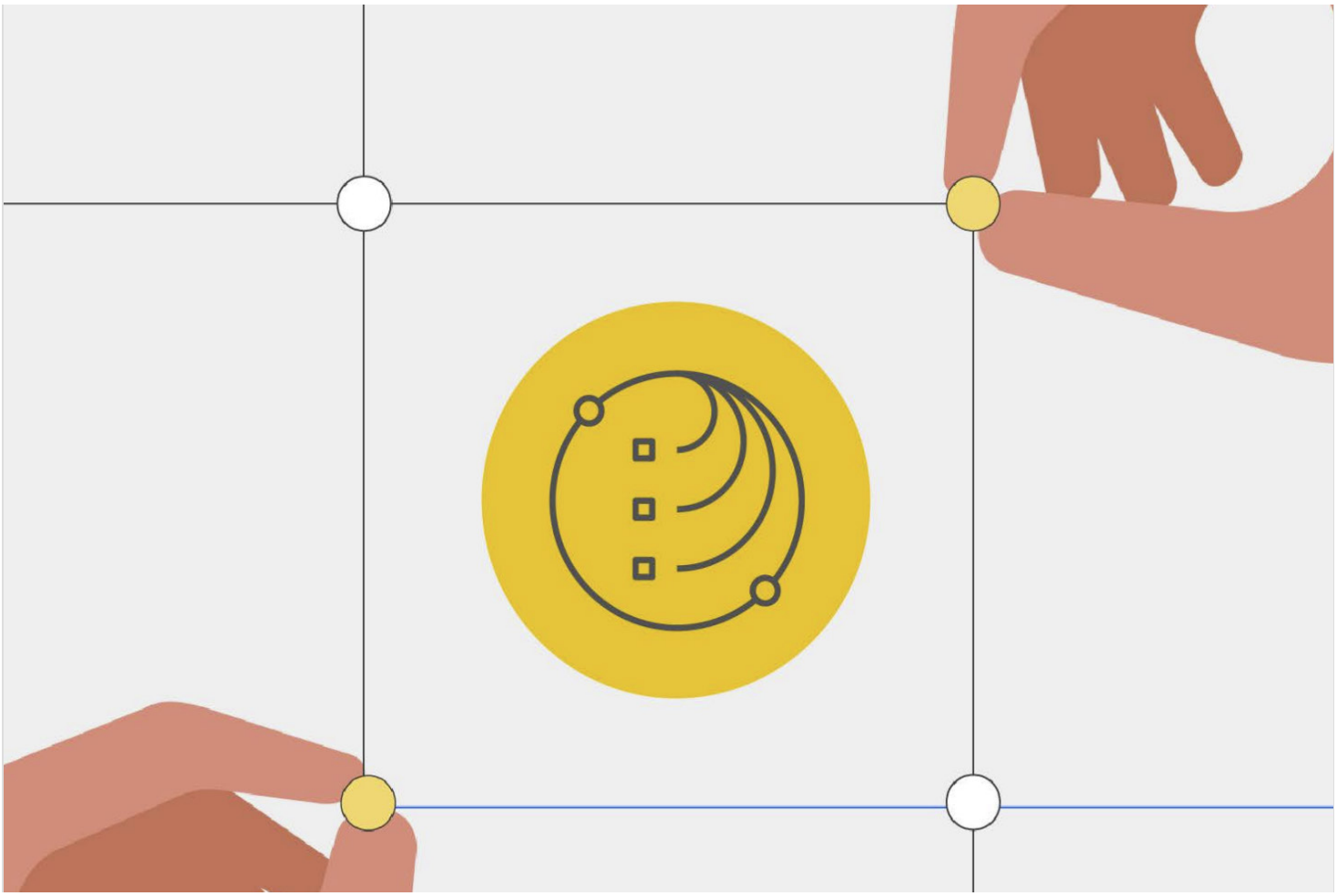
Agentes de IA:
Oportunidades, riesgos y mitigaciones



Atribuciones

Agradecemos a:

Los copresidentes del Comité de Ética de la IA y patrocinadores ejecutivos del flujo de trabajo: Christina Montgomery, Francesca Rossi, Kush Varshney, Manish Bhide y Rob Parkin

Colaboradores: Saishruthi Swaminathan, Ambrish Rawat, Chan Nasseb, Christopher Noessel, Daniel Karl I. Weidele, David Piorkowski , Heloísa Candello, Jamie VanDodick, John Richards, Matt Bellio, Michael Hind, Michael Muller, Mihaela Bornea, Milena Pribic, Phaedra Boinodris, Rogerio Abreu de Paula, Sara E Berger y Simon Rogers

Agradecimientos

Los copresidentes, patrocinadores ejecutivos y colaboradores desean agradecer a los siguientes miembros por sus comentarios y feedback: Alina Glaubitz, Amit Dhurandhar, Christopher Hay, Jill Maguire, Katherine Fick, Kevin Black, Lauren Quigley, Manish Goyal, Maryam Ashoori, Monica Patel, Pin-Yu Chen, Ryan Hagemann y Wouter Oosterbosch

Índice

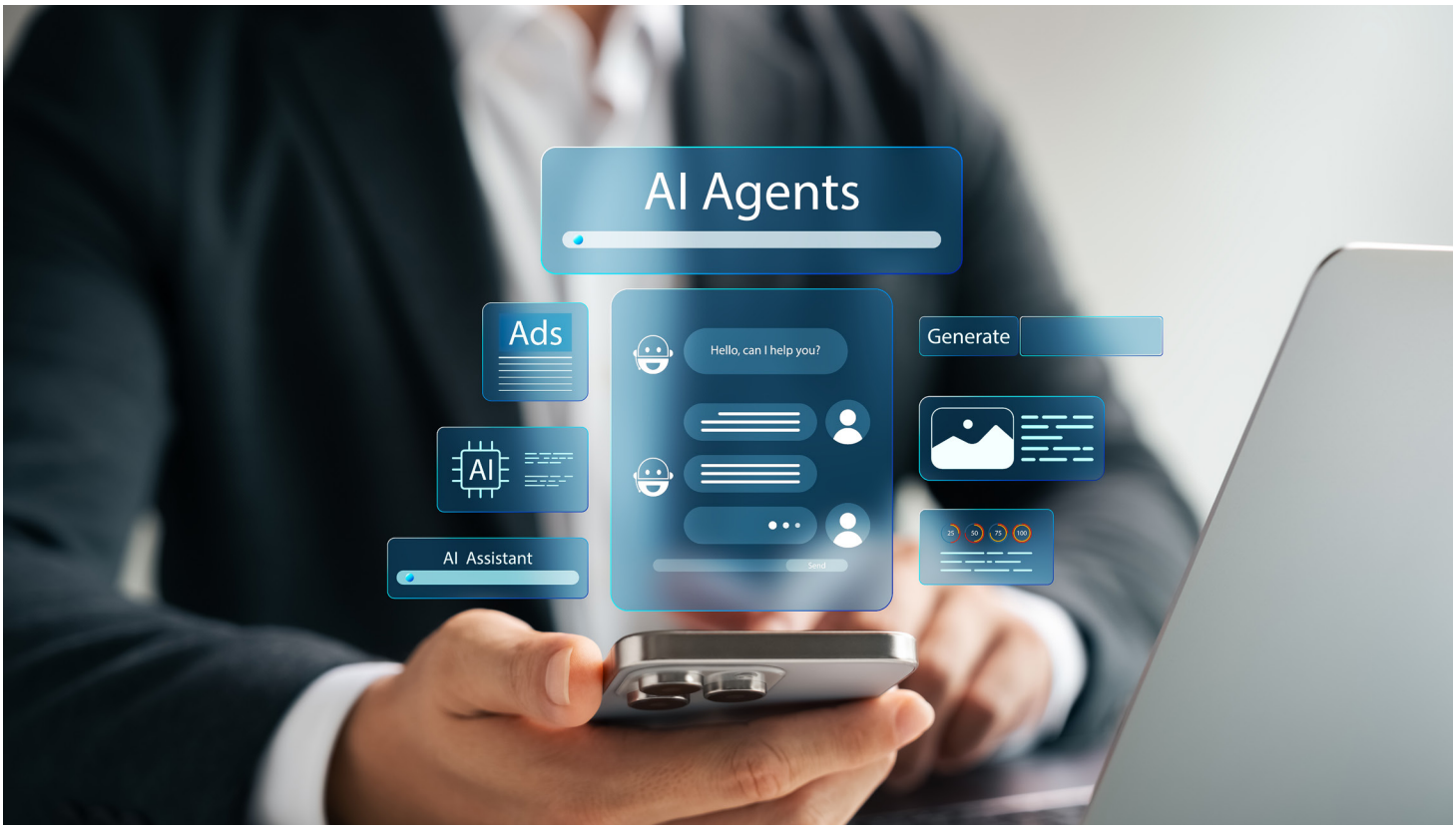
Introducción

La inteligencia artificial (IA) está revolucionando la forma en que las personas interactúan con el mundo que las rodea. A medida que la tecnología de IA continúa evolucionando, genera nuevas oportunidades y ayuda a resolver problemas complejos en diversos campos, como las finanzas, la salud, el medio ambiente, la educación y el deporte. Ahora estamos entrando en una era de IA en la que las empresas están adoptando agentes de IA para transformar sus procesos, maximizar el impacto empresarial en todas las funciones y acelerar el tiempo de creación de valor. Un agente de IA es una entidad de software que usa técnicas de IA y tiene capacidad para actuar en su entorno con base en objetivos establecidos, lo que significa que puede decidir qué acciones realizar y tiene la capacidad de ejecutarlas. Los agentes de IA han existido durante mucho tiempo, desde los agentes de IA basados en reglas hasta los recientes agentes de IA que usan modelos de lenguaje grandes y demuestran un gran potencial en varios sectores. Los sistemas de IA agéntica son sistemas de software que aprovechan los agentes de IA (junto con otros componentes como herramientas, planificadores, memoria y conjunto de datos), buscan alcanzar metas y pueden operar de forma autónoma. Aunque los agentes de IA y los sistemas de IA agéntica son software, se pueden usar para controlar el hardware.

***Nota**: En este documento, los términos "agentes de IA" y "sistema de IA agéntica" se usan indistintamente.*

IBM es una empresa de nube híbrida e IA; mediante el uso de la fuerza de nuestros equipos de investigación, productos y consultoría, junto con socios externos y el Comité de Ética de IA, IBM ayuda a llevar el poder de los agentes de IA a nuestros clientes. Algunos ejemplos incluyen el agente de ingeniería de software de IBM® Research, los servicios de integración de IA de IBM® Consulting, Granite BeeAI Agent Framework, IBM watsonx Orchestrate® y watsonx.ai® de IBM Technology.

Debido a que los rápidos avances en tecnología de IA, los agentes de IA se están volviendo más poderosos y sofisticados, lo que plantea preguntas sobre sus ramificaciones éticas y su impacto social. En este documento se describe el punto de vista actual de IBM sobre la ética y la gobernanza de los agentes de IA. Es la primera versión, y en iteraciones futuras se ampliarán varios aspectos del enfoque de ética y gobernanza de IBM con respecto a sus agentes de IA.



Beneficios

Los agentes de IA pueden mejorar en gran medida la eficacia con la que los humanos realizan tareas y logran resultados empresariales. Estos beneficios incluyen:

Ampliar la inteligencia humana

Los agentes de IA pueden integrarse en los flujos de trabajo para ayudar a reducir la cantidad de tiempo que se dedica a completar tareas y aumentar el rendimiento humano. Por ejemplo, el sistema IBM AI Agent SWE-1.0 consta de un agente de localización y un agente de edición que se integran en flujos de trabajo de GitHub para ayudar a los desarrolladores a reducir la cantidad de tiempo que los desarrolladores dedican a encontrar errores, desarrollar y probar arreglos a defectos de software.

Automation

Los agentes de IA pueden automatizar tareas rutinarias o que consumen mucho tiempo para permitir que se tenga un mayor enfoque en la innovación y el trabajo estratégico. Por ejemplo, AI Digital Assistant AskHR de IBM usa agentes de IA para automatizar procesos comunes de RR. HH., como la asistencia a los empleados y su incorporación. Los agentes de IA están desarrollados sobre watsonx Orchestrate e impulsados con IA generativa. Con esta automatización, AskHR de IBM gestiona ya el 94 % de las consultas de los empleados y resuelve alrededor de 10.1 millones de interacciones al año, lo que permite al equipo de RR. HH. de IBM centrarse más en labores tan importantes como la planeación estratégica.

Mejor eficiencia y productividad

Los agentes de IA pueden operar de forma continua, gestionar múltiples tareas de manera simultánea y enfrentar retos complejos. Esto puede ayudar a acelerar la velocidad de entrega, aumentar la productividad y mejorar la eficiencia de las operaciones empresariales. Por ejemplo, IBM está ampliando su asociación con Salesforce para proporcionar agentes de IA predefinidos que estarán disponibles para los clientes todo el tiempo, con el fin de brindar apoyo a las ventas y los servicios. IBM aprovechará watsonx Orchestrate para crear agentes de IA para Agentforce, la suite de agentes autónomos de Salesforce, con el fin de ayudar a las empresas a mejorar la productividad, al tiempo que se mantiene la seguridad

Toma de decisiones y calidad mejoradas de las respuestas

Los agentes de IA pueden conectarse con múltiples recursos externos, herramientas y otros agentes para ayudar a mejorar su toma de decisiones y la calidad de sus respuestas. También pueden proporcionar respuestas completas y personalizadas para el usuario, lo que resulta en una mejor experiencia del cliente. Por ejemplo, IBM Consulting trabajó con una empresa global de ciencias biológicas para recopilar una serie de agentes de IA con el fin de acelerar la generación de documentación técnica basada en hechos y rastreable.

Riesgos, desafíos e Impactos sociales

Al igual que todas las tecnologías que están avanzando rápidamente, los agentes de IA pueden generar riesgos además de los beneficios. Diversas acciones legales y regulatorias pueden estar relacionadas con algunos de estos riesgos, así como consecuencias reputacionales y operativas. En general, los riesgos plantean cuestiones sociales y técnicas, por lo que deben abordarse y mitigarse mediante métodos sociales y técnicos, incluidas herramientas informáticas, procesos de evaluación de riesgos, marcos éticos de IA, mecanismos de gobernanza, consultas con múltiples stakeholders, normas y regulación.

Método

Los agentes de IA pueden realizar tres tipos de acciones:

- Realizar acciones que impacten al mundo (físico o digital).
- Consultar recursos y usar herramientas.
- Decidir qué proceso elegir en la selección de recursos, herramientas u otros agentes de IA y seleccionarlos.

Los agentes de IA pueden realizar estas acciones de forma autónoma, es decir, sin supervisión humana continua.

Por su naturaleza, los agentes de IA y los sistemas de IA agéntica pueden tener las siguientes características:

- Opacidad debido a la visibilidad limitada de cómo operan los agentes de IA, incluido su funcionamiento interno e interacciones.
- Apertura en la selección de recursos, herramientas u otros agentes de IA para ejecutar acciones. Esto puede aumentar la posibilidad de ejecutar acciones inesperadas.
- La complejidad surge como consecuencia de la apertura y se combina con el escalamiento de la apertura.
- La irreversibilidad como consecuencia de tomar acciones que podrían impactar al mundo.

Estas características, así como el rango de niveles de autonomía que pueden tener los agentes de IA y sistemas de IA agéntica, se traducen en varios riesgos, desafíos e impactos sociales, como se muestra en las siguientes tablas. Nos basamos en los riesgos y desafíos identificados para los modelos fundacionales y, por lo tanto, solo incluimos en las tablas a continuación los riesgos, desafíos e impactos sociales relacionados con los agentes de IA que son nuevos o que amplifican los de la tabla de riesgos del modelo fundacional.

Nota: En las siguientes tablas, el agente de IA se usa para describir tanto el agente de IA como el sistema de IA agéntica.

Riesgos

Grupo	Riesgo	Indicador de riesgo con el motivo
Alineación de valores	**Acciones desalineadas:** los agentes de IA puede tomar acciones que no estén alineadas con los valores humanos, las consideraciones éticas, las directrices y las políticas relevantes. Las acciones desalineadas pueden suceder de diferentes maneras, tales como: • Aplicar los objetivos aprendidos de forma inapropiada a situaciones nuevas o imprevistas. • Usar los agentes de IA para un propósito o meta que no sea parte de su uso previsto. • Seleccionar recursos o herramientas con sesgo. • Usar tácticas engañosas para lograr el objetivo mediante el desarrollo de la capacidad de • realizar esquemas con base en instrucciones dadas dentro de un contexto específico. • Poner en riesgo los valores del agente de IA para trabajar con otro agente de IA o herramienta para realizar la tarea.	Ampliado Motivo: • Autonomía del agente de IA para tomar acciones

Imparcialidad	**Acciones discriminatorias:** los agentes de IA puede tomar acciones en las que un grupo de humanos se ve injustamente favorecido sobre otro debido a las decisiones del modelo. Esto puede deberse a sesgos de los agentes de IA en las acciones que impactan al mundo, en los recursos consultados y en el proceso de selección de recursos. Por ejemplo, un agente de IA puede generar código con sesgo.	Ampliado
		
		Motivo:
		• Autonomía del agente de IA para tomar acciones
		  • Tomar acciones que impacten al mundo
		  • Consultar recursos con sesgo
		  • Proceso de selección de recursos con sesgo

	Sesgo de datos: las acciones específicas del agente de IA, como modificar un conjunto de datos o una base de datos, pueden introducir sesgos en el recurso que es utilizado por otros o por sí mismo para realizar acciones.	Nuevo Motivo: • Los agentes de IA toman acciones que impactan al mundo. Aquí, el sesgo se debe a una acción específica • Apertura
Confianza mal depositada	**Dependencia excesiva o insuficiente:** la dependencia, es decir, la disposición a aceptar el comportamiento de un agente de IA, depende de cuánto confía un usuario en ese agente y para qué lo está usando. La dependencia excesiva se produce cuando un usuario confía demasiado en un agente de IA, ya que acepta el comportamiento de un agente de IA incluso cuando probablemente no sea deseado. La dependencia insuficiente es lo contrario, cuando el usuario no confía en el agente de IA, pero debería. El aumento de la autonomía (para tomar medidas, seleccionar y consultar recursos o herramientas) de los agentes de IA y la posibilidad de opacidad y apertura aumentan la variabilidad y la invisibilidad del comportamiento de los agentes, lo que dificulta la calibración de la confianza y posiblemente contribuya tanto a la dependencia excesiva como a la insuficiente.	Ampliado Motivo: • Las acciones de los agentes de IA dificultan la evaluación de la confianza

Grupo	Riesgo	Indicador de riesgo con el motivo
Ineficiencia informática	**Acciones redundantes:** los agentes de IA pueden ejecutar acciones que no son necesarias para alcanzar la meta. Estas acciones podrían causar un desperdicio de recursos informáticos, reducir la eficiencia del agente de IA en el logro de la meta y llevar a resultados potencialmente perjudiciales. En un caso extremo, los agentes de IA podrían entrar en un ciclo de ejecución de las mismas acciones repetidamente sin ningún progreso. Esto podría suceder debido a condiciones inesperadas en el entorno, la incapacidad del agente de IA para reflexionar sobre su acción, errores de razonamiento y planeación del agente de IA o la falta de conocimiento del agente de IA sobre el problema. Esto podría impedir que el agente de IA alcance la meta y agote los recursos informáticos.	Nuevo Motivo: • Los agentes de IA pueden actuar
Robustez	**Ataque a los recursos externos de los agentes de IA:** los atacantes crean vulnerabilidades de forma intencionada o explotan vulnerabilidades existentes en recursos externos (herramientas, bases de datos, aplicaciones, servicios y otros agentes) de los que los agentes de IA dependen para ejecutar sus acciones previstas o para alcanzar sus metas. Los recursos externos en riesgo podrían afectar el rendimiento del agente de IA de diferentes maneras, tales como: • Manipular a los agentes de IA para tratar de alcanzar una meta diferente. Ejemplo: cambiar la meta de los agentes de IA para agregar comentarios positivos a un producto de la preferencia del atacante cuando la meta original del usuario es agregar consultas sobre el producto. • Manipular a los agentes de IA para ejecutar acciones no deseadas. Ejemplo: engañar a los agentes de IA para descargar malware. • Capturar y retransmitir interacciones de agentes de IA a actores maliciosos. • Lograr que los agentes de IA compartan información personal o confidencial.	Nuevo Motivo: • Los agentes de IA pueden actuar • Apertura y complejidades debidas a la apertura • Acceder a más recursos
	Uso no autorizado: si los atacantes pueden obtener acceso al agente de IA y sus componentes, pueden realizar acciones que podrían causar diferentes niveles de daño dependiendo de las capacidades del agente y la información a la que tiene acceso. Además, los atacantes pueden ejecutar acciones que conduzcan a la degradación del sistema, como agotar los recursos disponibles y afectar el rendimiento. Ejemplos: • Utilizar información personal almacenada para imitar la identidad o suplantar la identidad con intención de engañar. • Manipular el comportamiento del agente de IA a través de feedback al agente de IA o corromper su memoria para cambiar su comportamiento. • Manipular la descripción del problema o la meta para que el agente de IA se comporte mal o ejecute comandos dañinos.	Ampliado Motivo: • Los agentes de IA pueden actuar • Los agentes de IA son más capaces • Característica de personalización de los agentes de IA
	Explotar la discordancia de la confianza: los atacantes podrían iniciar ataques de inyección para eludir el límite de confianza, que es un punto distinto o una línea conceptual en la que cambia el nivel de confianza en un sistema, aplicación o red. Esto podría tener como consecuencia que los límites de confianza no coincidan (esperados frente a realizados) y podría causar un uso no deseado de herramientas, una agencia excesiva y un escalamiento de privilegios. Además, la ejecución en segundo plano en entornos multiagente aumenta el riesgo de canales encubiertos si la validación de entrada o resultados es deficiente.	Ampliado Motivo: • Apertura • Complejidad
	Alucinación en llamada de funciones: los agentes de IA podrían cometer errores al generar llamadas de funciones (llamadas a herramientas para ejecutar acciones). Esas llamadas de funciones podrían generar acciones incorrectas, innecesarias o dañinas. Ejemplos: generar funciones incorrectas o parámetros incorrectos para las funciones.	Nuevo Motivo: • Los agentes de IA pueden consultar recursos y herramientas • Los agentes de IA pueden actuar

Grupo	Riesgo	Indicador de riesgo con el motivo
Privacidad y propiedad intelectual	**Compartir información de IP, personal y confidencial con el usuario:** los agentes de IA con acceso sin restricciones a recursos o bases de datos o herramientas podrían almacenar y compartir información de IP, personal y confidencial con los usuarios del sistema al realizar sus acciones.	Ampliado Motivo: • Naturaleza multicomponente con capacidad para realizar acciones
	Compartir información de IP, personal y confidencial con herramientas: los agente de IA con acceso sin restricciones a recursos o bases de datos o herramientas podrían almacenar y compartir información de IP, personal y confidencial con otras herramientas o agentes al realizar sus acciones.	Nuevo Motivo: • Naturaleza multicomponente con capacidad para realizar acciones
Explicabilidad y transparencia	**Acciones inexplicables e imposibles de rastrear:** las explicaciones, la información de linaje y rastreo, y la atribución de fuentes para las acciones del agente de IA pueden ser difíciles, imprecisas o inalcanzables.	Ampliado Motivo: • Difícil de rastrear: causa y efecto o influencia de diferentes componentes, incluidos los LLM, en las acciones finales
	Falta de transparencia: la falta de transparencia se debe a la documentación insuficiente del proceso de diseño, desarrollo y evaluación del agente de IA, a la ausencia de insights sobre su funcionamiento interno y a la interacción con otros agentes, herramientas o recursos.	Ampliado Motivo: • Dependencia de otros documentos disponibles para otras herramientas o agentes

Desafíos

Desafío	Indicador de desafío con motivo
Evaluación: desafío en la evaluación del rendimiento y exactitud de los agentes de IA debido a la complejidad del sistema y su apertura.	Ampliado Motivo: • Complejidad de los sistemas agénticos
Mitigación y mantenimiento: desafío para saber en qué parte del sistema algo está fallando y cómo arreglarlo o qué podría necesitar protocolos de mantenimiento.	Ampliado Motivo: • Complejidad y apertura de los sistemas agénticos
Reproducibilidad: desafío en la reproducción del comportamiento o resultados del agente debido a la falta de disponibilidad o cambios en las herramientas o recursos utilizados para la ejecución de las acciones.	Nuevo Motivo: • Apertura
Rendición de cuentas: desafío en la asignación de responsabilidad por una acción realizada por un sistema de IA agéntica.	Ampliado Motivo: • Complejidad y apertura. El componente puede ser de diferentes proveedores
Cumplimiento normativo: desafío para determinar el cumplimiento normativo debido a que los agentes de IA son complejos y es posible que no haya suficiente información para comprender si todo el sistema de IA agéntica es compatible.	Ampliado Motivo: • Apertura • Complejidad • Falta de transparencia

Impactos sociales

Impacto social	Indicador de impacto social
Impacto en la dignidad humana: si los trabajadores humanos perciben que los agentes de IA hacen su trabajo mejor que ellos, podrían experimentar una disminución en su autoestima y bienestar.	Ampliado
Impacto en la agencia humana: la naturaleza autónoma de los agentes de IA en tareas o acciones podría afectar la capacidad de los individuos para participar en el pensamiento crítico, tomar decisiones y actuar de manera independiente.	Ampliado
Impacto en los puestos de trabajo: la adopción generalizada de agentes de IA para realizar tareas complejas podría conducir a una automatización generalizada de roles y tener como consecuencia que se eliminen puestos de trabajo.	Ampliado
Impacto en el medio ambiente: la complejidad de las tareas y la posibilidad de que los agentes de IA realicen acciones redundantes podría llevar a ineficiencias informáticas y aumentar el impacto ambiental.	Ampliado

Mitigación y gobernanza

A medida que la IA acelera la transformación del negocio, la confianza se ha vuelto imperativa para afrontar un entorno empresarial competitivo. El compromiso de IBM con la confianza se plasma en nuestros Principios de confianza y transparencia y en los Pilares de la IA confiable, y en este documento se destaca el enfoque holístico de IBM hacia la cultura, los procesos y las herramientas y cómo los equipos de investigación, producto y consultoría de IBM se unen al Comité de Ética de IA para construir soluciones responsables de IA agéntica.

Comité de ética de IA

IBM ha establecido una cultura que respalda el desarrollo, el despliegue y el uso responsable de la IA. En el centro de nuestra gobernanza organizacional se encuentra el Comité de ética de IA multidisciplinario, que es responsable del proceso de gobernanza y toma de decisiones para las políticas y prácticas de ética de IA. El Comité de ética de IA ha estado en funciones durante más de cinco años. Entre sus muchas actividades, el Comité de ética de IA trabaja con Puntos Focales dentro de cada unidad de negocio de IBM para evaluar los casos de uso de IA y alinearlos con los valores fundamentales de IBM.

Gobernanza integrada

El Programa de gobernanza integrada (IGP) es un enfoque unificado de responsabilidad y cumplimiento. Al crear una visión holística de principio a fin de los datos y los modelos construidos y utilizados, el IGP escala los flujos de trabajo de gobernanza en torno a la privacidad de los datos y la IA sin interrumpir la innovación y los procesos de negocio. El IGP ha potenciado a IBM para desplegar estándares internos de datos uniformes, lo que permite la transparencia de los datos y el desarrollo de IA confiable, al tiempo que permite la innovación a escala.

Productos y ofertas

IBM watsonx.governance® permite a las organizaciones impulsar una IA responsable, transparente y explicable. Proporciona capacidades completas de gobernanza del ciclo de vida en toda la IA, incluida la IA agéntica, desde la solicitud de caso de uso hasta el despliegue, incluida la evaluación inicial de riesgos y la evaluación de riesgos para ayudar a identificar los riesgos en las primeras etapas del proceso. Cuenta con generación aumentada por recuperación (RAG) y métricas de evaluación de IA agéntica como la fidelidad, la relevancia del contexto y la similitud de las respuestas para ayudar a confirmar si los agentes de IA están actuando adecuadamente. IBM watsonx.governance tendrá métricas adicionales diseñadas para monitorear y mejorar el rendimiento de los agentes. También tendrá capacidades para detectar la alucinación de llamadas de herramientas, gestionar riesgos, ayudar al cumplimiento normativo y capturar metadatos y otros detalles sobre los agentes de IA en una hoja informativa.

IBM watsonx.ai simplifica, unifica y optimiza la gestión del ciclo de vida de los agentes (conocida como AgentOps), mediante transparencia, trazabilidad y flexibilidad para descubrir, gestionar, monitorear y optimizar el agente de IA.

IBM watsonx Orchestrate pone a trabajar la IA al ayudar a los usuarios a crear, desplegar y gestionar poderosos agentes de IA que automatizan tareas con IA generativa. Orchestrate proporcionará transparencia en el razonamiento del agente de IA sobre la selección de herramientas y/o la colaboración del agente para que un usuario pueda comprender cómo se completó una tarea o flujo de trabajo.

IBM® Guardium AI Security ayuda a los clientes a monitorear continuamente los controles de sus modelos de IA generativa en producción y permite un despliegue seguro y responsable.

La oferta de estrategia y gobernanza de IA de IBM Consulting potencia a las organizaciones para que aprovechen el potencial transformador de la IA curada de manera responsable, desde la tradicional hasta la agéntica, arraigada en sus datos empresariales para avanzar y remodelar su estrategia comercial. Aconsejamos a los clientes que "apliquen la IA de manera correcta" al desbloquear el retorno de la inversión (ROI) en IA y que "creen la IA correcta" al permitir la rendición de cuentas de los modelos de IA que construyen y compran, para que puedan escalar de manera responsable y reducir el riesgo de sus inversiones. Esta infraestructura holística permite a los clientes enfrentar los desafíos cambiantes de IA agéntica, descubrir nuevas oportunidades, acelerar la innovación y mejorar la toma de decisiones organizacionales. A continuación, se destacan las prácticas implementadas y recomendadas por la oferta para ayudar a las empresas a mitigar los riesgos de la IA agéntica:

- Permitimos la observabilidad para entender qué acciones realizó el agente con qué entradas para determinar la atribuibilidad de los resultados correspondientes.
- Examinamos el seguimiento completo para determinar si el agente de IA ha progresado hacia la meta global con cada acción realizada, para depurar el flujo e identificar cuellos de botella en el diseño del sistema que hagan aflorar repetidamente acciones redundantes o innecesarias.
- Creamos ontologías para mejorar la precisión y la confiabilidad, y proporcionar el linaje y la procedencia de datos.
- Realizamos pruebas funcionales que implican probar rigurosamente cada componente del sistema agéntico de forma aislada (es decir, herramientas y agentes), las interacciones entre los componentes, la capacidad de seleccionar las herramientas adecuadas con parámetros y valores correctos, así como las medidas de protección para mitigar las vulnerabilidades.
- Configuramos los agentes de IA para mejorar la eficiencia informática al detectar cuándo los agentes de IA ingresan a posibles escenarios de bucle infinito mediante el seguimiento, por tarea, el tiempo transcurrido, el token total consumido o el número de iteraciones fallidas.
- Creamos agentes de IA hiperenfocados y les proporcionamos solo las herramientas adecuadas para completar sus tareas, y nos aseguramos de que la ejecución siempre se realice con el uso del contexto de autorización del usuario que accede.
- Definimos medidas de seguridad a nivel de modelo para detectar y mitigar el contenido de HAP, jailbreaking, intentos de inyección de instrucción, divulgación de información confidencial no autorizada y alucinaciones.

Modelos

Los modelos Granite Guardian son una suite sólida de medidas de seguridad diseñadas para detectar riesgos tanto en la instrucción como en las respuestas. En los casos de uso de RAG, los modelos de Guardian evalúan la relevancia del contexto, la fundamentación y la relevancia de la respuesta. También cuenta con detectores para la alucinación de llamada de funciones dentro de los flujos de trabajo agénticos. Esto incluye la evaluación de la validez de las llamadas a funciones y la detección de información fabricada, especialmente durante la traducción de consultas.

Participación de humanos en el proceso y supervisión humana

La supervisión y los comentarios de humanos pueden ayudar a identificar riesgos y corregir errores. La validación y el feedback humano ayudan a garantizar que las acciones tomadas por los agentes de IA sean precisas, relevantes y alineadas. Como parte del proceso de evaluación ética de los casos de uso de IA de IBM, se considera la cuestión sobre participación de humanos en el proceso y se implementa la supervisión humana adecuada. Augmenting Human Intelligence POV de IBM describe ejemplos de casos de uso, indicadores clave de rendimiento y mejores prácticas para mejorar la inteligencia humana con IA y empoderar a las personas para transitar por un entorno empresarial competitivo en asociación con IA.

Educación

IBM proporciona educación sobre la ética y la gobernanza de la IA a los equipos, los clientes y la comunidad. IBM SkillsBuild, el canal de YouTube de IBM Technology, los cursos de IBM Developer watsonx, así como IBM AI Academy son algunos de los recursos a través de los cuales IBM proporciona capacitación a las personas sobre la ética y la gobernanza de la IA.

A continuación, se destacan las investigaciones y herramientas de vanguardia de IBM para ayudar a los usuarios a lo largo del ciclo de vida del agente de IA y construir soluciones responsables de IA agéntica.

Técnicas y métodos

- Técnicas como forzar a los humanos a tomar más tiempo para pensar (estrategia de desanclaje basada en el tiempo) ayudan a lograr una colaboración óptima entre humanos y IA y reducen el sesgo de anclaje, donde los humanos confían ciegamente en la decisión de la IA. Otra técnica se basa en la el marco colaborativo IA-humano basado en valores que introduce fricción al incentivar a los humanos con recomendaciones de decisión, cuando es necesario, en la interacción humano-IA cuando el humano es el responsable de tomar la decisión final.
- Métodos como la explicación multinivel para modelos de lenguaje generativo y las explicaciones contrastivas para grandes modelos de lenguaje ayudan en la explicabilidad y la atribución de fuentes.
- Métodos como la colaboración adversarial, en la que la IA examina la base de la decisión humana en lugar de ofrecer recomendaciones alternativas, ayudan a reducir el impacto de la automatización y la reducción de la agencia en la dignidad humana.
- Attack Atlas, una taxonomía intuitiva y organizada de vectores de ataque de entrada de un solo turno, proporciona a la comunidad un punto de partida unificado en el campo de rápido crecimiento de la seguridad de la IA generativa y el red-teaming.

Herramientas y puntos de referencia

- Los enfoques de ajuste de modelos, como los que utiliza IBM Alignment Studio, ayudan a mitigar los riesgos de alineación de valor.
- El kit de herramientas de código abierto de IBM IA Fairness 360 ayuda a mitigar los riesgos de equidad.
- ITBench, un conjunto de puntos de referencia, ofrece a los profesionales de la IA una forma de medir la eficacia de los agentes que están creando para resolver problemas reales y cómo se comparan sus agentes con otros en las tareas que las empresas realizan todos los días.
- Carbon for IA, un sistema de diseño de código abierto de IBM, aprovecha un icono interactivo de IA para promover la explicabilidad y una comprensión más clara de las capacidades de IA.

Alfonso Nápoles Gandara 3111
Col. Parque corporativo de Peña Blanca
C.P. 01210
México D.F.
IBM Corporation
New Orchard Road
Armonk, NY 10504

Elaborado en los
Estados Unidos de América
Marzo de 2025