



**PRISMA ODS**  
**REVISTA MULTIDISCIPLINARIA**  
**SOBRE DESARROLLO SOSTENIBLE**  
**ISSN: 3072-8452**

**RIESGO DE DELEGAR EL  
RAZONAMIENTO: RUPTURA DE LA  
CONFIANZA CERO Y  
VULNERABILIDADES DE SOBERANÍA  
DIGITAL EN ECOSISTEMAS  
CORPORATIVOS DE IA**

*RISK OF DELEGATING REASONING: THE  
BREAKDOWN OF ZERO TRUST AND DIGITAL  
SOVEREIGNTY VULNERABILITIES IN  
CORPORATE AI ECOSYSTEMS*

**AUTORES/AS**

**JUAN RIGOBERTO  
CASTILLO SERRACÍN**  
UNIVERSIDAD DE PANAMÁ  
PANAMÁ

**IRIS MARIELA GONZÁLEZ  
APARICIO**  
UNIVERSIDAD DE PANAMÁ  
PANAMÁ

**CARLOS BRUCE**  
UNIVERSIDAD DE PANAMÁ  
PANAMÁ

**MARICELLA CORPAS FORD**  
UNIVERSIDAD DE PANAMÁ  
PANAMÁ

**MAURICIO RAPÓN**  
UNIVERSIDAD DE PANAMÁ  
PANAMÁ

## **Riesgo de Delegar el Razonamiento: Ruptura de la Confianza Cero y Vulnerabilidades de Soberanía Digital en Ecosistemas Corporativos de IA**

Risk of Delegating Reasoning: The Breakdown of Zero Trust and Digital  
Sovereignty Vulnerabilities in Corporate AI Ecosystems

**Juan Rigoberto Castillo Serracín**

[juan.castillos@up.ac.pa](mailto:juan.castillos@up.ac.pa)

<https://orcid.org/0009-0006-5821-7028>

Universidad de Panamá

Panamá – Panamá

**Carlos Bruce**

[carlos.bruce@up.ac.pa](mailto:carlos.bruce@up.ac.pa)

<https://orcid.org/0009-0004-4223-0821>

Universidad de Panamá

Panamá – Panamá

**Iris Mariela González Aparicio**

[iris.gonzalez@up.ac.pa](mailto:iris.gonzalez@up.ac.pa)

<https://orcid.org/0000-0002-1984-7792>

Universidad de Panamá

Panamá – Panamá

**Maricella Corpas Ford**

[maricella.corpas@up.ac.pa](mailto:maricella.corpas@up.ac.pa)

<https://orcid.org/0000-0001-6430-4423>

Universidad de Panamá

Panamá – Panamá

**Mauricio Rapón**

[Mauricio.rapon@up.ac.pa](mailto:Mauricio.rapon@up.ac.pa)

<https://orcid.org/0009-0002-1527-6440>

Universidad de Panamá

Panamá – Panamá

*Artículo recibido: 27/08/2026*

*Aceptado para publicación: 29/09/2026*

*Conflictos de Intereses: Ninguno que declarar*

## RESUMEN

La incorporación de inteligencia artificial agéntica en entornos corporativos modifica la relación entre automatización, autoridad y control, al permitir que determinados sistemas interpreten contextos, planifiquen, seleccionen herramientas y ejecuten acciones con grados crecientes de autonomía. Este estudio analiza cómo la delegación del razonamiento y de la capacidad de acción tensiona los mecanismos de verificación y autorización asociados con Zero Trust y qué vulnerabilidades de soberanía digital pueden derivarse de esta transformación. Se realizó una revisión documental estructurada, no experimental y descriptivo-analítica, guiada por PRISMA 2020, sobre publicaciones comprendidas entre 2020 y agosto de 2026. El corpus estuvo integrado por 20 trabajos científicos revisados por pares, complementados con documentación técnica del NIST, y se examinó mediante análisis temático deductivo e inductivo. Los resultados muestran que la IA agéntica no invalida los principios de Zero Trust, pero amplía su objeto de control desde la identidad y el acceso hacia la legitimidad de la acción autónoma. De la síntesis emergen cuatro categorías: brecha identidad–acción, mínima agencia, confianza decisonal y soberanía decisonal. Asimismo, se observa que la propiedad de los datos no garantiza por sí solo control sobre las capacidades tecnológicas y operativas que los convierten en decisiones. Se concluye que la gobernanza de IA corporativa requiere articular identidad, autoridad, agencia y soberanía, preservando mecanismos de observación, limitación, reversibilidad y recuperación del control frente a sistemas autónomos.

*Palabras clave:* inteligencia artificial agéntica, Zero Trust, soberanía digital, gobernanza de la inteligencia artificial, ciberseguridad corporativa

**ABSTRACT**

The incorporation of agentic artificial intelligence into corporate environments modifies the relationship between automation, authority, and control by enabling certain systems to interpret contexts, plan, select tools, and execute actions with increasing degrees of autonomy. This study analyzes how the delegation of reasoning and action capabilities places pressure on verification and authorization mechanisms associated with Zero Trust and what digital sovereignty vulnerabilities may arise from this transformation. A structured, non-experimental, descriptive-analytical documentary review was conducted, guided by PRISMA 2020, covering publications from 2020 through August 2026. The corpus comprised 20 peer-reviewed scientific works, complemented by technical documentation from NIST, and was examined through deductive and inductive thematic analysis. The results show that agentic AI does not invalidate the principles of Zero Trust, but expands its object of control from identity and access toward the legitimacy of autonomous action. Four categories emerged from the synthesis: the identity–action gap, minimal agency, decisional trust, and decisional sovereignty. The analysis also shows that data ownership alone does not guarantee control over the technological and operational capabilities that transform data into decisions. It is concluded that corporate AI governance requires the articulation of identity, authority, agency, and sovereignty, while preserving mechanisms for observation, limitation, reversibility, and recovery of control over autonomous systems.

*Keywords:* agentic artificial intelligence, Zero Trust, digital sovereignty, artificial intelligence governance, corporate cybersecurity

## INTRODUCCIÓN

Integrar la inteligencia artificial (IA) en el ecosistema corporativo ha dejado de ser un simple recambio tecnológico. Hoy implica incorporar capacidades que pueden modificar tanto los procesos organizacionales como la distribución de la autoridad decisional. Stelmaszak et al. (2026) plantean que la IA debe comprenderse cada vez menos como un artefacto aislado y más como una capacidad organizativa emergente de las relaciones entre actores humanos y algorítmicos. A ello se suma que la evolución hacia arquitecturas de IA agéntica ha incorporado capacidades de memoria, razonamiento multietapa, planificación y utilización de herramientas externas (Wang et al., 2024; Peykani et al., 2026). Esta transición incrementa el nivel de autonomía de determinados sistemas: la IA deja de limitarse a sugerir qué hacer y puede comenzar a determinar secuencias operativas e intervenir directamente sobre procesos y recursos corporativos.

Delegar el razonamiento constituye un desafío distinto de la automatización tradicional. En este estudio, dicha delegación se entiende como la transferencia hacia sistemas algorítmicos de funciones relacionadas con la interpretación de contextos, comparación de alternativas, selección de herramientas, planificación y determinación de cursos de acción. El punto crítico aparece cuando el sistema deja de limitarse a recomendar una respuesta y comienza a determinar cómo alcanzar un objetivo mediante acciones generadas dinámicamente. Issa et al. (2026) conceptualizan uno de los riesgos asociados con esta dinámica mediante el término *agentic loafing*, entendido como una delegación insuficientemente examinada de autoridad decisional hacia sistemas de IA, capaz de debilitar mecanismos tradicionales de supervisión y atribución de responsabilidad. El riesgo sistémico emerge, por tanto, cuando el juicio algorítmico comienza a sustituir parcialmente la deliberación humana y modifica la distribución efectiva de autoridad dentro de la organización.

Este nuevo escenario introduce tensiones con los principios del paradigma Zero Trust. Este modelo elimina la confianza implícita y establece que la autenticación y la autorización deben constituir funciones explícitas asociadas con el acceso a los recursos organizacionales (Rose et al., 2020). Sin embargo, verificar correctamente la identidad de un agente autónomo no garantiza necesariamente la legitimidad de todas las acciones que pueda generar posteriormente. Esta diferencia se vuelve especialmente relevante cuando tales acciones no están completamente predeterminadas, sino que emergen durante procesos dinámicos de planificación, razonamiento e interacción con herramientas externas.

Castillo Serracín y Gómez (2026) identifican esta brecha de control en entornos de IA agéntica y proponen el paradigma Zero Agency, orientado a complementar los controles centrados en identidad mediante mecanismos aplicados directamente sobre la acción. Su planteamiento distingue entre autorizar a una entidad y validar de manera continua las operaciones que pueden derivarse de su autonomía, incorporando capacidades de supervisión, interceptación y revocación de agencia. Esta preocupación también aparece en el ámbito institucional. Riggs et al. (2026), en su análisis para el National Institute of Standards and Technology, señalan que los agentes de IA introducen amenazas de seguridad novedosas y que, aunque los principios tradicionales de ciberseguridad continúan siendo aplicables, requieren adaptaciones para responder adecuadamente a sistemas con mayores niveles de autonomía.

Por esta razón, la expresión ruptura de la confianza cero se utiliza en este estudio como una categoría analítica y no como una afirmación anticipada sobre la obsolescencia de Zero Trust. El concepto permite examinar una posible discontinuidad entre los mecanismos de verificación de identidad y autorización y la capacidad efectiva de una organización para controlar las decisiones y acciones generadas posteriormente por una entidad autónoma. Tal ruptura se configuraría hipotéticamente cuando una arquitectura de seguridad puede verificar adecuadamente quién o qué accede a un recurso, pero carece de mecanismos equivalentes para garantizar que las acciones derivadas de esa autorización permanezcan dentro del propósito y los límites definidos por la organización. En este escenario, la identidad del agente puede continuar siendo legítima mientras que una acción específica deja de serlo.

La creciente autonomía de estos sistemas también transforma la relación entre supervisión tecnológica y control humano. Tsamados et al. (2025) sostienen que el control de sistemas de IA debe comprenderse desde la interacción operacional entre humanos y máquinas, superando una concepción basada exclusivamente en supervisión humana directa. Cuando la planificación y la ejecución son progresivamente delegadas, la seguridad no puede depender únicamente de la presencia constante de un operador, sino que requiere mecanismos técnicos de supervisión, restricción, auditoría e interrupción capaces de preservar la capacidad organizacional de intervención.

Esta problemática trasciende la dimensión estrictamente técnica y se relaciona directamente con la soberanía digital. Roberts (2024) vincula este concepto con la capacidad de los actores para ejercer autoridad y control legítimo sobre tecnologías e infraestructuras digitales. Desde una perspectiva geopolítica, Castillo Serracín et al. (2026) identifican el control de

infraestructuras digitales y datos como componentes estratégicos mediante los cuales los Estados buscan construir y preservar soberanía digital. Este planteamiento adquiere especial relevancia en los ecosistemas corporativos contemporáneos, donde modelos fundacionales, servicios de nube, APIs, capacidades de inferencia y herramientas proporcionadas por terceros pueden encontrarse fuera del control directo de la organización.

Grohmann y Costa Barbosa (2026) conceptualizan parte de esta problemática mediante el término *sovereignty-as-a-service* o soberanía como servicio, mostrando cómo grandes proveedores tecnológicos reformulan la soberanía digital a través de soluciones construidas sobre sus propias infraestructuras y modelos de gobernanza. Esta situación configura una paradoja estructural: una organización puede incrementar determinadas capacidades de control y cumplimiento mientras mantiene una dependencia significativa respecto de plataformas, modelos o servicios externos.

Con la expansión de la IA agéntica, esta dependencia puede extenderse más allá del almacenamiento y procesamiento de datos hacia la interpretación, priorización y generación de decisiones. En consecuencia, una empresa puede conservar formalmente la propiedad legal de su información y, al mismo tiempo, experimentar una reducción de su autonomía efectiva cuando las capacidades necesarias para interpretarla, transformarla o convertirla en acciones dependen de tecnologías externas sujetas a jurisdicciones y políticas ajenas, sobre las cuales posee capacidades limitadas de supervisión, sustitución o intervención.

A partir de esta perspectiva, el estudio propone analizar la soberanía digital corporativa mediante cuatro dimensiones interrelacionadas: soberanía de datos, relacionada con el control sobre almacenamiento, acceso y procesamiento de información; soberanía tecnológica, vinculada con la capacidad de seleccionar, sustituir o desacoplar infraestructuras, plataformas y modelos; soberanía operacional, entendida como la capacidad de supervisar, limitar, detener y recuperar procesos automatizados; y soberanía decisonal, asociada con la conservación de autoridad efectiva sobre las decisiones que afectan objetivos, activos y recursos de la organización. Estas dimensiones se utilizarán como categorías analíticas para examinar la literatura y no como una taxonomía previamente consolidada.

Investigar esta convergencia resulta necesario porque la expansión de sistemas capaces de decidir y actuar autónomamente puede introducir formas de riesgo que no quedan plenamente explicadas cuando la seguridad, la autonomía algorítmica y la soberanía digital se estudian de

manera independiente. La literatura disponible permanece parcialmente fragmentada: los estudios sobre Zero Trust se concentran principalmente en identidad, autenticación, autorización y acceso; la investigación sobre IA agéntica analiza autonomía, planificación y utilización de herramientas; mientras que los trabajos sobre soberanía digital enfatizan infraestructura, dependencia tecnológica, regulación y concentración de poder. Aunque aproximaciones recientes, como el paradigma Zero Agency propuesto por Castillo Serracín y Gómez (2026) y el riesgo de delegación decisonal descrito por Issa et al. (2026), comienzan a establecer puentes entre estas dimensiones, permanece insuficientemente articulada la relación entre la delegación del razonamiento, los límites de los mecanismos tradicionales de autorización y la soberanía digital corporativa.

A partir de esta brecha, el objetivo de este estudio es analizar críticamente cómo la delegación de funciones de razonamiento a sistemas de IA corporativa tensiona los principios de Zero Trust y puede generar vulnerabilidades de soberanía digital. La pregunta central que guía la investigación es: ¿de qué manera la delegación del razonamiento y de la capacidad de acción a sistemas corporativos de IA tensiona los mecanismos de verificación y autorización asociados con Zero Trust y qué vulnerabilidades de soberanía digital pueden derivarse de esta transformación? Para responderla, se desarrolló una revisión documental estructurada del estado del arte correspondiente al periodo 2020–2026, sustentada en trabajos científicos revisados por pares seleccionados mediante criterios explícitos de elegibilidad y complementada con documentación técnica del National Institute of Standards and Technology. El propósito final es integrar estos campos en un marco conceptual que permita evaluar la seguridad de la IA corporativa no solamente desde el control de identidad y acceso, sino también desde la preservación de la capacidad de intervención, la autoridad organizacional y la soberanía digital.

## **METODOLOGÍA**

### **Diseño de la investigación**

El estudio se desarrolló mediante un diseño no experimental, documental y descriptivo-analítico, orientado a construir un estado del arte sobre los riesgos asociados con la delegación del razonamiento a sistemas de inteligencia artificial en ecosistemas corporativos y sus implicaciones para Zero Trust y la soberanía digital. Debido a que no se realizaron encuestas, entrevistas, experimentos ni intervenciones con participantes humanos, la evidencia estuvo

constituida exclusivamente por literatura científica y documentación técnica previamente publicada.

El proceso de identificación, cribado, elegibilidad e inclusión de documentos se organizó siguiendo las recomendaciones de PRISMA 2020 (Page et al., 2021), utilizadas como guía para garantizar transparencia y trazabilidad en la selección documental. El periodo de análisis comprendió publicaciones entre enero de 2020 y agosto de 2026, intervalo seleccionado por coincidir con la consolidación de Zero Trust, la expansión de la IA generativa y el desarrollo reciente de arquitecturas de IA agéntica con capacidades de planificación, memoria, razonamiento y utilización autónoma de herramientas.

### **Unidades de análisis y procedimiento de selección**

Las unidades de análisis estuvieron constituidas por 20 trabajos científicos revisados por pares, entre artículos publicados en revistas académicas y trabajos completos presentados en congresos internacionales arbitrados de alta relevancia. Se priorizaron publicaciones indexadas en Scopus, Web of Science e IEEE Xplore y trabajos procedentes de revistas de alto impacto, preferentemente Q1 y Q2, así como contribuciones presentadas en conferencias de reconocido prestigio en ciencias de la computación y ciberseguridad, entre ellas NeurIPS y USENIX Security.

No se estableció previamente un número máximo de estudios. El corpus científico quedó determinado por los documentos que cumplieron los criterios de elegibilidad después de las etapas de identificación, cribado y evaluación a texto completo.

Como apoyo conceptual y normativo se incorporaron dos documentos técnicos del National Institute of Standards and Technology (NIST) relacionados con arquitectura Zero Trust y seguridad de agentes de inteligencia artificial. Estas fuentes se analizaron como documentación técnica complementaria y no se contabilizaron dentro del corpus científico principal. Asimismo, Page et al. (2021) se utilizó exclusivamente como referencia metodológica para la aplicación de PRISMA 2020.

La selección documental se realizó mediante criterios explícitos de elegibilidad. Se incluyeron trabajos publicados entre 2020 y agosto de 2026, en inglés o español, revisados por pares y relacionados directamente con al menos uno de los siguientes ejes: IA agéntica, autonomía

algorítmica, delegación de decisiones, Zero Trust, identidad y autorización, control humano, seguridad de agentes autónomos, dependencia tecnológica o soberanía digital. Se excluyeron documentos duplicados, publicaciones divulgativas, noticias, blogs, contenido comercial, trabajos sin relación directa con la pregunta de investigación y documentos que no aportaran evidencia suficiente para el análisis.

La selección se efectuó en cuatro etapas: identificación de registros, eliminación de duplicados, cribado mediante título y resumen, y evaluación a texto completo. Estas etapas se organizaron siguiendo las recomendaciones de PRISMA 2020 con el propósito de mantener transparencia y trazabilidad en el proceso de selección documental.

### **Técnicas de recolección de datos**

La búsqueda se realizó principalmente en Scopus, Web of Science Core Collection e IEEE Xplore, utilizando términos predominantemente en inglés debido a su mayor presencia en la literatura científica internacional. Las expresiones se adaptaron a la sintaxis específica de cada fuente y se organizaron en tres ejes temáticos:

#### **IA agéntica y seguridad**

("agentic AI" OR "AI agents" OR "autonomous AI agents" OR "LLM agents") AND ("zero trust" OR cybersecurity OR security OR authentication OR authorization OR "least privilege")

#### **IA y soberanía digital**

("artificial intelligence" OR "agentic AI" OR "AI agents") AND ("digital sovereignty" OR "data sovereignty" OR "technological sovereignty" OR "digital autonomy" OR "vendor dependence")

#### **Delegación decisonal y control**

("agentic AI" OR "AI agents" OR "autonomous agents") AND ("decision delegation" OR "decision-making" OR "human control" OR oversight OR autonomy OR governance)

La información de los documentos seleccionados se registró en una matriz de análisis documental que incluyó autores, año de publicación, objetivo, tipo de IA estudiada, nivel de autonomía, mecanismos de identidad y autorización, relación con Zero Trust, riesgos

identificados, mecanismos de supervisión, dependencia tecnológica, implicaciones para la soberanía digital, principales hallazgos y limitaciones.

### **Procedimiento de análisis**

La evidencia se examinó mediante análisis temático de contenido con orientación deductiva e inductiva. La fase deductiva se estructuró alrededor de cinco dimensiones previamente definidas:

- Autonomía y delegación del razonamiento.
- Identidad, autorización y mínimo privilegio.
- Control y seguridad agéntica.
- Dependencia y soberanía digital.
- Gobernanza, supervisión y responsabilidad.

La fase inductiva permitió identificar relaciones emergentes entre las categorías analizadas, particularmente la distinción entre confianza de identidad y confianza decisional, así como las dimensiones de soberanía de datos, tecnológica, operacional y decisional. De esta integración surgieron posteriormente los conceptos de brecha identidad–acción, mínima agencia, confianza decisional y soberanía decisional, utilizados para estructurar la síntesis de los resultados.

Debido a la heterogeneidad de los diseños, objetivos y enfoques de los estudios incluidos, no se realizó metaanálisis. Los resultados se integraron mediante una síntesis narrativa y comparativa orientada a identificar convergencias, diferencias, tendencias y vacíos en la literatura. La trazabilidad del proceso se sostuvo mediante criterios explícitos de elegibilidad, verificación bibliográfica, lectura a texto completo y registro sistemático de la evidencia en la matriz documental.

### **Consideraciones éticas**

Por su naturaleza documental, el estudio no involucró participantes humanos ni tratamiento de datos personales. El análisis se realizó exclusivamente sobre literatura científica y documentación técnica previamente publicada, respetando los principios de atribución, trazabilidad de las fuentes e integridad académica mediante la citación y referencia de los documentos utilizados.

## ***RESULTADOS Y DISCUSIÓN***

El análisis documental reveló una convergencia entre tres campos que, si bien están relacionados, habían avanzado históricamente mediante agendas de investigación separadas: la arquitectura de confianza cero (Zero Trust), la autonomía de los sistemas de inteligencia artificial y la soberanía digital. Los trabajos examinados coinciden en que el salto desde modelos puramente generativos hacia arquitecturas agénticas altera no solo las capacidades técnicas de los sistemas, sino el tipo de control corporativo que estos exigen. Mientras la seguridad tradicional se ha ocupado principalmente de auditar qué identidad solicita acceso, bajo qué contexto, a qué recurso y con qué privilegios, un agente de IA sigue operando después de superar ese filtro. Lo hace mediante procesos de interpretación, planificación, selección de herramientas y ejecución de acciones que rara vez estaban delimitadas de forma exhaustiva al momento de otorgar la autorización. El problema de seguridad, en consecuencia, se desplaza parcialmente: ya no basta con validar la legitimidad de la entidad; también resulta necesario evaluar la legitimidad de su comportamiento posterior.

Este desplazamiento se manifestó de manera recurrente al contrastar las cinco dimensiones de nuestra metodología: autonomía y delegación del razonamiento; identidad, autorización y mínimo privilegio; control y seguridad agéntica; dependencia y soberanía digital; y gobernanza, supervisión y responsabilidad. Frente a especulaciones sobre el fin de Zero Trust, la literatura no respalda que la IA agéntica invalide por sí misma estos fundamentos. Al contrario, ratifica la vigencia de principios como la verificación explícita y el mínimo privilegio, aunque advierte que la autonomía introduce una capa adicional que no puede resolverse únicamente mediante controles de acceso. Paralelamente, los estudios sobre soberanía digital subrayan que el control organizacional ya no puede medirse solo por la propiedad o residencia física de los datos, dado que los modelos y las APIs que transforman esos datos en acciones pueden residir fuera del dominio directo de la empresa. Al integrar estas posturas, se identificaron cuatro categorías analíticas que vertebran los resultados: brecha identidad–acción, mínima agencia, confianza decisional y soberanía decisional.

**Tabla 1.** Síntesis de los principales resultados de la revisión documental

| <b><i>DIMENSIÓN ANALÍTICA</i></b>       | <b><i>RESULTADO INTEGRADO</i></b>                                                                                    | <b><i>TENSIÓN IDENTIFICADA</i></b>                                                                  | <b><i>FUENTES REPRESENTATIVAS</i></b>                                                                                      |
|-----------------------------------------|----------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
| Autonomía y delegación del razonamiento | La IA evoluciona desde recomendación y generación hacia planificación, uso de herramientas y ejecución autónoma.     | La organización puede delegar no solo una tarea, sino parte de la determinación del procedimiento.  | Wang et al. (2024); Adam et al. (2024); Köbis et al. (2025); Issa et al. (2026)                                            |
| Identidad y autorización                | Una entidad correctamente autenticada puede producir posteriormente una operación contraria al propósito autorizado. | Identidad legítima y acción legítima dejan de ser categorías equivalentes.                          | Rose et al. (2020); Debenedetti et al. (2024); Riggs et al. (2026); He et al. (2026)                                       |
| Seguridad agéntica                      | Memoria, herramientas, APIs y datos externos amplían la superficie de ataque.                                        | Una manipulación semántica puede propagarse hasta una acción técnicamente autorizada.               | Liu et al. (2024); Liu et al. (2025); Tang et al. (2026)                                                                   |
| Gobernanza y supervisión                | La autonomía puede disminuir el control humano efectivo y redistribuir responsabilidad.                              | Existencia formal de supervisión no garantiza capacidad efectiva de intervención.                   | Adam et al. (2024); Tsamados et al. (2025); Köbis et al. (2025); Xu et al. (2026)                                          |
| Soberanía digital                       | La dependencia puede desplazarse desde infraestructura y datos hacia capacidades de interpretación y decisión.       | La propiedad de los datos no garantiza control sobre los procesos que los convierten en decisiones. | Roberts (2024); Fratini et al. (2024); van Drunen (2025); Grohmann y Costa Barbosa (2026); Castillo Serracín et al. (2026) |

**Fuente:** Elaboración propia con base en la síntesis temática de la evidencia documental analizada. Las fuentes citadas corresponden a trabajos representativos de cada dimensión analítica.

### **Delegar el razonamiento: de ejecutar instrucciones a determinar cursos de acción**

La primera constante observada en la literatura remite a la distancia creciente entre automatización y delegación del razonamiento. En los esquemas tradicionales, la empresa diseña la secuencia y el sistema ejecuta operaciones previamente definidas. Las arquitecturas agénticas modifican este patrón al ser capaces de recibir metas expresadas en un nivel elevado de abstracción, descomponerlas, analizar el contexto, seleccionar herramientas y corregir su

curso sobre la marcha. Wang et al. (2024) mapearon esta dinámica mediante componentes de memoria, planificación y acción, mientras trabajos posteriores reafirman que la combinación de razonamiento multietapa y uso de APIs amplía considerablemente el espacio operacional del algoritmo. El cambio crítico, por tanto, no es únicamente el volumen de tareas que la IA puede ejecutar, sino el margen de indeterminación que permanece entre la orden formulada y la acción finalmente realizada.

Desde los estudios experimentales surgen argumentos de peso para tratar esta cuestión como un problema de control y no de mera eficiencia. En Decision Support Systems, Adam et al. (2024) encontraron que cuando el sistema inicia la delegación, los usuarios perciben una mayor amenaza cuando anticipan una pérdida de control sobre la operación. Poco después, Köbis et al. (2025) mostraron en Nature que la delegación mediante objetivos generales puede favorecer comportamientos no deseados; ante determinadas solicitudes poco éticas, los agentes de IA mostraron mayor disposición que los participantes humanos a ejecutar dichas instrucciones. Si bien la incorporación de salvaguardas redujo estas desviaciones, el riesgo no desapareció en todos los escenarios.

La lectura conjunta de estos estudios conduce a una advertencia relevante. Cuando la organización establece un objetivo en lugar de definir de manera exhaustiva cada paso, cede un margen de decisión que el sistema debe completar dinámicamente. Esa flexibilidad es precisamente una de las características que otorgan valor a la IA agéntica, pero implica una transferencia parcial de autoridad operacional. El debate no consiste en prohibir esta delegación, sino en determinar cuánto margen interpretativo puede concederse antes de que la autonomía incremente de manera significativa el riesgo. Issa et al. (2026) refuerzan esta idea mediante el concepto de *agentic loafing*, mientras Xu et al. (2026) muestran que, en determinados contextos organizacionales, las personas pueden recurrir a la IA para desplazar parte de la responsabilidad asociada con decisiones difíciles. La autonomía del software y la responsabilidad humana deben, por tanto, analizarse conjuntamente.

Frente a esto, el primer resultado integrado de la revisión puede plantearse en términos de una transición fundamental: el perfil de riesgo cambia cuando la organización deja de delegar únicamente la ejecución y comienza a delegar también la determinación del curso de acción.

## De Zero Trust a la brecha identidad–acción

El segundo hallazgo conecta directamente con el núcleo de la pregunta de investigación. Zero Trust exige eliminar la confianza implícita, de modo que cada decisión de acceso dependa de la identidad, el contexto y las políticas asociadas con recursos específicos (Rose et al., 2020). Estos principios conservan plena relevancia. Las recientes iniciativas del NIST confirman, además, que proteger las identidades y credenciales de entidades no humanas constituye un requisito fundamental. La revisión muestra, sin embargo, que autorizar correctamente el acceso no garantiza por sí solo que el comportamiento posterior de la entidad continúe siendo legítimo.

Es posible que un agente posea credenciales válidas, opere desde un entorno permitido y disponga de accesos legítimos a una base de datos, pero termine utilizando esos privilegios de una forma contraria al propósito organizacional. Este escenario no requiere necesariamente suplantación de identidad ni escalamiento de privilegios; puede originarse en un prompt manipulado, una interpretación incorrecta del contexto, información externa adversa o una salida errónea incorporada posteriormente a la planificación. Desde la perspectiva de los controles tradicionales de acceso, la operación puede seguir apareciendo como técnicamente válida.

Los resultados de AgentDojo, publicados por Debenedetti et al. (2024) en NeurIPS, aportan evidencia empírica relevante al respecto. Al evaluar 97 tareas realistas y 629 casos de seguridad en diferentes dominios, mostraron que la incorporación de información manipulada mediante fuentes externas podía inducir a los agentes a realizar acciones favorables a un atacante. Este hallazgo resulta especialmente pertinente porque evidencia que la amenaza no depende necesariamente de comprometer la identidad del agente, sino de manipular la información que este interpreta y posteriormente traduce en acciones mediante herramientas legítimas.

La literatura especializada en inyección de instrucciones converge en esa dirección. Liu et al. (2024), en USENIX Security, formalizaron este vector de ataque y mostraron limitaciones en las defensas disponibles para aplicaciones integradas con modelos de lenguaje. Como respuesta a este problema, He et al. (2026) propusieron un mecanismo defensivo basado en analizar la causalidad de las invocaciones de herramientas. Su enfoque no se limita a determinar si una entrada es benigna o maliciosa, sino que evalúa si una acción concreta se encuentra realmente respaldada por la intención del usuario o ha sido inducida por contenido externo no confiable. El significado conceptual de este cambio es importante: algunas de las líneas defensivas más

recientes comienzan a desplazarse desde la evaluación del contenido hacia la justificación de la acción.

A la luz de estas evidencias se formula el concepto de brecha identidad–acción, entendido como la discontinuidad que puede surgir entre una entidad correctamente autenticada y la legitimidad organizacional de las acciones que esa misma entidad ejecuta posteriormente.

Esta relación puede sintetizarse en la siguiente proposición:

**P1. Identidad autorizada no implica automáticamente acción autorizada.**

Esta premisa permite precisar a qué se refiere el estudio con la expresión ruptura de la confianza cero. No se trata de un colapso doctrinal de la arquitectura. La discontinuidad es funcional y aparece cuando el sistema verifica satisfactoriamente la identidad y autoriza el acceso, pero no dispone de mecanismos equivalentes para garantizar que las decisiones generadas posteriormente permanezcan dentro del propósito autorizado.

Esta diferencia resulta más clara al observar dos secuencias concatenadas. La primera representa el dominio clásico del control de acceso: identidad → autenticación → contexto → autorización → recurso. La segunda comprende el espacio agéntico que comienza después: objetivo → interpretación → razonamiento → planificación → selección de herramienta → acción → consecuencia. El riesgo se concentra en la articulación entre ambas cadenas. En el software tradicional, el comportamiento suele encontrarse fuertemente delimitado por las funciones de la aplicación; en un entorno agéntico, la entidad autorizada dispone de mayor capacidad para construir dinámicamente su comportamiento. NIST reconoce que esta autonomía introduce amenazas novedosas y requiere adaptar los principios existentes de ciberseguridad (Riggs et al., 2026). Desde una perspectiva complementaria, Kshetri (2025) señala la necesidad de reconsiderar los marcos de seguridad frente a sistemas capaces de automatizar actividades de decisión y respuesta.

**La superficie de ataque se desplaza del modelo al flujo agéntico**

Una tercera constante emerge al examinar dónde se origina y cómo se propaga el riesgo. La superficie de ataque ya no se encuentra únicamente en el modelo fundacional. La memoria persistente, las bases de datos conectadas, las APIs, las herramientas externas y la interacción continua con el entorno amplían las posibles vías de compromiso. Tang et al. (2026), en

Information Fusion, concluyen que los agentes enfrentan una superficie de ataque considerablemente más amplia que los modelos utilizados mediante inferencias convencionales, especialmente cuando combinan razonamiento multietapa e interacción persistente con herramientas.

El alcance práctico de esta expansión se observa también en evaluaciones empíricas. Mediante AgentFuzz, Liu et al. (2025) analizaron vulnerabilidades en veinte aplicaciones agénticas de código abierto e identificaron 34 vulnerabilidades de alto riesgo, varias relacionadas con inyección de código; 23 de ellas recibieron posteriormente identificadores CVE. Estos resultados muestran que la interacción entre lenguaje natural, funciones sensibles y herramientas externas genera rutas de explotación observables en implementaciones reales.

El ciclo de riesgo puede representarse así: entrada → interpretación → razonamiento → planificación → herramienta → ejecución → consecuencia. Una manipulación introducida en una etapa inicial puede propagarse hacia etapas posteriores y aumentar su impacto conforme el agente adquiere capacidad operacional. En una aplicación conversacional, una respuesta incorrecta puede limitarse a la salida textual; cuando el agente administra repositorios cloud, correo corporativo, infraestructura o bases de datos, la misma desviación puede transformarse en una acción sobre recursos reales. Surge así un escenario particularmente relevante para la seguridad: una acción puede ser técnicamente autenticada y formalmente permitida, pero resultar organizacionalmente perjudicial.

Este escenario obliga a reconsiderar el alcance práctico del mínimo privilegio. Restringir los accesos de una identidad continúa siendo indispensable, pero no determina completamente cómo puede utilizarse dicho acceso. Un agente de infraestructura puede requerir privilegios administrativos para gestionar determinados servidores y un agente financiero puede necesitar conectividad con APIs contables; en ambos casos, el privilegio puede resultar legítimo, aunque una utilización particular del mismo sea inadecuada. De aquí se desprende la segunda proposición fundamental:

## **P2. Mínimo privilegio no implica automáticamente mínima agencia.**

En este estudio se propone la mínima agencia como el menor nivel de autonomía decisional y operacional necesario para cumplir un objetivo determinado. Mientras el privilegio determina qué recursos puede utilizar la entidad, la agencia delimita qué puede hacer autónomamente con

ellos, hasta dónde puede actuar y bajo qué condiciones. Ambas dimensiones son complementarias.

Reconocer esta distinción abre la posibilidad de establecer controles proporcionales. Puede concederse a un agente amplia autonomía para analizar o representar información, autonomía condicionada para realizar modificaciones reversibles y aprobación adicional para alterar bases de datos críticas, infraestructura o recursos financieros. El paradigma Zero Agency de Castillo Serracín y Gómez (2026) avanza en esta dirección mediante mecanismos de interceptación y revocación de agencia. La investigación reciente ofrece además respaldo experimental a controles aplicados directamente sobre acciones: He et al. (2026), por ejemplo, evalúan causalmente invocaciones de herramientas para determinar si se corresponden con la intención legítima del usuario.

### **Control humano, confianza decisonal y responsabilidad**

Al evaluar la cuarta dimensión, la literatura muestra que mantener a una persona formalmente dentro del proceso de aprobación no asegura por sí solo control efectivo. Tsamados et al. (2025) cuestionan una visión del control basada exclusivamente en supervisión humana directa y proponen comprender la interacción entre humanos y sistemas de IA mediante formas de agencia colaborativa. La dificultad aparece cuando el volumen, la velocidad o la complejidad de las operaciones algorítmicas hacen impracticable que un operador evalúe individualmente cada acción antes de su ejecución.

La literatura sobre delegación aporta matices complementarios. Adam et al. (2024) muestran que la disposición a delegar se relaciona con el grado de control que los usuarios esperan conservar posteriormente. Köbis et al. (2025) evidencian que objetivos formulados en términos generales dejan al sistema mayor libertad para determinar el procedimiento y pueden producir comportamientos difíciles de anticipar. Xu et al. (2026), mediante ocho experimentos, encontraron además que las personas pueden utilizar sistemas de IA como mecanismo para desplazar parte de la responsabilidad asociada con decisiones difíciles que afectan a terceros. En conjunto, estos trabajos muestran que la pérdida o redistribución del control no procede exclusivamente de la arquitectura técnica; también puede emerger de prácticas organizacionales de delegación.

Por ello, resulta necesario diferenciar entre supervisión nominal y capacidad efectiva de intervención. La primera implica la existencia formal de un actor responsable; la segunda exige disponer de información, tiempo y mecanismos suficientes para comprender, detener, revertir o escalar una operación. Una organización puede implementar formalmente un modelo human-in-the-loop y conservar, sin embargo, una capacidad limitada de control si las alertas se reciben una vez ejecutada la acción o si la complejidad del proceso impide reconstruir oportunamente lo ocurrido.

A partir de esta diferencia se propone el concepto de confianza decisional, definido como el grado de evidencia disponible para considerar que una acción generada por una entidad autónoma resulta coherente con el objetivo delegado, el contexto, las políticas organizacionales y el nivel de riesgo permitido. Se diferencia de la confianza de identidad: la primera se refiere a la legitimidad de la acción; la segunda, a la legitimidad de la entidad y sus privilegios.

Un caso ilustrativo permite clarificar esta diferencia. Supóngase que un sistema interno posee credenciales legítimas para depurar una base de clientes. Si interpreta incorrectamente el criterio de duplicidad y elimina registros válidos, los mecanismos de autenticación y control de acceso pueden haber funcionado correctamente; el problema se encuentra en la relación entre objetivo y acción. La confianza decisional requiere, en consecuencia, controles que exceden la autenticación: observabilidad, evaluación de impacto, límites transaccionales, trazabilidad, reversibilidad y mecanismos de revocación.

La implicación de esta sección no consiste en exigir que un operador humano valide cada acción, sino en reconocer que delegar la ejecución no debe equivaler a una transferencia irreversible de la autoridad organizacional.

### **Soberanía digital: de la propiedad de los datos al control de las capacidades**

El análisis de soberanía digital amplía el problema más allá de la capa estrictamente técnica. Roberts (2024) relaciona la soberanía digital con autoridad y control legítimo sobre las tecnologías, evitando reducir el concepto a la ubicación geográfica de los datos. Frattini et al. (2024), a partir de un análisis amplio de la literatura, muestran que las aspiraciones de autonomía digital deben considerar las dependencias derivadas de infraestructuras, mercados y capacidades tecnológicas concentradas en actores privados.

Para los ecosistemas corporativos de IA, esta dimensión adquiere especial relevancia porque un mismo proceso puede distribuirse entre APIs, servicios de nube, modelos fundacionales e infraestructura de terceros. Van Drunen (2025), al examinar un contexto sectorial específico, muestra cómo la dependencia de infraestructuras controladas por empresas de IA puede limitar la autonomía de las organizaciones y destaca la transparencia, interoperabilidad y disponibilidad de alternativas como factores relevantes para reducir dicha dependencia. Trasladado al ámbito corporativo, esto implica que la capacidad de sustituir o desacoplar tecnologías representa una dimensión importante del control efectivo.

Grohmann y Costa Barbosa (2026) profundizan esta problemática mediante el concepto de *sovereignty-as-a-service*, mostrando cómo determinados proveedores ofrecen soluciones de soberanía construidas sobre infraestructuras que continúan bajo su propio control. Castillo Serracín et al. (2026), desde una perspectiva geopolítica, identifican igualmente infraestructura y datos como componentes estratégicos para la construcción de soberanía digital. La externalización tecnológica no implica automáticamente pérdida de soberanía, pero puede condicionar el margen de decisión cuando sustituir un proveedor, trasladar un servicio o interrumpir una operación resulta técnica, económica u operacionalmente complejo.

La evolución de la IA agéntica introduce una dimensión adicional porque parte de la dependencia se desplaza hacia las capacidades de interpretación y decisión. Una organización puede conservar la propiedad formal de sus bases de datos y, al mismo tiempo, depender de un modelo externo para interpretarlas, priorizarlas y determinar qué acciones ejecutar. La soberanía sobre los datos y la soberanía sobre el proceso decisional dejan así de ser categorías equivalentes. De esta relación surge la tercera proposición central:

### **P3. Propiedad de los datos no implica automáticamente soberanía decisional.**

Esta proposición expresa que la titularidad o control formal sobre los datos no garantiza, por sí sola, que la organización conserve autoridad efectiva sobre los procesos que los interpretan y los transforman en decisiones; en otros términos, la propiedad de los datos no equivale necesariamente a soberanía decisional.

Para analizar esta diferencia, la soberanía digital corporativa se organiza en cuatro dimensiones interdependientes. La soberanía de datos comprende el control sobre almacenamiento, acceso, procesamiento y transferencia de información. La soberanía tecnológica representa la

capacidad de seleccionar, sustituir o desacoplar proveedores, modelos e infraestructura sin comprometer la continuidad operacional. La soberanía operacional se refiere a la capacidad de observar, limitar, detener y recuperar procesos automatizados. Finalmente, la soberanía decisional expresa la capacidad de la organización para conservar autoridad efectiva sobre decisiones que afectan sus recursos, objetivos e intereses estratégicos.

Estas dimensiones se condicionan mutuamente. Una elevada dependencia de un modelo propietario puede reducir la sustituibilidad tecnológica; una baja sustituibilidad puede limitar la capacidad operacional para modificar o detener determinados componentes; y ambas condiciones pueden afectar progresivamente la capacidad de la organización para gobernar decisiones críticas.

### **Marco integrador: identidad, autoridad, agencia y soberanía**

Integrar los resultados anteriores permite construir un marco explicativo unificado. El proceso comienza cuando la organización delega a un sistema una tarea que exige interpretación o razonamiento. A medida que aumenta su autonomía, el agente selecciona información, planifica y utiliza herramientas. Puede operar mediante credenciales legítimas y, aun así, generar una acción que exceda o se aparte del propósito autorizado, configurando la brecha identidad–acción. Si, además, la organización carece de mecanismos suficientes para observar, limitar o revocar esa acción debido a dependencias tecnológicas u operacionales, el incidente puede trascender el control de acceso y convertirse en una vulnerabilidad de soberanía.

El ciclo puede visualizarse de la siguiente manera:

Delegación del razonamiento → autonomía → planificación dinámica → selección de herramientas → acceso legítimo → acción emergente → brecha identidad–acción → reducción del control efectivo → vulnerabilidad de soberanía operacional y decisional.

Este modelo permite observar que la tensión con Zero Trust no surge por el mero uso de inteligencia artificial. Requiere la convergencia de tres condiciones: una identidad y privilegios legítimos, autonomía suficiente para generar acciones no completamente especificadas al momento de la autorización y ausencia de controles proporcionales que permitan observar, validar, limitar o revocar dichas acciones antes de que produzcan consecuencias relevantes.

Las tres proposiciones derivadas de la revisión pueden sintetizarse de la siguiente manera:

- P1. Identidad autorizada no implica automáticamente acción autorizada.
- P2. Mínimo privilegio no implica automáticamente mínima agencia.
- P3. Propiedad de los datos no implica automáticamente soberanía decisonal.

Estas proposiciones permiten ordenar la gobernanza de IA en cuatro niveles. El nivel de identidad determina quién o qué es el agente. El nivel de autoridad delimita a qué recursos puede acceder. El nivel de agencia establece qué margen decisonal y operacional puede ejercer mediante esos privilegios. Finalmente, el nivel de soberanía determina si la organización conserva capacidad efectiva para observar, limitar, sustituir, detener y recuperar el control.

Lejos de sustituir a Zero Trust, esta estructura extiende su lógica hacia el comportamiento. El principio never trust, always verify conserva su vigencia, pero el objeto de verificación se amplía desde identidad y acceso hacia la relación entre objetivo, acción e impacto. Los trabajos recientes sobre atribución causal de invocaciones de herramientas (He et al., 2026) y los resultados de AgentDojo (Debenedetti et al., 2024) aportan evidencia de que esta transición desde el control de entrada hacia el control de acción constituye ya una línea activa de investigación en seguridad.

No todas las operaciones requieren el mismo nivel de supervisión. Resumir un correo electrónico no presenta el mismo riesgo que reconfigurar un servidor o autorizar una transferencia financiera. El principio práctico que se desprende de la revisión es la necesidad de establecer proporcionalidad entre nivel de agencia, privilegio, reversibilidad e impacto operacional, de modo que el grado de autonomía concedido no exceda la capacidad organizacional de gobernarlo.

### **Respuesta a la pregunta de investigación y discusión general**

Los hallazgos permiten responder la pregunta central del estudio. La delegación del razonamiento y de la capacidad de acción tensiona los mecanismos de verificación y autorización asociados con Zero Trust cuando una entidad correctamente autenticada y autorizada dispone de suficiente autonomía para producir acciones que no estaban completamente determinadas en el momento de conceder sus privilegios. La tensión no necesariamente se origina en una contraseña comprometida o en una escalada de privilegios,

sino en la discontinuidad operacional entre autorizar a una entidad para utilizar un recurso y garantizar que las acciones generadas posteriormente permanezcan dentro del propósito autorizado. Esta discontinuidad constituye la brecha identidad–acción.

Esta constatación amplía el escenario tradicional de amenaza. Buena parte de la ciberseguridad corporativa ha sido diseñada para enfrentar identidades ilegítimas que intentan obtener acceso a recursos protegidos. Los sistemas agénticos introducen una segunda posibilidad: una identidad legítima que genera una acción no deseada dentro del espacio que le fue correctamente autorizado. La evidencia aportada por estudios sobre prompt injection, uso de herramientas y vulnerabilidades en agentes muestra que este escenario no es meramente conceptual. En consecuencia, gobernar el acceso sigue siendo necesario, pero debe complementarse con mecanismos capaces de evaluar objetivos, herramientas, acciones e impacto y de revocar oportunamente la autonomía concedida.

El segundo eje de la pregunta se resuelve al analizar la relación entre control y soberanía. Las vulnerabilidades de soberanía pueden aparecer cuando una organización externaliza simultáneamente componentes esenciales del ecosistema algorítmico —datos, infraestructura, modelos de inferencia, APIs y mecanismos de ejecución— y pierde capacidad práctica para sustituirlos o intervenir sobre ellos sin afectar significativamente su continuidad operacional. Esta configuración no implica automáticamente pérdida de soberanía, pero introduce variables como sustituibilidad, interoperabilidad, observabilidad, reversibilidad y capacidad efectiva de intervención, que deben incorporarse al análisis del riesgo tecnológico.

El principal aporte de la síntesis consiste, entonces, en integrar seguridad agéntica, Zero Trust y soberanía digital como dimensiones relacionadas de un mismo problema de gobernanza tecnológica. Delegar razonamiento implica transferir parcialmente capacidad para determinar acciones; transferir capacidad de acción exige gobernar la agencia; y depender de terceros para proporcionar las capacidades que sustentan dicha agencia puede condicionar la autonomía tecnológica, operacional y decisional de la organización. Bajo esta perspectiva, la arquitectura corporativa debe evolucionar para responder no solo quién accede a los recursos, sino también qué puede decidir, qué puede ejecutar, bajo qué límites y mediante qué mecanismos puede recuperarse el control.

Estos resultados deben interpretarse considerando que la literatura sobre IA agéntica continúa siendo emergente y metodológicamente heterogénea. La evidencia analizada integra estudios

experimentales, evaluaciones de seguridad, modelos conceptuales y revisiones científicas, complementados con documentación institucional, mientras la literatura sobre soberanía digital mantiene todavía una orientación importante hacia Estados, plataformas e infraestructuras. En consecuencia, conceptos propuestos en este estudio, como brecha identidad–acción, mínima agencia, confianza decisional y soberanía decisional, deben entenderse como resultados de una síntesis conceptual fundamentada que requieren posterior validación empírica en entornos corporativos reales.

Debido a esta limitación, sería prematuro afirmar que cualquier implementación de IA agéntica produce una ruptura de Zero Trust o una pérdida de soberanía. El nivel de riesgo dependerá de la amplitud de los privilegios, el grado de autonomía delegado, la criticidad de las acciones, la observabilidad, la dependencia tecnológica, la sustituibilidad de los componentes y la capacidad efectiva de intervención. En conjunto, la evidencia revisada sugiere que la delegación del razonamiento no invalida la confianza cero, pero amplía el objeto de verificación desde la identidad y el acceso hacia la acción autónoma; de manera paralela, la soberanía digital corporativa deja de limitarse a la propiedad o localización de los datos y pasa a incorporar la capacidad efectiva de conservar autoridad sobre los procesos que transforman esos datos en decisiones y acciones.

## **CONCLUSIÓN**

La delegación del razonamiento algorítmico representa una transformación relevante en la gobernanza tecnológica de las organizaciones. Los resultados de esta revisión permiten sostener que la adopción de arquitecturas de inteligencia artificial agéntica amplía uno de los problemas centrales de la ciberseguridad corporativa: proteger adecuadamente el acceso a los recursos continúa siendo necesario, pero puede resultar insuficiente cuando una entidad autorizada posee capacidad para interpretar contextos, seleccionar herramientas, construir secuencias operativas y ejecutar acciones que no estaban completamente determinadas al momento de conceder sus privilegios. En estos entornos, el riesgo deja de concentrarse exclusivamente en la intrusión de actores ilegítimos y se extiende hacia el comportamiento autónomo que puede producirse mediante identidades y accesos legítimos.

La principal contribución conceptual del estudio radica en la articulación de la brecha identidad–acción, entendida como la discontinuidad potencial entre la legitimidad de una entidad correctamente autenticada y la legitimidad organizacional de las acciones que esta

genera posteriormente. La literatura revisada confirma la vigencia de los principios de Zero Trust, pero muestra la necesidad de ampliar su alcance cuando la verificación de identidad y acceso no resulta suficiente para gobernar la autonomía posterior del sistema. En este contexto, las tres proposiciones derivadas del análisis —P1. una identidad autorizada no implica automáticamente una acción autorizada; P2. el mínimo privilegio no implica automáticamente mínima agencia; y P3. la propiedad de los datos no implica automáticamente soberanía decisional— permiten integrar seguridad agéntica, gobernanza y soberanía digital dentro de un mismo marco analítico. Este marco amplía el control desde identidad y autoridad hacia dos niveles adicionales: la agencia, referida al margen decisional y operacional concedido al sistema, y la soberanía, entendida como la capacidad efectiva de la organización para observar, limitar, sustituir, detener y recuperar el control.

El análisis también permite ampliar la comprensión de la soberanía digital en el ámbito corporativo. La propiedad o localización de los datos constituye una dimensión importante, pero no garantiza por sí sola que la organización conserve capacidad efectiva sobre los modelos, infraestructuras, APIs y servicios que los interpretan y los transforman en decisiones y acciones. Cuando funciones críticas dependen de componentes externos difícilmente sustituibles, la dependencia tecnológica puede trasladarse hacia los planos operacional y decisional. De este modo, la soberanía decisional emerge como una dimensión complementaria de la soberanía digital corporativa, vinculada con la capacidad de determinar qué decisiones pueden delegarse, bajo qué límites, mediante qué tecnologías y en qué condiciones debe recuperarse la autoridad sobre el proceso.

Estos hallazgos deben interpretarse considerando el carácter emergente y metodológicamente heterogéneo de la literatura sobre IA agéntica. Las categorías propuestas —brecha identidad—acción, mínima agencia, confianza decisional y soberanía decisional— constituyen resultados de una síntesis conceptual fundamentada y requieren posterior validación empírica. En consecuencia, futuras investigaciones deberían desarrollar métricas capaces de evaluar el grado de confianza decisional, analizar experimentalmente la eficacia de mecanismos de revocación dinámica y control a nivel de acción, y determinar cómo variables como sustituibilidad tecnológica, observabilidad, reversibilidad y dependencia de proveedores afectan la capacidad organizacional para intervenir frente a comportamientos autónomos no deseados.

En conjunto, la evidencia revisada sugiere que el principal desafío de la IA agéntica no consiste en sustituir los fundamentos existentes de ciberseguridad, sino en extenderlos hacia una nueva

unidad de control: la acción autónoma. La evolución desde sistemas que ejecutan instrucciones hacia sistemas capaces de determinar cursos de acción exige que las organizaciones gobiernen no solo quién accede y con qué privilegios, sino también qué puede decidir un agente, qué puede ejecutar, bajo qué condiciones y cómo puede revocarse esa autonomía. Bajo esta perspectiva, delegar razonamiento y ejecución no debería implicar renunciar a la capacidad organizacional de conservar y recuperar el control.

## REFERENCIAS

- Adam, M., Diebel, C., Goutier, M., & Benlian, A. (2024). Navigating autonomy and control in human-AI delegation: User responses to technology- versus user-invoked task allocation. *Decision Support Systems*, 180, 114193. <https://doi.org/10.1016/j.dss.2024.114193>
- Castillo Serracín, J. R., & Gómez, J. (2026). Determinantes de seguridad y brecha de control en la era agéntica: Hacia un marco operativo de Zero Agency validado por expertos. *Revista Científica de Salud y Desarrollo Humano*, 7(2), 961–976. <https://doi.org/10.61368/r.s.d.h.v7i2.1047>
- Castillo Serracín, J. R., Aranda, O., Farnum Castro, F., & Gómez Solís, J. M. (2026). Digital frontiers: Cybersecurity, sovereignty, and the new paradigms of geopolitical power. *Revista Iberoamericana de educación*, 10(2), 238–251. <https://doi.org/10.31876/rie.v10i2.372>
- Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems*, 37, 82895–82920. <https://doi.org/10.52202/079017-2636>
- Fratini, S., Hine, E., Novelli, C., Roberts, H., & Floridi, L. (2024). Digital sovereignty: A descriptive analysis and a critical evaluation of existing models. *Digital Society*, 3, Article 59. <https://doi.org/10.1007/s44206-024-00146-7>
- Grohmann, R., & Costa Barbosa, A. (2026). Sovereignty-as-a-service: How Big Tech companies co-opt and redefine digital sovereignty. *Media, Culture & Society*, 48(2), 416–424. <https://doi.org/10.1177/01634437251395003>
- He, Y., Zhu, H., Li, Y., Shao, S., Yao, H., Liu, Z., & Qin, Z. (2026). AttriGuard: Defeating indirect prompt injection in LLM agents via causal attribution of tool invocations. In *35th USENIX Security Symposium (USENIX Security 26)*. USENIX

Association.

<https://www.usenix.org/conference/usenixsecurity26/presentation/he-yu>

- Issa, H., Petani, F. J., & Glavas, D. (2026). *Agentic loafing: An AI decision delegation risk*. *Risk Analysis*, 46(8), e70306. <https://doi.org/10.1111/risa.70306>
- Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). *Delegation to artificial intelligence can increase dishonest behaviour*. *Nature*, 646, 126–134. <https://doi.org/10.1038/s41586-025-09505-x>
- Kshetri, N. (2025). *Transforming cybersecurity with agentic AI to combat emerging cyber threats*. *Telecommunications Policy*, 49(6), 102976. <https://doi.org/10.1016/j.telpol.2025.102976>
- Liu, F., Zhang, Y., Luo, J., Dai, J., Chen, T., Yuan, L., Yu, Z., Shi, Y., Li, K., Zhou, C., Chen, H., & Yang, M. (2025). *Make agent defeat agent: Automatic detection of taint-style vulnerabilities in LLM-based agents*. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 3767–3786). USENIX Association. <https://www.usenix.org/conference/usenixsecurity25/presentation/liu-fengyu>
- Liu, Y., Jia, Y., Geng, R., Jia, J., & Gong, N. Z. (2024). *Formalizing and benchmarking prompt injection attacks and defenses*. In *Proceedings of the 33rd USENIX Security Symposium* (pp. 1831–1847). USENIX Association. <https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupe>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews*. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Peykani, P., Ghanidel, S., Javadi-Sisi, I., Snasel, V., & Mirjalili, S. (2026). *A holistic review of agentic AI frameworks, applications, and research trajectories*. *Archives of Computational Methods in Engineering*. Advance online publication. <https://doi.org/10.1007/s11831-026-10675-8>
- Riggs, J., Hamin, M., Perry, N., Edelman, B., & Cihon, P. (2026). *Summary analysis of responses to the request for information regarding security considerations for AI agents (NIST Trustworthy and Responsible AI 800-5)*. National Institute of Standards and Technology. <https://www.nist.gov/publications/summary-analysis-responses-request-information-regarding-security-considerations-ai>

- Roberts, H. (2024). *Digital sovereignty and artificial intelligence: A normative approach*. *Ethics and Information Technology*, 26, Article 70. <https://doi.org/10.1007/s10676-024-09810-5>
- Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture (NIST Special Publication 800-207)*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-207>
- Stelmaszak, M., Joshi, M., & Constantiou, I. (2026). *Artificial intelligence as an organizing capability arising from human-algorithm relations*. *Journal of Management Studies*, 63(2), 335–365. <https://doi.org/10.1111/joms.70003>
- Tang, Y., Liu, Y., Lan, J., Yan, Z., & Gelenbe, E. (2026). *Security of LLM-based agents regarding attacks, defenses, and applications: A comprehensive survey*. *Information Fusion*, 127, 103941. <https://doi.org/10.1016/j.inffus.2025.103941>
- Tsamados, A., Floridi, L., & Taddeo, M. (2025). *Human control of AI systems: From supervision to teaming*. *AI and Ethics*, 5(2), 1535–1548. <https://doi.org/10.1007/s43681-024-00489-4>
- van Drunen, M. Z. (2025). *Safeguarding media freedom from infrastructural reliance on AI companies: The role of EU law*. *Telecommunications Policy*, 49(7), 102990. <https://doi.org/10.1016/j.telpol.2025.102990>
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). *A survey on large language model based autonomous agents*. *Frontiers of Computer Science*, 18, 186345. <https://doi.org/10.1007/s11704-024-40231-1>
- Xu, L., Tian, H., Zhang, Y., & Yu, F. (2026). *Shifting accountability to artificial intelligence: Delegating challenging decisions to AI for responsibility avoidance*. *Journal of Business Research*, 215, 116313. <https://doi.org/10.1016/j.jbusres.2026.116313>

© Los autores. Este artículo se publica en Prisma ODS bajo la Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Esto permite el uso, distribución y reproducción en cualquier medio, incluidos fines comerciales, siempre que se otorgue la atribución adecuada a los autores y a la fuente original.



 <https://doi.org/10.65011/prismaods.v5.i5.350>

**Cómo citar este artículo (APA 7ª edición):**

Castillo Serracín, J. R., Bruce, C., González Aparicio, I. M., Corpas Ford, M., & Rapón, M. (2026). Riesgo de Delegar el Razonamiento: Ruptura de la Confianza Cero y Vulnerabilidades de Soberanía Digital en Ecosistemas Corporativos de IA

*Prisma ODS: Revista Multidisciplinaria Sobre Desarrollo Sostenible*, 5(5), 165-192. <https://doi.org/10.65011/prismaods.v5.i5.350>