu para 3%. A redução observada foi utilizada como evidência para a transição entre o nível de execução com aprovação prévia e o nível de execução autônoma com auditoria posterior.

O procedimento não considera que uma taxa de divergência específica seja universalmente aceitável para qualquer tarefa. O limite adequado depende das características, consequências e riscos da operação analisada.

Consequentemente, o percentual de 3% observado no estudo não deve ser interpretado como um limite geral de segurança para agentes autônomos. Ele constitui uma evidência específica do caso analisado e demonstra a utilidade de estabelecer métricas antes da promoção de uma tarefa.

3.7 Indicadores operacionais

Dois indicadores foram utilizados para caracterizar os resultados da implementação.

O primeiro foi a taxa de divergência entre as decisões do agente e as decisões humanas durante o período de calibração e estabilização.

O segundo foi o tempo médio de resolução das solicitações.

Antes da implementação, o tempo médio de resolução informado para o atendimento executado pela equipe de suporte era de aproximadamente 48 horas. Após a implementação da execução autônoma, o tempo médio observado passou para aproximadamente 10 minutos a partir da abertura do chamado pelo usuário.

Também foi registrada uma economia superior a um milhão de reais no orçamento da área como consequência da automação dos processos analisados e de outros processos relacionados.

Embora esse resultado econômico seja relevante para caracterizar o impacto organizacional da implementação, ele não é utilizado como indicador isolado de segurança ou de adequação da autonomia. Ganhos financeiros não substituem critérios de governança, segurança, auditabilidade ou responsabilização.

3.8 Considerações sobre validade e limitações metodológicas

A metodologia apresenta limitações inerentes ao desenho de estudo de caso único.

A primeira limitação está relacionada à ausência de generalização estatística. Os resultados observados em uma organização brasileira de médio porte não podem ser automaticamente extrapolados para empresas de outros setores, tamanhos ou níveis de maturidade.

A segunda limitação está relacionada à possibilidade de dependência do contexto tecnológico específico da organização. Características da infraestrutura, qualidade dos dados, sistemas integrados e políticas internas podem influenciar significativamente o desempenho dos agentes.

A terceira limitação decorre da própria natureza da medida de divergência. A decisão humana utilizada como referência também pode conter inconsistências ou limitações. Portanto, a redução da divergência não significa necessariamente que todas as decisões do agente sejam objetivamente corretas.

A quarta limitação é temporal. Agentes autônomos operam em ambientes sujeitos a mudanças. Alterações em sistemas integrados, políticas internas, perfis de usuários ou requisitos regulatórios podem alterar a adequação de um critério anteriormente estabilizado.

Essas limitações reforçam a necessidade de interpretar o modelo como uma estrutura para governança e avaliação contínua, e não como um mecanismo de certificação definitiva da autonomia.

4. CRITÉRIOS DE DELEGABILIDADE

A principal proposição deste estudo é que o grau admissível de autonomia deve ser determinado pelas características da tarefa delegada. Para operacionalizar essa proposição, são definidos cinco critérios de delegabilidade.

Os critérios não devem ser utilizados como uma classificação puramente mecânica. Uma tarefa pode apresentar características favoráveis em alguns critérios e desfavoráveis em outros. O objetivo é permitir uma avaliação estruturada da combinação desses fatores.

4.1 Determinismo do critério

O primeiro critério é o grau de determinismo da regra ou do critério que orienta a decisão.

Uma tarefa apresenta elevado determinismo quando sua decisão pode ser descrita por regras explícitas, consistentes e auditáveis. Quanto menor a dependência de interpretação contextual, maior tende a ser a possibilidade de delegação.

Por exemplo, uma regra que determine que determinado perfil de usuário deve receber um conjunto específico de permissões em função de seu cargo pode apresentar elevado grau de determinismo quando os dados necessários estiverem corretamente estruturados.

Em contraste, uma decisão que dependa da interpretação de circunstâncias excepcionais, de interesses conflitantes ou de informações que não possam ser representadas adequadamente por critérios definidos apresenta menor determinismo.

O determinismo, portanto, não significa que o processo seja simples. Uma tarefa pode envolver grande quantidade de dados e ainda apresentar critério altamente determinístico.

A questão relevante é se o critério decisório pode ser explicitado e submetido a auditoria.

Quanto maior o determinismo, maior tende a ser o nível de autonomia admissível.

4.2 Verificabilidade do resultado

O segundo critério diz respeito à capacidade de verificar se a execução produzida pelo agente foi correta.

Uma tarefa é mais adequada à autonomia quando existe um mecanismo objetivo para confirmar o resultado após sua execução.

A verificabilidade permite que o controle seja deslocado parcialmente da aprovação prévia para a auditoria posterior.

Por exemplo, se um agente executa uma alteração de acesso e o sistema permite verificar imediatamente se o usuário passou a possuir exatamente as permissões previstas para sua função, existe uma condição favorável à verificação posterior.

Por outro lado, quando o resultado somente pode ser avaliado muito tempo depois, depende de julgamento subjetivo ou não pode ser objetivamente reconstruído, a aprovação prévia torna-se mais importante.

A verificabilidade também deve considerar a qualidade do mecanismo utilizado para confirmar o resultado. Uma verificação superficial não oferece o mesmo grau de segurança que uma verificação independente e tecnicamente confiável.

4.3 Reversibilidade do efeito

O terceiro critério é a possibilidade de desfazer uma ação incorreta.

A reversibilidade reduz o impacto potencial de erros porque permite restaurar o estado anterior.

Uma operação de concessão de acesso pode, em determinadas arquiteturas, ser revertida mediante remoção da permissão concedida. Uma ação que exclua definitivamente informações, produza uma comunicação externa ou gere uma obrigação jurídica pode apresentar características muito diferentes.

A reversibilidade também possui uma dimensão econômica e operacional. Não basta que uma ação seja teoricamente reversível. Deve-se avaliar quanto tempo, esforço e recursos são necessários para realizar a reversão e quais consequências podem ocorrer enquanto ela não é concluída.

Quanto mais reversível for o efeito e menor for o custo da reversão, maior tende a ser a possibilidade de controle posterior.

4.4 Consequência regulatória

O quarto critério considera a possibilidade de a ação produzir consequências jurídicas ou regulatórias.

Determinadas ações não devem ser tratadas como simples operações técnicas porque produzem efeitos perante titulares de dados, clientes, terceiros, autoridades ou órgãos reguladores.

A existência de consequência regulatória direta reduz o grau de autonomia admissível, especialmente quando a ação envolve declaração formal, comunicação externa, cumprimento de obrigação legal ou exercício de uma decisão que exige responsabilização humana.

Nesse contexto, a confiabilidade técnica do agente não elimina a necessidade de decisão humana.

O princípio proposto é que a autoridade jurídica e regulatória não deve ser implicitamente transferida para um sistema autônomo apenas porque a execução técnica pode ser automatizada.

4.5 Custo do erro em escala

O quinto critério considera o impacto de erros sistemáticos.

Um agente autônomo pode executar uma tarefa centenas ou milhares de vezes. Essa capacidade constitui uma das principais fontes de eficiência da tecnologia, mas também amplia o potencial de impacto de um erro de critério.

Se o agente aplicar incorretamente uma regra a um único caso, o impacto pode ser limitado. Se a mesma regra inadequada for aplicada automaticamente a milhares de usuários, o problema assume outra dimensão.

O custo do erro em escala deve, portanto, considerar não apenas a gravidade de um erro individual, mas a quantidade potencial de operações afetadas antes que o erro seja identificado e corrigido.

Esse critério reforça a importância de monitoramento contínuo, detecção de padrões de divergência e mecanismos independentes de interrupção.

Uma tarefa de alto volume pode ser excelente candidata à automação, mas justamente por isso precisa apresentar controles capazes de identificar rapidamente erros sistemáticos.

5. MODELO DE CINCO NÍVEIS DE DELEGAÇÃO

Com base nos critérios apresentados, este estudo propõe um modelo de cinco níveis para classificar a delegação de trabalho de governança de segurança da informação a agentes autônomos.

Os níveis representam graus crescentes de autoridade operacional.

5.1 Nível 0 — Execução humana

No Nível 0, o agente não possui autoridade para executar ações sobre o processo.

Sua utilização restringe-se a atividades de consulta, organização de informações, pesquisa, apoio analítico ou outras funções que não alterem diretamente o estado do processo.

A decisão e a execução permanecem integralmente sob responsabilidade humana.

Esse nível é adequado para tarefas de elevada complexidade contextual, alta consequência regulatória, baixa verificabilidade ou efeitos pouco reversíveis.

O Nível 0 também pode representar o ponto inicial de avaliação de uma nova aplicação de IA antes da concessão de qualquer autoridade operacional.

5.2 Nível 1 — Preparação assistida

No Nível 1, o agente passa a participar diretamente da preparação da decisão.

Ele pode coletar informações, consultar sistemas, correlacionar dados, identificar padrões e preparar uma recomendação ou conjunto de ações.

A decisão permanece humana e a execução também é realizada por uma pessoa.

Esse nível é especialmente adequado para tarefas nas quais a coleta e organização das informações são trabalhosas, mas o julgamento final depende de contexto.

A principal vantagem do Nível 1 é permitir ganhos de produtividade sem transferir autoridade operacional ao agente.

5.3 Nível 2 — Execução com aprovação prévia

No Nível 2, o agente possui capacidade para preparar a ação completa e, potencialmente, executar tecnicamente a operação, mas a execução depende de aprovação humana explícita.

Esse nível constitui o ponto de entrada recomendado para processos produtivos nos quais a organização pretende testar a capacidade do agente antes de conceder autonomia posterior.

A aprovação humana funciona como uma barreira de controle antes da alteração do estado do sistema.

Durante esse estágio, a organização pode medir a divergência entre as decisões propostas pelo agente e as decisões que seriam tomadas por humanos.

O Nível 2 é particularmente adequado para o período de calibração e aprendizagem operacional.

5.4 Nível 3 — Execução com auditoria posterior

No Nível 3, o agente executa a tarefa autonomamente sem aprovação humana individual para cada operação.

O controle humano é deslocado para a etapa posterior, por meio de auditoria por amostragem, análise de exceções e monitoramento dos indicadores definidos para o processo.

Esse nível é adequado para tarefas que apresentam critérios relativamente determinísticos, resultados verificáveis e efeitos suficientemente reversíveis.

A passagem do Nível 2 para o Nível 3 não deve ocorrer simplesmente porque o agente apresentou resultados satisfatórios durante determinado período. É necessário que os indicadores definidos pela organização demonstrem estabilidade dentro de limites previamente estabelecidos.

O estudo de caso analisado neste artigo encontra-se nessa condição.

5.5 Nível 4 — Execução autônoma com controle por exceção

O Nível 4 representa o maior grau de autonomia previsto pelo modelo.

Nesse nível, o agente executa as operações normalmente e somente encaminha para avaliação humana os casos que violam condições previamente definidas.

O controle deixa de ocorrer predominantemente por amostragem posterior e passa a depender de regras de exceção capazes de identificar situações fora do padrão.

Esse nível deve ser reservado a tarefas de alto volume, com critérios altamente estáveis, elevada verificabilidade, efeitos plenamente reversíveis e baixo potencial de consequência regulatória direta.

Mesmo nesse nível, a autonomia não significa ausência de governança.

A organização deve manter identidade própria do agente, privilégio mínimo, trilhas de auditoria, monitoramento de desempenho e mecanismo externo de interrupção.

5.6 Progressão entre níveis

A progressão entre os níveis constitui uma decisão de governança e deve ser baseada em evidências.

O modelo proposto estabelece a seguinte lógica:

Nível 0 → Nível 1 → Nível 2 → medição → estabilização → Nível 3 → monitoramento contínuo → eventual Nível 4.

A passagem de um nível para outro deve considerar:

- taxa de divergência;
- estabilidade da taxa de divergência;
- natureza das divergências identificadas;
- gravidade dos erros;
- verificabilidade dos resultados;
- reversibilidade das ações;
- alterações no ambiente tecnológico;
- alterações nas políticas internas;
- alterações regulatórias;
- capacidade de intervenção humana.

A autonomia, portanto, não deve ser tratada como um estado permanente conquistado pelo agente. Ela deve permanecer condicionada à manutenção das condições que justificaram sua concessão.

Se essas condições deixarem de existir, a tarefa deve poder retornar a um nível inferior de autonomia.

Esse mecanismo de **degradação controlada de autonomia** é uma consequência direta da abordagem proposta: se a promoção depende de evidências, a manutenção da autonomia também deve depender da continuidade dessas evidências.

6. CAMADA DE RESPONSABILIZAÇÃO

A delegação de trabalho a agentes autônomos não elimina a necessidade de responsabilização humana. Ao contrário, quanto maior o grau de autonomia concedido a um sistema, maior é a necessidade de estabelecer claramente quem autorizou a delegação, quais limites foram definidos e quem responde pelo processo.

A premissa central desta seção é que **a execução pode ser delegada, mas a responsabilidade pelo processo não é automaticamente transferida ao agente**.

O agente não constitui, por si só, um sujeito responsável pelo processo organizacional. Sua atuação ocorre dentro de uma estrutura de governança definida pela organização e deve estar vinculada a responsáveis humanos identificáveis.

Para que uma delegação seja defensável perante auditoria, investigação de incidente ou eventual escrutínio regulatório, quatro condições fundamentais devem ser preservadas.

6.1 Identidade própria e distinta para o agente

Cada agente que possua capacidade de executar ações em sistemas corporativos deve possuir uma identidade própria e distinta das identidades utilizadas por pessoas.

Essa identidade deve permitir determinar, de maneira inequívoca, que determinada ação foi executada pelo agente, e não diretamente por um usuário humano.

A utilização de uma credencial pessoal para permitir que um agente execute operações cria uma ruptura na cadeia de responsabilização. O registro do sistema pode indicar que determinada pessoa realizou uma ação quando, na realidade, a operação foi executada por um mecanismo automatizado.

A identidade própria também permite aplicar políticas de acesso específicas ao agente, limitar suas permissões e acompanhar sua atividade de maneira independente.

O agente deve possuir apenas os privilégios necessários para desempenhar a função para a qual foi autorizado.

6.2 Proprietário humano do processo

Cada processo delegado deve possuir um responsável humano formalmente identificado.

Esse responsável não precisa executar pessoalmente cada operação realizada pelo agente, mas deve responder pela definição dos critérios, pela autorização da delegação, pelo acompanhamento dos indicadores e pela revisão das condições que justificam a manutenção da autonomia.

Essa estrutura evita que a automação produza uma situação de responsabilidade difusa, na qual nenhum indivíduo consiga explicar por que determinada tarefa foi delegada, quais limites foram estabelecidos ou quem deveria agir diante de uma falha.

A responsabilidade deve permanecer associada ao processo e não à capacidade técnica do agente.

Assim, mesmo em um processo operando no Nível 4, a organização deve ser capaz de identificar o responsável humano pela governança daquela atividade.

6.3 Trilha de auditoria e reconstrução da decisão

A auditoria de um agente autônomo não pode limitar-se ao registro da ação final.

Para que uma decisão possa ser reconstruída adequadamente, a trilha de auditoria deve permitir identificar, na medida tecnicamente possível, o contexto que originou a operação, as informações utilizadas, o critério aplicado, a autorização disponível, a ação executada e o resultado produzido.

O objetivo não é necessariamente registrar todo o funcionamento interno do modelo de inteligência artificial, mas preservar evidências suficientes para reconstruir o processo decisório operacional.

Em uma investigação posterior, deve ser possível responder perguntas como:

- qual agente executou a operação?
- quando a operação ocorreu?
- sobre qual objeto ou usuário ela incidiu?
- quais informações estavam disponíveis?
- qual regra ou política orientava a operação?
- quais permissões o agente possuía?
- qual ação foi executada?
- qual resultado foi produzido?
- quem era o responsável humano pelo processo?

A capacidade de responder a essas perguntas constitui elemento fundamental da auditabilidade.

6.4 Mecanismo independente de interrupção

Todo processo delegado deve possuir mecanismo de interrupção que não dependa exclusivamente do próprio agente.

Essa condição é particularmente importante em situações nas quais o agente apresenta comportamento inadequado, erro sistemático ou degradação silenciosa.

Se o mesmo sistema responsável pela execução for também o único mecanismo capaz de interromper sua própria atividade, a organização pode enfrentar uma situação de dependência operacional incompatível com uma estrutura adequada de controle.

O mecanismo de interrupção deve, portanto, ser operacionalmente independente do agente e estar disponível para os responsáveis pelo processo.

Sempre que tecnicamente possível, deve existir também mecanismo de reversão capaz de restaurar o estado anterior ou limitar os efeitos de uma operação inadequada.

6.5 Responsabilidade e autonomia

Os quatro elementos anteriores permitem estabelecer uma distinção fundamental entre autonomia operacional e responsabilidade organizacional.

Um agente pode operar autonomamente sem que a organização deixe de possuir controle sobre o processo.

Autonomia significa que a execução de determinada operação não exige intervenção humana individual.

Responsabilidade significa que existe uma estrutura humana e organizacional capaz de responder pela definição, autorização, monitoramento e revisão dessa operação.

Essa distinção impede que o aumento da autonomia seja interpretado como uma transferência de responsabilidade para o sistema.

7. ESTUDO DE CASO E RESULTADOS EMPÍRICOS

7.1 Caracterização do caso

O estudo de caso analisado corresponde à implementação de agentes autônomos em uma organização brasileira de médio porte, cuja identidade é preservada por acordo de confidencialidade.

A solução desenvolvida envolveu agentes e sistemas proprietários aplicados a processos de tecnologia da informação, com destaque para atendimento de primeiro nível e gestão de identidades e acessos.

No domínio de gestão de identidade e acesso, a solução incluiu provisionamento automatizado, aplicação de perfis baseados em cargo, revisão periódica de acessos e integração com o diretório corporativo e o sistema de chamados.

A arquitetura permitiu que os agentes recebessem solicitações, consultassem informações relevantes, aplicassem critérios previamente estabelecidos e executassem operações nos sistemas corporativos dentro das permissões concedidas.

No momento da análise, o atendimento de primeiro nível do suporte técnico e a totalidade dos processos de gestão de acessos abrangidos pela implementação eram executados autonomamente, sob regime de auditoria contínua por amostragem.

7.2 Volume de operações

Entre março de 2026 e o período de redação do estudo, mais de mil solicitações foram processadas de forma autônoma.

O volume é relevante para a análise porque permite observar o comportamento do sistema em condições operacionais reais, em vez de limitar a avaliação a testes controlados ou demonstrações de laboratório.

A utilização em produção também permitiu identificar situações relacionadas à qualidade dos dados, alterações nos sistemas integrados e comportamento das regras em diferentes tipos de solicitação.

O volume processado, entretanto, não deve ser interpretado isoladamente como prova de segurança ou confiabilidade. A quantidade de operações demonstra exposição operacional suficiente para avaliação, mas a qualidade da delegação depende da combinação entre volume, natureza da tarefa, desempenho e mecanismos de controle.

7.3 Divergência durante a calibração

O principal indicador utilizado para acompanhar a evolução do agente foi a divergência entre a decisão proposta pelo sistema e a decisão tomada pelo processo humano de referência.

Durante o período de calibração, a divergência observada foi de 14%.

Esse resultado indicou que a operação ainda não apresentava estabilidade suficiente para justificar a retirada da aprovação humana individual.

O agente permaneceu, portanto, em regime de operação assistida enquanto os critérios eram analisados e ajustados.

Após a estabilização dos critérios, a taxa de divergência caiu para 3%.

A redução representa uma diferença de 11 pontos percentuais entre o período inicial e o período estabilizado.

Mais importante do que o valor absoluto é a trajetória observada.

A experiência demonstrou que a avaliação da autonomia pode ser realizada de maneira progressiva: primeiro o agente propõe, depois suas decisões são comparadas às decisões humanas e, somente após a observação de estabilidade, a autoridade operacional pode ser ampliada.

Esse resultado sustenta empiricamente a proposição de que a promoção entre níveis de autonomia deve ser condicionada à medição do desempenho.

7.4 Redução do tempo de atendimento

Outro resultado observado foi a redução do tempo médio de resolução das solicitações.

No modelo anterior, executado pela equipe de suporte, o tempo médio informado era de aproximadamente 48 horas.

Após a implementação dos agentes autônomos, o tempo médio de resolução passou para aproximadamente 10 minutos a partir da abertura do chamado pelo usuário.

A diferença demonstra um ganho operacional significativo.

Entretanto, a redução do tempo de atendimento não constitui, isoladamente, justificativa para ampliar a autonomia.

O ganho de velocidade precisa ser analisado em conjunto com a qualidade das decisões, a reversibilidade das operações, a existência de mecanismos de auditoria e os demais critérios de delegabilidade.

Essa distinção é importante porque uma automação capaz de executar rapidamente uma regra inadequada pode aumentar a velocidade de produção do erro.

7.5 Economia operacional

A automação dos processos analisados e de outros processos relacionados resultou em economia superior a um milhão de reais no orçamento da área.

O resultado econômico demonstra que a delegação de trabalho pode produzir impacto financeiro relevante além da redução do tempo de atendimento.

A economia, entretanto, é tratada neste estudo como consequência organizacional da implementação, e não como critério suficiente para determinar o grau de autonomia.

Uma tarefa não deve receber maior autoridade simplesmente porque sua automação produz economia. A decisão de delegação deve permanecer vinculada às características do processo e aos controles necessários.

7.6 Posicionamento do caso no modelo proposto

Após o período prolongado de operação assistida e a estabilização da taxa de divergência, o atendimento de primeiro nível e a gestão de acessos passaram a operar sem aprovação humana prévia para cada solicitação.

Ao mesmo tempo, foi mantida auditoria posterior por amostragem.

Essa configuração corresponde ao **Nível 3 — Execução com auditoria posterior** do modelo proposto.

A classificação é relevante porque demonstra que autonomia operacional não significa ausência de controle.

O agente possui autoridade para executar as operações previstas, mas sua atuação permanece sujeita à verificação posterior e à possibilidade de intervenção.

O caso não é classificado como Nível 4 porque a execução ainda depende de mecanismos de auditoria por amostragem, e não exclusivamente de controle baseado em exceções.

7.7 Relação entre divergência e progressão de autonomia

A evolução de 14% para 3% constitui o principal resultado empírico relacionado ao mecanismo de progressão proposto.

Durante a fase inicial, a divergência indicava que a aplicação dos critérios pelo agente ainda não apresentava estabilidade suficiente.

Após o período de calibração, a redução e estabilização da divergência forneceram evidência para a mudança do Nível 2 para o Nível 3.

O resultado não demonstra que 3% seja um limite universalmente adequado.

O que o caso demonstra é a utilidade de estabelecer um limite antes da promoção, medir o desempenho em relação a esse limite e condicionar a ampliação da autonomia à estabilidade observada.

Essa distinção é essencial para evitar que o modelo seja interpretado como uma fórmula matemática universal.

O percentual aceitável deve ser determinado considerando a natureza, o impacto e a reversibilidade da tarefa.

8. DISCUSSÃO

8.1 A tarefa como unidade de decisão de governança

Os resultados do estudo de caso reforçam a proposição de que a tarefa deve constituir a unidade principal para decisão sobre autonomia.

O mesmo agente pode atuar em diferentes processos com níveis distintos de autoridade.

Essa abordagem é mais adequada do que classificar uma organização inteira ou um agente inteiro como “autônomo” ou “não autônomo”.

Uma tecnologia pode ser utilizada para preparar decisões em uma atividade de elevado risco e, simultaneamente, executar autonomamente operações rotineiras e reversíveis em outro processo.

Essa granularidade permite que a organização capture ganhos de produtividade sem transformar a adoção de agentes em uma concessão generalizada de autoridade.

8.2 Autonomia como função de risco e controle

Os resultados também indicam que autonomia não deve ser entendida exclusivamente como uma propriedade técnica.

Um agente pode apresentar elevada capacidade de processamento, acesso a múltiplos sistemas e desempenho satisfatório em determinada tarefa. Ainda assim, isso não significa que seja apropriado conceder-lhe autoridade para executar qualquer ação.

A autonomia admissível resulta da relação entre características da tarefa e controles disponíveis.

Uma tarefa altamente determinística, verificável e reversível apresenta condições favoráveis à execução autônoma.

Uma tarefa dependente de julgamento contextual, pouco verificável, irreversível ou associada a consequência regulatória direta exige maior participação humana.

O modelo, portanto, estabelece uma relação entre risco e autoridade.

Quanto maior o potencial de consequência de uma decisão, maior deve ser a exigência de controle prévio ou de intervenção humana.

8.3 A importância da medição

A principal evidência empírica do caso está relacionada à transição de 14% para 3% na divergência.

Essa evolução demonstra a utilidade de transformar a progressão de autonomia em um processo mensurável.

Em abordagens baseadas exclusivamente em julgamento subjetivo, uma organização pode ampliar a autonomia porque considera o sistema "maduro", "confiável" ou "estável".

Essas avaliações podem ser úteis, mas são insuficientes para uma decisão de governança.

A utilização de métricas permite estabelecer uma condição verificável.

O modelo proposto não determina uma taxa universal de divergência. Em vez disso, determina que a organização estabeleça previamente o limite considerado aceitável para determinada tarefa e verifique se o comportamento do agente permanece dentro desse limite.

Essa abordagem aproxima a governança da autonomia de um processo de controle contínuo.

8.4 O problema do erro sistemático

A análise também reforça a importância de diferenciar erro aleatório de erro sistemático.

A auditoria por amostragem é útil para identificar desvios, mas pode ser insuficiente quando o agente aplica uma regra inadequada de forma consistente.

Se uma determinada classe de solicitações for tratada incorretamente em todos os casos, uma amostra pequena pode não revelar imediatamente a existência do problema.

Por essa razão, o controle da autonomia deve combinar auditoria por amostragem com indicadores capazes de detectar mudanças de comportamento, concentrações anormais de divergência e alterações no ambiente de execução.

Essa necessidade é especialmente importante em sistemas integrados, nos quais mudanças em um sistema de origem podem alterar o significado dos dados utilizados pelo agente sem produzir uma falha técnica evidente.

8.5 Degradação silenciosa

Um dos limites observados no estudo é a possibilidade de degradação silenciosa.

O agente pode continuar executando tecnicamente suas operações enquanto o contexto no qual os critérios foram definidos se modifica.

Uma alteração em uma política interna, em uma estrutura organizacional, em um sistema integrado ou em um cadastro pode fazer com que uma regra anteriormente adequada deixe de produzir resultados adequados.

Esse tipo de problema é particularmente perigoso porque não necessariamente gera erros técnicos.

O sistema continua funcionando, os chamados continuam sendo processados e as operações continuam sendo registradas.

O problema aparece no conteúdo das decisões, e não na disponibilidade do sistema.

Essa característica reforça a necessidade de revisão periódica dos critérios e de monitoramento do ambiente no qual a autonomia opera.

8.6 Qualidade dos dados como condição de delegabilidade

O caso também evidencia que a qualidade dos dados de origem constitui condição importante para a delegação.

Um agente não elimina automaticamente inconsistências existentes nos sistemas corporativos.

Quando uma decisão depende de informações incorretas, incompletas ou desatualizadas, a automação pode reproduzir e ampliar o problema.

Essa característica reforça a ideia de que a autonomia deve ser avaliada não apenas em relação ao agente, mas em relação ao processo completo.

Uma tarefa que seria adequada à autonomia em uma organização com dados confiáveis pode não ser adequada em outra organização que apresente cadastros desestruturados.

A qualidade da informação, portanto, deve ser tratada como condição prévia para a delegação de determinadas atividades.

8.7 Preservação da competência humana

Outro limite identificado é o risco de perda de competência da equipe.

Quando um processo é integralmente delegado por longos períodos, profissionais humanos podem deixar de praticá-lo regularmente.

Isso pode reduzir a capacidade da equipe de executar manualmente a atividade em caso de indisponibilidade do agente e também pode dificultar a avaliação da qualidade das decisões automatizadas.

Existe, portanto, um paradoxo da automação: quanto mais eficiente o sistema se torna, menor pode ser a exposição humana ao processo e, consequentemente, menor pode ser a competência residual disponível para supervisioná-lo.

A governança de agentes autônomos deve considerar esse risco.

A supervisão humana não pode depender de profissionais que desconhecem o processo que estão supervisionando.

8.8 Responsabilização como condição permanente

A experiência analisada também reforça a necessidade de manter a responsabilização independentemente do nível de autonomia.

A mudança do Nível 2 para o Nível 3 não eliminou a responsabilidade humana.

O agente passou a executar sem aprovação individual, mas continuou sujeito a auditoria e permaneceu vinculado a uma estrutura de responsabilidade.

Essa característica diferencia autonomia de abandono de controle.

A autonomia operacional deve ser acompanhada por mecanismos de identificação, autorização, registro, supervisão e intervenção.

Nesse sentido, a responsabilização não deve ser vista como uma etapa adicional do processo de automação, mas como uma camada permanente da arquitetura de governança.

8.9 Implicações para organizações

O modelo apresenta implicações práticas para organizações que pretendem utilizar agentes autônomos em processos de segurança da informação.

Primeiro, a organização deve realizar um inventário das tarefas que pretende delegar, em vez de avaliar a autonomia apenas no nível do sistema.

Segundo, cada tarefa deve ser avaliada segundo os cinco critérios propostos.

Terceiro, a organização deve iniciar processos novos em nível de autonomia compatível com o risco, preferencialmente utilizando operação assistida quando houver incerteza relevante.

Quarto, a progressão deve ser baseada em métricas.

Quinto, cada agente deve possuir identidade própria, privilégios mínimos e responsável humano.

Sexto, a organização deve possuir mecanismos de interrupção e reversão independentes.

Por fim, a autonomia deve ser revisada sempre que ocorrerem mudanças relevantes no processo, nos sistemas integrados, nas políticas internas ou nas exigências regulatórias.

9. LIMITAÇÕES DO ESTUDO

Este estudo apresenta limitações que precisam ser consideradas na interpretação dos resultados.

A principal limitação é o desenho de estudo de caso único. A análise de uma organização não permite afirmar que os resultados observados ocorrerão da mesma forma em organizações de diferentes setores, tamanhos, arquiteturas tecnológicas ou níveis de maturidade.

A segunda limitação é a dependência do contexto específico da implementação. A qualidade dos dados, a estrutura dos processos, as políticas internas e os sistemas utilizados podem ter contribuído para os resultados observados.

A terceira limitação está relacionada ao indicador de divergência. A comparação entre decisões do agente e decisões humanas não representa necessariamente uma comparação

entre uma decisão errada e uma decisão correta. O processo humano utilizado como referência também pode apresentar inconsistências.

A quarta limitação é o período de observação. Embora mais de mil solicitações tenham sido processadas e o sistema tenha sido utilizado em ambiente real, períodos mais longos de acompanhamento são necessários para avaliar adequadamente fenômenos de degradação silenciosa, mudanças de política e alterações na distribuição das solicitações.

A quinta limitação é a ausência de comparação sistemática com múltiplas organizações. A replicação do estudo em outros ambientes seria necessária para avaliar a robustez do modelo e identificar quais critérios apresentam maior poder explicativo em diferentes contextos.

Essas limitações não invalidam a proposta, mas delimitam o alcance das conclusões. O modelo deve ser considerado uma proposição inicial sustentada por evidências empíricas de um caso aplicado, cuja validade externa ainda requer investigação adicional.

10. CONCLUSÃO

A adoção de agentes autônomos de inteligência artificial em processos de segurança da informação cria uma oportunidade significativa de redução de trabalho operacional, aumento de velocidade e melhoria de eficiência. Ao mesmo tempo, a capacidade desses sistemas de interpretar informações e executar ações diretamente sobre ambientes corporativos introduz uma questão de governança que não pode ser respondida simplesmente pela distinção entre processos manuais e automatizados.

Este artigo propôs que a decisão sobre delegação deve ser realizada no nível da tarefa e não exclusivamente no nível da tecnologia.

Para operacionalizar essa proposta, foram definidos cinco critérios: determinismo do critério, verificabilidade do resultado, reversibilidade do efeito, consequência regulatória e custo do erro em escala.

Esses critérios permitem avaliar se uma determinada tarefa possui características compatíveis com maior ou menor autonomia.

A partir deles, foi proposto um modelo de cinco níveis de delegação, variando da execução integralmente humana à execução autônoma com controle por exceção.

O modelo estabelece que a autonomia deve crescer progressivamente e permanecer condicionada à evidência de desempenho.

A principal evidência empírica obtida no estudo de caso foi a redução da divergência entre decisões do agente e decisões humanas de 14% durante a calibração para 3% após a estabilização dos critérios.

Essa redução sustentou a transição do Nível 2, caracterizado por execução condicionada à aprovação humana, para o Nível 3, caracterizado por execução autônoma com auditoria posterior.

O caso também demonstrou ganhos operacionais expressivos. O tempo médio de resolução passou de aproximadamente 48 horas para cerca de 10 minutos, enquanto a

automação dos processos analisados e de outros processos relacionados resultou em economia superior a um milhão de reais no orçamento da área.

Esses resultados demonstram que a delegação de trabalho pode produzir benefícios significativos, mas também reforçam que eficiência não deve ser utilizada isoladamente como justificativa para ampliação da autonomia.

A velocidade de execução não substitui critérios de segurança, auditabilidade e responsabilização.

A análise também identificou limitações importantes. A qualidade dos dados de origem influencia diretamente a qualidade das decisões automatizadas. Erros sistemáticos podem ser reproduzidos em escala. Alterações em sistemas, políticas ou dados podem produzir degradação silenciosa. E a delegação prolongada pode reduzir a competência humana necessária para supervisionar ou assumir novamente o processo.

Esses fatores demonstram que a autonomia não deve ser tratada como um estado permanente.

A manutenção de determinado nível de autonomia deve depender da continuidade das condições que justificaram sua concessão. Quando essas condições deixam de existir, a organização deve ser capaz de reduzir o nível de autonomia ou interromper a execução automatizada.

A responsabilização constitui o segundo elemento central da proposta.

A execução pode ser delegada ao agente, mas a responsabilidade pelo processo permanece vinculada à organização e a um responsável humano identificável.

Para que a delegação seja defensável, é necessário preservar identidade própria do agente, privilégio mínimo, proprietário humano do processo, trilha de auditoria e mecanismo independente de interrupção e reversão.

A contribuição central do estudo consiste, portanto, em propor uma estrutura que permita tratar a autonomia como uma decisão de governança mensurável e proporcional ao risco.

Em vez de perguntar simplesmente se uma organização deve ou não utilizar agentes autônomos, a estrutura proposta permite perguntar quais tarefas podem ser delegadas, em qual nível, mediante quais evidências e sob quais condições de controle.

Essa abordagem também permite superar uma possível falsa dicotomia entre automação e trabalho humano.

A questão não é substituir integralmente pessoas por agentes, mas distribuir funções de maneira proporcional às características da tarefa.

Tarefas determinísticas, verificáveis e reversíveis podem admitir maior autonomia. Tarefas dependentes de julgamento contextual, com efeitos irreversíveis ou consequências regulatórias diretas devem preservar maior participação humana.

O estudo de caso oferece evidência inicial em favor dessa abordagem, especialmente ao demonstrar que a promoção da autonomia acompanhada por medição objetiva pode ser realizada de maneira progressiva.

Entretanto, por se tratar de estudo de caso único, os resultados não devem ser considerados generalizáveis sem investigação adicional.

Pesquisas futuras podem aplicar o modelo em organizações de diferentes setores, comparar diferentes tipos de agentes e avaliar longitudinalmente a estabilidade dos critérios de delegação.

Também será relevante investigar métodos quantitativos para definição dos limites de divergência, mecanismos de detecção de erros sistemáticos e estratégias para preservar competência humana em ambientes altamente automatizados.

Assim, a principal conclusão deste estudo é que **autonomia não deve ser concedida ao agente em função de sua capacidade tecnológica, mas à tarefa em função de suas características, dos riscos envolvidos e da existência de controles capazes de preservar a responsabilização**.

A delegação defensável não é aquela que elimina a intervenção humana, mas aquela que estabelece claramente onde a intervenção deixa de ser necessária, quais evidências justificam essa decisão e quais mecanismos permanecem disponíveis para recuperar o controle quando as condições se alterarem.

REFERÊNCIAS

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). *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

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

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

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.

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

BIOGRAFIA DO AUTOR

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.*