rgp.

Perspectivas del seminario web de julio

Navegando el riesgo en la era de la IA.

Resumen ejecutivo

A medida que los sistemas de IA se integran a una velocidad vertiginosa en operaciones críticas, la conversación sobre gobernanza ha pasado de ser opcional a esencial. Con la IA llegan nuevos retos: dilemas éticos, incertidumbres operativas y una creciente presión regulatoria.

Las empresas se encuentran ahora en un punto de inflexión crucial. La tarea que se avecina no es simplemente escalar la IA, sino hacerlo de forma responsable, incluso cuando las reglas globales aún están en constante cambio. La gobernanza eficaz de la IA debe evolucionar hacia un marco dinámico que permita la innovación mientras proteja frente a consecuencias no deseadas.

Marcos de gobernanza sólidos hacen que los sistemas de IA sean más explicables para los reguladores, más fiables para los socios y más valiosos para los usuarios finales. Fomentan la confianza, un requisito previo para una adopción generalizada, y reducen la exposición legal mientras mejoran el rendimiento del modelo. Lo más importante es que, estos marcos, crean las condiciones para una ventaja competitiva sostenible.

Los riesgos sistémicos únicos de la IA

Gobernar la IA es especialmente complejo porque la IA en sí misma se comporta de forma diferente al software tradicional. A diferencia de los sistemas heredados que siguen reglas explícitamente definidas, la IA, especialmente los grandes modelos de lenguaje y otras formas de aprendizaje automático, generan resultados de forma probabilística, basándose en patrones en vastos conjuntos de datos. Esto los hace poderosos y flexibles, pero también impredecibles y opacos.

Esa imprevisibilidad introduce riesgos sistémicos: desafíos difíciles de detectar, difíciles de modelar y capaces de extenderse en cascada a toda la organización. Cuanto más profundamente está integrada la IA en funciones esenciales, más exigen estos riesgos una atención estratégica seria.

Los pilares de la IA confiable

Los Cinco Pilares de la IA confiable han surgido como un consenso global en el ámbito académico y industrial, y organismos gubernamentales. Estos principios pueden convertirse en la referencia para cada política posterior y revisión. Es mejor considerarlos como un estándar compartido y en evolución, más que como un marco propietario. Estos pilares reflejan un amplio acuerdo interdisciplinario sobre los requisitos esenciales para construir sistemas de IA en los que la gente y los reguladores puedan confiar. En otras palabras, ofrecen los principios fundamentales que cualquier programa de gobernanza de IA debe abordar:

1. **Explicabilidad:** Los resultados deben ser lo suficientemente claros para que los expertos del sector puedan verificarlos. "Si el resultado no puede explicarse, no se puede confiar", enfatizó.

2. **Equidad:** Los modelos deben ser evaluados para asegurar que no perpetúan discriminación ni prejuicios.

3. **Transparencia:** Las organizaciones necesitan visibilidad sobre los datos de entrenamiento y la lógica de sus modelos.

4. **Robustez:** Los sistemas deben estar envueltos en un entorno de control seguro capaz de soportar manipulaciones y ataques.

5. **Privacidad:** La información sensible debe protegerse contra filtraciones o mal uso.

Aquí tienes un repaso a algunos de los mayores riesgos:

Opacidad y falta de explicabilidad
Muchos sistemas modernos de IA funcionan como cajas negras: ofrecen resultados estadísticamente precisos sin ofrecer una visión clara de cómo se toman las decisiones. Esta opacidad es especialmente problemática en sectores como la salud, la justicia penal y las finanzas, donde los reguladores y otros actores exigen transparencia y rendición de cuentas.

Violaciones de privacidad y filtración de datos
Los modelos de IA, especialmente los grandes modelos generativos, memorizan partes de sus datos de entrenamiento, lo que podría exponer información personal sensible como historiales médicos, detalles financieros o nombres. Esto plantea serias preocupaciones regulatorias, especialmente bajo leyes de privacidad como el Reglamento General de Protección de Datos (RGPD) promulgado por la Unión Europea en 2018, que regula cómo deben gestionarse los datos personales, y la Ley de Privacidad del Consumidor de California (CCPA), una ley estatal de privacidad a nivel estadounidense promulgada en California en 2020 que otorga a los consumidores derechos respecto a la información personal que las empresas recopilan sobre ellos.

Sesgo y discriminación
Los sistemas de IA reflejan y perpetúan frecuentemente los sesgos presentes en datos de entrenamiento históricos o recopilados en la web. Estos sesgos pueden manifestarse de formas que perjudican desproporcionadamente a grupos específicos, como mujeres, minorías o personas de entornos socioeconómicos más bajos, especialmente en ámbitos como la contratación, el préstamo o la aplicación de la ley.

Degradación y fragilidad del modelo
Los sistemas de IA pueden degradarse con el tiempo a medida que el entorno externo cambia, debido a cambios en el comportamiento del usuario, las condiciones del mercado o tácticas de la competencia. Estas degradaciones pueden pasar desapercibidas hasta que el sistema falla en escenarios de alto riesgo como la optimización de la cadena de suministro, el control autónomo o la detección de fraudes.

Exceso de dependencia de modelos de fundación centralizados
Un riesgo sistémico emergente en la IA proviene de la amplia dependencia de un conjunto limitado de modelos fundamentales, especialmente los grandes modelos de lenguaje (LLMs). Esto crea un único punto de fallo: una vulnerabilidad o fallo en un modelo comúnmente utilizado puede propagarse rápidamente a través de una empresa.

Vulnerabilidades de seguridad
A medida que los modelos de IA se vuelven más capaces y accesibles, son cada vez más objetivo de actores maliciosos. Los vectores de ataque comunes incluyen pruebas de adversarios o actores maliciosos, inyección de prompts, intoxicación de datos y extracción de modelos, cada uno de los cuales puede minar la integridad del modelo o filtrar información propietaria o sensible.

Pérdida de la supervisión humana
A medida que los sistemas de IA automatizan cada vez más las decisiones, existe el riesgo de que los humanos confíen demasiado en la tecnología o no intervengan cuando realmente importa. Esto puede provocar incumplimientos éticos, errores operativos o la falta de detección de casos límite.

Estos riesgos únicos exigen enfoques igualmente novedosos en la gobernanza. Los controles informáticos tradicionales, aunque necesarios, son insuficientes por sí solos. Una gobernanza eficaz de la IA requiere un enfoque integrado que abarque salvaguardas técnicas, prácticas operativas y supervisión estratégica.

Creación de marcos de gobernanza de IA

¿Entonces, cómo empiezan las empresas? Empieza definiendo su umbral de riesgo en IA: una comprensión clara de lo que es aceptable y lo que no. Este proceso requiere comprender los objetivos estratégicos de la organización, el entorno regulatorio y los principios éticos de la organización, y luego establecer umbrales claros para el comportamiento y el rendimiento de la IA.

Aclarar los objetivos de la implementación de la IA es fundamental. ¿Qué espera lograr la organización con la IA? ¿Y qué riesgos y consecuencias se consideran razonables en la búsqueda de esos resultados? Estas respuestas ayudan a determinar qué riesgos merecen la pena asumir y cuáles requieren gestión.

Esta tolerancia al riesgo debe estar alineada con el enfoque más amplio de la empresa hacia el riesgo empresarial y sus políticas de riesgo. No basta con que la gobernanza de la IA opere de forma aislada; debe integrarse en los marcos existentes de toma de decisiones, supervisión y cumplimiento. Una vez clarificado el apetito por el riesgo, la organización puede comenzar a mapear los posibles peligros, como el mal uso de datos, el sesgo sistémico, las infracciones regulatorias o la degradación del rendimiento, y evaluar cuán probables e impactantes pueden ser esos riesgos.

Para apoyar este trabajo, herramientas de evaluación de riesgos como la modelización de escenarios, la puntuación de impacto y las simulaciones de amenazas pueden ayudar a las empresas a cuantificar los compromisos y priorizar la acción en consecuencia. Una dimensión crítica de esta evaluación es determinar qué sistemas son de alto riesgo y, por tanto, requieren los niveles más altos de control.

Alto escrutinio: Donde el bienestar humano, los derechos o la estabilidad sistémica pueden verse afectados.

- ¿Esta IA afecta a los derechos individuales, la seguridad o los medios de vida?
- ¿Podrían los resultados sesgados resultar en resultados injustos o discriminatorios?
- ¿Se utiliza esto en sectores críticos como la salud, las finanzas, la justicia penal o la contratación?
- ¿El fracaso de la IA causaría daño a individuos o al público?
- ¿Está la IA tomando decisiones autónomas sin supervisión humana?

Ejemplos: aprobación de préstamos, diagnóstico médico, reconocimiento facial para las fuerzas del orden

Escrutinio moderado: Donde los resultados importan, pero no cambian la vida.

- ¿Necesita esta IA explicar su salida a los usuarios o a los reguladores?
- ¿Se utiliza en el apoyo a la toma de decisiones (en vez de la toma de decisiones)?
- ¿Podrían los errores afectar la confianza o el rendimiento operativo, pero no la seguridad ni los derechos?
- ¿Se utiliza la IA para influir (no para decidir) cosas como marketing, previsión o atención al cliente?

Ejemplos: chatbots, previsión de demanda, puntuación de compromiso de empleados

Bajo escrutinio: Donde el impacto en las personas es mínimo o indirecto.

- ¿Se utiliza la IA principalmente para mejorar la eficiencia administrativa u operativa?
- ¿Los errores tendrían un impacto limitado o nulo en las personas?
- ¿Este caso de uso es orientado hacia el interior o limitado a la automatización de tareas rutinarias?

Ejemplos: procesamiento de facturas, clasificación de documentos, gestión de inventarios

Árbol de Decisiones Rápidas

¿Es orientado al consumidor o al público?

- **Sí:** Probablemente un mayor escrutinio
- **No:** Posiblemente bajo escrutinio

¿Necesita ser explicable o auditable?

- **Sí:** Escrutinio moderado
- **No:** Puede que sea poca escrutinio

¿Puede afectar a la salud, las finanzas, el trabajo o la situación legal de alguien?

- **Sí:** Alto escrutinio
- **No:** Proceder al siguiente

¿Podrían los errores causar pequeñas ineficiencias o daños importantes?

- **Menor:** Bajo
- **Mayor:** Alto

Traduciendo la gobernanza en la práctica

Una vez definidos los niveles de riesgo, las empresas pueden empezar a traducir principios abstractos en políticas accionables. Esto suele comenzar con marcos internos que definen requisitos de transparencia, estándares de documentación y estructuras de rendición de cuentas. Por ejemplo, las organizaciones pueden crear políticas de acceso escalonadas para diferentes tipos de modelos, imponer protocolos de clasificación de riesgos o formar juntas de revisión ética interfuncional para supervisar casos de uso de alto impacto.

Varios marcos globales pueden apoyar esta transición:

- El **Marco de Gestión de Riesgos de IA del NIST** ofrece un enfoque basado en el ciclo de vida para identificar, gestionar y mitigar los riesgos de IA antes de que ocurra un daño.
- **ISO/IEC 42001,** la primera norma internacional para la gobernanza de la IA, ofrece una guía completa para alinear las operaciones de IA con las normas legales y éticas.
- **Las Directrices de Auditoría de IA de ISACA capacitan** a los equipos de TI y cumplimiento para evaluar la integridad y la alineación ética de los sistemas de IA.
- La **Ley de IA de la UE**, aprobada en 2024, es la primera regulación integral de IA del mundo. Establece requisitos estrictos para las organizaciones que operan dentro de la UE o le venden en ella, marcando un cambio decisivo de la autorregulación a una rendición de cuentas aplicable.

Más allá de los marcos, las empresas deberían integrar prácticas responsables de IA en su cultura operativa. Eso incluye la formación de equipos de gobernanza diversos que reúnen conocimientos de las áreas legal, técnica, de cumplimiento y de negocio. Significa abordar de forma proactiva las consideraciones éticas—equidad, transparencia, privacidad desde el diseño, a lo largo del ciclo de vida de la IA. Las auditorías de desempeño continuas, los programas de formación de empleados y el diálogo interno abierto son esenciales para garantizar un uso responsable. Y, en ámbitos complejos o de alto riesgo, las organizaciones no deberían dudar en buscar orientación externa de profesionales de ética en IA, asesores regulatorios y auditores externos.

Mirando hacia el futuro: la gobernanza como ventaja estratégica

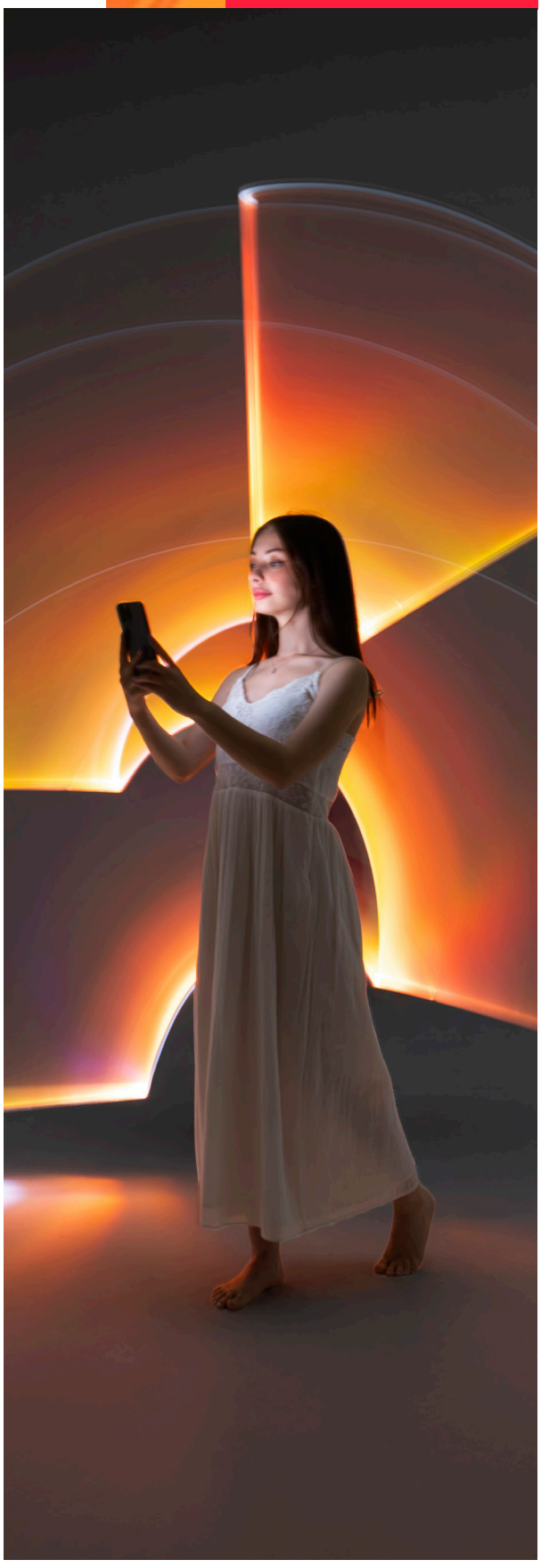
En última instancia, la gobernanza de la IA no consiste en ralentizar el progreso. Se trata de asegurar que la innovación se desarrolle de manera responsable, resiliente y alineada con los valores humanos. Cuando se aborda de forma estratégica, la gobernanza de la IA se convierte en el catalizador de la confianza, no de la restricción. Garantiza que los sistemas de IA sean responsables ante los reguladores, responsables de los socios y estén alineados con las expectativas del usuario final. La gobernanza sienta las bases para la transparencia y la confianza, que son fundamentos de una amplia aceptación, minimizando el riesgo y reforzando el rendimiento.

En esencia, permite la experimentación responsable y posiciona a la IA como una ventaja estratégica a largo plazo.

A medida que la IA se convierte en una parte indispensable de los negocios y la sociedad, las organizaciones que prosperen serán aquellas que traten la gobernanza no como una carga, sino como un modelo para el éxito a largo plazo.

> **A medida que la IA se convierte en una parte indispensable de los negocios y la sociedad, las organizaciones que prosperen serán aquellas que traten la gobernanza no como una carga, sino como un modelo para el éxito a largo plazo.**

Ver grabación del seminario web →



Sobre RGP

RGP es una firma global de servicios profesionales con casi tres décadas de experiencia ayudando a las principales organizaciones del mundo — desde empresas Fortune 50 hasta startups de alta evolución— a resolver los complejos problemas empresariales actuales. Socios de confianza para CFOs y líderes financieros, ofrecemos las soluciones de talento, consultoría y servicios externalizados que necesitas para crecer más rápido, trabajar de forma más inteligente y mantenerte al día con los cambios, todo ello a través de un modelo flexible y una red global de expertos.