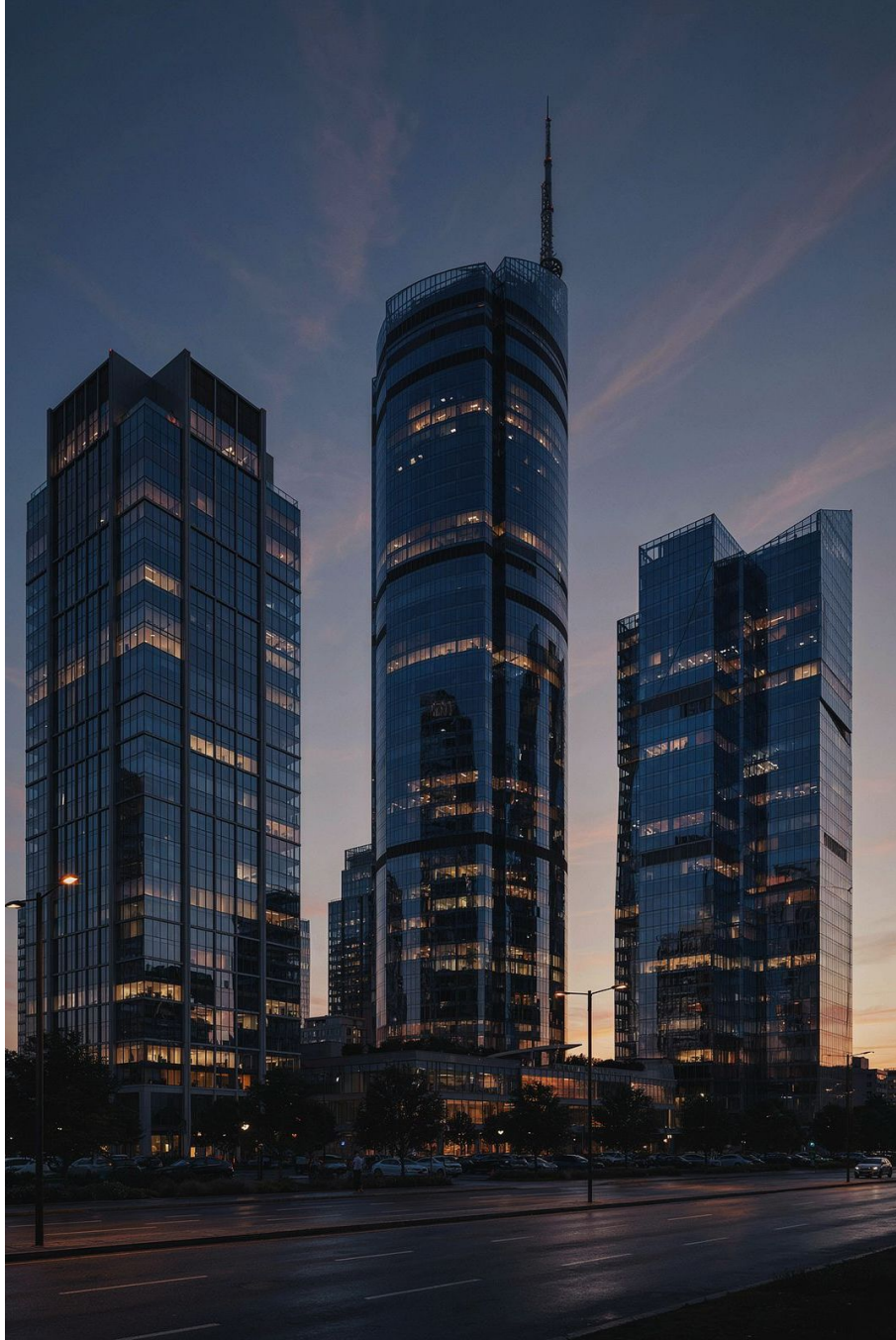
Gestión del Riesgo en Soluciones de Inteligencia Artificial y Machine Learning

8.º Seminario Latinoamericano en Gestión del Riesgo Operacional — **ALGERisk Corporation**

ECON. YURI VLADIMIR LUNA - GERENTE DE CONSULTORIA GESTION DE RIESGOS NO FINANCIEROS

PRIME CONSULTING INTERNATIONAL LLC

Risk. Governance. Continuity. Compliance.



La IA ya está en el corazón de las operaciones

Las soluciones de inteligencia artificial participan hoy en una amplia variedad de procesos financieros, desde la originación de créditos hasta la supervisión regulatoria.

Decisiones Crediticias

Evaluación crediticia, originación, scoring, gestión de cobranzas y decisiones de inversión.

Seguridad y Fraude

Prevención de fraude, monitoreo de transacciones, detección de operaciones inusuales y ciberseguridad.

Operaciones y Clientes

Segmentación de clientes, asistentes virtuales, análisis documental y procesamiento de siniestros.

Control y Supervisión

Automatización de controles, generación de reportes y supervisión regulatoria continua.

La pregunta central de esta conferencia

¿Cómo pueden las instituciones financieras utilizar la inteligencia artificial sin perder el control sobre sus decisiones, sus datos, sus procesos y sus responsabilidades?

Principales factores de riesgo:

Deriva del modelo

Una solución puede mostrar excelentes resultados durante el piloto y deteriorarse cuando cambian los clientes, los datos o el entorno económico.

Opacidad

Un modelo puede ser técnicamente preciso y, al mismo tiempo, producir resultados que la institución no puede explicar.

Pérdida de autonomía

Un agente puede comenzar recomendando acciones y, tras varias integraciones, terminar ejecutándolas sin supervisión efectiva.

La IA como sistema multi-dimensional

Gestionar el riesgo de inteligencia artificial no consiste únicamente en validar un algoritmo. Consiste en ***gobernar un sistema*** que integra múltiples dimensiones interdependientes.

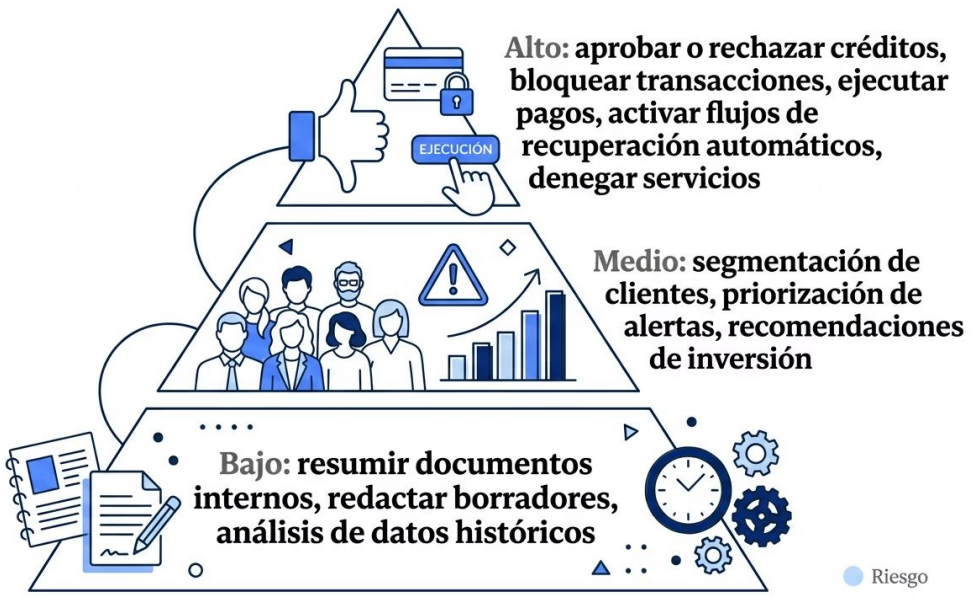
No toda inteligencia artificial presenta la misma exposición al riesgo

El primer error es gestionar todas las soluciones de IA con los mismos requisitos. La intensidad del gobierno debe responder al riesgo real y a la materialidad del caso de uso.

Un IA Generativa que responde a preguntas de Clientes en un ChatBot	**BAJO**
Un IA Generativa que clasifica incidentes de TI y de seguridad	**MEDIO**
Un IA Generativa que analiza solicitudes de créditos y aprueba	**ALTO**

El espectro de impacto determina los controles

En función a la exposición al riesgo y/o la criticidad del mismo, se debe evaluar qué mitigantes se debe incorporar



El principio rector: los controles deben ser proporcionales al impacto potencial, no a la popularidad de la tecnología.

Criterios de materialidad

Para determinar la materialidad de un caso de uso, la institución debe evaluar sistemáticamente los siguientes factores antes de definir el nivel de gobierno requerido.

Impacto y autonomía

- Criticidad del proceso
- Impacto sobre clientes
- Nivel de autonomía del sistema
- Reversibilidad de los resultados

Datos y regulación

- Sensibilidad de los datos
- Relevancia y exposición regulatoria
- Posibilidad de discriminación
- Dependencia de terceros

Consecuencias

- Consecuencias financieras
- Volumen de operaciones
- Grado de explicabilidad
- Impacto sobre servicios críticos
- Reputacionales

Cuatro conceptos que hay que diferenciar

Antes de evaluar riesgos, es necesario distinguir con precisión entre:

- Inteligencia artificial,
- Machine learning,
- IA generativa y
- Agentes de IA.

Cada uno presenta un perfil de riesgo diferente.

IA, ML, IA Generativa y Agentes

Inteligencia Artificial

Concepto amplio. Sistemas capaces de generar predicciones, recomendaciones, computer vision, videos, contenido o decisiones para determinados objetivos.

Machine Learning

Técnicas mediante las cuales un modelo identifica patrones a partir de datos y los utiliza para producir predicciones o clasificaciones.

IA Generativa

Produce contenido nuevo: texto, imágenes, código, resúmenes, análisis. Su comportamiento puede ser menos determinista y puede inventar información con apariencia convincente.

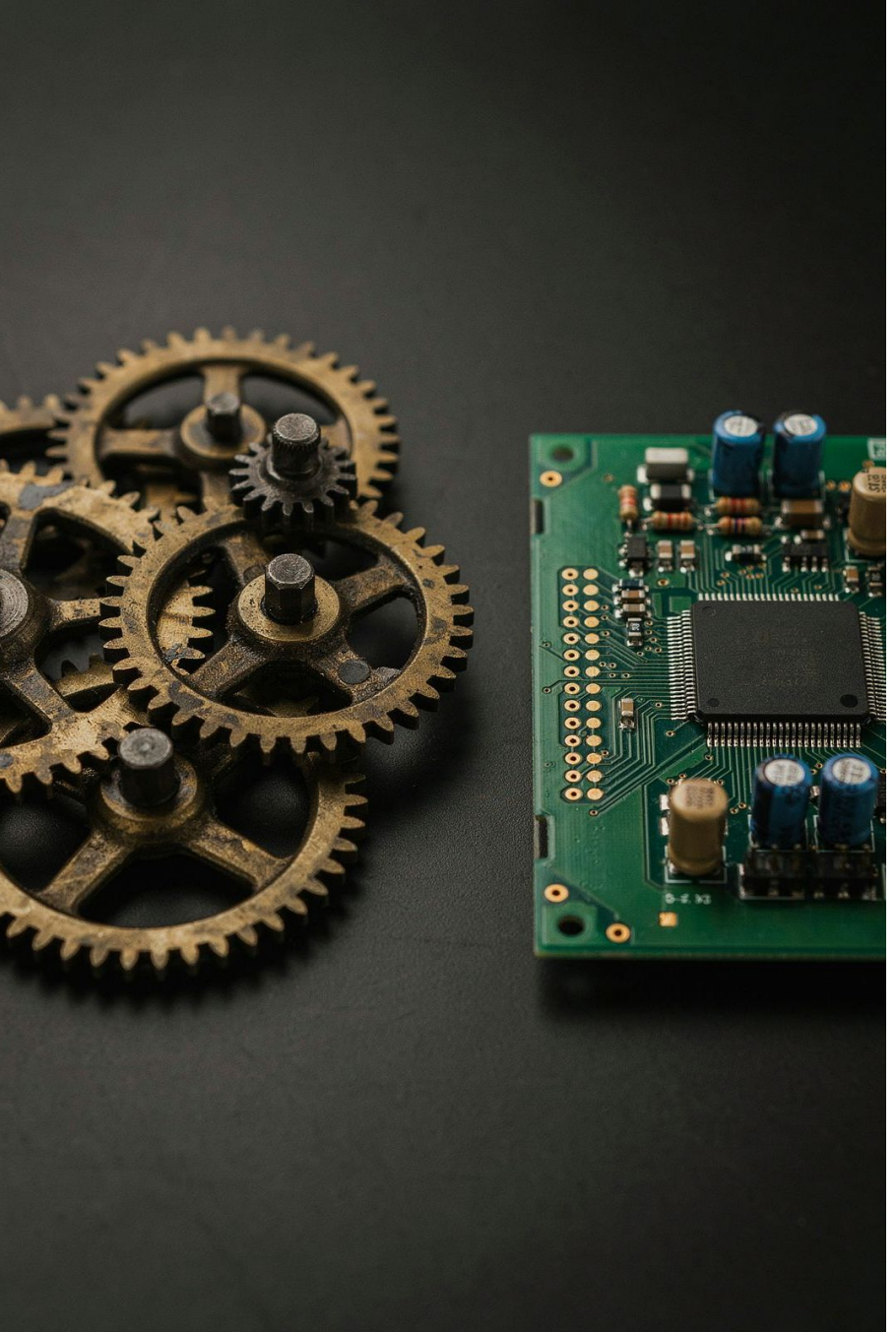
Agentes de IA

Planifican tareas, consultan fuentes, acceden a sistemas, ejecutan acciones y coordinan otros agentes. Cuando aumenta la autonomía, también aumenta la exposición.

¿Hasta dónde puede actuar la solución?

No basta con registrar que la organización "utiliza inteligencia artificial". Debe conocerse qué tipo de capacidad emplea y hasta dónde puede actuar de manera autónoma.

Cada nivel de autonomía requiere un nivel superior de gobierno, controles y supervisión humana efectiva.



¿Por qué la IA no puede gestionarse como software tradicional?

En el software tradicional, la organización programa reglas explícitas.

En machine learning, el comportamiento emerge de patrones aprendidos. Esto introduce fuentes de variabilidad que los controles tradicionales no cubren.

Sistemas deterministas (soft tradicional) vs Sistemas probabilísticos (IA)

Nuevas fuentes de riesgo

¿Cuándo puede cambiar el resultado?

- Cambian los datos de entrada
- Cambia el comportamiento del cliente
- Se modifica el entorno económico
- Aparecen casos no representados en el entrenamiento
- El proveedor actualiza el modelo
- Se ajustan los parámetros o se incorporan nuevas fuentes

Controles adicionales necesarios

- Gobierno del dato
- Gestión del riesgo de modelo
- Supervisión humana efectiva
- Monitoreo continuo de deriva
- Gestión de proveedores
- Trazabilidad completa
- Pruebas adversariales
- Mecanismos de interrupción
- Auditoría algorítmica

Los controles tradicionales de desarrollo y pruebas son necesarios, pero insuficientes para gobernar sistemas de IA.

El gobierno de IA comienza por el inventario de soluciones

Una institución no puede gestionar lo que no conoce. El inventario IA debe abarcar soluciones desarrolladas internamente, adquiridas a terceros, e incorporadas dentro de otros procesos o productos sin que la organización lo reconozca explícitamente.

Una vez inventariadas, se requiere clasificarlas y determinar criticidad.

El riesgo de la IA invisible (Shadow AI)

Una categoría especialmente crítica es la IA que se incorpora de forma silenciosa a través de actualizaciones de proveedores o funcionalidades embebidas en plataformas existentes.

También esta el uso de IA que los usuarios ya pueden estar incorporando en sus procesos diarios sin la autorización de la empresa.

Actualizaciones no reconocidas

Un proveedor puede agregar una funcionalidad de IA mediante una actualización sin que la organización la identifique como un nuevo caso de uso.

Herramientas embebidas

Una herramienta ofimática puede incorporar asistentes inteligentes. Una solución antifraude puede modificar su motor analítico silenciosamente.

APIs y plataformas cloud

Un proveedor cloud puede habilitar capacidades adicionales de IA mediante APIs que amplían el alcance de la solución sin aprobación formal.

Gobierno IA y responsabilidades

La gestión efectiva del riesgo de IA requiere responsabilidades claramente definidas en todos los niveles de la organización.

La función de Riesgos no debe convertirse en propietaria de los riesgos de IA.

El Comité de Riesgos debe incorporar el análisis y evaluación de los riesgos derivados de la implementación de soluciones IA.

Responsabilidades por nivel

1

Directorio y Alta Gerencia

El Directorio comprende los casos materiales, aprueba principios de uso y supervisa exposiciones agregadas.

La Alta Gerencia convierte los principios en recursos, estructuras y límites ejecutables.

2

Propietario del proceso y Tecnología

El propietario responde por la finalidad del caso de uso y sus consecuencias.

Tecnología responde por arquitectura, infraestructura, integraciones y gestión de cambios.

3

Ciencia de Datos y Riesgo de Modelo

Ciencia de Datos diseña, entrena, documenta y mantiene el modelo.

Riesgo de Modelo evalúa metodología, desempeño, supuestos, limitaciones y explicabilidad de forma independiente.

4

Riesgo Operacional, Seguridad y Datos

Riesgo Operacional evalúa escenarios de falla, vulnerabilidades e impactos.

Seguridad protege identidades, accesos e infraestructura.

Gobierno del Dato asegura calidad, linaje y representatividad.

Funciones especializadas de control

Privacidad, Legal y Cumplimiento

Evalúan finalidad, consentimiento, transparencia, protección del cliente, riesgo de discriminación y obligaciones regulatorias aplicables a cada caso de uso.

Seguridad de Informacion

Gestión de identidades y accesos, Protección de datos de información, Desarrollo seguro y MLOps, Seguridad de APIs e integraciones, Controles específicos para IA generativa.

Ciberseguridad

Protección del pipeline MLOps, Gestión de identidades y privilegios, Gestión segura de claves API, Seguridad del algoritmo y dependencias, Protección de datos de entrenamiento y producción.

Continuidad del Negocio

Asegura que el proceso pueda mantenerse o recuperarse cuando el modelo, la infraestructura o el proveedor no estén disponibles.

Auditoría Interna

Proporciona aseguramiento independiente sobre gobierno, inventario, controles, validaciones y cumplimiento. Es la última línea de defensa para confirmar que el sistema funciona según principios definidos.

Auditoría debe verificar

- Existencia de una política institucional de IA.
- Aprobación por el órgano competente.
- Definición del apetito y tolerancias.
- Responsabilidades del Directorio y la Alta Gerencia.
- Comités y mecanismos de escalamiento.
- Segregación entre desarrollo, validación, aprobación y operación.
- Integración de la IA con riesgo operacional, tecnológico, de modelo, seguridad, privacidad y cumplimiento.

Principio clave: El propietario del proceso siempre continúa siendo responsable por la utilización y sus resultados.

Principales escenarios de riesgo

No debe registrarse un riesgo genérico denominado "fallas de inteligencia artificial". Deben formularse escenarios concretos, evaluados con base en su probabilidad, impacto y capacidad de detección.

Ejemplo de escenario de riesgo de IA

Proceso: Otorgamiento de créditos
Solución: Modelo de machine learning para calificación crediticia

Escenario de riesgo:

> Aprobación o rechazo incorrecto de solicitudes de crédito debido al deterioro no detectado del modelo, datos de entrada incompletos o sesgados y ausencia de supervisión humana efectiva, generando pérdidas financieras, trato inequitativo a clientes, incumplimientos regulatorios y daño reputacional.

Vulnerabilidades

Deterioro del desempeño

El modelo produce resultados incorrectos, inconsistentes o menos precisos de lo esperado. *Ejemplo: un modelo de fraude genera demasiados falsos positivos y bloquea operaciones legítimas.*

Cambio de Concepto

La relación entre los datos y el resultado cambia con el tiempo. Un modelo desarrollado en período estable puede deteriorarse cuando cambian empleo, inflación o comportamiento de pago.

Deterioro de calidad de datos e integridad

Los datos pueden ser incompletos, incorrectos, desactualizados, no representativos, obtenidos sin autorización o utilizados para una finalidad diferente. La calidad del modelo no puede superar sostenidamente la calidad del dato que lo alimenta.

Sesgo algorítmico

El modelo puede producir resultados diferenciados entre grupos aunque la variable sensible no se use directamente. Variables aparentemente neutrales pueden actuar como aproximaciones de ubicación, edad, género u origen.

Eventos operativos, de seguridad y de terceros

Explicabilidad

La institución puede no ser capaz de explicar por qué se produjo un resultado, qué variables influyeron o qué supervisión humana existió.

En procesos materiales, esto genera posibilidad de incumplimiento y posible daño de reputación.

Alucinación

En IA generativa, el sistema puede inventar hechos, referencias u obligaciones.

El problema no es solo que se equivoque: puede hacerlo usando un lenguaje altamente convincente.

Seguridad

La solución puede estar expuesta a accesos no autorizados, manipulación de datos, alteración de instrucciones, fuga de información e integraciones vulnerables.

Terceros y concentración

La institución puede depender de modelos, datos e infraestructura controlados por un proveedor, limitando transparencia, auditoría y continuidad.

Una falla sistémica puede generar impactos simultáneos en múltiples instituciones.

Legal y regulatorio

Incumplimientos en protección de datos, transparencia, protección del consumidor, prevención de discriminación y gestión de externalización.

Continuidad

La entidad puede perder capacidad operativa cuando el modelo falla, la API se interrumpe o el personal ya no sabe ejecutar el proceso manual.

Caso práctico

Una entidad implementa un asistente para analizar normativa, contratos y reportes de cumplimiento. El asistente resume documentos y propone conclusiones.

Con el tiempo, los usuarios comienzan a copiar directamente las respuestas en informes regulatorios.

IA generativa en la Gestión del Cumplimiento

Eventos de Riesgos IA identificados

- Interpretaciones incorrectas presentadas con confianza
- Referencias inexistentes en normativa
- Uso de normativa desactualizada
- Pérdida de confidencialidad documental
- Falta de evidencia sobre las fuentes
- Dependencia excesiva del usuario
- Ausencia de revisión independiente

Controles mínimos recomendados

- Fuentes autorizadas y actualizadas
- Recuperación con referencias verificables
- Separación entre borrador y aprobación
- Revisión obligatoria por personal competente
- Restricción de datos sensibles en el prompt
- Registro de instrucciones y respuestas
- Prohibición de envío automático al supervisor
- Monitoreo de errores materiales

Agentes GPT con capacidad de ejecución: controles obligatorios

Cuando un Agente GPT pasa de recomendar a ejecutar procesos, los controles deben aumentar proporcionalmente. La autonomía sin límites verificables representa el escenario de mayor exposición.

Identidad y privilegio

Identidad propia del agente, mínimo privilegio, lista explícita de acciones permitidas.

Supervisión humana

Separación entre recomendación y ejecución. Aprobación humana para acciones materiales con límites de monto, alcance y frecuencia.

Kill switch y reversión

Mecanismo de interrupción probado, procedimiento manual de contingencia documentado y capacidad de reversión ante errores.

Gestión del riesgo durante el ciclo de vida de Soluciones IA

La evaluación de riesgo no debe realizarse una sola vez al inicio. Debe acompañar todo el ciclo de vida del modelo, desde la definición del caso de uso hasta el retiro controlado de la solución.

Las ocho etapas del ciclo de vida

1. Definición

Aprobación del caso de uso y evaluación inicial de impacto.

2. Adquisición

Requisitos de riesgo y control incorporados desde el diseño.

3. Datos

Aprobación formal del dataset y documentación de transformaciones.

4. Entrenamiento

Reproducibilidad: documentar parámetros, variables, supuestos y decisiones.

5. Validación

Prohibición de pasar a producción sin aprobación o excepción formal.

6. Implementación

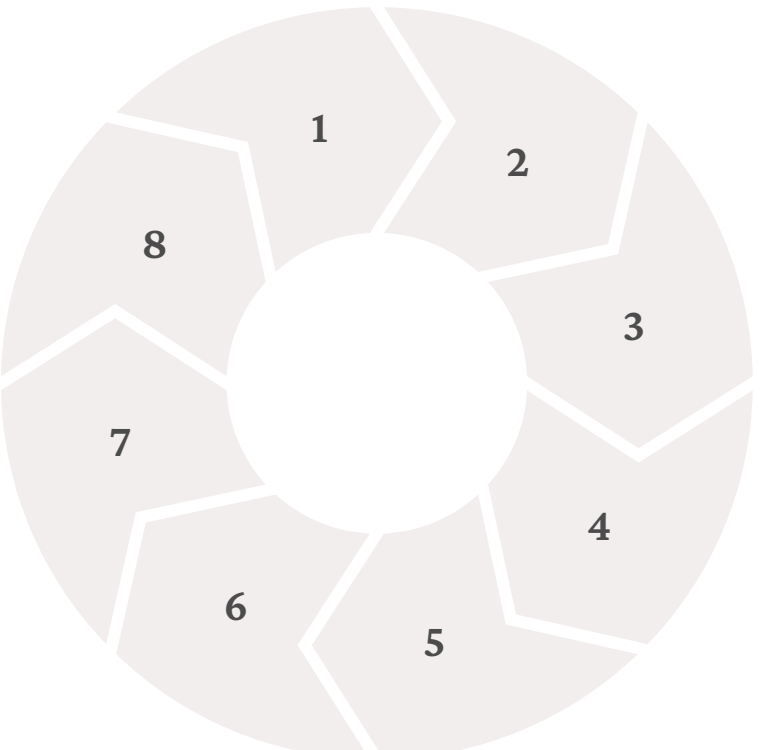
Autorización de producción con criterios de aceptación definidos.

7. Operación

Umbrales de desempeño y acciones asociadas al monitoreo continuo.

8. Retiro

Estrategia de salida, cierre controlado y conservación de evidencia.



Monitoreo continuo en operación

La etapa de operación es donde más frecuentemente fallan los controles. El monitoreo debe ser continuo, automatizado en lo posible, y con umbrales que activen respuestas concretas.

Desempeño del modelo

Exactitud, falsos positivos, falsos negativos, deriva estadística y comparación con modelo de referencia.

Calidad de datos

Cambios en distribución de variables de entrada, nuevas fuentes, datos faltantes y representatividad.

Comportamiento del usuario

Incidentes, excepciones, quejas de clientes, intervenciones humanas y señales de automatización por complacencia.

Validación independiente

Una validación efectiva no consiste únicamente en repetir las métricas del desarrollador. Tampoco que el área de desarrolle valide lo que ellos han diseñado.

La validación debe ser independiente (tercero o Auditioría). Debe responder preguntas y validar funciones que el propio equipo de desarrollo puede no estar incentivado a formular.

Dimensiones de la validación independiente

Preguntas que debe responder

- ¿El modelo es adecuado para el propósito declarado?
- ¿Los datos representan la población real de aplicación?
- ¿Cómo se comporta por segmentos de clientes?
- ¿Qué ocurre fuera de las condiciones normales?
- ¿Puede explicar sus resultados de forma comprensible?
- ¿Existe riesgo de uso fuera del alcance aprobado?
- ¿Los controles compensatorios son efectivos?

Autoridad de la validación

La función validadora debe tener autoridad formal para:

- Aprobar sin condiciones
- Aprobar con condiciones documentadas
- Solicitar ajustes al modelo
- Restringir el alcance de uso
- Rechazar el modelo en casos de alto impacto
- Recomendar incluso la suspensión

Supervisión humana efectiva

Muchas instituciones utilizan la expresión "human in the loop" como si fuera suficiente. Pero la supervisión humana puede ser únicamente formal: aprobar cientos de recomendaciones sin tiempo, información ni autoridad para cuestionarlas.

¿Qué hace real a la supervisión humana?

La pregunta no es si existe una aprobación humana. La pregunta es: **¿puede esa persona detectar, cuestionar y detener una decisión incorrecta?**

Comprensión y capacitación

La persona debe comprender la finalidad del modelo, conocer sus limitaciones y haber recibido capacitación específica.

Información y tiempo

Debe disponer de información suficiente sobre el caso y tiempo real para revisar, no solo aprobar de forma mecánica.

Autoridad y protección

Debe poder rechazar el resultado, registrar el motivo, escalar situaciones dudosas y no ser penalizada por cuestionar al sistema.

Monitoreo de la supervisión

La institución debe monitorear si la supervisión humana es real o si se está produciendo automatización por complacencia.

Controles mínimos institucionales

Una institución financiera debería considerar como mínimo los siguientes controles, adaptados al tamaño, complejidad y materialidad de cada caso.

Gobierno

- Política institucional de IA
- Inventario completo de casos de uso
- Clasificación por riesgo y materialidad
- Propietario formal asignado

Control técnico

- Documentación técnica y funcional
- Validación independiente
- Gestión de accesos y segregación
- Pruebas de seguridad adversariales

Operación continua

- Supervisión humana proporcional
- Monitoreo de desempeño y deriva
- Registro de decisiones y acciones
- Gestión formal de cambios

Resiliencia y salida

- Plan de continuidad del proceso
- Mecanismo de interrupción probado
- Estrategia de salida documentada
- Revisión periódica del riesgo residual

Indicadores para el Comité de Riesgos

El Comité necesita una visión agregada, no métricas aisladas de Tecnología. Los indicadores deben relacionarse con clientes, productos, procesos críticos, pérdidas, cumplimiento y apetito de riesgo.

Dashboard de riesgo de IA para el Comité

Inventario y cobertura

- Casos identificados con brechas vs. aprobados
- % de procesos críticos con Soluciones IA
- Modelos sin propietario formal
- Casos sin evaluación de riesgo vigente

Validación y desempeño

- Modelos sin validación independiente
- Modelos con deterioro sobre umbral
- Cambios no evaluados formalmente
- Excepciones de validación vencidas

Incidentes y decisiones

- Incidentes y eventos cercanos de IA
- Decisiones materiales con intervención humana
- Quejas relacionadas con decisiones automatizadas
- Casos sin alternativa manual probada

Agentes y riesgo residual

- Agentes con capacidad de ejecución con brechas vigentes
- Agentes sin kill switch probado
- Concentración por proveedor de IA
- Riesgos residuales fuera del apetito

Gestión de incidentes de IA

La institución necesita reconocer y registrar los incidentes de IA como una categoría diferenciable, con su propio proceso de detección, contención y aprendizaje.

¿Qué es un incidente de IA?

- Deterioro significativo del desempeño
- Decisión material incorrecta
- Sesgo no previsto identificado
- Fuga o uso no autorizado de información
- Acción ejecutada fuera de los límites
- Imposibilidad de explicar una decisión
- Cambio del proveedor no evaluado

Respuesta al incidente

- Detección, clasificación y contención
- Suspensión cuando corresponda
- Activación del proceso alternativo
- Preservación de evidencia
- Notificación regulatoria si aplica
- Análisis de causa raíz
- Revalidación y lecciones aprendidas

El incidente no debe cerrarse únicamente porque la solución volvió a funcionar. Debe determinarse si el riesgo residual continúa siendo aceptable.

Conclusiones ejecutivas

La responsabilidad es compartida

No pertenece exclusivamente a Tecnología ni a Ciencia de Datos. El propietario del proceso responde por la finalidad y la utilización. Las funciones especializadas aportan control, supervisión y aseguramiento.

Toda solución material necesita continuidad y capacidad de interrupción

La trazabilidad completa y el mecanismo de kill switch no son técnicos opcionales: son requisitos de gobierno.

Evalúe también lo que ocurre cuando falla

La inteligencia artificial no debe evaluarse únicamente por lo que puede hacer cuando funciona correctamente. También debe evaluarse por lo que puede ocasionar cuando se equivoca, se deteriora, es manipulada o actúa fuera de los límites previstos.

Conclusiones ejecutivas

La IA es una capacidad empresarial, no solo tecnológica

Debe gobernarse como un sistema que puede modificar decisiones, controles y procesos críticos.

La materialidad determina los controles

No todas las soluciones necesitan los mismos controles. La autonomía e impacto deben determinar la intensidad del gobierno.

La evaluación no termina con la implementación

Los modelos se deterioran, los datos cambian y los usuarios amplían el alcance de forma no autorizada.

Validación independiente y monitoreo son controles esenciales

No opcionales: son la base del gobierno efectivo de cualquier modelo material.

Marcos de referencia clave

ISO/IEC 42001:2023

Establece requisitos para implementar y mejorar un sistema de gestión de IA. Incorpora gobierno, políticas, responsabilidades, evaluación, control y mejora continua.

ISO/IEC 23894:2023

Orientación para integrar la gestión del riesgo de IA en las actividades y funciones de la organización de manera coherente con otros marcos de riesgo.

NIST AIRMF

Organiza la gestión alrededor de cuatro funciones: Govern, Map, Measure y Manage. Busca incorporar atributos de confianza durante el diseño, desarrollo, uso y evaluación.

FSB — Junio 2026

Doce prácticas para la adopción responsable de IA por instituciones financieras: gobierno institucional, gestión en el ciclo de vida, y riesgos cibernéticos, tecnológicos y de terceros.

Una inteligencia artificial confiable no es aquella que nunca se equivoca. Es aquella cuyos límites, riesgos y decisiones pueden ser comprendidos, monitoreados y controlados.

Mucgas gracias

Econ. Yuri Luna C.

Gerente de Consultoría de Riesgos No Financieros