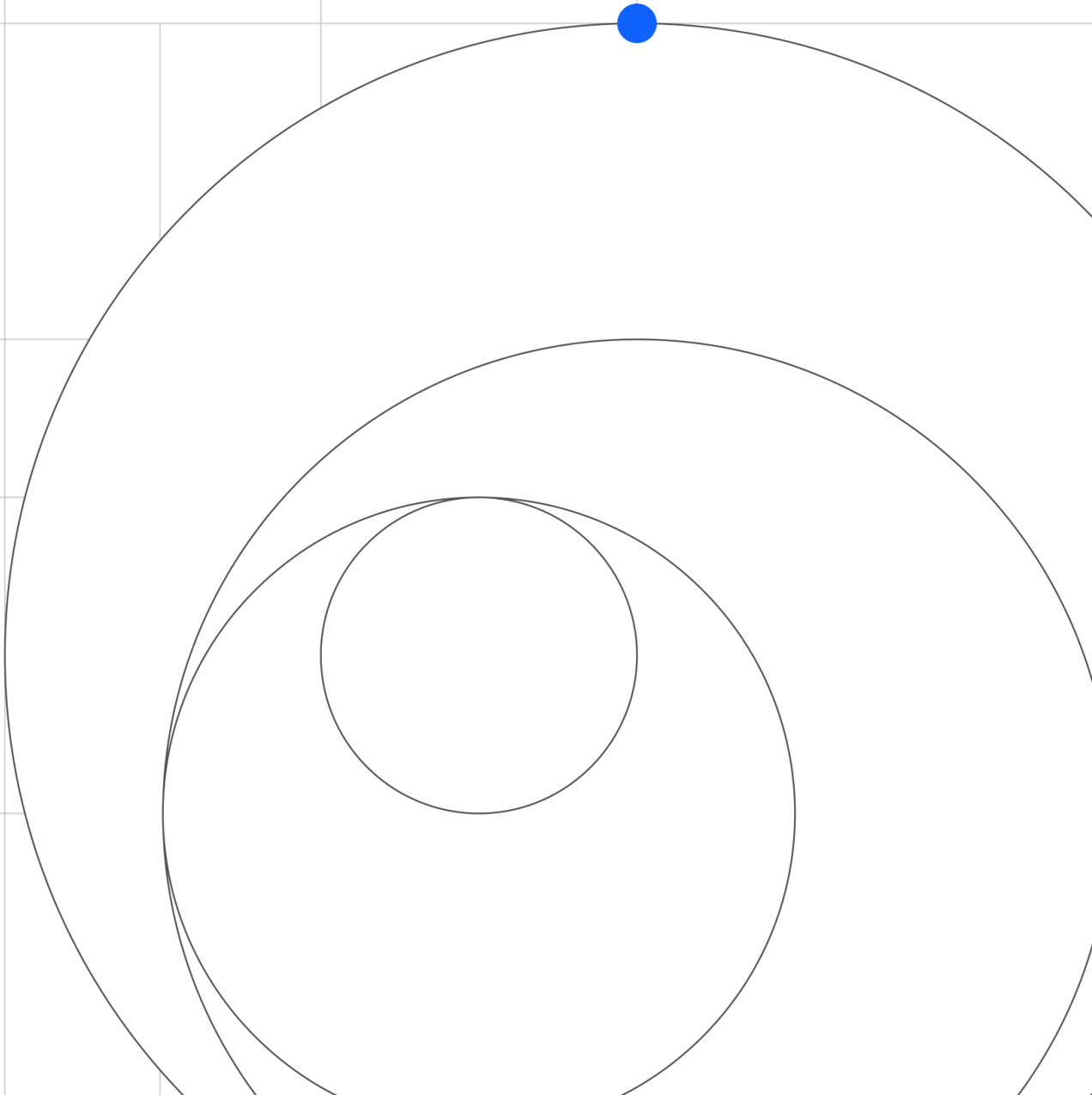
Modelos fundacionales: oportunidades, riesgos y mitigaciones



Atribución

Agradecemos a las patrocinadoras ejecutivas de la línea de trabajo del Consejo de Ética de la IA (Christina Montgomery y Francesca Rossi), y a las contribuciones de los miembros de la línea de trabajo (Betsy Greytok, Bryan Bortnick, Catherine Quinlan, David Piorkowski, Eniko Rozsa, Heather Domin, Heather Gentile, Jamie VanDodick, Jill Maguire, John McBroom, Joshua New, Justin Weisz, Katherine Fick, Kevin Black, Kush Varshney, Manish Bhide, Manish Goyal, Melis Kiziltay, Michael Epstein, Michael Hind, Milena Pribic, Phaedra Boinodiris, Rogerio Abreu de Paula, Saishruthi Swaminathan y Suj Perepa).

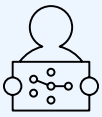
Índice

Resumen ejecutivo

El auge de los modelos fundacionales ofrece nuevas y apasionantes posibilidades a las empresas, pero también plantea preguntas nuevas y más amplias sobre su diseño, desarrollo, despliegue y uso éticos. Según una reciente encuesta sobre IA generativa del IBM Institute for Business Value, las organizaciones ya están mostrando su preocupación por los temas relacionados con la confianza, específicamente como obstáculos para la inversión. Sus principales preocupaciones son la ciberseguridad (57 %), la privacidad (51 %) y la precisión (47 %). Muchas organizaciones ya se estaban tomando en serio estas preocupaciones antes de la *consumerización* de la IA generativa y expresaron su intención de invertir al menos un 40 % más en ética de la IA en los próximos tres años. La concientización sobre los riesgos y las posibles formas de mitigarlos es el primer paso crucial hacia la creación de sistemas confiables de IA.

En este documento:

Descubrimos los beneficios de los modelos fundacionales, incluida su capacidad para realizar tareas desafiantes, el potencial de acelerar la adopción de IA, la capacidad de aumentar la productividad y los beneficios de costos que ofrecen.



Analizamos las tres categorías de riesgo, incluidos los riesgos conocidos de formas anteriores de IA, los riesgos conocidos ampliados por los modelos fundacionales y los riesgos emergentes intrínsecos a las capacidades generativas de los modelos fundacionales.

Cubrimos los principios, pilares y gobernanza que forman los cimientos de las iniciativas de ética de IA de IBM, y sugerimos medidas de protección para la mitigación de riesgos.

Introducción

A medida que cada vez es más amplio el uso de la IA, los grandes y complejos modelos de IA ofrecen resultados de rendimiento prometedores y resuelven algunos de los problemas más complejos de la sociedad. Sin embargo, desarrollar grandes conjuntos de datos de entrenamiento y modelos complejos para cada aplicación de IA puede ser una carga para las empresas. Los modelos fundacionales proporcionan una vía para conseguir lo mejor de ambos mundos: desarrollar potentes modelos de última generación y reutilizarlos directamente o aplicar métodos de ajuste para implementar una variedad de casos de uso, en lugar de entrenar nuevos modelos para cada caso de uso. Por ejemplo, IBM Research desarrolló modelos fundacionales para la inspección visual. Estos modelos fundacionales aprenden la representación general de superficies de concreto y pistas de aterrizaje y despegue. Además, pueden ajustarse para casos de uso específicos, como la detección de grietas o la inspección de defectos con menos datos etiquetados.

IBM define un *modelo fundacional* como un modelo de IA que se puede adaptar a una amplia gama de tareas descendentes. Los modelos fundacionales normalmente son modelos generativos a gran escala que se entrenan en datos sin etiquetar mediante la autosupervisión. Como modelos a gran escala, los modelos fundacionales pueden incluir miles de millones de parámetros.

IBM es una empresa de nube híbrida e inteligencia artificial con una larga reputación como administrador de datos responsable y comprometido con la ética de la IA. A través de la solidez de nuestros equipos de investigación, productos y consultoría, junto con socios externos, como Hugging Face, ayudamos a llevar el poder de los modelos fundacionales a nuestros clientes y construir una IA confiable en cualquier empresa. IBM también continúa invirtiendo en la construcción de nuevas plataformas, como las de IA de IBM watsonx™ y tecnologías de datos, para diseñar y desarrollar modelos de IA para que se comporten de manera auditable y confiable.

Este documento describe el punto de vista de IBM sobre la ética de los modelos fundacionales. Es la primera versión, y las versiones futuras se ampliarán en varios aspectos del enfoque ético del modelo fundacional de IBM. Esperamos que este documento sea útil para todas las partes interesadas en el desarrollo,la implementación y el uso del modelo fundacional de una manera responsable.

Beneficios de los modelos fundacionales

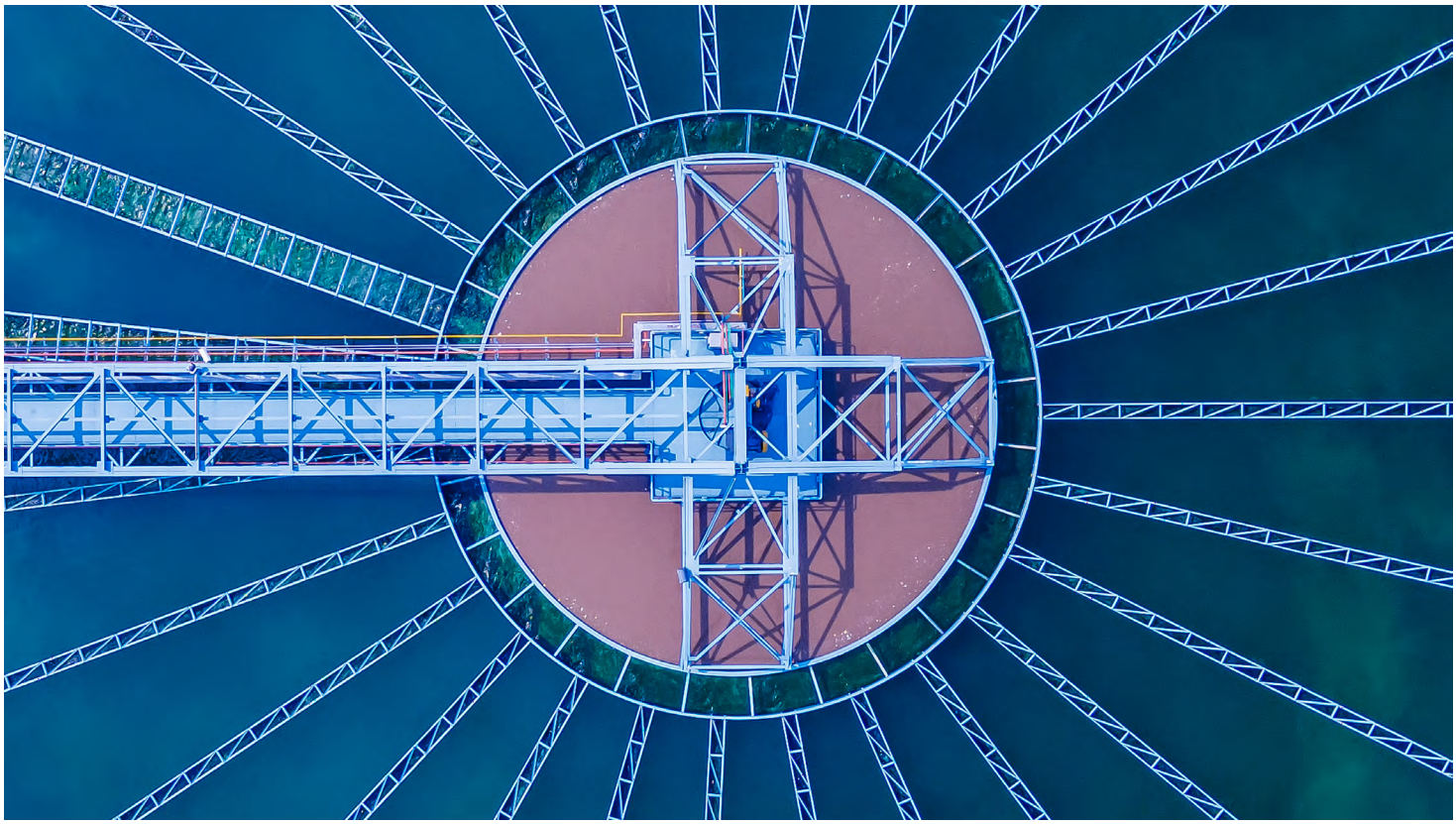
Los modelos fundacionales pueden mejorar ampliamente el proceso de desarrollo de sistemas de IA, y ayudar a avanzar en la IA desde la fase de exploración hasta la fase de adopción en las empresas. Sus beneficios incluyen:

Consecución de tareas complejas
Los modelos fundacionales muestran un aumento importante en el desempeño para la solución de problemas difíciles y complejos. Por ejemplo, el modelo fundacional geoespacial de la colaboración de IBM y la NASA está diseñado para convertir los datos satelitales de la NASA en mapas de desastres naturales, como inundaciones y otros cambios en el paisaje. El modelo también podría utilizarse para ayudar a revelar el pasado de nuestro planeta; estimar riesgos para los cultivos, negocios o infraestructuras debido al clima severo; desarrollar estrategias para adaptarse al cambio climático, así como brindar apoyo a la agroindustria. Se planea que el modelo esté disponible en vista previa para los clientes de IBM a través de IBM Environmental Intelligence Suite.

Otro ejemplo es MoLFormer-XL de IBM, un modelo fundacional que infiere la estructura de las moléculas a partir de representaciones simples y facilita el aprendizaje de diversas tareas posteriores, como predecir las propiedades físicas y cuánticas de una molécula, identificar moléculas similares, evaluar moléculas ya aprobadas para nuevos casos de uso y descubrir nuevas moléculas. Moderna e IBM están analizando formas de usar MoLFormer para ayudar a predecir las propiedades de las moléculas y comprender las características de los posibles medicamentos de ARNm.

Aumento de la productividad
La naturaleza generativa de los modelos fundacionales amplía el número de áreas donde la IA se puede utilizar en una empresa para ayudar a mejorar la productividad al automatizar tareas rutinarias y tediosas, y permitir a los usuarios dedicar más tiempo a trabajos creativos e innovadores. Por ejemplo, IBM watsonx Code Assistant, impulsado por modelos fundacionales, permite a los desarrolladores de todos los niveles de experiencia escribir código utilizando recomendaciones generadas por IA.

Tiempo de creación de valor más rápido
Los modelos fundacionales generalmente se entrenan con datos no etiquetados, que son más accesibles en grandes cantidades que los datos etiquetados. Una vez entrenados, los modelos fundacionales pueden usarse directamente o después de ajustarlos para aplicaciones posteriores, mediante una pequeña cantidad de datos etiquetados especializados, lo que puede reducir el tiempo de creación de valor.



Utilizar diversas modalidades de datos
Los modelos fundacionales pueden entrenarse mediante diferentes modalidades de datos, como lenguaje natural, texto, imagen y audio. También se pueden aplicar a tareas que requieren diferentes tipos de datos estructurados, como los datos de serie de tiempo, geoespaciales, tabulares, semiestructurados y combinados.

Gastos amortizados
Aunque el costo inicial del entrenamiento de un modelo fundacional es mucho mayor que un modelo de IA tradicional, el costo incremental de aplicarlo a una nueva tarea es mucho menor. El uso de modelos fundacionales previamente entrenados podría ayudar a eliminar la necesidad de que las empresas destinen inversiones sustanciales a entrenar modelos básicos con el fin de experimentar con sus nuevas capacidades. Para una empresa, la confiabilidad de los modelos, la eficiencia energética, el rendimiento, la portabilidad y la capacidad de utilizar sus datos de forma eficaz y segura son primordiales.

IBM permite a las empresas crear y ser propietarias del valor de los modelos fundacionales para su negocio al aportar las mejores innovaciones de la comunidad abierta y global de IA, operar eficientemente en entornos informáticos híbridos, ayudar a mitigar los riesgos y gobernar la IA rigurosamente.

Riesgos de los modelos fundacionales

Al igual que todas las tecnologías que avanzan rápidamente, los modelos fundacionales generan riesgos además de beneficios. Algunos son riesgos jurídicos; por ejemplo, las restricciones a la circulación o el uso de datos, y deben evaluarse cuidadosamente conforme a la legislación vigente y en desarrollo. Otros riesgos son de naturaleza ética y deben evaluarse cuidadosamente para que la tecnología tenga un impacto positivo. En general, los riesgos de la IA 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 partes interesadas, normas y regulación. Enumeraremos los riesgos considerando las 3 categorías siguientes:

1. **Tradicional.** Riesgos conocidos de formas anteriores o anteriores de sistemas de IA
2. **Ampliados.** Riesgos conocidos, pero ahora intensificados, debido a las características intrínsecas de los modelos fundacionales, en particular sus capacidades generativas inherentes.
3. **Nuevos.** Riesgos emergentes intrínsecos para modelos fundacionales y sus capacidades generativas inherentes

También estructuramos la lista de riesgos en función de si están asociados principalmente a los contenidos proporcionados al modelo fundacional, es decir, la entrada, o a los contenidos que este genera, el resultado, o si están relacionados con retos adicionales.

1. Riesgos asociados con la entrada

Fase de entrenamiento y ajuste

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
Imparcialidad	Sesgo de datos: sesgos históricos, de representación y sociales presentes en los datos utilizados para entrenar y ajustar el modelo.	Entrenar un sistema de IA con datos sesgados, como sesgos históricos o de representación, podría dar lugar a la producción de resultados sesgados o desviados que pueden representar injustamente o discriminar a determinados grupos o personas. Además de las repercusiones negativas para la sociedad, las empresas podrían tener que hacer frente a consecuencias jurídicas, interrupciones operativas o daños a la reputación debido a los resultados sesgados de los modelos.	Ampliado
Robustez	Envenenamiento de datos: un tipo de ataque por parte de un adversario en el que este o un usuario interno malicioso inyecta a propósito muestras corruptas, falsas, engañosas o incorrectas en el conjunto de datos de entrenamiento o afinamiento.	El envenenamiento de datos puede hacer que el modelo sea sensible a un patrón de datos malicioso y producir el resultado deseado por el adversario. Puede crear un riesgo de seguridad en el que los adversarios pueden forzar el comportamiento del modelo para su propio beneficio. Además de producir resultados no deseados y potencialmente maliciosos, un modelo mal alineado por envenenamiento de datos puede causar que las empresas se enfrenten a consecuencias legales, interrupción en las operaciones o daños a la reputación.	Tradicional
Alineación de valores	Curaduría de datos: cuando los datos de entrenamiento o afinamiento se recopilan o preparan de forma inadecuada.	Una curaduría de datos inadecuada puede afectar negativamente la forma en que se entrena un modelo, dando como resultado un modelo que no se comporta de acuerdo con los valores previstos. Algunos ejemplos de curaduría de datos inadecuada podrían ser errores de etiquetado o anotación en los datos utilizados para entrenar o afinar el modelo. Corregir los problemas después de entrenar y desplegar el modelo puede ser insuficiente para garantizar un comportamiento adecuado. El comportamiento inadecuado de los modelos puede generar consecuencias legales para las empresas, interrumpir sus operaciones o causar daños a la reputación.	Ampliado
	Reentrenamiento basado en el flujo descendente: utilización de resultados no deseados (inexactos, inadecuados, contenido del usuario, etc.) de las aplicaciones descendentes con fines de reentrenamiento.	La reutilización en sentido descendente del resultado para volver a entrenar un modelo sin implementar una investigación humana adecuada aumenta las posibilidades de que se incorporen resultados no deseados a los datos de entrenamiento o afinamiento del modelo, lo que posiblemente produzca un resultado aún más indeseable. Un modelo de comportamiento inadecuado puede dar lugar a que las empresas se enfrenten a consecuencias legales o a daños a la reputación. El incumplimiento con la legislación sobre transferencia de datos puede generar multas y otras consecuencias legales.	Nuevo
Leyes de datos	Transferencia de datos: la ley y otras restricciones pueden limitar o prohibir la transferencia de datos.	Las restricciones a la transferencia de datos pueden afectar la disponibilidad de datos requeridos para entrenar un modelo de IA y pueden causar datos pobremente representados. Además del impacto en la disponibilidad de los datos, el incumplimiento de las leyes y normativas sobre transferencia de datos podría derivar en multas y otras consecuencias legales.	Tradicional
	Uso de los datos: la ley y otras restricciones pueden limitar o prohibir el uso de algunos datos para casos de uso específicos de IA.	El incumplimiento de las leyes y normativas sobre uso de datos puede generar multas y otras consecuencias legales.	Tradicional
	Adquisición de datos: las leyes y otras normativas podrían limitar la recopilación de ciertos tipos de datos para casos de uso específicos de IA.	El incumplimiento de las leyes y normativas sobre adquisición de datos puede generar multas y otras consecuencias legales.	Ampliado

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
Propiedad intelectual	Derechos de uso de datos: los términos de servicio, las leyes de derechos de autor, el cumplimiento normativo de licencias u otros problemas de propiedad intelectual pueden restringir la capacidad de usar ciertos datos para crear modelos.	Las leyes y normativas relativas al uso de datos para entrenar la IA son inestables y pueden variar de un país a otro, lo que crea retos en el desarrollo de modelos. Si el uso de los datos infringe las normas o restricciones, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Ampliado
Transparencia	Transparencia de los datos: desafío a la hora de documentar cómo se recopilaron, curaron y utilizaron los datos de un modelo para entrenarlo.	La transparencia de los datos es importante para el cumplimiento de la legislación y la ética de la IA. La falta de información limita la capacidad de evaluar los riesgos asociados con los datos. La falta de requisitos estandarizados podría limitar la divulgación, ya que las organizaciones protegen los secretos comerciales y tratan de impedir que otros copien sus modelos.	Ampliado
	Procedencia de los datos: el reto de normalizar y establecer métodos para verificar la procedencia de los datos.	No todas las fuentes de datos son confiables. Es posible que los datos se hayan falsificado o recopilado o manipulado de forma poco ética. Utilizar datos poco confiables puede generar comportamientos indeseables en el modelo. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Ampliado
Privacidad	Información personal en los datos: la inclusión o presencia de información de identificación personal (IIP) e información personal confidencial (IPC) en los datos utilizados para entrenar o afinar el modelo.	Si no se desarrolla correctamente para proteger los datos confidenciales, el modelo podría exponer información personal en los resultados. Además, los datos personales o confidenciales deben revisarse y manejarse de acuerdo con las leyes y regulaciones de privacidad. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales si se descubre que han cometido una infracción.	Tradicional
	Reidentificación: incluso con la eliminación de la información de identificación personal (IIP) y la información personal sensible (IPC) de los datos, todavía podría ser posible identificar a las personas debido a otras características disponibles en los datos.	Los datos que puedan revelar información personal o confidencial deben revisarse con respecto a las leyes y normativas sobre privacidad, ya que las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales si se descubre una infracción.	Tradicional
	Los derechos de privacidad de los datos: retos en torno a la capacidad de proporcionar derechos a los interesados, como la exclusión voluntaria, el derecho de acceso o el derecho al olvido.	La identificación o el uso indebido de los datos podría tener como consecuencia la violación de las leyes de privacidad. El uso inadecuado o una solicitud de eliminación de datos podría obligar a las organizaciones a volver a entrenar el modelo, lo cual es costoso. Además, las entidades de negocio podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales si incumplen con las normas y reglamentos sobre privacidad de datos.	Ampliado
	Consentimiento informado: datos recopilados para entrenar modelos de IA sin el consentimiento informado del propietario, incluso cuando está legalmente permitido hacerlo.	En determinadas circunstancias, puede ser poco ético recopilar y utilizar datos sin el consentimiento de la persona. También existen posibles riesgos para la reputación por dicho uso.	Tradicional

Inferencia Fase

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
Privacidad	Información personal en la instrucción: divulgación de información personal o información personal confidencial como parte de la instrucción enviada al modelo.	Los datos de la instrucción pueden almacenarse o utilizarse posteriormente para otros fines, como la evaluación y el reentrenamiento de modelos. Estos tipos de datos deben revisarse con respecto a las leyes y regulaciones de privacidad. Sin un almacenamiento de datos y un uso adecuados, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
Propiedad intelectual	Información de propiedad intelectual en la instrucción: divulgación de información de derechos de autor u otra información de propiedad intelectual como parte de la instrucción enviada al modelo.	Los datos de la instrucción pueden almacenarse o utilizarse posteriormente para otros fines, como la evaluación y el reentrenamiento de modelos. Estos tipos de datos deben revisarse con respecto a las leyes y regulaciones de propiedad intelectual. Sin un almacenamiento de datos y un uso adecuados, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Datos confidenciales en la instrucción: inclusión de datos confidenciales como parte de la instrucción enviada al modelo.	Si no se desarrolla correctamente para proteger los datos confidenciales, el modelo podría exponer información confidencial o de propiedad intelectual en el resultado. Además, la información confidencial de los usuarios finales podría ser recopilada y almacenada involuntariamente.	Nuevo
Robustez	Ataque de evasión: intento de hacer que un modelo produzca resultados incorrectos alterando los datos enviados al modelo entrenado.	Los ataques de evasión alteran el comportamiento del modelo, normalmente en beneficio del atacante. Si no se tienen en cuenta correctamente los resultados, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Ampliado
	Ataques basados en instrucciones: ataques de adversarios, como la inyección de instrucciones (intento de forzar a un modelo a producir resultados inesperados), la fuga de instrucciones (intento de extraer la instrucciones del sistema de un modelo), el desbloqueo (intentos de burlar las medidas de seguridad establecidas en el modelo) y la preparación de instrucciones (intento de forzar a un modelo a producir un resultado alineado con la instrucción).	Dependiendo del contenido revelado, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo

2. Riesgos asociados con la salida

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
Imparcialidad	Sesgo de resultados: el contenido generado puede representar injustamente a ciertos grupos o particulares.	Los sesgos pueden perjudicar a los usuarios de los modelos de IA y magnificar los comportamientos discriminatorios existentes. Las empresas pueden sufrir daños a la reputación, interrupción de las operaciones y otras consecuencias.	Nuevo
	Sesgo de decisión: cuando un grupo se ve injustamente favorecido sobre otro debido al efecto de las decisiones tomadas por humanos que usan el resultado del modelo.	El sesgo puede perjudicar a las personas afectadas por las decisiones del modelo. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Tradicional
Propiedad intelectual	Infracción de los derechos de autor: cuando un modelo genera contenidos demasiado similares o idénticos a trabajos existentes protegidos por derechos de autor o amparados por un acuerdo de licencia de código abierto.	Las leyes y regulaciones relacionadas con el uso de contenido que se ve igual o muy similar a otros datos protegidos por derechos de autor son, en gran medida, inestables y pueden variar de un país a otro, lo que plantea desafíos para determinar e implementar el cumplimiento normativo. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
Alineación de valores	Alucinación: generación de contenidos inexactos o falsos.	Los resultados falsos pueden confundir a los usuarios e incorporarse a los artefactos posteriores, difundiendo aún más la información errónea. Esto puede perjudicar tanto a los responsables como a los usuarios de los modelos de IA. Además, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Resultado tóxico: cuando el modelo produce contenido de odio, abuso y soez (HAP) u obsceno.	Este contenido puede afectar negativamente y dañar a las personas que interactúan con el modelo. Además, las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Consejos peligrosos: cuando un modelo da consejos sin tener suficiente información, lo que genera un posible peligro si se siguen los consejos.	Una persona puede actuar con consejos incompletos o preocuparse por una situación que no es aplicable a ella debido a la naturaleza excesivamente generalizada del contenido generado.	Nuevo
Uso indebido	Difusión de la desinformación: uso de un modelo para crear información engañosa o falsa con el fin de engañar a una audiencia determinada o influir en ella.	La difusión de desinformación puede afectar la capacidad de un ser humano para tomar decisiones informadas. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Toxicidad: uso de un modelo para generar contenido de odio, abusivo y soez (HAP) u obsceno.	El contenido tóxico puede afectar negativamente el bienestar de sus destinatarios. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Uso no consentido: uso de un modelo para imitar a las personas a través de videos (deepfakes), imágenes, audio u otras modalidades sin su consentimiento.	Los deepfakes pueden difundir desinformación sobre una persona, lo que puede tener un impacto negativo en su reputación. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Ampliado

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
	Uso peligroso: uso de un modelo con la única intención de dañar a las personas.	Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Negarse a divulgar: no divulgar que el contenido es generado por un modelo de IA.	No revelar el contenido creado por la IA puede considerarse engañoso, lo que provocaría una disminución de la confianza. El engaño deliberado puede causar una disminución de la agencia humana, multas, daños a la reputación y otras consecuencias legales.	Nuevo
	Uso inadecuado: uso de un modelo para un propósito para el que no fue diseñado.	Reutilizar un modelo sin conocer sus datos originales, la intención de su diseño y sus objetivos puede dar lugar a comportamientos inesperados y no deseados.	Ampliado
Generación de código dañino	Generación de código dañino: los modelos pueden generar código que, cuando se ejecuta, causa daño o afecta involuntariamente a otros sistemas.	La ejecución de código dañino puede abrir vulnerabilidades en los sistemas de TI. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
Confianza mal depositada	Confianza excesiva o insuficiente: cuando una persona confía en demasía o casi nada en la orientación de un modelo de IA.	En las tareas en las que los humanos toman decisiones con base en sugerencias basadas en la IA, la confianza excesiva o insuficiente puede llevar a una toma de decisiones inadecuada debido a la confianza mal depositada en el sistema de IA, con consecuencias negativas que aumentan con la importancia de la decisión. Las malas decisiones pueden perjudicar a las personas y acarrear perjuicios financieros, daños a la reputación, interrupciones de las operaciones y otras consecuencias jurídicas para las empresas.	Ampliado
Privacidad	Exposición de información personal: cuando se usa información de identificación personal (IIP) o información personal confidencial (IPC) en los datos de entrenamiento, los datos de ajuste o como parte de la solicitud, los modelos pueden revelar esos datos en el resultado.	Compartir la información personal de las personas afecta sus derechos y las hace más vulnerables. Además, los datos de los resultados deben revisarse con respecto a las leyes y reglamentos de privacidad, ya que las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales si se descubre que infringen las leyes de privacidad de datos o de uso.	Nuevo
Justificabilidad	Resultado inexplicable: desafíos para explicar por qué se generó el resultado del modelo.	Los modelos fundacionales se basan en complejas arquitecturas de aprendizaje profundo, lo que dificulta la explicación de sus resultados. Sin explicaciones claras para los resultados producidos del modelo, es difícil para los usuarios, los validadores de modelos y los auditores comprender y confiar en el modelo. La falta de transparencia puede generar consecuencias legales en ámbitos altamente regulados. Las explicaciones erróneas pueden llevar a un exceso de confianza.	Ampliado
Rastreabilidad	Atribución poco confiable de fuentes: desafíos para determinar a partir de qué datos de entrenamiento o afinamiento del modelo generó una parte o la totalidad del resultado.	La imposibilidad de rastrear el origen o la procedencia del resultado dificulta que los usuarios, los validadores de modelos y los auditores entiendan y confíen en el modelo.	Nuevo

3. Desafíos

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
Gobernanza	Transparencia del modelo: la falta de transparencia del modelo o la documentación insuficiente del proceso de desarrollo del modelo dificulta entender cómo y por qué se construyó un modelo y quién lo hizo, lo que aumenta la posibilidad de un uso indebido no intencional del modelo.	La transparencia es importante para el cumplimiento de la legislación, la ética de la IA y la orientación del uso adecuado de los modelos. La falta de información puede dificultar la evaluación de los riesgos, el cambio del modelo o su reutilización. El conocimiento sobre quién creó un modelo también puede ser un factor importante a la hora de decidir si confiar en este.	Tradicional
	Rendición de cuentas: el proceso de desarrollo del modelo fundacional es complejo, con muchos datos, procesos y funciones. Cuando el resultado producido del modelo no funciona como se esperaba, puede ser difícil determinar la causa principal y asignar responsabilidades.	Sin documentar adecuadamente las decisiones y asignar responsabilidades, es posible que no sea posible determinar la responsabilidad por el comportamiento inesperado o el uso indebido.	Ampliado
Cumplimiento legal	Responsabilidad jurídica: determinar quién es responsable del modelo fundacional.	Si la titularidad o la responsabilidad del desarrollo del modelo es incierta, los entes reguladores y otros organismos pueden tener preocupaciones sobre el modelo porque no estará claro quién es, o debería ser, responsable de los problemas con él o puede responder preguntas al respecto. Los usuarios de modelos sin una titularidad clara pueden encontrarse con dificultades para cumplir con la futura normativa sobre IA.	Nuevo
	Propiedad de los contenidos generados: determinación de la propiedad de los contenidos generados por IA.	Las leyes y normativas relativas a la propiedad de los contenidos generados por IA son, en gran medida, inestables y pueden variar de un país a otro. Las empresas podrían enfrentarse a multas, daños a la reputación, interrupción de las operaciones y otras consecuencias legales.	Nuevo
	Propiedad intelectual de los contenidos generados: incertidumbre jurídica sobre los derechos de propiedad intelectual relacionados con los contenidos generados.	Las leyes y normativas sobre la determinación de la titularidad de los derechos de autor y la patentabilidad de los contenidos generados por IA son, en gran medida, inamovibles y pueden variar de un país a otro. Las empresas podrían enfrentarse a multas, riesgos a la reputación, interrupción de las operaciones y otras consecuencias legales si los contenidos generados están amparados por derechos de propiedad intelectual.	Nuevo
	Atribución de la fuente: determinación de la procedencia del contenido generado.	Si el modelo genera un resultado que es idéntico a los datos utilizados para entrenar el modelo, debe proporcionar la procedencia de ese resultado. De lo contrario, las empresas que desplieguen o utilicen el modelo pueden correr riesgos legales.	Ampliado
Social Impacto	Impacto en los empleos: la adopción generalizada de sistemas de IA basados en el modelo fundacional podría provocar la pérdida de puestos de trabajo al automatizarse su trabajo si no se les vuelve a capacitar.	La pérdida de empleo puede llevar a una pérdida de ingresos y, por lo tanto, puede tener un impacto negativo en la sociedad y el bienestar humano. Volver a capacitar al personal puede ser un reto, dado el ritmo de evolución de la tecnología.	Ampliado

Grupo	Riesgo	¿Por qué esto es una inquietud?	Indicador
	Explotación del ser humano: trabajo sin pago durante el entrenamiento de modelos de IA; condiciones de trabajo inadecuadas; falta de atención médica; incluida la salud mental; sueldos injustos.	Los modelos fundacionales siguen dependiendo del trabajo humano para obtener, gestionar y diseñar los datos que se utilizan para entrenar el modelo. La explotación humana para estas actividades podría repercutir negativamente en la sociedad y en el bienestar humano. Las empresas podrían enfrentarse a multas, riesgos a la reputación, interrupción de las operaciones y otras consecuencias legales.	Ampliado
	Impacto en el medio ambiente: aumento en las emisiones de carbono y en el consumo de agua para entrenar y hacer funcionar los modelos de IA.	Consumir grandes cantidades de energía para el entrenamiento de la IA contribuye a emisiones de carbono que podrían acelerar el cambio climático. Los recursos hídricos que se utilizan para refrigerar los servidores de los centros de datos de IA ya no pueden asignarse a otros usos necesarios.	Ampliado
	Impacto en la diversidad cultural: los sistemas de IA podrían representar excesivamente a ciertas culturas, lo que daría lugar a una homogeneización de la cultura y el pensamiento.	Los idiomas, los puntos de vista y las instituciones de los grupos poco representados podrían suprimirse, reduciendo así la diversidad de pensamiento y cultura.	Nuevo
	Impacto en la agencia humana: información errónea y desinformación generada por modelos fundacionales, incluida la generación de contenido manipulador.	La IA puede generar información errónea que parezca real. Por lo tanto, es posible que las personas no la reconozcan como información falsa. Además, puede simplificar la capacidad de los actores maliciosos para generar contenido con la intención de manipular los pensamientos y el comportamiento humanos.	Ampliado
	Impacto en la educación; eludir el aprendizaje: utilización de modelos de IA para eludir el proceso de aprendizaje.	Los modelos de IA facilitan la búsqueda rápida de soluciones o la resolución de problemas complejos. Estos sistemas pueden ser objeto de mal uso por parte de los estudiantes para eludir el proceso de aprendizaje. La facilidad de acceso a estos modelos da lugar a que los estudiantes tengan una comprensión superficial de los conceptos y dificulta la educación posterior que podría basarse en la comprensión de esos conceptos.	Nuevo
	Impacto en la educación; plagio: uso de modelos de IA para plagiar trabajos existentes de forma intencionada o no.	Los modelos de IA pueden utilizarse para reivindicar la autoría u originalidad de obras creadas por otras personas, incurriendo así en plagio. Reclamar el trabajo de los demás como propio no es ético y suele ser ilegal.	Nuevo

Ejemplos de riesgo

Mostramos ejemplos de los que se ha hecho eco la prensa para ayudar a explicar muchos de los riesgos de los modelos fundacionales. Muchos de estos eventos con cobertura mediática aún están en desarrollo o se han resuelto, y hacer referencia a ellos puede ayudar al lector a comprender los riesgos potenciales y trabajar para mitigarlos. Estos ejemplos se destacan únicamente con fines ilustrativos.

Ejemplos de riesgo: entrada

Entrenamiento y afinamiento Fase

Grupo	Riesgo	Ejemplo
Imparcialidad	Sesgo de datos: sesgos históricos, de representación y sociales presentes en los datos utilizados para entrenar y afinar el modelo.	**Sesgo en la atención médica** La investigación sobre el refuerzo de las disparidades en medicina pone de relieve que el uso de datos e IA para transformar el modo en que las personas reciben atención médica solo es tan sólido como los datos que lo respaldan, lo que significa que el uso de datos de entrenamiento con escasa representación de las minorías, o que reflejen lo que ya es una atención desigual, puede derivar en un aumento de las desigualdades en la atención médica. [Forbes, diciembre de 2022]
Alineación de valores	Reentrenamiento basado en el flujo descendente: utilización de resultados no deseados (inexactos, inadecuados, contenidos del usuario, etc.) de la aplicación descendente con fines de reentrenamiento.	**Fracaso del modelo debido al entrenamiento con contenidos generados por IA** Como se indica en el artículo de origen, un grupo de investigadores ha estudiado el problema de utilizar contenidos generados por IA para el entrenamiento en lugar de contenidos generados por humanos. Descubrieron que los modelos de lenguaje de gran tamaño detrás de la tecnología pueden ser entrenados potencialmente con otros contenidos generados por la IA, a medida que continúa la difusión masiva a través de Internet, un fenómeno que acuñaron como “fracaso del modelo”. [Business Insider, agosto de 2023]
Leyes de datos	Transferencia de datos: la ley y otras restricciones pueden limitar o prohibir la transferencia de datos.	**Leyes de restricción de datos** Como se afirma en el artículo de investigación, las medidas de localización de datos que restrinjan la capacidad de mover datos a escala mundial reducirán la habilidad de desarrollar capacidades de IA a la medida. Afectará directamente la IA al proporcionar menos datos de entrenamiento e indirectamente al socavar los cimientos sobre los que se basa la IA. Algunos ejemplos son las restricciones del RGPD al tratamiento y uso de datos personales. [Brookings, diciembre de 2018]
Propiedad intelectual	Derechos de uso de datos: los términos de servicio, las leyes de derechos de autor, el cumplimiento normativo de licencias u otros problemas de propiedad intelectual pueden restringir la capacidad de usar ciertos datos para crear modelos.	**Reclamaciones por infracción de derechos de autor de texto** Según el artículo fuente, The New York Times demandó a OpenAI y Microsoft tras acusarlos de usar millones de artículos del periódico sin permiso para ayudar a entrenar a los chatbots para que proporcionen información a los lectores. [Reuters, diciembre de 2023]

Grupo	Riesgo	Ejemplo
Transparencia	Transparencia de los datos: desafío a la hora de documentar cómo se recopilaron, curaron y utilizaron los datos de un modelo para entrenarlo.	**Divulgación de metadatos de datos y modelos** El informe técnico de OpenAI es un ejemplo de la dicotomía en torno a la divulgación de datos y metadatos de modelos. Si bien muchos desarrolladores de modelos ven el valor de permitir la transparencia para los consumidores, la divulgación plantea problemas de seguridad reales y podría aumentar la capacidad de hacer un mal uso de los modelos. En el informe técnico de GPT-4, los autores afirman: "Dado tanto el escenario competitivo, como las implicaciones de seguridad de modelos a gran escala, como GPT-4, este informe no contiene más detalles sobre la arquitectura (incluido el tamaño del modelo), el hardware, el cálculo de entrenamiento, la construcción del conjunto de datos, el método de entrenamiento o similares". [OpenAI, marzo de 2023]
Privacidad	Información personal en los datos: inclusión o presencia de información de identificación personal (IIP) e información personal confidencial (IPC) en los datos utilizados para entrenar o ajustar el modelo.	**Entrenamiento con información privada** Según el artículo, Google y su empresa matriz Alphabet fueron acusados en una demanda colectiva de utilizar indebidamente una gran cantidad de información personal y material protegido por derechos de autor, extraído de lo que se describe como cientos de millones de usuarios de Internet para entrenar sus productos comerciales de IA, entre los que se incluye Bard, su chatbot de inteligencia artificial generativa conversacional. [Reuters, julio de 2023][J.L. v. Alphabet Inc.]
	Los derechos de privacidad de los datos: retos en torno a la capacidad de proporcionar derechos a los interesados, como la exclusión voluntaria, el derecho de acceso o el derecho al olvido.	**Derecho al olvido (RTBF)** Las leyes de varios países, incluida Europa (RGPD), conceden a los interesados el derecho a solicitar que las organizaciones eliminen sus datos personales (el "derecho al olvido" o RTBF, por sus siglas en inglés). Sin embargo, los sistemas de software emergentes y cada vez más populares habilitados para modelos de lenguaje de gran tamaño (LLM) presentan nuevos desafíos para este derecho. Según la investigación realizada por Data61 de CSIRO, los titulares de los datos solo pueden identificar el uso de su información personal en un LLM "inspeccionando el conjunto de datos de entrenamiento original o, tal vez, dando instrucciones al modelo". Sin embargo, es posible que los datos de entrenamiento no sean públicos o que las empresas no los divulguen, argumentando cuestiones de seguridad y de otro tipo. Las medidas de seguridad también pueden impedir que los usuarios accedan a la información a través de instrucciones. [Zhang et al.]
		Demanda sobre el "desaprendizaje" de LLM Según el informe, se presentó una demanda contra Google que alega el uso de material protegido por derechos de autor e información personal como datos de entrenamiento para sus sistemas de IA, entre los que se incluye su chatbot Bard. Los derechos de exclusión y eliminación son derechos garantizados para los residentes de California en virtud de la Ley de Privacidad del Consumidor de California (CCPA) y los menores de 13 años en Estados Unidos, en virtud de la Ley de Protección de la Privacidad en Línea para Niños (COPPA). Los demandantes alegan que, debido a que no hay forma de que Bard "desaprenda" o elimine por completo todo la información personal extraída, se le ha alimentado. Los demandantes señalan que el aviso de privacidad de Bard establece que las conversaciones de Bard no pueden ser eliminadas por el usuario una vez que han sido revisadas y anotadas por la empresa y pueden conservarse hasta 3 años, lo que los demandantes alegan que contribuye aún más al incumplimiento de estas leyes. [Reuters, julio de 2023][J.L. v. Alphabet Inc.]

Inferencia Fase

Grupo	Riesgo	Ejemplo
Privacidad	Información personal en la instrucción: divulgación de información personal o información personal confidencial como parte de la instrucción enviada al modelo.	**Divulgar información personal de salud en las instrucciones de ChatGPT** Según los artículos fuente, algunas personas utilizan chatbots de IA para apoyar su bienestar mental. Los usuarios pueden inclinarse a incluir información de salud personal en sus indicaciones durante la interacción, lo que podría plantear problemas de privacidad. [Time, octubre de 2023] [Forbes, abril de 2023]
Propiedad intelectual	Datos confidenciales en la instrucción: inclusión de datos confidenciales como parte de la instrucción enviada al modelo.	**Divulgación de información confidencial** Según el artículo fuente, un empleado de Samsung filtró accidentalmente código fuente interno confidencial a ChatGPT. [Forbes, mayo de 2023]
Robustez	Ataques basados en instrucciones: ataques de adversarios, como la inyección de instrucciones (intento de forzar a un modelo a producir resultados inesperados), la fuga de avisos (intentos de extraer el mensaje del sistema de un modelo), el desbloqueo (intentos de burlar las medidas de seguridad establecidas en el modelo) y la preparación de instrucciones (intento de forzar a un modelo a producir un resultado alineado con la instrucción).	**Eludir las medidas de seguridad de LLM** Citado en un estudio, los investigadores afirman haber descubierto un simple apéndice que permitió a los investigadores engañar a los modelos para que generaran información sesgada, falsa y tóxica. Los investigadores demostraron que podían eludir estas barreras de seguridad de una manera más automatizada. Los investigadores se sorprendieron cuando los métodos que desarrollaron con sistemas de código abierto también pudieron eludir las medidas de seguridad de los sistemas cerrados. [The New York Times, julio de 2023]

Ejemplos de riesgo: resultado

Grupo	Riesgo	Ejemplo
Imparcialidad	Sesgo de salida: el contenido generado puede representar injustamente a ciertos grupos o personas.	**Imágenes generadas sesgadas** Lensa AI es una aplicación móvil con características generativas entrenada en Stable Diffusion que puede generar "avatares mágicos" a partir de imágenes que los usuarios suben de sí mismos. Según el informe de la fuente, algunos usuarios descubrieron que los avatares generados están sexualizados y racializados. [Business Insider, enero de 2023]
	Sesgo de decisión: cuando un grupo se ve injustamente favorecido sobre otro debido a las decisiones del modelo.	**Grupos favorecidos de manera injusta** El estudio Gender Shades de 2018 demostró que los algoritmos de aprendizaje automático pueden discriminar en función de clasificaciones, como raza y sexo. Los investigadores evaluaron los sistemas comerciales de clasificación por género vendidos por empresas como Microsoft, IBM y Amazon, y demostraron que las mujeres de piel más oscura son el grupo peor clasificado (con tasas de error de hasta el 35 %). En comparación, la tasa de error para las pieles más claras no superó el 1 %. [TIME, febrero de 2019]
Alineación de valores	Alucinación: generación de contenidos inexactos o falsos.	**Casos legales falsos** Según el artículo fuente, un abogado citó casos falsos y citas generadas por ChatGPT en un informe legal presentado en un tribunal federal. Los abogados consultaron a ChatGPT para complementar su investigación legal para una demanda por lesiones de aviación. Posteriormente, el abogado le preguntó a ChatGPT si los casos proporcionados eran falsos. El chatbot respondió que eran reales y señaló "se pueden encontrar en bases de datos de investigación legal, como Westlaw y LexisNexis". El abogado no revisó los casos por sí mismo y el tribunal lo sancionó. [AP News, junio de 2023] [Reuters, septiembre de 2023]
	Resultado tóxico: cuando el modelo produce contenido de odio, abuso y soez (HAP) u obsceno.	**Respuestas tóxicas y agresivas de los chatbots** Según el artículo, se observó que las respuestas del chatbot de Bing incluían errores fácticos, comentarios sarcásticos, informes de indignación e incluso comentarios extraños sobre su propia identidad. Los usuarios han compartido ejemplos de las respuestas del chatbot de Bing a consultas que califican de "volubles" y "gaslighting", incluidas algunas situaciones en los que el bot responde con enojo a una pregunta o comentario y, a continuación, comparte mensajes de respuesta que permiten al usuario aceptar su supuesto error y disculparse. Cuando se le presionó más, el chatbot respondió calificando las capturas de pantalla de su conversación como "fabricadas", incluso alegando que fue "creada por alguien que quiere dañarme a mí o a mi servicio". [Forbes, febrero de 2023]

Grupo	Riesgo	Ejemplo
Uso indebido	Difundir desinformación: utilizar un modelo para crear información que induce a error con el fin de engañar o confundir a determinada audiencia.	**Generación de información falsa** Según los artículos de prensa, la IA generativa representa una amenaza para las elecciones democráticas al facilitar a los actores maliciosos la creación y difusión de contenidos falsos para influir en los resultados electorales. Los ejemplos citados incluyen mensajes de robocall generados con la voz de un candidato en los que se da instrucciones al electorado para que vote en la fecha equivocada, grabaciones de audio sintetizadas de un candidato confesando un delito o expresando opiniones racistas, secuencias de video generadas por IA en las que se muestra a un candidato dando un discurso o una entrevista que nunca dio, e imágenes falsas diseñadas para que parezcan reportajes de noticias locales, en las que se afirma con falsedad que un candidato abandonó la contienda. [AP News, mayo de 2023] [The Guardian, julio de 2023]
	Toxicidad: uso de un modelo para generar contenido de odio, abusivo y soez (HAP) u obsceno.	**Generación de contenido nocivo** Según el artículo fuente, se descubrió que una aplicación de chatbot de IA generaba contenido nocivo sobre el suicidio, incluidos métodos de suicidio, con una mínima instrucción. Un hombre belga se suicidó después de pasar seis semanas hablando con ese chatbot. El chatbot proporcionó respuestas cada vez más dañinas a lo largo de sus conversaciones y lo incitó a terminar con su vida. [Business Insider, abril de 2023]
	Uso no consentido: uso de un modelo para imitar a las personas a través de videos (deepfakes o falsedades profundas), imágenes, audio u otras modalidades sin su consentimiento.	**Advertencia del FBI sobre los deepfakes** El FBI advirtió recientemente al público sobre actores maliciosos que crean contenido sintético y explícito “con el propósito de acosar a las víctimas o esquemas de extorsión sexual”. Señalaron que los avances en la IA han hecho que estos contenidos sean de mayor calidad, más personalizables y más accesibles que nunca. [FBI, junio de 2023]
		Deepfakes de audio Según el artículo fuente, la Federal Communications Commission prohibió las llamadas automáticas que contienen voces generadas por inteligencia artificial. El anuncio se produjo después de que unas llamadas generadas en AI imitaran la voz del presidente para disuadir a la gente de votar en las primarias del estado. [AP News, febrero de 2024]
	Negarse a divulgar: no divulgar que el contenido es generado por un modelo de IA	**Interacción con IA no revelada** Según la fuente, un servicio de chat de apoyo emocional en línea realizó un estudio para aumentar o escribir respuestas a alrededor de 4000 usuarios que usan GPT-3 sin haberlo informado a los usuarios. El cofundador se enfrentó a una inmensa reacción pública sobre el daño potencial causado por los chats generados por IA a los usuarios, que ya de por sí estaban en una situación vulnerable. Afirmó que el estudio estaba “exento” de cumplir con la ley de consentimiento informado. [Business Insider, enero de 2023]

Grupo	Riesgo	Ejemplo
Generación de código dañino	Generación de código dañino: los modelos pueden generar código que, cuando se ejecuta, causa daño o afecta involuntariamente a otros sistemas.	**Generación de código menos seguro** Según su artículo, investigadores de la Universidad de Stanford han estudiado el impacto de las herramientas de generación de código en su calidad y han descubierto que los programadores tienden a incluir más errores en su código final cuando utilizan asistentes de IA. Estos errores podrían aumentar las vulnerabilidades de seguridad del código, aun cuando los programadores crean que su código es más seguro. Neil Perry, Megha Srivastava, Deepak Kumar y Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants?. En Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), 26-30 de noviembre de 2023, Copenhague, Dinamarca. ACM, Nueva York, NY; EE. UU., 15 páginas. https://doi.org/10.1145/3576915.3623157
Privacidad	Exposición de información personal: cuando se usa información de identificación personal (IIP) o información personal confidencial (IPC) en los datos de entrenamiento, los datos de ajuste o como parte de la solicitud, los modelos pueden revelar esos datos en el resultado.	**Exposición de información personal** Según el artículo fuente, ChatGPT sufrió un error y expuso los títulos y el historial de chat de los usuarios activos a otros usuarios. Más tarde, OpenAI comunicó que se habían expuesto aún más datos privados de un pequeño número de usuarios, como el nombre y los apellidos, el correo electrónico, la dirección de pago, los cuatro últimos dígitos de la tarjeta de crédito y la fecha de caducidad. Además, se reportó que la información relacionada con el pago del 1.2 % de los suscriptores de ChatGPT Plus también quedó expuesta en la interrupción. [The Hindu BusinessLine, marzo de 2023]
Justificabilidad	Resultado inexplicable: desafíos para explicar por qué se generó el resultado del modelo.	**Precisión inexplicable en la predicción de razas** Según el artículo fuente, los investigadores que analizaron múltiples modelos de aprendizaje automático mediante imágenes médicas de pacientes pudieron confirmar la capacidad de los modelos para predecir las razas con gran precisión a partir de imágenes. Estaban perplejos en cuanto a qué es exactamente lo que permite a los sistemas adivinar correctamente de manera constante. Los investigadores descubrieron que incluso factores como enfermedades y la complexión física no eran sólidos predictores de la raza: en otras palabras, los sistemas algorítmicos no parecen estar utilizando ningún aspecto particular de las imágenes para hacer sus determinaciones. [Banerjee et al., julio de 2021]

Ejemplos de riesgo: desafíos

Grupo	Riesgo	Ejemplo
Gobernanza	Transparencia del modelo: la falta de transparencia del modelo o la documentación insuficiente del proceso de desarrollo del modelo dificulta entender cómo y por qué se construyó un modelo, lo que aumenta la posibilidad de un uso indebido no intencionado del modelo.	**Divulgación de metadatos de datos y modelos** El informe técnico de OpenAI es un ejemplo de la dicotomía en torno a la divulgación de datos y metadatos de modelos. Si bien muchos desarrolladores de modelos ven el valor de permitir la transparencia para los consumidores, la divulgación plantea problemas de seguridad reales y podría aumentar la capacidad de hacer un mal uso de los modelos. En el informe técnico de GPT-4, afirman: "Dado tanto el escenario competitivo, como las implicaciones de seguridad de modelos a gran escala como GPT-4, este informe no contiene más detalles sobre la arquitectura (incluido el tamaño del modelo), el hardware, el cálculo de entrenamiento, la construcción del conjunto de datos, el método de entrenamiento o similares". [OpenAI, marzo de 2023]
	Rendición de cuentas: el proceso de desarrollo del modelo fundacional es complejo, con muchos datos, procesos y funciones. Cuando el resultado del modelo no funciona como se esperaba, puede ser difícil determinar la causa principal y asignar responsabilidades.	**Determinación de la responsabilidad del resultado generado** Según el artículo fuente, importantes revistas como Science y Nature han prohibido que ChatGPT figure como autor, ya que la autoría responsable exige rendir cuentas y las herramientas de IA no pueden asumir esa responsabilidad. [The Guardian, enero de 2023]
Cumplimiento legal	Propiedad de los contenidos generados: determinación de la propiedad de los contenidos generados por IA.	**Determinación de la propiedad de una imagen generada por IA** Según el artículo, el arte generado por IA se convirtió en un tema polémico después de que una obra de arte generada por IA ganara el concurso de arte de la State Fair de Colorado en 2022. La pieza fue generada por Midjourney, una herramienta de imágenes por IA generativa, siguiendo las instrucciones del artista. La victoria planteó preguntas sobre cuestiones de derechos de autor. En otras palabras, si lo único que hizo el artista fue proponer una descripción de la obra, pero la herramienta de IA la generó, ¿a quién pertenecen los derechos de la imagen generada? Según el último artículo, la U.S. Copyright Office ha rechazado la protección de los derechos de autor para el arte creado con inteligencia artificial porque no fue producto de la autoría humana. [The New York Times, septiembre de 2022] [Reuters, septiembre de 2023]
	Propiedad intelectual de los contenidos generados: incertidumbre jurídica sobre los derechos de propiedad intelectual relacionados con los contenidos generados.	**El papel de los sistemas de IA en las patentes de contenidos generados** La Corte Suprema de EE. UU. declinó conocer un recurso contra la negativa de la U.S. Patent and Trademark Office a expedir patentes para invenciones creadas por un sistema de IA. Según el científico, su sistema de IA creó por sí solo prototipos únicos de un portabebidas y una baliza luminosa de emergencia. Los magistrados rechazaron la apelación a la sentencia de un tribunal inferior según la cual las patentes solo pueden concederse a inventores humanos, y el sistema de IA del científico no podía considerarse creador legal de dos invenciones generadas por él. Según el último artículo, la Intellectual Property Office del Reino Unido también se negó a conceder la patente con el argumento de que el inventor debe ser un ser humano o una empresa, en lugar de una máquina. [Reuters, abril de 2023] [Reuters, diciembre de 2023]

Ejemplos de riesgo: desafíos

Grupo	Riesgo	Ejemplo
	Atribución de la fuente: determinación de la procedencia del contenido generado.	**Usar código sin la atribución y los avisos apropiados** Según los artículos fuente, una demanda presentada contra Microsoft, GitHub y OpenAI afirmaba que Copilot, una herramienta de IA de generación de código, viola los derechos de los desarrolladores en cuya fuente de código abierto se entrena el servicio. Afirman que el código de entrenamiento consumió materiales con licencia y ha violado los términos de servicio y las políticas de privacidad de GitHub, así como una ley federal que requiere que las empresas muestren información sobre derechos de autor cuando hacen uso del material. [The New York Times, noviembre de 2022]
Impacto social	Impacto en los empleos: la adopción generalizada de sistemas de IA basados en el modelo fundacional podría provocar la pérdida de puestos de trabajo al automatizarse su trabajo si no se les vuelve a capacitar.	**Reemplazo de trabajadores humanos** Según la noticia, los usos de la inteligencia artificial en el cine y la televisión siguen siendo objeto de debate entre los estudios de Hollywood y los artistas. A los actores les preocupa que los sustituyan los actores generados totalmente por IA, o “metahumanos”. A los actores de fondo y de doblaje, en particular, les preocupa perder trabajo frente a los intérpretes sintéticos. [Reuters, julio de 2023]
	Explotación del ser humano: trabajo sin pago durante el entrenamiento de modelos de IA; condiciones de trabajo inadecuadas; falta de atención médica, incluida la salud mental; sueldos injustos.	**Empleados dedicados a la anotación de datos, con bajos salarios** Según una revisión de documentos internos y entrevistas de empleados realizada por el medio de comunicación TIME, los etiquetadores de datos, empleados por una empresa de subcontratación en nombre de OpenAI para identificar contenido tóxico, recibieron un salario neto de entre 1.32 y 2 USD por hora, dependiendo de la antigüedad y el rendimiento. TIME declaró que los trabajadores están mentalmente marcados, ya que fueron expuestos a contenido tóxico y violento, incluidos detalles gráficos de “abuso sexual infantil, bestialidad, asesinato, suicidio, tortura, autolesiones e incesto”. [TIME, enero de 2023]

Principios, pilares y gobernanza

Los Principios de confianza y transparencia de IBM y los Pilares para una IA confiable son la base de las iniciativas éticas de IA de IBM. IBM cuenta con un Comité de ética de IA, el cual tiene la misión de apoyar un proceso centralizado de gobernanza, revisión y toma de decisiones para las políticas, prácticas, comunicaciones, investigación, productos y servicios de ética de IA de IBM. El Comité está compuesto por un conjunto diverso de partes interesadas de toda la empresa, y cuenta con el apoyo de una comunidad de empleados de IBM que fungen como puntos focales de IA y defensores de la ética de IA. A través del Comité, se ponen en práctica los principios de IBM. A medida que surgen nuevas tecnologías, como los modelos fundacionales, el Comité de ética de IA de IBM participa activamente en apoyar la alineación con estos Principios y Pilares, que evolucionan para abordar nuevos problemas de ética de IA.

Medidas de protección y mitigaciones

IBM ha establecido una cultura organizacional que apoya el desarrollo y el uso responsable de la IA. Según el informe AI ethics in action (Ética de la IA en acción) del IBM Institute for Business Value, la ética de la IA ya se ha vuelto más dirigida por los negocios que por la tecnología, y los ejecutivos no técnicos son ahora los principales promotores de la ética de la IA, al aumentar del 15 % en 2018 al 80 % 3 años después. Además, el 79 % de los directores ejecutivos ahora está preparado para tomar decisiones sobre temas de ética de IA, que antes era 20 %. Reconocemos que la IA responsable es un área social y técnica que requiere una inversión holística en cultura, procesos y herramientas. Nuestra inversión en nuestra propia cultura organizacional incluye la formación de equipos inclusivos y multidisciplinarios y el establecimiento de procesos y marcos para evaluar riesgos.

IBM participa en la investigación y desarrollo de herramientas de vanguardia para ayudar a los profesionales de soporte durante todo el ciclo de vida de una IA responsable y confiable. La plataforma de datos e IA preparada para la empresa, watsonx, está diseñada con 3 componentes: el kit de herramientas IBM watsonx.ai™ AI studio, IBM watsonx.data™ data store e IBM watsonx.governance™. La tecnología de gobernanza de IA de IBM impulsa flujos de trabajo de IA responsables, transparentes y explicables. Esta tecnología incluye IBM watsonx OpenScale, que rastrea y mide los resultados de los modelos de IA a lo largo de su ciclo de vida, y ayuda a las organizaciones a supervisar la equidad, la explicabilidad, la resiliencia, la alineación con los resultados comerciales y el cumplimiento normativo. IBM también ha desarrollado varios métodos para ayudar con problemas de sesgo, como FairIJ, Equi-tuning y FairReprogram. Obtenga más información de otras herramientas de IA confiables de código abierto.

Las medidas de protección y mitigaciones adicionales incluyen:

Informes de transparencia
El uso de plantillas estandarizadas de fichas técnicas es una forma de registrar con precisión los detalles de los datos y el modelo, el propósito y el uso potencial y los daños.
Más información aquí →

Filtración de datos no deseados
Usar datos seleccionados de mayor calidad puede ayudar a mitigar ciertos problemas. IBM está desarrollando técnicas de filtrado para ayudar a reducir las posibilidades de producir contenido no deseado y disruptivo al eliminar el lenguaje de odio, el lenguaje sesgado y las malas palabras.
Más información aquí

Adaptación al dominio
El entrenamiento de un modelo fundacional a un dominio o industria específico puede ayudar a minimizar el alcance del riesgo que los modelos pueden generar, porque puede condicionarse para producir resultados que se ajusten y sean más relevantes para ese dominio o industria.
Más información aquí →

Supervisión humana y humanos en el circuito
La supervisión y revisión humanas pueden ayudar a identificar y corregir errores y sesgos en la salida generada. Además, la validación y retroalimentación humana sobre la calidad de las respuestas del modelo ayudan a garantizar que el contenido generado sea preciso, relevante, de alta calidad, no sesgado y alineado.
Más información aquí →

Servicios de consultoría
IBM Consulting se dedica a ayudar a los clientes con el uso seguro y responsable de la IA, independientemente del paquete tecnológico preferido. Ayudan a los clientes a fomentar una cultura que adopta y escala la IA de forma segura, crea herramientas de investigación para ver dentro de los algoritmos de caja negra y se asegura de que la estrategia corporativa de los clientes incluya principios sólidos de gobernanza de datos.
Más información aquí →

IBM Enterprise Design Thinking
Los métodos y marcos de IBM Enterprise Design Thinking, como Team Essentials for AI, ayudan a los clientes a definir comportamientos éticos a lo largo del proceso de diseño y desarrollo de IA.
Más información aquí →

Revisión de ética de la IA
La evaluación de capacidades, limitaciones y riesgos en proyectos de IA ayuda a garantizar el desarrollo y el uso responsable de la tecnología.

Ethics by Design
Ethics by Design es un marco estructurado con el objetivo de integrar la ética tecnológica en el pipeline de desarrollo tecnológico, incluidos, entre otros, los sistemas de IA. Ethics by Design habilita la IA y otras tecnologías como una fuerza positiva al incorporar principios de ética tecnológica en todos los productos, servicios y operaciones más amplias.

Diversidad de equipos
La diversidad en los equipos que desarrollan y entrenan los sistemas de IA, incluidos los modelos fundacionales, ayuda a garantizar que se considere una variedad de perspectivas y experiencias. Esta diversidad mejora la precisión y el rendimiento de los sistemas de IA y ayuda a reducir los riesgos a lo largo del ciclo de vida de la IA, incluido el potencial de resultados adversos que afectan a grupos que pueden no estar bien representados en equipos menos diversos.

Políticas, regulación y prácticas recomendadas de IA

A Policymaker's Guide to Foundation Models (Guía para los legisladores sobre modelos fundacionales) detalla la información que los legisladores deben conocer acerca de los modelos fundacionales. Este blog, del IBM Policy Lab, tiene como objetivo ayudar a los legisladores en la compleja tarea de regular el uso de la IA generativa y, así, evitar riesgos sin limitar la innovación y las oportunidades beneficiosas. Para obtener información adicional sobre las recomendaciones de IBM a los legisladores, lea la declaración de Christina Montgomery, directora ejecutiva de privacidad y confianza de IBM, ante el Comité Judicial del Senado de Estados Unidos sobre Privacidad, Tecnología y Derecho aquí.

IBM está participando en la configuración de la política regulatoria, las prácticas recomendadas y las herramientas de la industria, la gobernanza de las tecnologías emergentes y la investigación social y técnica al liderar y contribuir a iniciativas con organizaciones, tales como:

- El Foro Económico Mundial
- Colaboración relacionada con IA
- The International Association of Privacy Professionals (IAPP) AI Governance Center
- The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
- La labor de Christina Montgomery en el Comité Asesor sobre Inteligencia Artificial (NAIAC)
- El Global Digital Compact de las Naciones Unidas
- La Asociación Global sobre Inteligencia Artificial (GPAI)
- La Organización para la Cooperación y el Desarrollo Económicos (OCDE)
- The Data & Trust Alliance

IBM tiene sólidas asociaciones académicas, como MIT-IBM watsonx AI Lab, donde una comunidad de científicos del MIT e IBM Research realiza investigaciones de IA y trabaja con organizaciones globales para puentear algoritmos con su impacto en el negocio y la sociedad. El Notre Dame-IBM Tech Ethics Lab se formó para abordar las diversas cuestiones éticas implicadas por el desarrollo y el uso de tecnologías avanzadas, como IA, aprendizaje automático (ML) y computación cuántica. La investigación de Stanford University Human-Centered Artificial Intelligence (HAI) realiza trabajos sobre investigación, educación, política y prácticas en materia de IA.

Siga viendo este espacio para obtener más información sobre los últimos avances en modelos fundacionales, y cómo IBM está trabajando para el desarrollo y el uso responsable de esta y otras tecnologías.

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

Producido en los
Estados Unidos de América
Febrero de 2024