Introducción a los riesgos algorítmicos

Repensando los modelos de seguridad y control en la era de la inteligencia artificial

DOI: 10.29236/sistemas.n177a6

Resumen

La acelerada incorporación de la inteligencia artificial (IA) y los agentes autónomos obliga a las organizaciones a revisar detalladamente los resultados de su implementación, ya que los modelos basados en *machine learning* generan respuestas que muchas veces no corresponden a una lógica esperada, sino a un patrón estadístico predeterminado. Esta nueva realidad ha dado lugar al riesgo algorítmico, un riesgo emergente definido como la posibilidad de daño, pérdida financiera o afectación de la reputación empresarial que surge del uso, despliegue o explotación maliciosa de sistemas de IA, que hace evidente las limitaciones propias de los modelos tradicionales de seguridad y control. Por tanto, este artículo propone adoptar una perspectiva ampliada de seguridad que incluye la autenticidad, la utilidad y la posesión como nuevos elementos a revisar, además de introducir el concepto de auditoría algorítmica como un proceso de aseguramiento crítico previo al despliegue de iniciativas con inteligencia artificial. Finalmente, este documento presenta un enfoque holístico de la seguridad denominado “Confianza por Diseño”, que opera como marco integrador que vincula la ética, la equidad y la explicabilidad, junto con la seguridad y la privacidad, como una postura proactiva para la gestión de riesgos empresariales en la era de la IA.

Palabras clave

Riesgo algorítmico, Confianza por diseño, Agentes autónomos, Integridad semántica, auditoría algorítmica

Introducción

La inteligencia artificial avanza de forma acelerada, y su incorporación en prácticamente todos los aspectos de la dinámica social, hace que tanto personas como organizaciones comiencen a revisar con detalle sus resultados y los efectos de su implementación, comoquiera que las respuestas de estos modelos (algoritmos basados en *machine learning*) muchas veces no corresponden ni a un contexto, ni a una lógica concreta y esperada, sino a un patrón estadístico previamente establecido en la programación inicial de la iniciativa (Janapa Reddi, 2025).

En este sentido, las iniciativas basadas en inteligencia artificial que se incorporan en la actualidad, no son capaces de distinguir entre datos contaminados o maliciosos y comandos legítimos, lo que no les permite interpretar la información de manera correcta, válida y conforme a la realidad o al contexto para el cual fueron diseñadas, creando una brecha de integridad semántica que genera necesariamente alucinaciones, pues el agente no puede verificar su propia integridad usando los mismos mecanismos potencialmente viciados (Raghavan & Schneier, 2025).

Frente a esta nueva realidad de riesgos algorítmicos, la postura tradicional de la seguridad basada en confidencialidad, integridad y disponibilidad, se queda corta para explicar y tratar de avanzar en el aseguramiento de las nuevas iniciativas basadas en inteligencia artificial, donde el reto como que puede observar no está en si se tiene o no acceso a la información o si está disponible o no, sino en la integridad de la información, más allá de que los datos estén completos y documentados sus cambios, sino asegurar que la interpretación de los resultados de la ejecución del modelo coincida con la intención humana o la realidad externa (Raghavan & Schneier, 2025).

En este sentido, plantear una estrategia de aseguramiento para los riesgos algorítmicos demanda retomar las reflexiones planteadas por Parker (1998), donde adicional a los principios básicos de seguridad de la información: confidencialidad, integridad y disponibilidad, se incluyan autenticidad, utilidad y posesión, los cuales resultan de especial interés en esta era de algoritmos y despliegue de agentes autónomos, que pueden terminar generando ataques no tradicionales, con efectos sistémicos que aún no se pueden determinar.

Por tanto, este artículo plantea una introducción al tratamiento de los nuevos riesgos algorítmicos, aquellos que se derivan del uso, despliegue o explotación maliciosa de sis-

temas de Inteligencia Artificial (IA), incluyendo la IA generativa (IA Gen) y los agentes autónomos (IA Agéntica), que permita retar los fundamentos actuales de la seguridad de la información, para plantear algunas propuestas de seguridad y control, y así abordar el desafío de aseguramiento que implica la acelerada implementación y puesta en operación de soluciones de agentes y algoritmos basados en *machine learning*.

Confidencialidad, integridad y disponibilidad: ¿por qué ya no son suficientes en la era de la IA?

Los fundamentos de seguridad de la información nacen en una era de flujos de información, control de acceso y tensiones internacionales por cuenta de la guerra fría. En esa época el reto se traducía en el ejercicio de confidencialidad: información correcta a personas correctas, con el fin de mantener la tranquilidad de las diferentes partes interesadas e involucradas (2011). En esa época, la información como fundamento del ejercicio del poder, configuraba las relaciones a nivel internacional para lograr una posición estratégica, que le permitía a las naciones y sus aliados tomar acciones de forma anticipada con la mayor afectación y la mínima capacidad de respuesta de su contraparte.

De igual forma, mantener un registro de la información obtenida en medio de las confrontaciones y la inteligencia entre los diferentes actores en disputa, genera la necesidad de asegurar su disponibilidad de forma oportuna y confiable, con el fin de elaborar los informes requeridos para la toma de decisiones, y cruzar diferentes tipos de fuentes de información. En este contexto, las bases de datos aparecen como elementos claves, no sólo para almacenamiento ordenado y efectivo de los datos, sino como la configuración de una fuente centralizada y con control de acceso estricto para generar los análisis de información por las personas autorizadas.

Si bien la integridad de la información en esa época era un elemento importante, muy pocos avances se dieron en aquellos momentos y aún hoy, la deuda con la integridad se sigue acumulando con la llegada y evolución acelerada de la inteligencia artificial. Hoy la integridad se entiende más allá de tener el conjunto de datos completo y asegurado, ahora en el escenario de la autenticidad, esto es el valor o significado extrínseco de la información, es decir, si es genuina y conforme a la realidad (Parker, 1998).

En términos prácticos, el objetivo de la integridad ahora es asegurar que la información y la lógica de decisión del agente de IA no sean alteradas o corrompidas, incluso si la entrada inicial fue maliciosa. Esto es, volver resistente la arquitectura donde funciona la iniciativa a ataques de inyección de *prompts* y

mantenerse ajustado al contexto de la organización, limitando su desconexión con las reglas del negocio (Díaz et al., 2025). Lo que supone incorporar un monitor de referencia, que media y controla todas las solicitudes o entradas, bien de los usuarios u otras IA, a los recursos del sistema, aplicando la política de seguridad del sistema para asegurar que solo se realicen operaciones autorizadas.

En estén sentido, Parker (1998) anticipa la necesidad de incorporar más allá de confidencialidad, integridad y disponibilidad, tres principios adicionales que se hacen necesarios para enmarcar la nueva era de los riesgos algoritmos y empezar una visión extendida de la comprensión y tratamiento de este riesgo. Los tres principios adicionales son la autenticidad, la utilidad y la posesión los cuales se describen en conjunto con los tradicionales con un ejemplo aplicado al despliegue de un agente de IA (Ver tabla No.1).

Tabla 1. Aplicación de los seis principios de seguridad de Parker

Elemento	Descripción para el Agente
Disponibilidad	El agente debe **acceder al inventario en tiempo real** a través de la API (*Application Program Interfase*) del sistema de gestión. Si la red falla o la base de datos se cae, la disponibilidad se pierde.
Utilidad	El agente necesita que los datos del inventario estén en un **formato legible y actual** (por ejemplo, un recuento numérico preciso, no un archivo de registro corrupto) para tomar decisiones de pedido. Si el agente solo puede recuperar la información de inventario de hace una semana o en un formato incomprensible (una pérdida de utilidad), la decisión de reabastecimiento será errónea.
Integridad	El agente debe asegurar que los documentos de origen, una vez recuperados de la base de conocimiento (RAG – *Retrieval-Augmented Generation*), estén **completos y que no hayan sido modificados** internamente desde su almacenamiento.
Autenticidad	El agente debe confirmar que los documentos recuperados son **genuinamente la última versión válida** firmada por la autoridad correcta (conformidad con la realidad). Si un atacante reemplaza un informe financiero real por uno falso, pero con el formato correcto, la autenticidad se pierde por **manipulación**, incluso si el archivo falso tiene una "integridad" interna asegurada (está completo).
Confidencialidad	El agente debe asegurar que la información personal contenida en el correo electrónico **no sea revelada** ni **observada** por partes no autorizadas. Esto se viola si el agente incluye accidentalmente esta información en una respuesta genérica que luego es vista por el público o por otros empleados sin autorización.
Posesión	El agente y, por extensión, la organización, deben mantener el **control físico y lógico** sobre la información del cliente. Si el agente es víctima de una **inyección de *prompt*** indirecta oculta en un documento, y es engañado para que *exfiltre* los datos sensibles a través de una URL maliciosa, se ha perdido la posesión de esos datos, ya que han pasado al control de un adversario. La pérdida de control (posesión) es un riesgo directo de las acciones deshonestas de los agentes.

Nota: Elaboración propia con ideas de Parker, 1998.

Riesgos algorítmicos. Una nueva frontera para la gestión de riesgos empresariales

Cuando se desarrolla la gestión de riesgos empresariales de forma tradicional dos retos posiblemente estén quedando fuera del radar de las organizaciones. Uno es entender que las organizaciones ahora se sitúan dentro de un ecosistema digital de negocios, por lo tanto no es sólo detallar y reconocer cómo la organización funciona de acuerdo con los retos y exigencias esperadas para cumplir con sus objetivos de negocios, sino cómo se relaciona y crea capacidades claves con sus terceros de confianza, para entregar experiencias distintas en sus clientes.

El otro es la emergencia de riesgos cognitivos, aquellos asociados con la capacidad de una persona o grupos de personas de crear contexto hiperrealistas, combinando información confiable con datos falsos para crear narrativas y situarlas en el imaginario de individuos o grupos sociales específicos aprovechándose de las vulnerabilidades sociales, cognitivas y tecnológicas (Bone & Lee, 2023), y de riesgos algorítmicos, previamente definidos.

El riesgo algorítmico es la posibilidad de un daño, pérdida financiera o afectación de la reputación de una organización que surge por: (Godhrawala, 2025)

- *Fallos internos del sistema de IA*: Sistemas que operan como "cajas negras" y cuyas decisiones son difíciles de interpretar y auditar. Estos fallos pueden generarse debido a comportamientos impredecibles (agentes autónomos que se desvían de los objetivos previstos), amplificación de sesgos presentes en los datos de entrenamiento que conducen a resultados discriminatorios, o errores que resultan en fallas operacionales y sobrecarga de recursos.

- *Explotación por actores maliciosos*: El uso de IA por parte de adversarios (cibercriminales, crimen organizado, personas inescrupulosas o actores no estatales y estatales) para automatizar, acelerar y escalar las amenazas cibernéticas. La IA Gen reduce drásticamente la barrera de entrada para los atacantes, permitiendo a actores menos calificados lanzar ataques sofisticados, ingeniería social avanzada y campañas de desinformación hiperpersonalizada con sólo elaborar los *prompts* adecuados.

El riesgo algorítmico introduce vulnerabilidades novedosas e inéditas que no existían en los sistemas tradicionales. La IA Agéntica y la IA Gen al operar en entornos inherentemente hostiles e interactuar con fuentes no confiables, introduce vulnerabilidades estructurales que generan retos en la interpretación de sus resultados. El problema fundamental radica en que la IA debe

comprimir la realidad en formas legibles para sus modelos, creando una brecha semántica que puede ser explotada por los adversarios (Raghavan & Schneier, 2025).

Para comprender mejor este escenario de riesgos algorítmicos se aplica el modelo OODA (Observar, Orientar, Decidir y Actuar) para los agentes de inteligencia artificial, particularmente autónomos como se observa en la tabla No.2. El modelo OODA fue introducido por el Coronel John Boyd de la Fuerza Aérea de los Estados Unidos de América hace décadas. Se concibió como un marco para que los pilotos de combate comprendieran la toma de decisiones continua en tiempo real. En este sentido, un agente de IA, al igual que un piloto, ejecuta este ciclo repetidamente para lograr sus objetivos dentro de

Tabla 2. Aplicación del modelo OODA para una IA Agéntica

Fase OODA	Definición	Riesgos de la IA Agéntica	Implicación de Seguridad Clave
Observar	Recopilación de información en tiempo real.	• Inyección de prompts. • Los atacantes proporcionan las observaciones y manipulan la salida.	• La capa de observación **carece de autenticación e integridad**. • Las instrucciones maliciosas ocultas en los datos pueden afectar el resultado.
Orientar	Construcción de la "visión del mundo" del agente basada en las observaciones.	• Envenenamiento de datos de entrenamiento • Manipulación de contexto • Puertas traseras semánticas[1]	La orientación del modelo puede ser influenciada meses antes del despliegue, activando comportamientos codificados con frases de activación.
Decidir	Formulación de un plan de acción.	• Corrupción Lógica mediante ataques de ajuste fino (*fine-tuning*) • Manipulación de recompensas • Desalineación de objetivos	El proceso de decisión probabilístico del LLM se convierte en la carga útil del ataque, y los modelos pueden ser manipulados para confiar en fuentes maliciosas preferentemente.
Actuar	Ejecución del plan de acción mediante el uso de herramientas.	• Manipulación de la salida, • Confusión de herramientas • Secuestro de Acciones	Cada llamada a una herramienta confía implícitamente en las etapas previas.

Nota: Basado en: Raghavan & Schneier, 2025

[1] Es una forma de compromiso de la integridad de un modelo de lenguaje grande (LLM) o agente de IA, que explota su proceso de aprendizaje para inyectar un comportamiento malicioso latente que solo se activa bajo condiciones o frases específicas conocidas como disparadores (*triggers*) (Chen et al., 2025)

un entorno en constante cambio. Los sistemas de IA Agéntica son sistemas diseñados para percibir su entorno, tomar decisiones y ejecutar acciones autónomas para alcanzar metas definidas por el usuario (Raghavan & Schneier, 2025).

Auditoría algorítmica. Nueva práctica de seguridad y control en la era de la IA

La Auditoría algorítmica (o de IA) previa al despliegue de una iniciativa de IA es un proceso de aseguramiento independiente y sistemático diseñado para verificar que los sistemas de Inteligencia Artificial (IA), modelos de lenguaje de gran escala (LLMs) o agentes autónomos (IA Agéntica) cumplen con los requisitos de negocio, legales, éticos y de seguridad antes de que entren en producción (Godhrawala, 2025).

El objetivo central es prevenir la materialización de riesgos algorítmicos (como sesgos, fallas operacionales, o inyección de instrucciones maliciosas) mediante la evaluación detallada del diseño, los datos de entrenamiento y los controles de gobierno y seguridad, antes de que el sistema comience a tomar decisiones de forma autónoma.

Para aplicar una auditoría algorítmica, es necesario enfocarse en los aspectos que aseguren la confianza de los resultados, la resiliencia frente a fallas y el cumplimiento regulatorio vigente a la fecha (por ahora fragmentado y poco claro): (Mckinsey, 2025).

1. Gobernanza y Responsabilidad (*Accountability*):

- Matriz de responsabilidad: Se debe evaluar si existe un equipo de gobernanza interdisciplinario y si se han desarrollado matrices de responsabilidad claras que definan la rendición de cuentas por las acciones de cada agente.

- Alineación estratégica: Asegurar que la estrategia de IA esté alineada con los objetivos de negocio y de Recursos Humanos, y que los procesos para mantener esta alineación existan y operen.

- Controles de terceros: Verificar que los procesos de adquisición de IA de terceros son robustos y que se ha negociado el derecho de auditar los modelos o servicios contratados.

2. Transparencia y Explicabilidad:

- Modelos interpretables: Confirmar que los sistemas, especialmente aquellos que producen resultados no matemáticos, despliegan modelos de IA Explicable que permiten interpretar y auditar la lógica de la decisión.

- Registros Inmutable: Asegurar que el sistema está diseñado para generar rastros de auditoría inmutables (no alterables) que capturen todas las acciones del

agente para asegurar la rendición de cuentas y la trazabilidad de los errores.

3. Riesgo de Datos y Sesgo:

- Calidad de datos: Auditar la calidad y curaduría de los conjuntos de datos de entrenamiento.

- Mitigación de sesgos: Auditar sistemáticamente los datos de entrenamiento para verificar su confiabilidad, y así limitar que la IA perpetúe sesgos presentes en los datos, que generen resultados discriminatorios.

- Protección de datos: Verificar que se hayan implementado políticas de clasificación de datos y que se proteja la propiedad intelectual y los datos propietarios utilizados en el entrenamiento.

4. Seguridad por Diseño y Pruebas:

- Analítica de comportamiento: Asegurar que existen capacidades para monitorear el comportamiento y los procesos de toma de decisiones de los modelos a lo largo de su ciclo de vida.

- Detección de novedad: Verificar la capacidad del modelo para detectar entradas que se encuentran fuera de su dominio de competencia o experiencia.

La auditoría algorítmica exige un conjunto de habilidades híbridas que combinan el conocimiento tecnológico profundo con la visión de gobernanza, riesgo y estrategia empresarial. En este sentido, el auditor moderno deberá desarrollar una serie de competencias claves para adelantar este nuevo tipo de auditorías, detalladas en la tabla No.3.

Riesgos algorítmicos. Repensar los marcos de seguridad y control

Los riesgos algorítmicos crean un escenario novedoso para las organizaciones, pues el reto no sólo es seguridad y privacidad por diseño, sino crear un entorno de confianza digital que ofrezca condiciones de ejecución confiables, y el adecuado tratamiento de la información, para asegurar la rendición de cuentas por uso y despliegue de agentes con IA.

Lo anterior pasa por temas como:

- Incorporar interruptores de parada y sistemas de verificación multi-agente.

- Asegurar un monitoreo continuo (24x7) y afinamiento del SIEM (*Security Information and Event Management*) para detectar anomalías en tiempo real y responder en segundos.

- Instalar un control de acceso estricto y múltiple factor de autenticación para todos los agentes y servicios.

- Incluir derechos de auditoría en contratos de proveedores de IA/SaaS (*Software as a Service*).

Tabla 3. Competencias claves para la auditoría algorítmica

Competencia	Requisitos clave para Auditoría Algorítmica	Aplicación Específica en IA
Fluidez técnica y de datos	Comprender en profundidad la tecnología y sus limitaciones.	• **Alfabetización en IA**: Entender LLMs, GenAI y la arquitectura de agentes. • **Explicabilidad de la AI**: Desplegar modelos interpretables para auditar lógicas de decisión opacas ("cajas negras"). • **Datos**: Asegurar gobernanza de datos, calidad y trazabilidad.
II. Gobernanza, Riesgo y Cumplimiento (GRC)	Asegurar marcos de control que se adapten a la toma de decisiones autónoma.	• **Gobernanza evolutiva**: Establecer controles para la trazabilidad, la observabilidad y la seguridad adaptativa. • **Rendición de cuentas**: Definir matrices de responsabilidad para agentes. • **Supervisión humana**: Requerir intervención humana en decisiones de alto riesgo.
III. Ciberseguridad y riesgos emergentes	Entender y anticipar el *crimen algorítmico* y amenazas disruptivas.	• **Inteligencia de amenazas basadas en IA**: Analizar riesgos de *deepfakes*, *phishing* generado por IA, y plataformas *Crime-as-a-Service*. • **Observabilidad**: Asegurar que las herramientas detecten anomalías conductuales de agentes. • **Switch de apagado**: Integrar desde el diseño el botón de apagado del agente.
IV. Habilidades humanas y estratégicas	Liderar la adaptabilidad y construir confianza frente a la incertidumbre.	• **Adaptabilidad:** Priorizar el aprendizaje continuo para igualar el ritmo de la amenaza. • **Anticipación:** Usar la planificación de escenarios para abordar riesgos sistémicos. • **Ética:** Actuar como "Arquitecto de la Confianza Digital", asegurando el uso ético y transparente de la IA.

Nota: Elaboración propia basado en: IIA, 2025; PwC, 2025

- Implementar módulos de IA Explicable y trazabilidad inmutable (registros de logs del procesos para llegar a los resultados no modificables) para asegurar la auditabilidad y el cumplimiento regulatorio.
- Simular ataques de *deepfake* y desinformación generada por IA en ejercicios de la junta directiva.
- Hacer una extensión del seguro cibernético para los posibles daños y afectaciones de los agentes desplegados.
- Promover formación en alfabetización digital sobre los agentes

de inteligencia artificial sus ventajas y limitaciones para el equipo directivo.

- Requerir la aprobación humana para decisiones de alto riesgo que generen los agentes de inteligencia artificial.

- Entender los componentes técnicos, cómo los modelos grandes de lenguaje (LLMs) y las arquitecturas de agentes perciben, orquestan y actúan de forma autónoma.

Como se puede observar, no es solamente situarse en los principios tradicionales de la seguridad, sino establecer una vista holística que integre la seguridad y la privacidad desde el diseño, la ética, la equidad y la explicabilidad, para abordar, no sólo la defensa del modelo de la IA y sus datos, sino el impacto social y la confiabilidad a largo plazo. En pocas palabras, se debe configurar una "confianza por diseño - CpD" que asegure la rendición de cuentas de la organización en el despliegue de sistemas complejos (ahora basados en agentes con IA) y la supervisión vigilante de su desempeño ético (Welle, 2025).

El CpD no sustituye a la seguridad por diseño o la privacidad por diseño, sino que actúa como un marco de gobernanza superior que los engloba y los extiende para abordar los desafíos de la IA.

A continuación un resumen consolidado de las tres perspectivas actuales alrededor de la seguridad y control con sus definiciones y limitaciones (Ver tabla Tabla 4).

Tabla 4. Perspectivas actuales de seguridad y control

Perspectiva	Definición	Limitaciones
Seguridad por diseño (SpD)	Protección de la infraestructura y el sistema contra amenazas. Integración de controles técnicos (cifrado, autenticación, etc).	No aborda el impacto social de un sistema técnicamente confiable. Un algoritmo puede ser confiable pero éticamente sesgado.
Privacidad por diseño (PpD)	Protección del dato personal. Se enfoca en los derechos del interesado sobre su información.	Es insuficiente cuando el riesgo proviene del uso del modelo (ej. decisiones discriminatorias) y no del dato en sí, especialmente si usa datos anónimos o no personales.
Confianza por diseño (CpD)	Enfoque holístico. Integra SpD, PpD, ética, equidad y explicabilidad. Aborda el impacto social y la confiabilidad a largo plazo.	Asegurar una vista integrada del reto de la implementación de la inteligencia artificial a nivel empresarial que incluya la alineación con el negocio, el apetito de riesgo cibernético empresarial y sus impactos (oportunidades y amenazas)

Nota: Elaboración propia con ideas: Behbahani, 2025

Conclusiones

La IA ya no es una promesa tecnológica, sino un motor de cambio que impulsa la automatización de flujos de trabajo complejos y la toma de decisiones. Dado que los sistemas de IA, especialmente la IA Agéntica, son capaces de tomar decisiones y ejecutar acciones de forma autónoma, y su complejidad puede hacerlos "cajas negras", la función de auditoría interna debe evolucionar para asegurar la confiabilidad, la transparencia y la rendición de cuentas tanto para la empresa como para el ecosistema digital donde opera.

En este sentido, el riesgo algorítmico se configura como un acelerador significativo de amenazas cibernéticas que enfrentan las organizaciones hoy, que hace evidente su característica sistémica.

La rápida adopción de la IA Generativa (IA Gen) y los agentes autónomos (IA Agéntica) está redefiniendo el panorama de riesgos, llevando la ciberseguridad de una comprensión exclusivamente técnica a un imperativo estratégico que afecta la estabilidad financiera, la reputación y la continuidad del negocio.

El marco tradicional de seguridad basado en confidencialidad, integridad y disponibilidad (CID) se considera incompleto y limitado para los desafíos modernos. Parker (1998) propone un nuevo marco de análisis que extiende los elementos mencionados de seguridad y control incluyendo ahora la utilidad, la posesión y la autenticidad, los cuales, para el momento actual con la incorporación de la IA Gen y sistema autónomos, resultan del mayor interés la combinación entre integridad y autenticidad.

Los sistemas de IA Generativa y los Agentes de IA son inherentemente no deterministas, lo que significa que la misma *entrada* puede generar una variedad de posibles *salidas*, haciéndolos difíciles de gestionar y vulnerables a errores. Los principales riesgos que amenazan directamente la autenticidad y la integridad incluyen: alucinaciones, inyección de *prompts*, envenenamiento de datos, divulgación de datos sensibles y degradación del modelo, los cuales hacen evidente el reto de la integridad semántica que no le permite al agente interpretar la información de manera correcta, válida y conforme a la realidad.

En resumen, es necesario gestionar el riesgo algorítmico en una realidad cambiante como la actual, donde los sistemas de IA configuran un ecosistema dinámico sujeto a degradación intrínseca y ataques continuos que explotan la falta de separación entre instrucciones y datos. Por tanto, la seguridad no es solo una característica que se añade al final de una iniciativa inteligente, sino una arquitectura que se elige desde el principio, ya que el enfoque de ser "rápido" e "inteli-

gente" sin verificación de la información (Raghavan & Schneier, 2025), sacrifica inherentemente la seguridad y la privacidad, afectando en el mediano y largo plazo el reto real de una empresa en el contexto digital con su cliente: la "confianza por diseño".

Referencias

Behbahani, A. (2025). *Why Trust Cannot Be an Afterthought*. TÜV SÜD. https://www.tuvsud.com/en-us/resource-centre/blogs/business-assurance/trust-by-design-the-value-driven-route-to-trusted-and-certified-ai-with-iso-iec-42001

Bone, J. & Lee, J. (2023). *Cognitive risk*. Boca Raton, Fl. USA. CRC Press.

Cankaya, E. C. (2011). *Bell-LaPadula confidentiality model.* En Encyclopedia of Cryptography and Security (pp. 71–74). Springer US.

Chen, K., Zhou, X., Lin, Y., Su, J., Yu, Y., Shen, L., & Lin, F. (2025). *A survey on data security in large Language Models.* En arXiv [cs.CR]. https://doi.org/10.48550/ARXIV.2508.02312

Díaz, S., Kern, C. & Olive, K. (2025). *Google's Approach for Secure AI Agents.* Google Research. https://research.google/pubs/an-introduction-to-googles-approach-for-secure-ai-agents

Godhrawala, A. (2025). *How Agentic AI can transform industries by 2028.* EY Agentic AI Series. https://www.ey.com/en_in/insights/ai/how-agentic-ai-can-transform-industries-by-2028

Harrison, M., Ruzzo, W. & Ullman, J. (1976). *Protection in operating systems*. Communications of ACM. *19*(8). https://doi.org/10.1145/360303.360333

IIA (2025). *Risk in focus. Hot topics for internal auditor 2026.* IIA. https://www.theiia.org/globalassets/site/foundation/latest-research-and-products/risk-in-focus/2026/2026-global-report-en-riskinfocus.pdf?v=0925202501

Janapa Reddi, V. (2025). *Introduction to machine learning systems. Principles and Practices of Engineering Artificially Intelligent Systems*. School of Engineering and Applied Sciences. Harvard University. https://www.mlsysbook.ai/assets/downloads/Machine-Learning-Systems.pdf

Mckinsey (2025). Agentic AI security: Risks & governance for enterprises. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/deploying-agentic-ai-with-safety-and-security-a-playbook-for-technology-leaders

Parker, D. (1998). *Fighting computer crime. A New Framework for Protecting Information. New York*, NY. USA: John Wiley & Sons.

PwC (2025). Global digital trust insights 2026. https://www.pwc.com/us/en/services/consulting/cybersecurity-risk-regulatory/library/global-digital-trust-insights.html

Raghavan, B., & Schneier, B. (2025). *Agentic AI's OODA Loop Problem*. IEEE security & privacy, 2–4. https://doi.org/10.1109/msec.2025.3604105

Welle, J. (2025). *From compliance to competitive advantage: Building trust by design*. Capgemini.

https://www.capgemini.com/insights/expert-perspectives/from-compliance-to-competitive-advantage-building-trust-by-design

Jeimy J. Cano M., Ph.D, CFE, CICA.
Ingeniero y Magíster en Ingeniería de Sistemas y Computación, Universidad de los Andes. Especialista en Derecho Disciplinario, Universidad Externado de Colombia; Ph.D en Business Administration, Newport University, CA. USA. y Ph.D en Educación, Universidad Santo Tomás. Profesional certificado como Certified Fraud Examiner (CFE), por la Association of Certified Fraud Examiners y Certified Internal Control Auditor (CICA) por The Institute of Internal Controls. Profesor Distinguido de la Facultad de Derecho, Universidad de los Andes. Es director de la Revista SISTEMAS de la Asociación Colombiana de Ingenieros de Sistemas –ACIS–.