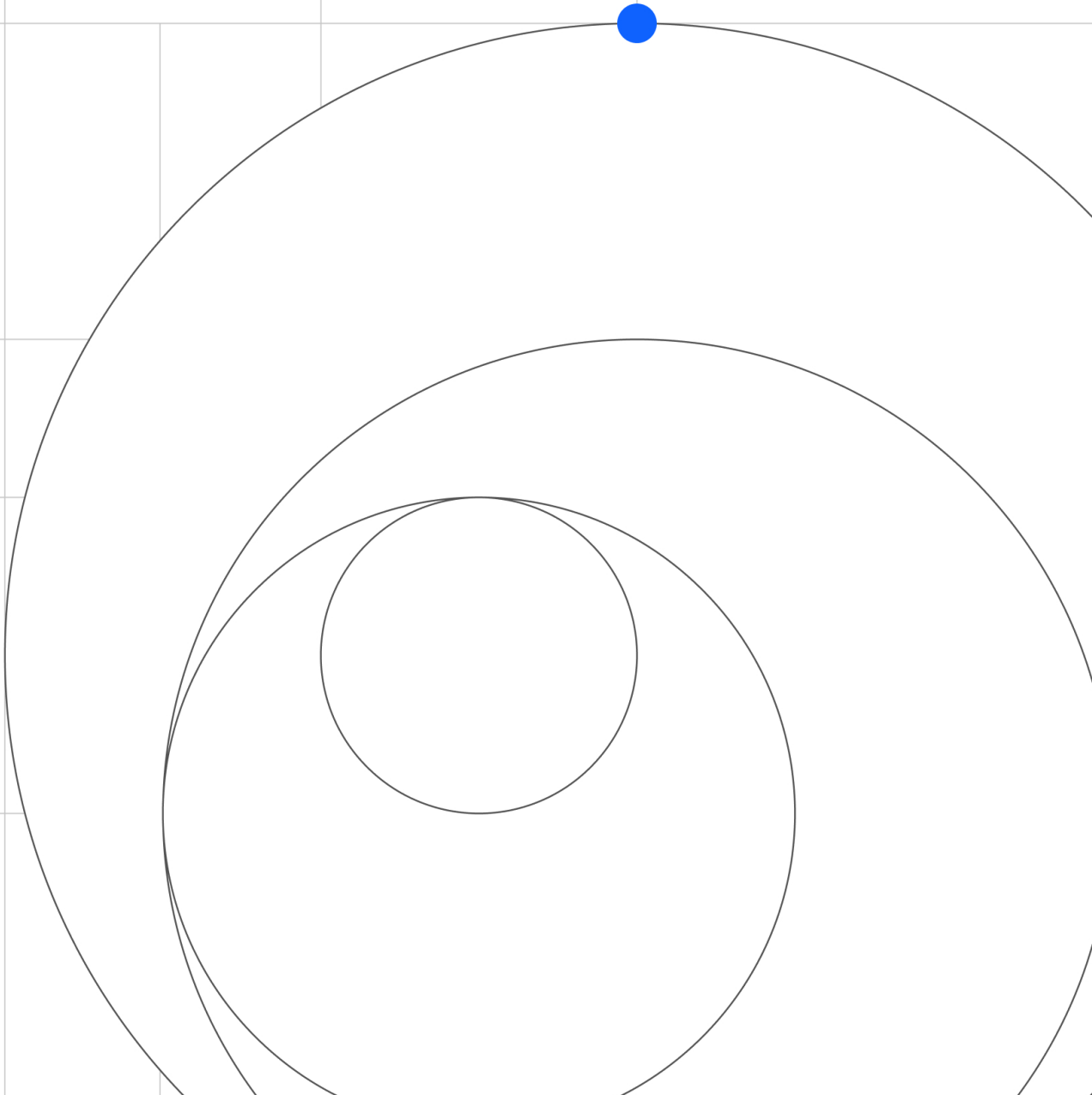
Modelos fundacionales: oportunidades, riesgos y mitigación



Atribución

Queremos expresar nuestro agradecimiento a las patrocinadoras ejecutivas del grupo de trabajo del Comité de Ética de la IA, Christina Montgomery y Francesca Rossi, y a las contribuciones de los miembros del grupo 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.

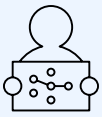
Índice

Resumen ejecutivo

El auge de los modelos fundacionales ofrece a las empresas nuevas posibilidades, pero también plantea y reformula preguntas sobre su diseño ético, desarrollo, implementación y uso. Según una encuesta reciente sobre IA generativa del IBM Institute for Business Value , las organizaciones ya expresan su preocupación por las cuestiones relacionadas con la confianza, en concreto como obstáculos a la inversión. Sus principales preocupaciones son la ciberseguridad (57%), la privacidad (51%) y la exactitud (47%). Muchas organizaciones ya se tomaban 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 concienciación sobre los riesgos y las posibles formas de mitigarlos es el primer paso crucial hacia la creación de sistemas de IA fiables.

En este documento, nosotros...

Exploramos las ventajas de los modelos fundacionales, incluida su capacidad para ejecutar tareas difíciles, su potencial para acelerar la adopción de la IA, su capacidad para aumentar la productividad y las ventajas económicas que aportan.



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

Cubrimos los principios, pilares y gobierno que forman la base de las iniciativas éticas de IA de IBM y sugerimos guardarraíles para la mitigación de riesgos.

Introducción

A medida que se extiende 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, construir 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 obtener lo mejor de ambos mundos: construir 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. Por ejemplo, IBM Research desarrolló modelos fundacionales para la inspección visual. Estos modelos fundacionales aprenden la representación general de las superficies de hormigón y de las pistas de aterrizaje y despegue, y pueden adaptarse a 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 posteriores. Los modelos fundacionales suelen ser modelos generativos a gran escala que se entrenan con datos no etiquetados utilizando autosupervisión. Como modelos a gran escala, los modelos fundacionales pueden incluir miles de millones de parámetros.

IBM es una empresa de cloud híbrido e inteligencia artificial con una larga trayectoria como administrador de datos responsable comprometido un uso ético de la IA. Al utilizar la fuerza 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 a construir una IA confiable en cualquier empresa. IBM también continúa invirtiendo en la construcción de nuevas tecnologías y plataformas de datos basadas en IA, como IBM watsonx, creada para diseñar y desarrollar modelos de IA 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, las versiones futuras abarcarán varios aspectos del enfoque ético del modelo fundacional de IBM. Esperamos que este documento sea útil para todas las partes interesadas en el desarrollo, implementación y uso del modelo fundacional de una manera responsable.

Beneficios de los modelos fundacionales

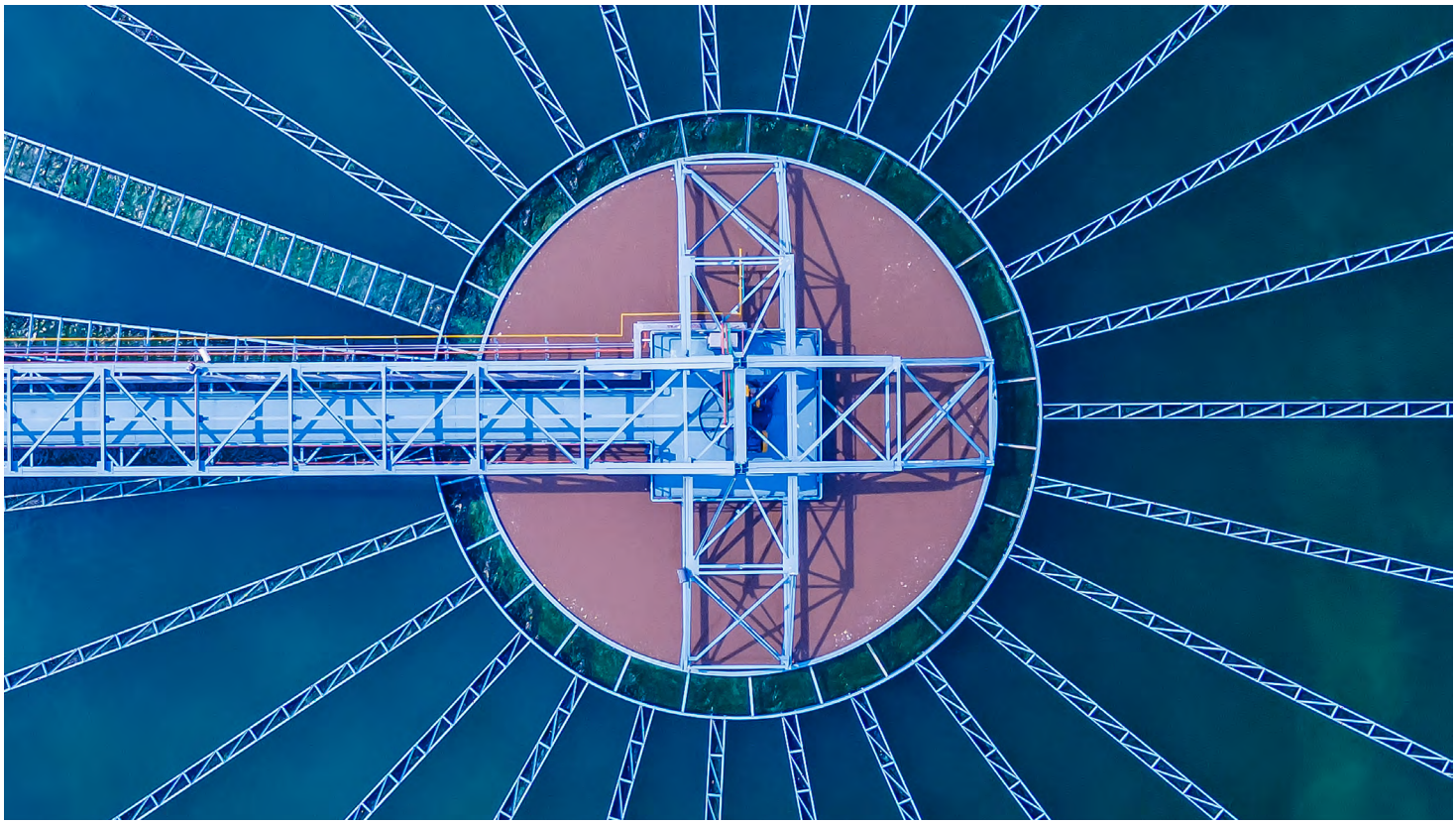
Los modelos fundacionales pueden mejorar significativamente el proceso de desarrollo de sistemas de IA y estimular su avance desde la fase de exploración hasta la fase de adopción en las empresas. Sus beneficios incluyen:

Realización de tareas complejas
Los modelos fundacionales muestran un aumento significativo del rendimiento en la resolución de problemas difíciles y complejos. Por ejemplo, el modelo geoespacial fundacional surgido de la colaboración entre 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 desvelar el pasado de nuestro planeta, estimar los riesgos que entrañan las inclemencias meteorológicas para los cultivos, las empresas o las infraestructuras, desarrollar estrategias de adaptación al cambio climático y ayudar a la agroindustria. Está previsto poner el modelo a disposición de los clientes de IBM en versión preliminar a través de la IBM Environmental Intelligence Suite.

Por poner otro ejemplo, el MoLFormer-XL de IBM es un modelo fundacional que deduce la estructura de las moléculas a partir de representaciones simples y facilita el aprendizaje de diversas tareas posteriores, como la predicción de las propiedades físicas y cuánticas de una molécula, la identificación de moléculas similares, la selección de moléculas ya aprobadas para nuevos casos de uso y el descubrimiento de nuevas moléculas. Moderna e IBM están estudiando formas de utilizar MoLFormer para ayudar a predecir las propiedades de las moléculas y comprender las características de los posibles medicamentos basados en ARNm.

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

Tiempo de valoración más rápido
Los modelos fundacionales generalmente están entrenados con datos sin etiquetar, que son más accesibles que los datos etiquetados cuando se trata de grandes cantidades. Una vez entrenados, los modelos fundacionales se pueden utilizar directamente o después de ajustarlos para aplicaciones en sentido descendente, utilizando una pequeña cantidad de datos etiquetados especializados, lo que puede reducir el tiempo de creación de valor.



Utilización de diversas modalidades de datos
Los modelos fundacionales pueden entrenarse utilizando diversas modalidades de datos, como lenguaje natural, texto, imagen y audio. También se pueden aplicar a tareas que requieran diferentes tipos de datos, como datos de series temporales, datos geoespaciales, datos tabulares, datos semiestructurados y datos de modalidad mixta, como el texto combinado con imágenes.

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

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

Riesgos de los modelos fundacionales

Al igual que todas las tecnologías que avanzan con rapidez, los modelos fundacionales tienen tanto riesgos como beneficios. Algunos son riesgos legales, como por ejemplo, las restricciones a la circulación o el uso de datos, y deben evaluarse cuidadosamente con arreglo a la legislación vigente y en evolución. Otros riesgos tienen una naturaleza ética y deben sopesarse con detenimiento para que la tecnología tenga un impacto positivo. En general, los riesgos de la IA plantean cuestiones sociotécnicas y deben abordarse y mitigarse mediante métodos sociotécnicos, incluidas herramientas informáticas, procesos de evaluación de riesgos, marcos éticos de IA, mecanismos de gobierno, consultas con las múltiples partes interesadas, normas y regulación. Vamos a enumerar los riesgos teniendo en cuenta las siguientes 3 categorías:

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

También estructuramos la lista de riesgos en función de su asociación con el contenido proporcionado al modelo fundacional (la entrada) o con el contenido generado por él (la salida), o de si guardan relación con desafíos adicionales.

1. Riesgos asociados con la entrada

Fase de formación y ajuste

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
Imparcialidad	Sesgos de los 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 con sesgos, como sesgos históricos o de representación, podría dar lugar a resultados sesgados o distorsionados que pueden representar injustamente o discriminar de otro modo a determinados grupos o individuos. Además de las repercusiones negativas en la sociedad, las entidades empresariales podrían enfrentarse a consecuencias legales, interrupciones en sus operaciones o daños a su reputación por los resultados sesgados de los modelos.	Amplificado
Robustez	Envenenamiento de los datos: un tipo de ataque adversario en el que un adversario o un usuario interno malicioso inyecta deliberadamente muestras corruptas, falsas, engañosas o incorrectas en el conjunto de datos de entrenamiento o de ajuste.	El envenenamiento de los datos puede hacer que el modelo sea susceptible a un patrón de datos malicioso y produzca 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 desajuste del modelo por envenenamiento de los datos puede hacer que las entidades empresariales se enfrenten a consecuencias legales, interrupción de las operaciones o daños a su reputación.	Tradicional
Alineación de valores	Conservación de los datos: cuando los datos de entrenamiento o ajuste se recopilan o preparan de forma inadecuada.	Una incorrecta conservación de los datos puede afectar negativamente al modo en que se entrena un modelo, dando como resultado un modelo que no se comporta de acuerdo con los valores previstos. Ejemplos de una incorrecta conservación de los datos podrían ser errores de etiquetado o anotación en los datos utilizados para el entrenamiento o el ajuste del modelo. Corregir los problemas después de entrenar e implementar el modelo puede no ser suficiente para garantizar un comportamiento adecuado. El comportamiento inadecuado del modelo puede dar lugar a que las entidades empresariales se enfrenten a consecuencias legales, a la interrupción de sus operaciones o a daños a su reputación.	Amplificado
	Reentrenamiento basado en aplicaciones en sentido descendente: uso de los resultados no deseados (inexactos, inadecuados, contenido del usuario, etc.) de las aplicaciones en sentido descendente con fines de reentrenamiento.	La reutilización de los datos de salida para volver a entrenar un modelo sin llevar a cabo un control humano adecuado aumenta las posibilidades de que se incorporen datos de salida no deseados a los datos de entrenamiento o ajuste del modelo, lo que posiblemente genere aún más datos de salida no deseados. Un comportamiento inadecuado del modelo puede hacer que las entidades empresariales se enfrenten a consecuencias legales o a daños a su reputación. El incumplimiento de las leyes de transferencia de datos puede acarrear multas y otras consecuencias legales.	Nuevo
Legislación sobre datos	Transferencia de datos: la ley y otras restricciones pueden limitar o prohibir la transferencia de datos.	Las restricciones en la transferencia de datos pueden afectar a la disponibilidad de los datos necesarios para entrenar un modelo de IA y pueden dar lugar a datos mal representados. Además del impacto en la disponibilidad de los datos, el incumplimiento de las leyes y normativas sobre transferencia de datos podría acarrear 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 la IA.	El incumplimiento de las leyes y normativas sobre el uso de datos podría acarrear multas y otras consecuencias legales.	Tradicional
	Adquisición de datos: las leyes y otras normativas pueden limitar la recopilación de ciertos tipos de datos para casos de uso específicos de la IA.	El incumplimiento de las leyes y normativas sobre la adquisición de datos podría acarrear multas y otras consecuencias legales.	Amplificado

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
Propiedad intelectual	Derechos de uso de los datos: las condiciones del servicio, las leyes de derechos de autor, el cumplimiento de licencias u otras cuestiones relacionadas con la propiedad intelectual pueden restringir la posibilidad de utilizar determinados datos para construir modelos.	Las leyes y normativas relativas al uso de datos para entrenar la IA no están establecidas 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 entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Amplificado
Transparencia	Transparencia de los datos: reto a la hora de documentar cómo se recopilaron, seleccionaron y utilizaron los datos para entrenar un modelo.	La transparencia de los datos es importante para el cumplimiento legal y la ética de la IA. La falta de información limita la capacidad de evaluar los riesgos asociados a los datos. La ausencia de requisitos estandarizados podría limitar la divulgación, ya que las organizaciones protegen los secretos comerciales e intentan impedir que otros copien sus modelos.	Amplificado
	Procedencia de los datos: reto en torno a la estandarización y el establecimiento de métodos para verificar la procedencia de los datos.	No todas las fuentes de datos son fiables. Los datos podrían haberse recopilado de forma poco ética, manipulado o falsificado. Utilizar datos poco fiables puede dar lugar a comportamientos no deseados en el modelo. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Amplificado
Privacidad	Información personal en los datos: inclusión o presencia de información de identificación personal (PII) e información personal confidencial (SPI) en los datos utilizados para el entrenamiento o el ajuste del modelo.	Si no se desarrolla adecuadamente para proteger los datos sensibles, el modelo podría exponer información personal en los resultados generados. Además, los datos personales o sensibles deben revisarse y tratarse de acuerdo con las leyes y normativas sobre privacidad. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus 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 (PII) y la información personal sensible (SPI) de los datos, aún podría ser posible identificar a las personas mediante otras características disponibles en los datos.	Los datos que puedan revelar información personal o sensible deben revisarse con respecto a las leyes y normativas sobre privacidad, ya que las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales si se descubre que han cometido una infracción.	Tradicional
	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 inadecuado de los datos podría dar lugar a la vulneración de las leyes de privacidad. Un uso inadecuado o una solicitud de eliminación de los datos podría obligar a las organizaciones a volver a entrenar el modelo, lo que resultaría muy caro. Además, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales si no cumplen las normas y reglamentos sobre privacidad de datos.	Amplificado
	Consentimiento informado: datos recopilados para entrenar modelos de IA sin el consentimiento informado del propietario, incluso cuando está legalmente permitido hacerlo.	En determinadas circunstancias, podría ser poco ético recopilar y utilizar datos sin el consentimiento de la persona. También existen posibles riesgos para la reputación derivados de dicho uso.	Tradicional

Inferencia Fase

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
Privacidad	Información personal en la instrucción: revelar información personal o información personal sensible 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 del modelo. Estos tipos de datos deben revisarse con respecto a las leyes y normativas sobre privacidad. Sin un almacenamiento y uso adecuados de los datos, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
Propiedad intelectual	Información de PI en la instrucción: revelar información sobre derechos de autor u otra información sobre 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 del modelo. Estos tipos de datos deben revisarse con respecto a las leyes y normativas sobre PI. Sin un almacenamiento y uso adecuados de los datos, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus 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 adecuadamente para proteger los datos confidenciales, el modelo podría exponer información confidencial o de propiedad intelectual en el resultado generado. Además, la información confidencial de los usuarios finales podría recopilarse y almacenarse de forma involuntaria.	Nuevo
Robustez	Ataque de evasión: intento de hacer que un modelo arroje resultados incorrectos perturbando los datos enviados al modelo entrenado.	Los ataques de evasión alteran el comportamiento del modelo, normalmente para beneficiar al atacante. Si los resultados no se contabilizan adecuadamente, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Amplificado
	Ataques basados en instrucciones: ataques adversarios como la inyección de instrucciones (intento de forzar a un modelo a producir un resultado inesperado), la filtración de instrucciones (intentos de extraer una instrucción del sistema de un modelo), el jailbreaking (intentos de quebrantar los límites de protección establecidos en el modelo) y la preparación de instrucciones o prompt priming (intento de forzar a un modelo a producir un resultado alineado con la instrucción).	Dependiendo del contenido revelado, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo

2. Riesgos asociados a la salida

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
Imparcialidad	Sesgo de salida: el contenido generado podría representar injustamente a ciertos grupos o individuos.	Los sesgos pueden perjudicar a los usuarios de los modelos de IA y magnificar los comportamientos discriminatorios existentes. Las entidades empresariales pueden enfrentarse a daños a su reputación, a la interrupción de sus operaciones y a otras consecuencias.	Nuevo
	Sesgo de decisión: cuando un grupo se ve injustamente favorecido frente a otro debido al efecto de las decisiones tomadas por el ser humano a partir de los resultados del modelo.	Los sesgos pueden perjudicar a las personas afectadas por las decisiones del modelo. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Tradicional
Propiedad intelectual	Infracción de los derechos de autor: cuando un modelo genera un contenido demasiado similar o idéntico a un trabajo existente protegido por derechos de autor o cubierto por un acuerdo de licencia de código abierto.	Las leyes y normativas relativas al uso de contenidos iguales o muy parecidos a otros datos protegidos por derechos de autor están en gran medida sin establecer y pueden variar de un país a otro, lo que plantea dificultades a la hora de determinar y aplicar su cumplimiento. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus 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 desinformación. Esto puede perjudicar tanto a los propietarios como a los usuarios de los modelos de IA. Además, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
	Resultado tóxico: cuando el modelo produce contenido reprobable, ofensivo y soez (HAP) u obsceno.	El contenido reprobable, ofensivo y soez u obsceno puede afectar negativamente y perjudicar a las personas que interactúan con el modelo. Además, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
	Consejos peligrosos: cuando un modelo proporciona un consejo sin disponer de suficiente información, lo que provoca un posible peligro si se sigue el consejo.	Una persona podría actuar siguiendo un consejo parcial o preocuparse por una situación que no le corresponde debido a la naturaleza excesivamente generalizada del contenido generado.	Nuevo
Uso indebido	Difundir la desinformación: utilizar un modelo para crear información engañosa o falsa con el fin de engañar o influir en un público determinado.	Difundir la desinformación puede afectar a la capacidad de un ser humano para tomar decisiones informadas. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
	Toxicidad: uso de un modelo para generar contenido reprobable, ofensivo y soez u obsceno.	El contenido tóxico podría afectar negativamente al bienestar de sus destinatarios. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
	Uso no consensuado: uso de un modelo para imitar a personas a través de vídeo (deepfakes), imágenes, audio u otras modalidades sin su consentimiento.	Los deepfakes pueden difundir desinformación sobre una persona, lo que puede repercutir negativamente en su reputación. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Amplificado

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
	Uso peligroso: uso de un modelo con la única intención de perjudicar a las personas.	Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
	No divulgación: no revelar que el contenido lo ha generado un modelo de IA.	No revelar el contenido creado por la IA puede considerarse engañoso y provocar una disminución de la confianza. El engaño intencionado puede dar lugar a una disminución de la acción humana, multas, daños en la reputación y otras consecuencias legales.	Nuevo
	Uso inadecuado: uso de un modelo para un fin para el que no ha sido 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 del modelo.	Amplificado
Generación de código malicioso	Generación de código malicioso: los modelos pueden generar código que, al ejecutarse, cause daños o afecte involuntariamente a otros sistemas.	La ejecución de código malicioso podría abrir vulnerabilidades en los sistemas de TI. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Nuevo
Confianza inapropiada	Exceso o falta de confianza: cuando una persona confía demasiado o muy poco en la orientación de un modelo de IA.	En las tareas en las que los humanos toman decisiones en función de sugerencias basadas en la IA, una confianza excesiva o insuficiente puede llevar a una toma de decisiones poco acertada debido a una confianza inapropiada 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 pueden acarrear perjuicios financieros, daños a la reputación, interrupción de las operaciones y otras consecuencias legales para las entidades empresariales.	Amplificado
Privacidad	Exposición de información personal: cuando se utiliza información de identificación personal (PII) o información personal confidencial (SPI) en los datos de entrenamiento, los datos de ajuste o como parte de la instrucción, los modelos podrían revelar esos datos en el resultado generado.	Compartir la IP de las personas afecta a sus derechos y las hace más vulnerables. Asimismo, los datos de salida deben revisarse con respecto a las leyes y normativas sobre privacidad, ya que las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales si se descubre que infringen las leyes sobre privacidad o uso de datos.	Nuevo
Explainability	Resultado inexplicable: retos a la hora de explicar por qué se generó la salida del modelo.	Los modelos fundacionales se basan en arquitecturas complejas de deep learning, lo que dificulta la explicación de sus resultados. Sin explicaciones claras de los resultados del modelo, es difícil que los usuarios, los validadores del modelo y los auditores lo entiendan y confíen en él. La falta de transparencia puede acarrear consecuencias legales en ámbitos muy regulados. Las explicaciones erróneas pueden conducir a un exceso de confianza.	Amplificado
Rastreabilidad	Atribución poco fiable de las fuentes: dificultades para determinar a partir de qué datos de entrenamiento o de ajuste el modelo generó una parte o la totalidad de sus resultados.	La incapacidad de rastrear el origen o la procedencia de los resultados dificulta la comprensión y la confianza en el modelo por parte de los usuarios, los validadores del modelo y los auditores.	Nuevo

3. Desafíos

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
Gobierno	Transparencia del modelo: la falta de transparencia del modelo o la documentación insuficiente del proceso de desarrollo del modelo dificulta la comprensión de cómo y por qué se construyó un modelo y quién lo construyó, lo que aumenta la posibilidad de un mal uso no intencionado del modelo.	La transparencia es importante para el cumplimiento legal, para la ética de la IA y para orientar el uso adecuado de los modelos. La falta de información puede dificultar la evaluación de riesgos, la modificación del modelo o su reutilización. Saber quién construyó un modelo también puede ser un factor importante a la hora de decidir si confiar en él.	Tradicional
	Responsabilidad: 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 raíz y asignar responsabilidades.	Si no se documentan adecuadamente las decisiones y se asigna la responsabilidad, puede que no sea posible determinar la responsabilidad por un comportamiento inesperado o un uso indebido.	Amplificado
Cumplimiento legal	Responsabilidad jurídica: determinar quién es responsable del modelo fundacional.	Si la propiedad o la responsabilidad del desarrollo del modelo es incierta, los reguladores y otras personas pueden tener dudas sobre el modelo porque no estará claro quién es (o debería ser) responsable de los problemas que surjan con él o puede responder a preguntas sobre el mismo. Los usuarios de modelos sin una propiedad 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 la IA están en gran medida sin definir y pueden variar de un país a otro. Las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus 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 los derechos de autor y la patentabilidad de los contenidos generados por la IA están en gran medida sin definir y pueden variar de un país a otro. Las entidades empresariales podrían enfrentarse a multas, riesgos para la reputación, interrupción de sus operaciones y otras consecuencias legales si los contenidos generados están amparados por derechos de propiedad intelectual.	Nuevo
	Atribución de la fuente: determinar la procedencia del contenido generado.	Si el modelo genera un resultado que es idéntico a los datos utilizados para entrenar el modelo, debe indicar la procedencia de ese resultado. No hacerlo puede poner en riesgo legal a las entidades empresariales que implementen o utilicen el modelo.	Amplificado
Social Impacto	Impacto en el empleo: la adopción generalizada de sistemas de IA basados en modelos fundacionales podría provocar la pérdida de empleo de las personas al automatizarse su trabajo si no se les vuelve a cualificar.	La pérdida de empleo puede conllevar una pérdida de ingresos y, por tanto, repercutir negativamente en la sociedad y en el bienestar humano. La recualificación puede ser un reto dado el ritmo de evolución de la tecnología.	Amplificado

Grupo	Riesgo	¿Por qué es esto motivo de preocupación?	Indicador
	Explotación humana: uso de trabajo fantasma en el entrenamiento de modelos de IA, condiciones de trabajo inadecuadas, falta de atención sanitaria, incluida la salud mental, e indemnizaciones injustas.	Los modelos fundacionales siguen dependiendo de la mano de obra humana para obtener, gestionar y diseñar los datos que se utilizan para entrenar el modelo. La explotación humana para estas actividades podría tener un impacto negativo en la sociedad y en el bienestar humano. Además, las entidades empresariales podrían enfrentarse a multas, daños a su reputación, interrupción de sus operaciones y otras consecuencias legales.	Amplificado
	Impacto en el medio ambiente: aumento de las emisiones de carbono y del uso de agua para entrenar y hacer funcionar los modelos de IA.	Consumir grandes cantidades de energía para el entrenamiento de la IA contribuye a las 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 la IA ya no pueden destinarse a otros usos necesarios.	Amplificado
	Impacto en la diversidad cultural: los sistemas de IA podrían representar excesivamente ciertas culturas, lo que daría lugar a una homogeneización de la cultura y el pensamiento.	Los idiomas, puntos de vista e instituciones de los grupos infrarrepresentados podrían suprimirse, lo que reduciría la diversidad de pensamiento y cultural.	Nuevo
	Impacto en la acción humana: información errónea y desinformación generadas por los modelos fundacionales, incluida la generación de contenidos manipuladores.	La IA puede generar información errónea que parezca real. Por lo tanto, es posible que la gente no la reconozca como información falsa. Además, puede simplificar la capacidad de los actores malintencionados para generar contenidos con la intención de manipular los pensamientos y el comportamiento humanos.	Amplificado
	Impacto en la educación – Eludir el aprendizaje: uso de los modelos de IA para evitar el proceso de aprendizaje.	Los modelos de IA facilitan la búsqueda rápida de soluciones o la resolución de problemas complejos. Estos sistemas pueden ser utilizados indebidamente por los alumnos para eludir el proceso de aprendizaje. La facilidad de acceso a estos modelos hace que los estudiantes tengan una comprensión superficial de los conceptos y dificulta la formación posterior que podría basarse en la comprensión de dichos conceptos.	Nuevo
	Impacto en la educación – Plagio: uso de modelos de IA para plagiar trabajos existentes de forma intencionada o involuntaria.	Los modelos de IA pueden utilizarse para reivindicar la autoría u originalidad de obras que fueron creadas por otras personas, incurriendo así en plagio. Reivindicar el trabajo de otros como propio es poco ético y a menudo ilegal.	Nuevo

Ejemplos de riesgo

Proporcionamos ejemplos cubiertos por la prensa para ayudar a explicar muchos de los riesgos de los modelos fundacionales. Muchos de estos sucesos cubiertos por la prensa aún están en evolución o se han resuelto, y hacer referencia a ellos puede ayudar al lector a comprender los posibles riesgos y a trabajar para mitigarlos. Estos ejemplos son solo ilustrativos.

Ejemplos de riesgo: entrada

Entrenamiento y ajuste Fase

Grupo	Riesgo	Ejemplo
Imparcialidad	Sesgos de los datos: sesgos históricos, de representación y sociales presentes en los datos utilizados para entrenar y afinar el modelo.	**Sesgos de la atención sanitaria** La investigación sobre el refuerzo de las desigualdades en medicina pone de relieve que el uso de datos e IA para transformar la forma en que las personas reciben atención sanitaria 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 conducir a un aumento de las desigualdades sanitarias. [Forbes, diciembre de 2022]
Alineación de valores	Reentrenamiento basado en aplicaciones en sentido descendente: uso de los resultados no deseados (inexactos, inadecuados, contenido del usuario, etc.) de las aplicaciones en sentido descendente con fines de reentrenamiento.	**Colapso del modelo debido al entrenamiento con contenidos generados por la IA** Como se indica en el artículo original, un grupo de investigadores ha estudiado el problema de utilizar contenidos generados por la IA para el entrenamiento en lugar de contenidos generados por humanos. Descubrieron que los grandes modelos de lenguaje que subyacen a la tecnología pueden entrenarse potencialmente con otros contenidos generados por la IA a medida que siguen propagándose en masa por Internet, un fenómeno que acuñaron como “colapso del modelo”. [Business Insider, agosto de 2023]
Legislación sobre 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 nivel mundial reducirán la capacidad de desarrollar capacidades de IA a medida. Afectará a la IA directamente al proporcionar menos datos de entrenamiento y de forma indirecta al socavar los cimientos sobre los que se construye 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 los datos: las condiciones del servicio, las leyes de derechos de autor, el cumplimiento de licencias u otras cuestiones relacionadas con la propiedad intelectual pueden restringir la posibilidad de utilizar determinados datos para construir modelos.	**Reclamaciones por infracción de los derechos de autor de textos** Según el artículo original, The New York Times demandó a OpenAI y Microsoft acusándoles de utilizar millones de artículos del periódico sin permiso para ayudar a entrenar chatbots que proporcionan información a los lectores. [Reuters, diciembre de 2023]

Grupo	Riesgo	Ejemplo
Transparencia	Transparencia de los datos: reto a la hora de documentar cómo se recopilaron, seleccionaron y utilizaron los datos para entrenar un modelo.	**Divulgación de datos y metadatos del modelo** El informe técnico de OpenAI es un ejemplo de la dicotomía existente en torno a la divulgación de los datos y los metadatos del modelo. Aunque muchos desarrolladores de modelos ven el valor de permitir la transparencia para los consumidores, la divulgación plantea verdaderos problemas de seguridad y podría aumentar la capacidad de hacer un mal uso de los modelos. En el informe técnico del GPT-4, los autores afirman: "Dado tanto el panorama competitivo como las implicaciones para la seguridad de los modelos a gran escala como el 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 (PII) e información personal confidencial (SPI) en los datos utilizados para el entrenamiento o el ajuste del modelo.	**Entrenamiento con información privada** Según el artículo, Google y su empresa matriz Alphabet fueron acusadas en una demanda colectiva de utilizar indebidamente una gran cantidad de información personal y material protegido por derechos de autor tomado 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.]
	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, por sus siglas en inglés)** Las leyes de múltiples lugares, incluida Europa (RGPD), conceden a los interesados el derecho a solicitar que las organizaciones eliminen sus datos personales ("Derecho al olvido", o RTBF). Sin embargo, los sistemas emergentes de software habilitados para grandes modelos de lenguaje (LLM), que son cada vez más populares, presentan nuevos retos en lo que respecta a este derecho. Según la investigación realizada por Data61 de CSIRO, los titulares de los datos solo pueden identificar el uso de sus datos personales en un LLM "inspeccionando el conjunto de datos de entrenamiento original o quizás dando instrucciones al modelo". Sin embargo, es posible que los datos de entrenamiento no sean públicos o que las empresas no los divulguen, alegando motivos de seguridad y de otro tipo. Los límites de protección también pueden impedir que los usuarios accedan a la información mediante la formulación de preguntas. [Zhang et al.] **Demanda sobre el desaprendizaje del 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 supresión están garantizados para los residentes en California en virtud de la CCPA y para los menores de 13 años en Estados Unidos en virtud de la COPPA. Los demandantes alegan que no hay forma de que Bard "desaprenda" o elimine por completo toda la IP desechada con la que ha sido alimentado. Los demandantes señalan que la declaración de confidencialidad 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: revelar información personal o información personal sensible como parte de la instrucción enviada al modelo.	**Revelar información sanitaria personal en las instrucciones de ChatGPT** Según los artículos originales, algunas personas utilizan chatbots de IA para contribuir a su bienestar mental. Los usuarios pueden tender a incluir información personal sobre su salud en sus instrucciones 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 original, un empleado de Samsung filtró accidentalmente código fuente interno sensible a ChatGPT. [Forbes, mayo de 2023]
Robustez	Ataques basados en instrucciones: ataques adversarios como la inyección de instrucciones (intento de forzar a un modelo a producir un resultado inesperado), la filtración de instrucciones (intentos de extraer una instrucción del sistema de un modelo), el jailbreaking (intentos de quebrantar los límites de protección establecidos en el modelo) y la preparación de instrucciones o prompt priming (intento de forzar a un modelo a producir un resultado alineado con la instrucción).	**Eludir los límites de protección del LLM** Citado en un estudio, los investigadores afirman haber descubierto un simple anexo de la instrucción que les permitía engañar a los modelos para que generasen información sesgada, falsa y tóxica. Los investigadores demostraron que podían sortear estos límites de forma más automatizada. Los investigadores se sorprendieron cuando los métodos que desarrollaron con sistemas de código abierto también pudieron sortear los límites de protección de los sistemas cerrados. [The New York Times, julio de 2023]

Ejemplos de riesgo: salida

Grupo	Riesgo	Ejemplo
Imparcialidad	Sesgo de salida: el contenido generado podría representar injustamente a ciertos grupos o individuos.	**Imágenes generadas de forma sesgada** Lensa AI es una aplicación para móvil con funciones generativas entrenadas con Stable Diffusion que puede generar “avatares mágicos” a partir de imágenes que los usuarios suben de sí mismos. Según el informe original, 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 injustamente favorecidos** El estudio Gender Shades de 2018 demostró que los algoritmos de machine learning pueden discriminar en función de clases como la raza y el 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, las tasas de error de las personas de piel clara no superaron el 1 %. [TIME, febrero de 2019]
Alineación de valores	Alucinación: generación de contenidos inexactos o falsos.	**Casos judiciales falsos** Según el artículo original, un abogado citó casos falsos y citas generadas por ChatGPT en un escrito legal presentado ante un tribunal federal. Los abogados consultaron ChatGPT para complementar su investigación jurídica para una demanda por lesiones de aviación. El abogado preguntó posteriormente a ChatGPT si los casos proporcionados eran falsos. El chatbot respondió que eran reales y que “pueden encontrarse en bases de datos de investigación jurídica como Westlaw y LexisNexis”. El abogado no comprobó los casos por sí mismo y el tribunal le sancionó. [AP News, junio de 2023] [Reuters, septiembre de 2023]
	Resultado tóxico: cuando el modelo produce contenido reprobable, ofensivo y soez (HAP) u obsceno.	**Respuestas tóxicas y agresivas de los chatbots** Según el artículo, se vio que las respuestas del chatbot de Bing incluían errores fácticos, comentarios sarcásticos, informes airados e incluso comentarios extraños sobre su propia identificación. Los usuarios han compartido ejemplos de las respuestas del chatbot de Bing a consultas que están calificando de “descabelladas” y “gaslighting”, incluyendo escenarios en los que el bot responde airadamente a una pregunta o comentario y luego 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 de “inventadas”, incluso alegando que habían sido “creadas por alguien que quiere perjudicarme a mí o a mi servicio”. [Forbes, febrero de 2023]

Grupo	Riesgo	Ejemplo
Uso indebido	Difundir la desinformación: utilizar un modelo para crear información engañosa con el fin de engañar o confundir a un público determinado.	**Generación de información falsa** Según los artículos de prensa, la IA generativa supone 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 a los votantes para que depositen su voto en la fecha equivocada, grabaciones de audio sintetizadas de un candidato confesando un delito o expresando opiniones racistas, secuencias de vídeo 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 falsamente que un candidato abandonó la carrera. [AP News, mayo de 2023] [The Guardian, julio de 2023]
	Toxicidad: uso de un modelo para generar contenido reprobable, ofensivo y soez u obsceno.	**Generación de contenido nocivo** Según el artículo original, se descubrió que una aplicación de chatbot de IA generaba contenidos nocivos sobre el suicidio, incluidos métodos de suicidio, con instrucciones mínimas. Un hombre belga se suicidó tras pasar seis semanas hablando con ese chatbot. El chatbot le proporcionó respuestas cada vez más nocivas a lo largo de sus conversaciones y le animó a acabar con su vida. [Business Insider, abril de 2023]
	Uso no consensuado: uso de un modelo para imitar a personas a través de vídeo (deepfakes), imágenes, audio u otras modalidades sin su consentimiento.	**Advertencia del FBI sobre los deepfakes** El FBI advirtió recientemente a la opinión pública de la existencia de actores maliciosos que crean contenidos sintéticos y explícitos “con el fin de acosar a las víctimas o de llevar a cabo esquemas de sextorsión”. Señalaron que los avances en 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 original, la Comisión Federal de Comunicaciones de Estados Unidos ilegalizó las llamadas robóticas o robocalls que contienen voces generadas por inteligencia artificial. El anuncio se produjo después de que robocalls generados por inteligencia artificial imitaran la voz del presidente para disuadir a la gente de votar en las primarias del primer estado del país. [AP News, febrero de 2024]
	No divulgación: no revelar que el contenido lo ha generado un modelo de IA.	**Interacción con la IA no divulgada** Según la fuente, un servicio de chat de apoyo emocional en línea realizó un estudio para incrementar o escribir respuestas a unos 4000 usuarios utilizando GPT-3 sin informar 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, ya de por sí vulnerables. Alegó que el estudio estaba “exento” de la ley de consentimiento informado. [Business Insider, enero de 2023]

Grupo	Riesgo	Ejemplo
Generación de código malicioso	Generación de código malicioso: los modelos pueden generar código que, al ejecutarse, cause daños o afecte involuntariamente a otros sistemas.	**Generación de código menos seguro** Según su artículo, investigadores de la Universidad de Stanford han analizado el impacto de las herramientas de generación de código en la calidad del mismo 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, aunque los programadores creían que su código era más seguro. Neil Perry, Megha Srivastava, Deepak Kumar y Dan Boneh. 2023. Do Users Write More Insecure Code with AI Assistants?. En Actas de la Conferencia ACM SIGSAC 2023 sobre seguridad informática y de las comunicaciones (CCS '23), 23-30 de noviembre de 2023, Copenhague, Dinamarca. ACM, Nueva York, EE. UU. , 15 páginas. https://doi.org/10.1145/3576915.3623157
Privacidad	Exposición de información personal: cuando se utiliza información de identificación personal (PII) o información personal confidencial (SPI) en los datos de entrenamiento, los datos de ajuste o como parte de la instrucción, los modelos podrían revelar esos datos en el resultado generado.	**Exposición de datos personales** Según el artículo original, ChatGPT sufrió un fallo y expuso los títulos y el historial de chat de los usuarios activos a otros usuarios. Más tarde, OpenAI compartió que incluso más datos privados de un pequeño número de usuarios quedaron expuestos, incluyendo nombre y apellidos del usuario activo, dirección de correo electrónico, dirección de pago, los últimos cuatro dígitos de su número de tarjeta de crédito y la fecha de caducidad de la tarjeta de crédito. Además, se informó de que los datos relacionados con el pago del 1,2 % de los suscriptores de ChatGPT Plus también quedaron expuestos en la interrupción. [The Hindu BusinessLine, marzo de 2023]
Explainability	Resultado inexplicable: retos a la hora de explicar por qué se generó la salida del modelo.	**Precisión inexplicable en la predicción de la raza** Según el artículo original, los investigadores que analizaron múltiples modelos de machine learning utilizando imágenes médicas de pacientes pudieron confirmar la capacidad de los modelos para predecir la raza con gran precisión a partir de imágenes. Se quedaron perplejos con qué es exactamente lo que permite a los sistemas acertar de forma sistemática. Los investigadores descubrieron que incluso factores como la enfermedad y la complexión física no eran fuertes predictores de la raza; es decir, los sistemas algorítmicos no parecen estar utilizando ningún aspecto particular de las imágenes para sacar sus conclusiones. [Banerjee et al., julio de 2021]

Ejemplos de riesgo: retos

Grupo	Riesgo	Ejemplo
Gobierno	Transparencia del modelo: la falta de transparencia del modelo o la documentación insuficiente del proceso de desarrollo del modelo dificulta la comprensión de cómo y por qué se construyó un modelo, lo que aumenta la posibilidad de un mal uso no intencionado del modelo.	**Divulgación de datos y metadatos del modelo** El informe técnico de OpenAI es un ejemplo de la dicotomía existente en torno a la divulgación de los datos y los metadatos del modelo. Aunque muchos desarrolladores de modelos consideran valioso permitir la transparencia para los consumidores, la divulgación plantea verdaderos problemas de seguridad y podría aumentar la capacidad de hacer un mal uso de los modelos. En el informe técnico del GPT-4, afirman: "Dado tanto el panorama competitivo como las implicaciones para la seguridad de modelos a gran escala como el 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]
	Responsabilidad: 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 raíz y asignar responsabilidades.	**Determinar la responsabilidad de los resultados generados** Según el artículo original, 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 de prensa, el arte generado por IA causó polémica después de que una obra de arte generada por IA ganara el concurso de arte de la Feria Estatal de Colorado en 2022. La obra fue generada por Midjourney, una herramienta de generación de imágenes por IA, siguiendo las indicaciones del artista. La victoria planteó cuestiones sobre derechos de autor. En otras palabras, si todo lo que hizo el artista fue proponer una descripción del arte, pero la herramienta de IA lo generó, ¿a quién pertenecen los derechos de la imagen generada? Según el último artículo, la Oficina de Derechos de Autor de Estados Unidos ha rechazado la protección de los derechos de autor para el arte creado mediante inteligencia artificial porque su autoría no es 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 la patente de contenidos generados** El Tribunal Supremo de EE. UU. se negó a admitir una demanda por la negativa de la Oficina de Patentes y Marcas de EE. UU. a expedir patentes para las invenciones creadas por un sistema de IA. Según el científico, su sistema de IA creó prototipos únicos para un portabebidas y una baliza luminosa de emergencia completamente solo. Los magistrados rechazaron la apelación a la sentencia de un tribunal inferior según la cual las patentes solo pueden concederse a inventores humanos y que el sistema de IA del científico no podía considerarse el creador legal de dos invenciones que generó. Según el último artículo, la Oficina de Propiedad Intelectual del Reino Unido también se negó a conceder la patente alegando que el inventor debe ser un ser humano o una empresa, y no una máquina. [Reuters, abril de 2023] [Reuters, diciembre de 2023]

Ejemplos de riesgo: retos

Grupo	Riesgo	Ejemplo
	Atribución de la fuente: determinar la procedencia del contenido generado.	**Uso de código sin la atribución y los avisos apropiados** Según los artículos originales, 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 con cuyo código de fuente abierta se entrena el servicio. Afirman que el código de entrenamiento utilizó materiales con licencia y han violado las condiciones de servicio y las políticas de privacidad de GitHub, así como una ley federal que obliga a las empresas a mostrar información sobre derechos de autor cuando hacen uso de material. [The New York Times, noviembre de 2022]
Impacto social	Impacto en el empleo: la adopción generalizada de sistemas de IA basados en modelos fundacionales podría provocar la pérdida de empleo de las personas al automatizarse su trabajo si no se les vuelve a cualificar.	**Reemplazo de trabajadores humanos** Según el artículo de prensa, los usos de la inteligencia artificial en el cine y la televisión siguen siendo objeto de debate entre los estudios y los actores de Hollywood. A los actores les preocupa que actores totalmente generados por IA, o “metahumanos”, les sustituyan. Los figurantes y los actores de doblaje, en particular, temen que los intérpretes sintéticos les quiten el trabajo. [Reuters, julio de 2023]
	Explotación humana: uso de trabajo fantasma en el entrenamiento de modelos de IA, condiciones de trabajo inadecuadas, falta de atención sanitaria, incluida la salud mental, e indemnizaciones injustas.	**Trabajadores con bajos salarios para la anotación de datos** Según una revisión de documentos internos y entrevistas a empleados realizadas por el medio TIME, los etiquetadores de datos empleados por una empresa de subcontratación en nombre de OpenAI para identificar contenidos tóxicos cobraban un salario aproximado de entre 1,32 y 2 dólares por hora, en función de la antigüedad y el rendimiento. TIME declaró que los trabajadores están mentalmente afectados, ya que estuvieron expuestos a contenidos tóxicos y violentos, incluidos detalles gráficos de “abusos sexuales a menores, zoofilia, asesinatos, suicidios, torturas, autolesiones e incesto”. [TIME, enero de 2023]

Principios, pilares y gobierno

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. La junta de ética de la IA de IBM tiene la misión de apoyar un proceso centralizado de gobierno, revisión y toma de decisiones en relación con las políticas, prácticas, comunicaciones, investigación, productos y servicios de ética de IA de IBM. Esta incluye un conjunto diverso de partes interesadas de toda la empresa y cuenta con el apoyo de una comunidad de empleados de IBM que sirven como puntos de referencia y defensores de la ética de IA. A través de la junta, los principios de IBM se ponen en práctica. A medida que surgen nuevas tecnologías, como los modelos fundacionales, la junta de ética de la IA de IBM participa activamente en apoyar la alineación con estos principios y pilares, que evolucionan para abordar nuevos problemas éticos de la IA.

Guardarraíles y mitigación

IBM ha establecido una cultura organizacional que apoya el desarrollo y uso responsable de la IA. Según el informe de ética de la IA en acción del IBM Institute for Business Value, la ética de la IA ya está más orientada a los negocios que a la tecnología, y los ejecutivos no técnicos son ahora los principales defensores de la ética de la IA, al incrementarse del 15 % en 2018 al 80 % 3 años después. Además, el 79 % de los CEO ahora están preparados para actuar sobre cuestiones éticas de IA, frente al 20 % anterior. Reconocemos que la IA responsable es un área sociotécnica 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 los riesgos.

Actualmente, IBM participa en investigaciones de vanguardia y desarrolla herramientas para ayudar a los profesionales de soporte a lo largo del ciclo de vida de la IA responsable y confiable. La plataforma de datos e IA preparada para uso empresarial watsonx está construida con 3 componentes: IBM™ watsonx.ai AI studio, IBM™ watsonx.data data store e IBM™ watsonx.governance toolkit. La tecnología de gobierno de IA de IBM permite a los usuarios impulsar flujos de trabajo de IA responsables, transparentes y explicables. Esta tecnología incluye IBM Watson 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 monitorizar la equidad, la explicabilidad, la resiliencia, la alineación con los resultados empresariales y el cumplimiento. IBM también ha desarrollado varios métodos para ayudar con problemas de sesgo como FairIJ, Equi-tuning y FairReprogram. Obtenga más información sobre herramientas adicionales de IA confiables de código abierto.

Los guardarraíles y la mitigación adicionales incluyen:

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

Filtrado de datos no deseados
El uso de 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 indeseable y mal alineado eliminando el lenguaje ofensivo, el lenguaje sesgado y las palabrotas de los datos.
Más información aquí →

Adaptación de dominio
Entrenar un modelo fundacional en un dominio o sector específico puede ayudar a minimizar el alcance del riesgo que los modelos pueden ocasionar, ya que puede condicionarse para generar resultados que se ajusten más a ese dominio o sector.
Más información aquí →

Supervisión y participación humanas
La supervisión y revisión humanas pueden ayudar a identificar y corregir errores y sesgos en el resultado generado. Además, la validación humana y la valoración sobre la calidad de las respuestas del modelo ayudan a garantizar que el contenido generado sea preciso, relevante y de alta calidad, no tergiversado ni alineado.
Más información aquí →

Compromiso de consultoría
IBM™ Consulting se dedica a ayudar a los clientes con el uso seguro y responsable de la IA, independientemente de la pila tecnológica preferente. Ayudan a los clientes a nutrir 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 gobierno de datos.
Más información aquí →

IBM Enterprise Design Thinking
Los métodos e infraestructuras de IBM Enterprise Design Thinking, como Team Essentials para IA, ayudan a los clientes a definir comportamientos éticos en todo el proceso de diseño y desarrollo de IA.
Más información aquí →

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

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

Diversidad en los equipos
La diversidad en los equipos que crean y entrenan sistemas de IA, incluidos los modelos fundacionales, ayuda a garantizar que se tengan en cuenta diversas perspectivas y experiencias. Esta diversidad mejora la precisión y el rendimiento de los sistemas de IA y ayuda a reducir los riesgos durante todo su ciclo de vida, incluidos los posibles resultados adversos que afectan a grupos poco representados en equipos menos diversos.

Políticas, regulación y buenas prácticas de IA

Guía de modelos fundacionales para responsables políticos que recoge todo lo que deben saber sobre los modelos fundacionales. Este blog, del IBM Policy Lab, tiene como objetivo ayudar a los responsables políticos en la compleja tarea de regular el uso de la IA generativa, con el objetivo de evitar riesgos sin limitar la innovación y las oportunidades beneficiosas. Para obtener más información sobre las recomendaciones de IBM a los responsables políticos, lea aquí el testimonio de Christina Montgomery, directora de privacidad y confianza de IBM, ante el Subcomité Judicial de Privacidad, Tecnología y Derecho del Senado de los Estados Unidos.

IBM ejerce cierta influencia sobre la definición de la política reguladora, las buenas prácticas y herramientas del sector, el gobierno de las tecnologías emergentes y la investigación sociotécnica al liderar y contribuir a iniciativas con organizaciones, como por ejemplo:

- El Foro Económico Mundial
- Asociación en IA
- El Centro de Gobierno de IA de la Asociación Internacional de Profesionales de la Privacidad (IAPP)
- La Iniciativa Mundial del IEEE sobre la ética de los sistemas autónomos e inteligentes
- El Servicio de Christina Montgomery en el Comité Asesor Nacional de Inteligencia Artificial (NAIAC)
- El Pacto Digital Mundial 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 el MIT-IBM Watson AI Lab, donde una comunidad de científicos de MIT e IBM Research lleva a cabo investigaciones de IA y trabaja con organizaciones globales para acercar los algoritmos a su impacto en las empresas y en la sociedad. El laboratorio tecnológico Notre Dame-IBM se formó para abordar las diversas cuestiones éticas que conllevan el desarrollo y el uso de tecnologías avanzadas, incluida la IA, machine learning (ML) y la computación cuántica. La investigación sobre la inteligencia artificial centrada en el hombre de la Universidad de Stanford (HAI) avanza la investigación, la educación, la política y las prácticas de la IA.

Siga atento a este espacio para obtener más información sobre los últimos avances en modelos fundacionales y sobre cómo IBM trabaja en pro del desarrollo y el uso responsables de esta y otras tecnologías.

IBM España, S.A.
Santa Hortensia, 26-28
28002 Madrid
IBM Corporation
New Orchard Road
Armonk, NY 10504

Producido en los
Estados Unidos de América
Febrero de 2024

Este documento se actualizó por última vez en la fecha inicial de publicación e IBM puede modificarlo en cualquier momento. No todas las ofertas están disponibles en todos los países en los que opera IBM.